scientific reports

Check for updates

OPEN

Chromosome-level reference genome assembly of the gyrfalcon (*Falco rusticolus*) and population genomics offer insights into the falcon population in Mongolia

Farooq Omar Al-Ajli[1,2,3]✉, Giulio Formenti[3], Olivier Fedrigo[3], Alan Tracey[4], Ying Sims[4], Kerstin Howe[4], Ikdam M. Al-Karkhi[5], Asmaa Ali Althani[6,7], Erich D. Jarvis[3,8], Sadequr Rahman[2,9] & Qasim Ayub[2,9,10]✉

The taxonomic classification of a falcon population found in the Mongolian Altai region in Asia has been heavily debated for two centuries and previous studies have been inconclusive, hindering a more informed conservation approach. Here, we generated a chromosome-level gyrfalcon reference genome using the Vertebrate Genomes Project (VGP) assembly pipeline. Using whole genome sequences of 49 falcons from different species and populations, including "Altai" falcons, we analyzed their population structure, admixture patterns, and demographic history. We find that the Altai falcons are genomic mosaics of saker and gyrfalcon ancestries, and carry distinct W and mitochondrial haplotypes that cluster with the lanner falcon. The Altai maternally-inherited haplotypes diverged 422,000 years before present (290,000–550,000 YBP) from the ancestor of sakers and gyrfalcons, both of which, in turn, split 109,000 YBP (70,000–150,000 YBP). The Altai W chromosome has 31 coding variants in 29 genes that may possibly influence important structural, behavioral, and reproductive traits. These findings provide insights into the question of Altai falcons as a candidate distinct species.

Keywords Genomics, Conservation, Altai falcon, Gyrfalcon, Saker falcon, Peregrine, W chromosome, Reference Genome

Falcons from the Altai Mountain regions in Central and East Asia have historically been referred to as *Falco altaicus* and have been the focus of a more than 200-year-old debate[1,2]. Their habitat in the Altai region is a wintering zone for both gyrfalcons (*F. rusticolus*) and saker (*F. cherrug*) falcons, which has contributed to the speculation that the Altai region falcons could represent a natural interspecific hybrid population[3,4] (Fig. 1A). In Mongolia, this falcon population is restricted to the western parts of the country including the Altai mountain range[5]. The currently accepted taxonomical classification of these falcons is either a saker population or a saker subspecies, with its distinctive dark morphology interpreted as a population-specific morph[4,6]. Indeed, morphologically some Altai region falcons resemble an intermediate form between sakers and gyrfalcons.

Despite the significance of resolving their status for conservation, some researchers consider the Altai region falcons as natural interspecific gyrfalcon-saker hybrids merely based on their resemblance to sakers and/or gyrfalcons[4,7]. Others treat these falcons as a subspecies of either the gyrfalcon or saker (i.e. *F. cherrug/rusticolus altaicus*)[8]. Some biologists and taxonomists consider them as a separate species, based on a different set of criteria, including their ecological niche (foraging, breeding, and wintering), as they occupy highly elevated mountains, which sakers usually avoid[2,7,9,10]. Alternatively, it has been suggested that Altai region falcons are natural hybrids of first or subsequent generations between the gyrfalcons and sakers, with the saker genes

[1]Qatar Falcon Genome Project, Doha, Qatar. [2]School of Science, Monash University, Subang Jaya, Malaysia. [3]Vertebrate Genome Laboratory, The Rockefeller University, NY, USA. [4]Wellcome Sanger Institute, Cambridge, UK. [5]Souq Waqif Falcon Hospital, Doha, Qatar. [6]Biomedical Research Center, Qatar University, Doha, Qatar. [7]Health Cluster, Qatar University, Doha, Qatar. [8]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. [9]Tropical Medicine and Biology Multidisciplinary Platform, School of Science, Monash University, Subang Jaya, Malaysia. [10]Genomics Platform, School of Science, Monash University, Subang Jaya, Malaysia. ✉email: farooq.alajli@katara.net; qasim.ayub@monash.edu

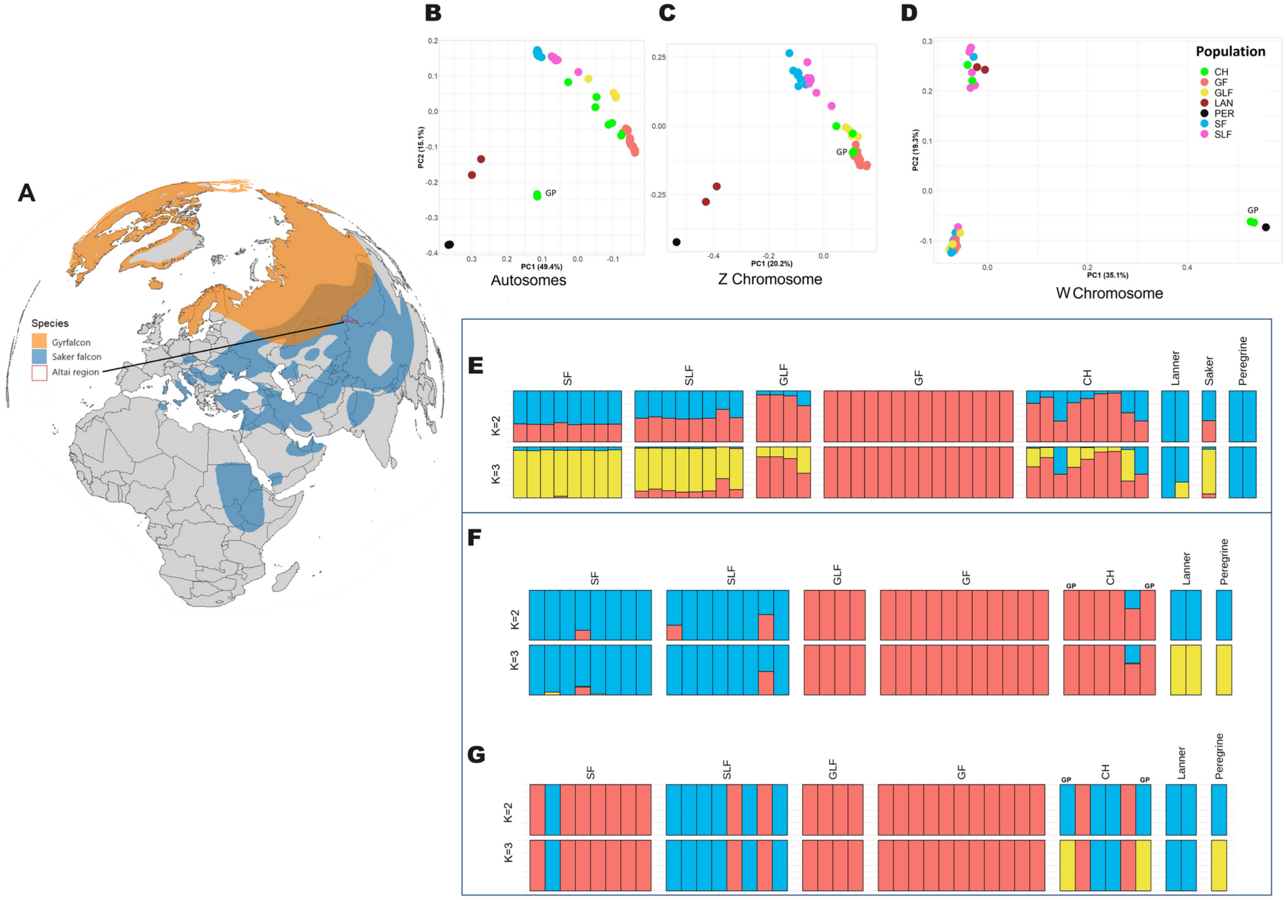

Fig. 1. Population structure analysis of 48 falcon genomes. (**A**) The distribution ranges of the gyrfalcon (Orange shade), saker (Blue shade), and Altai region (Red outline). (**B**), (**C**) and (**D**) show Principal Component Analysis (PCA), based on autosomal, Z and W-chromosomal SNP datasets, respectively. Individual falcons are represented by a circle and color coded by population. **(E), (F)** and **(G)** are ADMIXTURE analyses, where each bar represents the ancestry of an individual falcon using autosomal (E), Z-chromosomal (F), and W-chromosomal (G) SNP datasets, respectively, at a specific cluster number (K value). A single color denotes a pure lineage, while admixed individuals show multiple colors based on their ancestral groupings. The populations are North American gyrfalcons (GF) and sakers (SF). Falcons from the Altai region are designated gyrfalcon-like falcons (GLF) or saker-like falcons (SLF). Commercial Hybrids (CH) including gyrfalcon-saker and gyrfalcon-peregrine (labeled as "GP"). The publicly available genomes of one saker, one lanner and two peregrines were added to the analysis, and labeled as such. See Table 3 for details on individual samples.

predominating[3,4]. Others speculated that these falcons are indeed hybrids, but of falconry bird escapees, tracing their origin back to the Mongolian emperor Kublai Khan's hunting expedition (ca 1290 A.D.), which included 10,000 falconers carrying a "vast number of gyrfalcons, peregrine falcons and sakers"[7,11]. Based on criteria that include head and plumage patterns, Altai region falcons are informally classified into saker-like and gyrfalcon-like falcons[12], which is the neutral definition we follow in this study, as it does not presume their taxonomic status. Nonetheless, without a robust and comprehensive classification of Altai falcons *(F. altaicus)*, conservation efforts would be impeded, misguided and sometimes detrimental[13–16]. Being conflated with the saker falcon, whose range extends from Western Asia to Eastern Europe, the Altai region falcons may lack the due conservation attention and informed management they require.

Previous studies, based on conventional genetic markers (microsatellites and partial mitochondrial sequences) have not conclusively resolved the phylogenetic relationship between the sakers and the gyrfalcons[17]. While the sequenced genome of the saker falcon has allowed a more detailed insight into its genetic and demographic history, it was not without limitations. Similar to the peregrine falcon, the saker's genome size was found to be approximately 1.2 Gbp in length, encoding about 16,200 genes[18]. However, these previous genome assemblies, along with the prairie falcon[19], were generated using short-read sequencing technologies, resulting in highly-fragmented contigs that are prone to misassemblies and fall short of capturing whole chromosomes and the other high-quality standards proposed by the Vertebrate Genomes Project (VGP) and the Earth BioGenome Project[20]. The resolution of the genomes of non-model organisms that have high heterozygosity between haplotypes has been a challenge for short-read sequencing technologies and assembly algorithms, contributing to high fragmentation and misassembly[21,22]. The resultant assemblies have a higher number of gaps and poor resolution of structural variations, which can have critical conservation, adaptive, and phylogenetic implications[23–25]. As a result, the annotated dataset contains many missing, partial and/or misassembled genes, coding sequences (CDSs), and gene boundaries[26–29]. Capturing such inaccessible genomic information by using long-read sequencing and other complementary technologies allows for more accurate comparative studies on population, hybridization and speciation[25], and informs the conservation and management strategies. Here we generated a high-quality, chromosome-level VGP assembly of the gyrfalcon, followed by a comprehensive population analysis of gyrfalcons, sakers, Altai region falcons, peregrines, lanners and their hybrids, and used it to address their demography and taxonomic status of the Altai region falcons.

Results

High-quality chromosome-level annotated gyrfalcon assembly

Short-read-based genome assembly projects often sequence homogametic individuals (i.e. males in birds and females in mammals) to avoid the issues that arise from attempting to assemble the highly repetitive heterogametic W or Y chromosomes[30–33]. To represent both W and Z chromosomes we chose a heterogametic female gyrfalcon and generated 62 Gbp of Pacific Biosciences continuous long reads (CLR; 51.6X-coverage) and 121 Gbp of Illumina short-read raw data (101.3X), 150.87 Gbp of linked reads using 10 × Genomics (125.7X), 120.83 Gbp of Arima Hi-C data (100.7X) and 345 Gbp of Bionano Optical map data (287.5X). The data were assembled following the VGP pipeline 1.6[20]. In summary, contigs were generated with FALCON unzip, which were then scaffolded and polished with 10 × Genomics linked reads, followed by further scaffolding with Bionano optical genome maps and Hi-C linked. Next, manual curation, guided by Hi-C maps, made 52 breaks and 132 joins in the scaffolds to correct the assembly, and removed 49 sequences (1.2 Mb) (0.1%) relating to false duplication from the primary assembly. This process produced 22 autosomes and the W and Z sex chromosomes (25,584,520 bp and 84,785,561 bp, respectively), with a total genome size of 1,195,847,496 bp (Table S1). The number of gaps in the assembly is 517. This resulted in a contig N50 of 15.8 Mbp and NG50 of 60.5 Mbp, and a scaffold N50 and NG50 of 91.1 Mbp of the primary haplotype (Table 1). The estimated K-mer-based Quality Value (QV) of the final reference assembly is 40.6 with K-mer-based completeness of 96.9%, and BUSCO completeness of 98.3% (2,543 Complete Universal Single-Copy Orthologs, with only 23 orthologs missing; Fig. S1). This reference assembly satisfies the VGP quality values[20], and has a ~ 500-fold increase in contiguity compared to all previous short-read genome assemblies of any falcon species (Table 1). During the preparation of this paper, a genome assembly of the gyrfalcon using PacBio continuous long read (CLR) technology and Bionano Optical Maps was published[34]. Despite using long-read technology, this assembly still falls short of capturing the near-complete lengths of chromosomes, particularly the sex chromosomes and microchromosomes[34], which is reflected on the lower scaffold N50. Our reference assembly achieved the highest possible scaffold N50, as well as assembled near-complete lengths of autosomes and sex chromosomes (Table S1). The contrast between the reference assembly of this study and the assembly presented by Zuccolo et al. can also be attributed by the difference in the assembly methods. The advantage of the Vertebrate Genomes Project (VGP) assembly pipeline used in this study lies in its combination of long reads, short reads, Hi-C data, optical mapping, and manual curation guided by Hi-C maps. In the case of the assembly presented in our study, manual curation fixed many issues that arose from the scaffolding step. In general, long-read assemblies based on PacBio CLR sequencing are prone to false duplications[29]. The VGP pipeline mitigates this problem by integrating multiple data types and applying careful manual curation based on Hi-C maps, which is shown to significantly improve the quality of genome assemblies[20,35]. Moreover, Zuccolo et al. employed two assemblers, Mecat2 and Canu, to generate two assemblies, but ultimately selected Canu's as the baseline assembly due to its higher quality. However, Canu is less effective at preventing false duplications compared to the FALCON-UNZIP assembler that was used in this study as part of the VGP assembly pipeline[29]. FALCON-UNZIP generates a primary assembly while separately producing contigs for the alternative haplotype, and both were deposited to NCBI in our case. Additionally, the VGP pipeline utilizes 'purge_dup,' a tool highly effective at removing false duplicates, including collapsed haplotypes from artifactually duplicated regions[29], which are associated with

Species	Common Name	Assembly Total size (Gbp)	Number of contigs (C), scaffolds (S)	N50 contig (C), scaffold (S)	Ref
Falco rusticolus	**Gyrfalcon**	**1.20**	**C: 746** **S: 37**	**C: 15.8 Mbp** **S: 91.1 Mbp**	**This study**
Falco cherrug	Saker	1.18	C: 56,956 S: 1,069	C: 31.2 kbp, S: 4.15 Mbp	Zhan et al. 2013
Falco peregrinus	Peregrine	1.17	C: 61,119 S: 1,092	C:28.6 kbp S: 3.89 Mbp	Zhan et al. 2013
Falco mexicanus	Prairie	1.17	C: not reported S: 2,181	S: 3.7 Mbp	Doyle et al. 2018
Falco rusticolus	Gyrfalcon	1.19	C: 1,219 S: 33	C: 48 Mbp S: 73.4 Mbp	Zuccolo et al. 2023

Table 1. Summary statistics of the assembly parameters for the gyrfalcon reference genome generated in this study compared to the published genome assemblies of four falcon species (saker, peregrine and prairie, gyrfalcon). Abbreviations: C = contigs and S = scaffolds.

Canu assemblies[36]. While Bionano optical maps are a valuable tool for improving the contiguity of assemblies, using them alone to scaffold assemblies has been shown to be less effective than Hi-C alone[37]. Therefore, it is essential to combine multiple technologies (long-reads, Hi-C maps, Bionano Optical maps, linked-reads) to enhance the quality and completeness of genome assemblies[38,39].

The ab initio annotation yielded 19,301 coding and non-coding genes, which are within the expected range of annotated genes in other published avian genomes, especially that of falcons[18,19]. Of these, 1,612 and 462 genes were on the Z and W chromosomes, respectively, including known sex-linked genes such as the *CHD1* (chromodomain-helicase-DNA binding) on both chromosomes and Z-linked *DMRT1* (Doublesex and mab-3 related transcription factor 1). The assembled size (~ 85 Mbp) and number of genes on the gyrfalcon's Z chromosome are in line with the expected size and gene count of the highly conserved avian Z chromosome[40,41]. The final annotation of the assembly was done using the National Center for Biotechnology Information (NCBI) Eukaryotic Genome Annotation Pipeline, with the assembly upgraded to NCBI RefSeq assembly (GCF_015220075.1). This analysis identified 15,894 protein-coding genes, of which 734 are Z-linked, and 186 are W-linked (GCA_015220075.1). Although the number of genes is comparable to short-read-based falcon assemblies, the high contiguity and base quality of the gyrfalcon reference genome presented here, coupled with the significant reduction in gaps, yielded more complete genic and intergenic sequences, including phased euchromatic sequences. For example, the NCBI Annotation Release 102 for the short read saker falcon assembly (accession: GCF_000337975.1) identified 15,025 protein-coding genes, out of which 1,319 partial coding sequences (CDS). In contrast, NCBI annotation of the gyrfalcon assembly identified 41,414 CDSs, of which only 105 are partial CDSs. The gyrfalcon's genome assembly was also scanned for transposable elements (TE) utilizing a long-read-based curated repeat library developed using high-quality genomes of three species of sparrows[42]. The total percentage of the gyrfalcon's genome spanned by interspersed repeats is 5.57% (Table 2). Among the identified repeats, retroelements constituted the majority, with 180,305 elements (4.94%) (Table 2). Within these retroelements, LINEs were predominant, totaling 113,890 elements (3.11%), including L2/CR1/Rex (31,982 elements, 1.16%). SINEs were less abundant, contributing 7,165 elements (0.07%). LTR elements comprised 59,250 elements (1.77%), predominantly retroviral in origin (57,540 elements, 1.73%). DNA transposons were comparatively rare, consisting of 5,921 elements (0.07%), including low-level occurrences of hobo-Activator

Class of Transposable Elements			**Number of elements**	**Length occupied (bp)**	**Percentage of the genome (%)**
Retroelements			180,305	59,078,776	4.94
	SINEs		7165	807,045	0.07
	Penelope		0	0	0
	LINEs		113,890	37,133,287	3.11
		CRE/SLACS	0	0	0
		L2/CR1/Rex	31,982	13,926,711	1.16
		R1/LOA/Jockey	0	0	0
		R2/R4/NeSL	57	15,449	0
		RTE/Bov-B	0	0	0
		L1/CIN4	43	7446	0
	LTR elements		59,250	21,138,444	1.77
		BEL/Pao	0	0	0
		Ty1/Copia	0	0	0
		Gypsy/DIRS1	0	0	0
		Retroviral	57,540	20,680,503	1.73
DNA transposons			5921	817,806	0.07
	hobo-Activator		630	169,237	0.01
	Tc1-IS630-Pogo		209	34,750	0
	En-Spm		0	0	0
	MULE-MuDR		9	3053	0
	PiggyBac		0	0	0
	Tourist/Harbinger		2363	242,411	0.02
	Other (Mirage, P-element, Transib)		0	0	0
Rolling-circles			3	202	0
Unclassified			38,916	6,714,298	0.56
Total interspersed repeats				66,610,880	5.57
Small RNA			1325	179,751	0.02
Satellites			385	61,671	0.01
Simple repeats			267,255	10,367,087	0.87
Low complexity			54,362	2,625,933	0.22

Table 2. Summary of Repetitive Element Composition and Abundance in the gyrfalcon genome.

(630 elements, 0.01%) and Tourist/Harbinger (2,363 elements, 0.02%). Other repeat categories, such as rolling-circles and unclassified elements, as well as small RNA, satellites, simple repeats, and low-complexity regions, collectively accounted for minor fractions of the genome (Table 2).

Admixture in wild falcon populations in Mongolia and the identification of a highly differentiated W chromosome

Population structure and admixture analyses were carried out using three whole genome sequence datasets of high-quality autosomal, Z- and W-linked single nucleotide polymorphisms (SNPs) (Table S2). Principal component analysis (PCA) using 381,247 autosomal SNPs separated the sakers (SF) and gyrfalcons (GF) (Fig. 1B). Both falcons from the Altai region, the saker-like (SLF) and gyrfalcon-like (GLF), clustered closer to the species they morphologically resembled.

For the sex chromosomes, the Z-linked dataset based on 6,466 SNPs (Table S2) reflected a population differentiation similar to the autosomal dataset (Fig. 1C). The most striking result was the distinct clustering on the W chromosomes (Fig. 1D). Only females were included in this analysis since males do not carry a W chromosome. The W-chromosome PCA (using the 132,877 high-quality SNP dataset, Table S2) revealed three demonstrably distinguished clusters. The first component separated the peregrines and gyrfalcon-peregrine captive-bred hybrids (Fig. 1D, labeled “GP”) from all other samples. The second component split the latter into two groups one at the bottom left comprising of saker falcon (SF), gyrfalcon (GF) and gyrfalcon-like Altai region falcons (GLF) and the other containing the majority of saker-like Altai falcons (SLF), along with the lanner.

The PCA results were corroborated by the ancestral composition results (ADMIXTURE) (Fig. 1, E to G). The autosomal dataset was best represented at K = 2, i.e. 2-population level, with a cross-validation (CV) value = 0.54, and clearly separated the gyrfalcons (GF) and sakers (SF). All GF individuals showed a maximal gyrfalcon ancestry fraction (Q_G = 0.99999), while all SF individuals showed zero gyrfalcon ancestry fraction (Q_G = 0.00001) (Table S3). Based on these findings, we consider both of these populations as representatives of their corresponding species. In contrast, at K = 2, SLF revealed notable levels of admixture between sakers and gyrfalcons (Q_G ~ 15%). The four GLF individuals also showed significantly higher gyrfalcon ancestry levels, up to 94% Q_G, admixed with saker (Fig. 1E). The lanner appears to share autosomal resemblance with both the sakers, saker-like Mongolian falcons, and the peregrines, but not with the gyrfalcons or gyrfalcon-like falcons. ADMIXTURE plots for both sex chromosomes were also found to be concordant with their corresponding PCA analyses. The population structure based on the Z chromosome (Fig. 1F) generally resembles the autosomal patterns. The Z chromosome of both the lanner and the peregrine appears to be highly similar to each other, while dissimilar to both the gyrfalcon’s and the saker’s Z chromosome (Fig. 1F). The W-chromosome’s ADMIXTURE patterns at K = 2 separate the peregrines from the rest of the falcons (Fig. 1G). At K = 3, the SLF population seems to predominantly carry a distinct W-chromosome haplotype, which is also carried by the lanner. This “Altai haplotype” is different from all except one saker falcon and all the gyrfalcon’s. At K = 3, two individuals, both gyrfalcon-peregrine captive-bred hybrids (Fig. 1G, GP), form a distinct haplogroup along with the peregrine, as they carry a peregrine’s W chromosome from the mother side (Fig. 1G), and the gyrfalcon’s Z chromosome from their father’s side (Fig. 1F). In addition, our analysis of the Z chromosome of female hybrids revealed that gyrfalcon-saker hybrids tend to have a gyrfalcon sire (Fig. 1F, labeled “CH”). This confirmed the reported common practice among captive-falcon breeders where they use gyrfalcons as sires to increase the phenotypic resemblance (i.e. commercially valuable traits such as size and build) of the hybrid offspring to the gyrfalcons. This phenomenon is referred to as the “paternal effect”, as falcon sires contribute more sex-linked alleles to the offspring than the falcon dams[43]. Runs of homozygosity (ROH) analysis, which evaluates inbreeding levels, were estimated for GF, GLF, SF, and SLF populations. The analysis revealed that SLF and SF have low to moderate mean ROH lengths of 4,477 kbp (SD = 6,075) and 8,320 kbp (SD = 5,659), respectively. GLF and GF, on the other hand, showed higher homozygosity levels, with mean ROH length of 62,663 kbp (SD = 33,859) and 101,555 kbp (SD = 99,422), respectively. Nucleotide diversity (π) of 20 kbp non-overlapping window was calculated for the four populations. Both SF and SLF showed a relatively higher per-population average nucleotide diversity ($\pi = 0.083 \pm 0.032$ and $\pi = 0.074 \pm 0.032$, respectively), whereas GF and GLF had lower per-population average nucleotide diversity ($\pi = 0.036 \pm 0.031$ and $\pi = 0.050 \pm 0.031$, respectively). Both the low levels of homozygosity and high π suggest that SF and SLF populations have large effective population sizes and lower inbreeding levels, which could be beneficial for their adaptability and resilience against environmental changes, diseases, and other selective pressures. For a more general look at the genomic relatedness of the falcon populations in this study, we have calculated F_{ST} estimates between each pair of the falcon populations in this study (GF, GLF, SF and SLF) (Fig. S2). The pairwise F_{ST} estimates were found to be comparable to the suggested relationships between the populations as reported by PCA and admixture results in this study (Fig. 1).

In order to examine the nature of the admixture of SLF and GLF and discern incomplete lineage sorting (ILS) from hybridization, we performed a D-statistic test (or BABA-ABBA test, Fig. 2). The D-statistic test is a parsimony-like method for detecting gene flow, differentiating it from Incomplete Lineage Sorting (ILS)[44]. The D-statistic classifies alleles as ancestral (‘A’) and derived (‘B’) across the genomes of four populations; two ancestral, one “admixed” and one outgroup. In this case, the two ancestral populations are the gyrfalcons (GF) and the sakers (SF), the potentially admixed populations are the Mongolian Altai region falcons (SLF or GLF) and the outgroup is the peregrine (P). In the case of ILS alone, the number of discordant alleles (SNPs) grouping the ancestral populations with admixed populations should be roughly the same, i.e. BABA ≃ ABBA. However, a significant difference ($|z|$ score > 3) between ABBA and BABA indicates that one of the ancestral populations shares more derived alleles with the potentially admixed population than expected by chance, implying introgression and gene flow between these two populations[45]. Using autosomal and Z-linked SNPs, both populations from the Mongolian Altai region show significant gyrfalcon introgression. However, GLF shares significantly more derived alleles with the pure gyrfalcons compared to SLF (Fig. 2). Moreover, our results suggest minimal gene

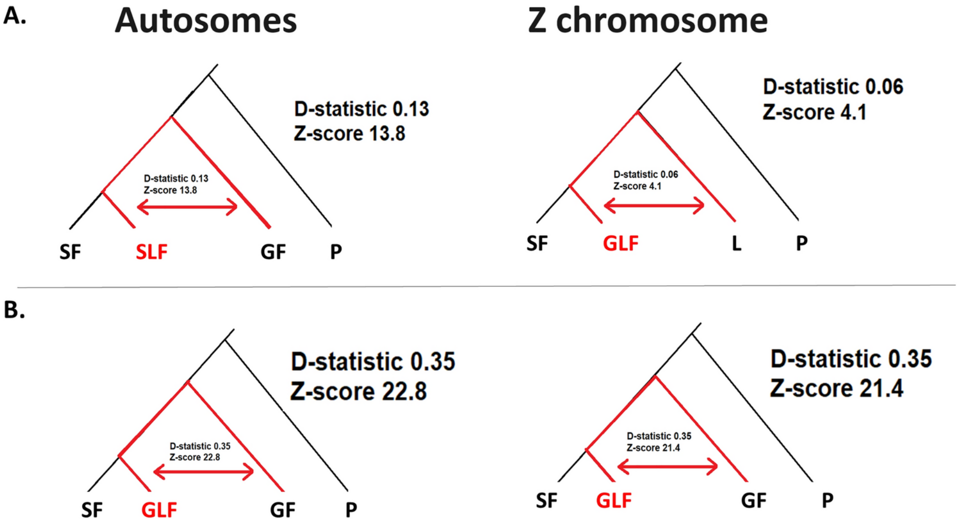

Fig. 2. D-statistics analysis for (**A**) Altai saker-like falcon, and (**B**) Altai gyrfalcon-like falcon estimated using two datasets; autosomes and Z-chromosome. The significance level is calculated using Z score. The populations are saker (SF), North American gyrfalcon (GF), Altai gyrfalcon-like (GLF), Altai saker-like (SLF) and Peregrines (P) as an outgroup. The red arrow represents the excess allele sharing.

flow between SF and lanner falcon compared to SLF and lanner ((((SF,SLF)LAN)P), D = 0.0127, Z-score = 5.66). Interestingly, our analysis reveals that SF shares more alleles with GF than it does with SLF ((((GF,SLF)SF)P), D = 0.0937, Z-score = 16.8). This observation adds another layer to the PCA and ADMIXTURE results, and lends further support to the suggestion of SLF as a distinct conservation unit.

Altai haplotype split 0.422 MYA from saker and gyrfalcon haplotypes

Next, to estimate the divergence time of the Altai haplotype from the other closely-related falcon species, we assembled complete de novo mitogenomes for 35 individuals and generated a time-calibrated phylogenetic tree (chronogram) (Fig. 3A). We also included seven additional falcon reference mitogenomes retrieved from NCBI (Accession numbers: common kestrel NC_011307.1, American kestrel NC_008547.1, merlin KM264304.1, saker NC_026715.1, peregrine 1 NC_000878.1, peregrine 2 JQ282801.1 and gyrfalcon KT989235.1). Moreover, we constructed a maximum-likelihood phylogenetic tree using an alignment of W-specific variants (3,007 SNPs) in 39 female falcons (Fig. 3B). The analyses of the mitochondrial phylogeny corroborated the distinct W haplotype observed in Mongolian saker-like falcons (Fig. 3, A and B), consistent with both being maternally-inherited in birds. Our results show that peregrines diverged from hierofalcons around 2.77 MYA, a result consistent with previous estimates[46,47]. Hierofalcon is a complex of species that includes the saker (*F. cherrug*), gyrfalcon (*F. rusticolus*), lanner (*F. biarmicus*), luggar (*F. jugger*) and the Australian black falcon (*F. subniger*)[17]. The hierofalcon group was further resolved into two main clades: one comprising the majority of SLF individuals along with the lanner falcon, and another comprising SF, GF and GLF. The two clades separated ~ 0.422 MYA (0.290–0.550 MYA) (Fig. 3A and Fig. S3). Two of the eight SLF individuals (SP42 and SP35) clustered with the sakers (SF), whereas SP32 clustered as an SLF. The divergence time between GF and GLF, on one hand, and SF on the other was much more recent and estimated to be around 0.109 MYA (0.070–0.150 MYA) (Fig. 3A and Fig. S3).

To uncover the dynamics of the demographic history of falcon populations, we performed pairwise sequentially Markovian coalescent (PSMC) utilizing the heterozygous autosomal sites identified in their genomes based on their autosomal data (Fig. 4). Our analysis inferred fluctuations in the effective population sizes (Ne) of four species of falcons along with saker-like Altai region falcons from 5 MYA to 10,000 years ago. We also identified a proto-hierofalcon as the most common recent ancestor of lanners, gyrfalcons and sakers, which appear to diverge from the ancestral peregrine falcons around 2–3 MYA (Fig. 4). This finding is consistent with the divergence time inferred using mitogenomes (i.e. 2.77 MYA, Fig. 3A). More specifically, the population of the proto-hierofalcon seems to expand until around 1 MYA, when the lanner splits from the rest of the hierofalcon group, and continued to expand well into the Last Glacial Period (LGP, 115,000—11,700 years ago; Fig. 4) when the global climate was colder. Meanwhile, the ancestral population of the gyrfalcons, sakers and Altai region falcons (SLF and GLF) appears to have experienced a bottleneck that lasted until around 200,000 years ago. After which, the saker population, like the lanner, demonstrates a drastic expansion, overlapping with LGP (Fig. 4). On the other hand, the gyrfalcon (both North American and GLF) appears to have never recovered from the bottleneck, maintaining a low, but steady effective population size until 10,000 years ago, even during the LGP. This may reflect an early adaptation of the gyrfalcons to cold climates, which also may have restricted the expansion of its population. Gyrfalcons are considered specialist predators in terms of diet and habitat. Unlike other falcons, gyrfalcons are a resident species in the Arctic tundra, adapted to the very cold weather, and specializing in preying on rock ptarmigan[48,49]. These factors may have contributed to the stabilization and limitation of the gyrfalcon population throughout its demographic history. The peregrine population appears to

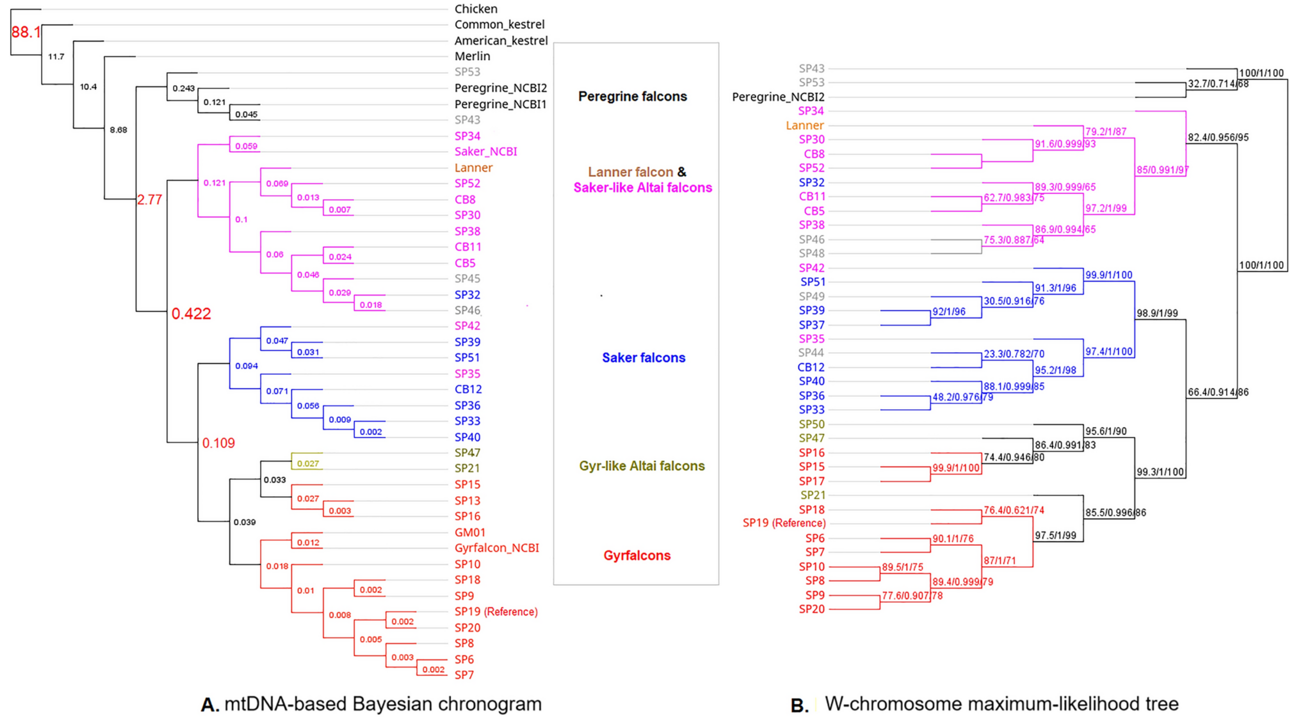

Fig. 3. Divergence times of multiple falcon species and populations. (**A**) Bayesian chronogram (time-calibrated phylogenetic tree) of 42 falcons based on whole mitogenome (mtDNA) sequences (excluding the control region, CR) with chicken mtDNA as an outgroup, generated using Beast2. The nodes are labeled with divergence time in million years ago (MYA), based on the estimated divergence time between chicken and falcons of 88 ± 10 MYA. A full tree with confidence intervals of the age estimates can be found in Fig. S3. (**B**) Maximum-likelihood phylogenetic tree using an alignment of W-specific variants (3,007 SNPs) in 39 female falcons. The tree was constructed using K2P + R2, which was the best-fit model according to BIC. Branch labels represent SH-aLRT support (%), aBayes support, and ultrafast bootstrap support (%), respectively. The red cluster denotes gyrfalcons (GF) and the green denotes Altai gyrfalcon-like falcons (GLF), the blue cluster represents sakers (SF), and the purple cluster represents saker-like Altai falcons (SLF). SP19 mitogenome is the reference mitogenome that was assembled in this project. Gray labels refer to hybrid individuals that were added to validate the W chromosome haplotype results. Sequences (labeled with the species name) were retrieved from NCBI (Accession numbers: common kestrel NC_011307.1, American kestrel NC_008547.1, merlin KM264304.1, saker NC_026715.1, peregrine 1 NC_000878.1, peregrine 2 JQ282801.1 and gyrfalcon KT989235.1).

have experienced a bottleneck after they diverged from the proto-hierofalcon, around 2 MYA, while maintaining a steady effective population size, until around 500,000, when the peregrine population seems to have expanded, before declining around 50,000 years ago (Fig. 4).

It has been suggested that the spike observed in effective population size (Ne) of an F1 hybrid represents the point at which gene flow ceased between two populations[50,51]. Using an F1 gyrfalcon-peregrine hybrid, the spike pattern indicates the maximum bound of divergence time between the most recent common ancestor of hierofalcons and ancestral peregrines. We estimated this spike between 2–3 MYA (Fig. 4), which supports the phylogenetic estimate previously reported[18,52], as well as the mitochondrial-based divergence time estimated in this study, i.e. 2.77 MYA (Fig. 3).

Insights into morphological and behavioral adaptations

To identify candidate regions under selection in Altai SLF population, genome-wide sliding-window analysis of the fixation index (F_{ST}) was employed. Our results identified 59 regions with significant differences in allele frequencies between SF and SLF populations. Only genomic windows corresponding to the top 0.1% of the empirical genome-wide F_{ST} distribution were considered as outliers and considered as potential candidate region under selection (variants above the red line ($F_{ST} > 0.227$) in Fig. 5A). These regions overlap with 40 genes (Table S6). In previous studies, these genes were associated with physiological pathways contributing to adaptation to altitude (*TMEM132C, OTUD7A, MSRB3, NIPBL, DHX36, E2F6)*[53–57], body size, weight and musculoskeletal development *(MSRB3, HMGA2, GREB1, ARHGEF26, EXOC4, MED15, SBF2, ZFYVE16, HNF1A, TMEM132D)*[58–67], immunological responses (*GATA3, OTUD7A, TBK1*)[68–70], and reproduction and embryogenesis (*GATA3, CCDC125, IPO11, YWHAH)*[71–74]. These genomic signatures draw attention to the phenotypic and genomic adaptive divergence of SLF that needs further in-depth studies. Moreover, it also offers an explanation to previously reported phenomena regarding the ecological distribution of these two falcon

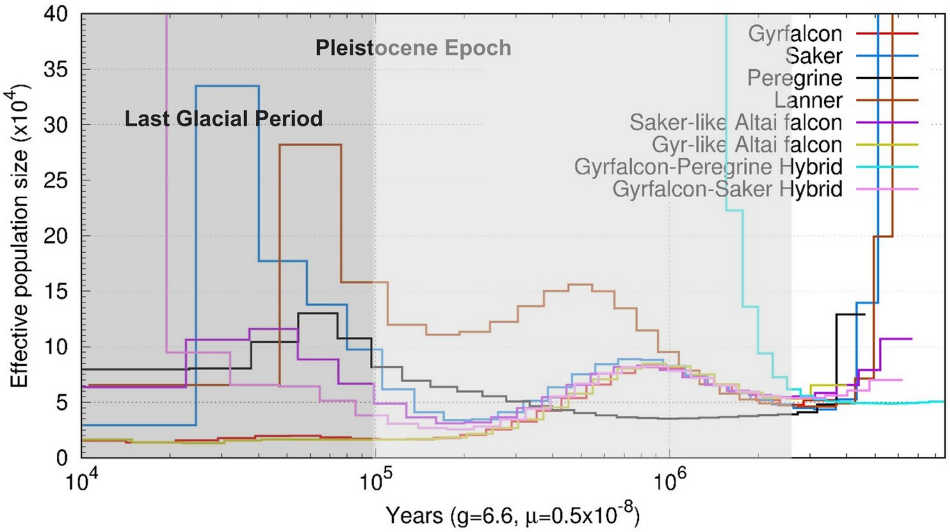

Fig. 4. Pairwise sequential Markovian coalescent (PSMC) analysis of the autosomal data based on the heterozygous sites, showing the demographic history of four falcon species, one Altai saker-like falcon, one Altai gyrfalcon-like falcon and two F1 hybrids. It demonstrates the fluctuation of effective population size from 6 million to 10,000 years ago. Time scale on the x-axis is calculated assuming a mutation rate of 0.5×10^{-8} per generation and generation time equal to 6.6 years.

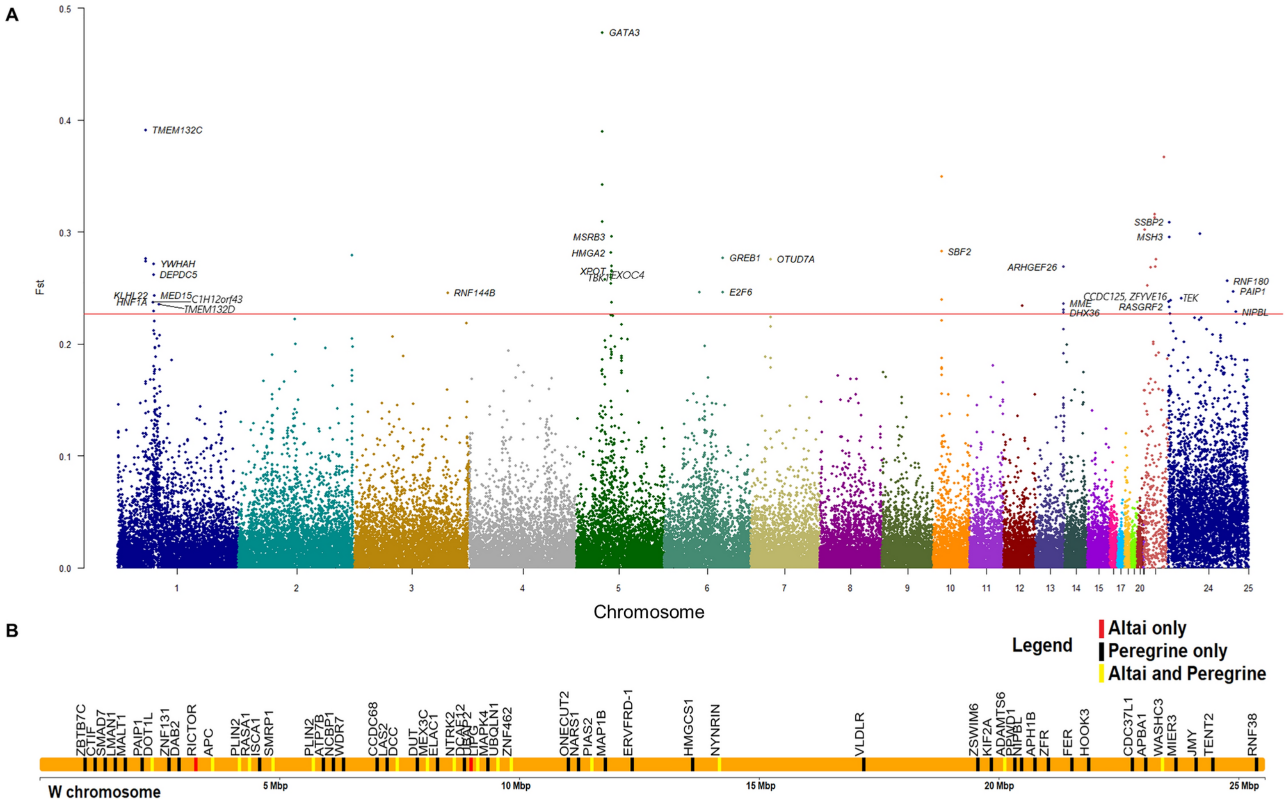

Fig. 5. Functional Analysis of the whole genome and W chromosome **(A)** Manhattan plot of the genomic differentiation between saker falcon (SF) and saker-like falcon (SLF) datasets. The F_{ST} value of 58,401 sliding 20 kbp genomic non-overlapping windows based on SNP datasets are plotted against the chromosome-level gyrfalcon reference genome. Regions above the red line (99.9th percentile of the data) are those exhibiting significant population differentiation and could potentially be under selection, $F_{ST} > 0.227$. For the complete list of the genes with their genomic coordinates and F_{ST} values see Table S6, **(B)** W-linked genes affected by moderate or high impact mutations exclusive to peregrine, Altai, or found in both. The reference alleles used in this analysis are the gyrfalcon's and saker's.

populations. For example, Altai falcons tend to occupy the high-altitude Altai mountain ranges, which sakers usually avoid[2].

W chromosome genes influence speciation

The gyrfalcon's W chromosome with a size of 25.6 Mbp is significantly larger compared to 9.1 Mbp in the chicken (VGP assembly, NCBI Accession: GCF_016699485.2) and larger than the 20 Mbp W chromosome of the zebra finch (VGP assembly, NCBI Accession: GCF_003957595.2). The expanded size of the falcon's W was also reflected in its relatively high gene content with 186 protein-coding genes (Table S4), compared with 90 found on the chicken's W chromosome and 165 protein-coding genes found on the zebra finch's W. Functional analysis shows that gyrfalcon's W-linked genes are involved in multiple reproductive and development pathways including the gonadotropin-releasing hormone receptor pathway, angiogenesis, fibroblast growth factor (FGF) and transforming growth factor β signaling pathways (Table S5). A few of these genes appear to be W-linked; *NEK5, ERVV-2, WASHC3, ERVW-1, ATP7B, VPS36, ALG11, SEC11C, RX2, ERVFRD-1* (2 copies) (Table S4). Some of these genes are also missing from the chicken and the zebra finch genomes, such as *ERVFRD-1,* which encodes syncytin-2, a protein associated with trophoblast development in humans[75]. Many other W-linked genes belong to the endogenous retroviruses (ERVs) family including *ERVFRD-1, ERVW-1, NYNRIN, ERVK-5, ERVV2.* The ERV family of genes is thought to play a role in reproduction, female-biased mutational load, and reproductive isolation[30,76]. Moreover, ERVs were found to function as species-specific enhancers in germline gene expression, influencing spermatogenesis and possibly contributing to speciation[77]. *NTRK2*, a gene associated with body weight in humans and mice[78], was found to have two copies on the gyrfalcon's W chromosome. While it has a gametolog on the Z chromosome, the two W-linked copies may suggest a possible overexpression in females that may contribute to the reverse sexual dimorphism (RSD) observed in falcons, where females are larger than males. Another W-linked gene associated with body size is *NIPBL*, which, along with *NTRK2*, has been implicated in RSD in Chinese tongue sole (*Cynoglossus semilaevis*)[79]. The presence of these genes on the gyrfalcon's W chromosome may provide an explanation for the RSD in female falcons.

As the extent of the adaptive information provided by mitogenome variants is limited, we wanted to gain a deeper insight into the genetic basis underlying the selective variance within the W chromosome, and what may have led to the emergence of the Altai falcon haplotype. We looked at candidate genes under selection that may have contributed to the overall speciation trajectory of peregrines from the common ancestor with hierofalcons. Peregrines are a well-resolved species with a postzygotic reproductive barrier with hierofalcons. Out of 186 W-linked protein-coding genes, 90 genes contained 115 nonsynonymous variants in the peregrine W with predicted moderate (e.g. missense) to high (e.g. nonsense, stop-gain) impact (Fig. 5B and Table S7). These genes are associated with a range of phenotypes that could have contributed to the development of the peregrine as a separate species. For example, *SPINW*, a gene transcribed most prominently in ovarian granulosa and theca cells, that has been implicated in sexual differentiation in female chickens[80–82], appears to gain a stop codon in peregrines, affecting one of five transcripts of the gene (c.28C > T|p.Arg10*). The same mutation is also an intronic variant for the other transcripts. Another gene with a missense variant affecting both of its transcripts in peregrines (allele frequency 3/3) is *SMAD7* (c.148A > G|p.Ser50Gly). In female chickens, *SPINW* and *SMAD7* were reported to be upregulated in feather-forming tissues[83], which is thought to contribute to their plumage sexual dimorphism. In contrast, sexual dimorphism in large falcons is restricted to size, where plumage is identical in both sexes, suggesting that large falcons may have lost the plumage-sexual dimorphism pathway, despite having genes involved in this pathway on the W chromosome. *THSD1*, an endocardial gene overexpressed in hearts in humans also has an alternative stop codon (c.1483 T > C|p.Ter495Arg) in the peregrine. Other genes including *PLIN2W*, *UBAP2LW*, *DCAF1*, *NEK3*, *HINT1W*, *ARRDC3W*, *ISCA1* and *PIAS2* have also been identified with high-impact variants (i.e. inactivating gene mutations) in the peregrine, along with the Altai-W haplotype in the case of the *LMAN1W* and *CHD1W* genes, that shifted the reading frame and introduced an alternative stop codon (Fig. 5B and Table S7). Notably, *HINT1W* is expressed in the developing gonads of female embryos at the critical time of sexual differentiation and has been suggested to play a role in avian sex determination[84].

Our analysis of the Altai region falcon W-specific variants revealed that 29 W-linked genes contain 31 variants of identical or similar impact as the peregrines (Fig. 5B and Table S7). This suggests a shared W haplotype ancestry. For example, both the Altai W-haplotype and peregrines share a fixed missense mutation (c.118C > T, p.Pro40Ser) in the only transcript of *ZNF462*, a gene associated with embryonic development and craniofacial and neurodevelopmental abnormalities in humans[85] and abnormal behavior in mice[86]. Another frameshift mutation (p.Gln771fs) in *CHD1W is* also shared between all the Altai-haplotype carriers and peregrine-W carriers with a predicted high impact (Table S7). Heterozygous missense variants in *CHD1* have been associated with developmental and facial structural changes[87]. Speciation in falcons has also included gains in body size. Younger clades, including the largest and one of the youngest falcons, gyrfalcon, tend to be bigger and heavier compared to older clades such as the kestrels and the merlins. This could be linked to the adaptation to new expansive ecologies and a wide range of prey, including larger ones such as the willow grouse (*Lagopus lagopus*) and ptarmigan (*L. mutus*)[88]. It is then expected that genes that are associated with body size, physiology, reproduction, and behavior to be under adaptive selection in a speciation trajectory[89]. Indeed, *NTRK2*, *WASH3C*, *PLIN2W*, *PIAS2* and *MEX3C* genes are predicted to be impacted with high and/or moderate impact variants in peregrines and Altai W-haplotype carriers (Fig. 5B and Table S7). These genes are associated with vital structural, reproductive, and physiological traits. For example, *WASHC3*, a gene associated with bill size in black-bellied seedcracker[90], shares a missense variant between Altai-haplotype carriers and peregrine-W carriers (Table S7). Another shared missense variant was found within one of *NTRK2* transcripts*,* a gene that is associated with the development of ovaries in mice[91] and litter size in sheep[92]. Missense variants with predicted moderate impact were also identified in multiple ERVs, mainly *NYNRIN* and *NYNRIN-like*, which may be associated with the female reproductive physiology and potentially the development of reproductive isolation[30,76]. Both copies of *PLIN2,* a gene associated with fat deposition in Peking ducks[93], chicken[94,95], and emu[96] are also impacted with a missense variant that is shared between Altai-haplotype carriers and peregrines. Moreover, the second *PLIN2* copy has a premature stop-gain variant (p.Leu437*) that is private to the peregrine W (Table S7). Perilipins, including *PLIN2*, also appear to play an important role in the physiological preparations for migration, including fat metabolism, in gray catbirds[97]. Another gene, *MEX3C*, which is also linked to fat deposition and muscle development and whose function appears to be highly conserved in multiple species[98–100] is also predicted to be impacted in both peregrines and Altai-haplotype carriers (Table S7). *PIAS2* is reportedly associated with the immune response to viral infections in ducks[101]. In *PIAS2* a missense variant was identified in Altai-haplotype carriers, and a nonsense mutation (stop-loss) was fixed in peregrines (Table S7). This gene being W-linked, and with variability among different species of falcons may contribute to a female-specific immunity. Significant sex-based variability in the immune system is also observed in many other species of birds[102–104]. As apex predators, adaptive immunity plays an important role in the expansion into new niches, where potential prey may harbor novel pathogens. While the functional impact of these coding variants cannot be conclusively determined from genomic data alone, the pathways and traits with which they are associated offer insights into a possible speciation process. In general, the accumulation of mutations, under selective pressure, in genes linked to such critical traits could lead to incompatibilities with closely related populations, which would eventually create reproductive barriers, and eventually the possible emergence of a new species[89,105]. Our analysis also revealed that gyrfalcons and sakers share the same derived alleles in these loci. This demonstrates that the W chromosome is functionally conserved between these two species of falcons.

Discussion

We used the high-quality chromosome-level genome of a falcon to help resolve the 200-year-old debate on the taxonomy of Mongolian falcons. As evident by the full assembly of the W chromosome and the mitochondrial genome, we have identified a distinct haplotype that seems to be predominant among Mongolian saker-like falcons. This Altai haplotype, along with the lanners, diverged more than 420,000 years from the common ancestor of the saker and the gyrfalcon which, in turn, forms a younger node that split around 100,000 years ago. The lanner falcon's current distribution range is limited to Africa, Southern Europe, and fragmented and scattered points in the Middle East[106–109]. Finding the lanner's maternally-inherited haplotype in the Altai population of falcons in Mongolia, which shares minimal physical resemblance with lanner, and with the lack of range contiguity to connect the two populations reinforces the idea of a complex demographic history of falcons. Moreover, we detected consistently similar patterns of autosomal admixture in Altai saker-like falcons, caused by a gene flow from both sakers and gyrfalcons. Both of these results suggest multiple hybridization events. The first event would be a lanner-like population (as evident by the W haplotype), hybridizing with the gyrfalcon-saker common ancestor. This would then be followed by a second event of consistent and repeated hybridization with sakers, which may explain the significantly higher saker-like ancestry, and saker-like Z chromosome detected in Altai saker-like falcons. If the scenario of the hybridization events were in reverse order, the gyrfalcons would have donated their mitochondrial and W haplotype along with their autosomal markers, a conclusion that our results do not support in the case of the Altai saker-like population. However, this reverse-order scenario seems to fit the admixture data of Altai gyrfalcon-like individuals. This alludes to the origin of gyrfalcon-like falcons as being indeed a gyrfalcon population that was introgressed by sakers, but maintained its gyrfalcon paternally- and maternally-inherited haplotypes, separating them from the Altai region saker-like falcons.

The results show that Mongolia appears to be a hybridization hotspot for falcons. With a larger sample size we expect to detect admixed populations with maternal haplotypes of gyrfalcons, sakers and Altai region falcons. We would also expect to find autosomally admixed individuals with different ancestral proportions, reflecting both old and new hybridizations. Our results offer a glimpse of that, with one saker individual carrying an Altai maternal haplotype, and two Altai saker-like falcons carrying saker haplotypes. We also observed one Altai gyrfalcon-like individual (SP50) with nearly 50% saker ancestry (Fig. 1E, K = 3), despite morphologically resembling gyrfalcons and carrying a gyrfalcon W haplotype (Fig. 1G and Fig. 3B). All things considered, however, these observations do not change the fact that there are at least three maternally distinct, yet interbreeding populations of falcons in Mongolia; the gyrfalcons, the sakers, and the distinct Altai-haplotype falcons. Morphological appearance alone does not provide conclusive evidence of the taxonomy of these closely related falcon populations. This further highlights the importance of considering the autosomal admixture patterns along with the maternally-inherited haplotypes in discerning the taxonomy of Altai region falcons. While the designation of the gyrfalcons (including GLF) and sakers species is in line with the maternally-inherited haplotypes they carry, the Altai W-haplotype carrying falcons may also be considered for a species status.

Proving homoploid hybrid speciation in animals is challenging, but with the aid of whole genome data, cases in birds and fish have been reported, where a hybrid lineage arises from the interbreeding of two divergent species[110,111]. In the case of the genomically mosaic Altai region falcons, the W chromosome appears to also play an important role in speciation, especially due to its maternally inherited nature. By looking at the divergence of the peregrine W chromosome from the gyrfalcon's and saker's, we gained an insight into some of the genes that underwent selection, and hence the physiological, morphological, and reproductive processes that contributed to the peregrine's speciation trajectory. We then compared these W-linked genes to the Altai W chromosome to assess how functionally divergent it is from the older peregrine W haplotype and younger gyrfalcon-saker W haplotype. This potential could only be utilized by examining a fully-assembled W chromosome, instead of relying solely on mitochondrial analysis. Our results showed that not only does the Altai W chromosome share some of the functional mutations with the peregrine, but also that the impacted genes appear to be associated with vital pathways that could play a potential role in speciation. Considering the shared W-linked variants between the Altai and the peregrine, based on the D-statistics, phylogenomics, and the closer examination of the shared functional variants, we find that Altai W is an older lineage and resembles the ancestral hierofalcon's W haplotype. On the other hand, the saker and gyrfalcon, which split later, and share a functionally similar W haplotype, are shown to accumulate advantageous derived alleles in multiple W-linked genes. This demonstrates the impact of the female-biased gene flow on the trajectory of Altai falcon development as a population.

The samples sequenced in our study attempt to capture the diversity among hierofalcons, and mainly the Altai falcons. As with any other protected and endangered species, acquiring high-quality blood samples that are suitable for high-throughput sequencing from five species and populations of falcons from multiple countries has proved to be time- and resource-consuming. However, our ongoing analysis will expand in the future to increase the sample size from additional geographic regions, including the lugger falcon and additional lanner falcons. Despite these limitations, our results suggest that the currently accepted falcon taxonomic classifications, mainly conflating the Altai falcons with sakers, may warrant revision. If both gyrfalcons and sakers maintain their status as two distinct species and conservation units, despite sharing highly similar sex-chromosome haplotypes and producing fertile offspring[6,43], then a fortiori the taxonomic and conservation status of Altai falcons may be considered as well. The Altai falcons appear to be a mosaic of multiple ancestries. They are predominantly of autosomally admixed origin, with generally consistent saker and gyrfalcon ancestral proportions. They also carry a W haplotype that clusters with the lanner falcon and is separate from the ancestor of the gyrfalcon and the saker. Any autosomal traces of the lanner-like ancestry seem to be lost to the multiple subsequent admixture events with sakers and the gyrfalcons. Moreover, not only does the Altai W-haplotype split 0.42 MYA, but it also contains coding variants with implications on reproductive and structural traits.

Conclusion

Since the suggested species status of Altai falcons is based on phylogenetic and functional bases, it offers an essential foundation to inform conservation and management efforts. The focus of falcon conservation resources and projects currently ongoing in Mongolia, where Altai falcons reside, may need to be re-evaluated based on the current population trends of the Altai falcon, which is currently conflated with the saker falcon. The sakers, whose range overlaps with the Altai region falcons and extends as far as Western Europe, may warrant their own separate conservation programs. We hope that our work will facilitate and enable a focused attention towards Altai falcons (*F. altaicus*) as a distinct unit of conservation, separate from sakers and gyrfalcons. Such designation of Altai falcons as a distinct unit of conservation should help inform and guide the ongoing conservation efforts of falcons in Asia, and maintain the genetic identity of Altai falcons by preventing accidental introduction of Central Asian saker falcons, via captive-breeding, into its ranges. One of the main implications of our results is the power of using high-quality reference genomes for population genomics, in addressing long-debated questions in biology, taxonomy, ecology, and conservation. The success of approaches, including breed-and-release, is critically dependent on taxonomic classifications, hence the need for a robust approach to delineating phylogenies. Defining units of conservation is the cornerstone of any conservation endeavor, without which unpredicted and sometimes harmful consequences can happen. The current work demonstrates the importance of expanding the use of W chromosomes beyond phylogeny, to include fingerprints of divergence and its vast implications in non-model species for conservation purposes. With the current mass extinction of species underway, a comprehensive and robust approach should be followed to determine the taxonomic status of populations, which in turn helps prioritize the conservation resources[112].

Materials and methods

Sampling and animal ethics

A total of 45 falcons were sequenced (Table 3). These included: 14 gyrfalcons from different populations and color morphs, eight sakers; 12 Altai falcons (originating from western Mongolia, and based on morphological traits designated as Altai-type falcons); one lanner falcon; and 10 commercial hybrids (eight gyrfalcon-saker individuals (including one purported hybrid with no confirmed pedigree) and two gyrfalcon-peregrine hybrids). The Canadian and Alaskan gyrfalcon individuals ($n = 13$) were captive-bred using pure-bred lineages, without intraspecific hybridization, i.e. crossing with other gyrfalcon populations; one individual (GM01) was labeled North American as its intra-specific captive-bred lineage was not precisely determined (Table 3). All the gyrfalcons had previously been exported to Qatar for breeding in captive-breeding facilities where they later were sampled. The sakers and the lanner were sampled in Qatar, while the Altai region falcons ($n = 12$) had been previously exported from Mongolia (classified as sakers following the accepted taxonomy) to Qatar, where they were sampled. The latter resembled the morphology of two distinct populations of "Altai-type falcons"; gyrfalcon-like and saker-like[12]. The commercial captive-bred hybrid falcons and their ancestry and filial generation were acquired from the breeders. Blood samples (< 1 ml, each) from each bird were collected in 1.5 ml tubes containing potassium EDTA anticoagulant, at the Souq Waqif Falcon Hospital, Doha, Qatar. Sample collection was carried out under the official approval of the Ministry of Municipality and Environment in Qatar (Reference No. 2017/283,748) and by trained veterinarians at the Falcon Hospital. The blood samples were only taken during routine diagnostics from falcons being checked for purposes unrelated to the research activity and the researcher had no influence on the timing or location of the medical check-up. These blood samples would have normally been discarded if the researcher had not asked for them. The Monash University Animal Ethics Committee (AEC) advised that such activity, considered as scavenging, did not require animal ethics approval.

DNA extraction, library preparation and resequencing

High molecular weight genomic DNA (HMW gDNA) was extracted from all of the blood samples that were collected (Table 3) using Qiagen's MagAttract Kit (Qiagen, Germany). Extracted DNA was quantified and its quality and integrity were assessed by using agarose gel electrophoresis and Qubit® 3.0 Fluorometer (Thermo

Sample ID	Species name	Common Name	Sex	Color	Provenance
North American Gyrfalcons (GF)					
SP19*	*F. rusticolus*	Gyrfalcon	F	White	Alaskan (CB)
SP6	*F. rusticolus*	Gyrfalcon	F	Black	Canadian (CB)
SP7	*F. rusticolus*	Gyrfalcon	F	Black	Canadian (CB)
SP8	*F. rusticolus*	Gyrfalcon	F	Black	Canadian (CB)
SP9	*F. rusticolus*	Gyrfalcon	F	Black	Canadian (CB)
SP10	*F. rusticolus*	Gyrfalcon	F	Black	Canadian (CB)
SP13	*F. rusticolus*	Gyrfalcon	**M**	White	Alaskan (CB)
SP14	*F. rusticolus*	Gyrfalcon	**M**	White	Alaskan (CB)
SP15	*F. rusticolus*	Gyrfalcon	F	White	Alaskan (CB)
SP16	*F. rusticolus*	Gyrfalcon	F	White	Alaskan (CB)
SP17	*F. rusticolus*	Gyrfalcon	F	White	Alaskan (CB)
SP18	*F. rusticolus*	Gyrfalcon	F	White	Alaskan (CB)
SP20	*F. rusticolus*	Gyrfalcon	F	White	Alaskan (CB)
GM01	*F. rusticolus*	Gyrfalcon	**M**	White	North American (CB)
Saker Falcons (SF)					
CB12	*F. cherrug*	Saker	F	Light brown	Qatar (W)
SP32	*F. cherrug*	Saker	F	Brown	Qatar (W)
SP33	*F. cherrug*	Saker	F	Brown	Qatar (W)
SP36	*F. cherrug*	Saker	F	Brown	Qatar (W)
SP37	*F. cherrug*	Saker	F	Brown	Qatar (W)
SP39	*F. cherrug*	Saker	F	Brown	Qatar (W)
SP40	*F. cherrug*	Saker	F	Brown	Qatar (W)
SP51	*F. cherrug*	Saker	F	Dark Brown	Qatar (W)
Saker-like Falcons from the Altai Region (SLF)					
CB8	*F. cherrug (?)*	Saker-like	F	Light brown	Mongolian (W)
CB11	*F. cherrug (?)*	Saker-like	F	Silver	Mongolian (W)
SP30	*F. cherrug (?)*	Saker-like	F	Black	Mongolian (W)
SP34	*F. cherrug (?)*	Saker-like	F	Black	Mongolian (W)
SP35	*F. cherrug (?)*	Saker-like	F	Brown	Mongolian (W)
SP38	*F. cherrug (?)*	Saker-like	F	Grey	Mongolian (W)
SP42	*F. cherrug (?)*	Saker-like	F	Grey	Mongolian (W)
SP52	*F. cherrug (?)*	Saker-like	F	Black	Mongolian (W)
Gyrfalcon-like Falcons from the Altai Region (GLF)					
SP50	*F. rusticolus (?)*	Gyrfalcon-like	F	Grey/Silver	Mongolian (W)
SP21	*F. rusticolus (?)*	Gyrfalcon-like	F	Brown	Mongolian (W)
SP47	*F. rusticolus (?)*	Gyrfalcon-like	F	Silver	Mongolian (W)
QFGP14	*F. rusticolus (?)*	Gyrfalcon-like	F	Silver	Mongolian (W)
Commercial Hybrids (CH)					
SP11	*Hybrid*	F1 Gyrfalcon-Saker	**M**	Black	Hybrid (CB)
SP12	*Hybrid*	F4 Gyrfalcon-Saker	**M**	Black	Hybrid (CB)
SP44	*Hybrid*	F4 Gyrfalcon-Saker	F	White	Hybrid (CB)
SP45	*Hybrid*	F4 Gyrfalcon-Saker	**M**	Black	Hybrid (CB)
SP46	*Hybrid*	F5 Gyrfalcon-Saker	F	Sliver	Hybrid (CB)
SP48	*Hybrid*	F4 Gyrfalcon-Saker	F	white	Hybrid (CB)
SP49	*Hybrid*	F3 Gyrfalcon-Saker	F	White	Hybrid (CB)
SP43	*Hybrid*	F1 Gyrfalcon-Peregrine	F	White	Hybrid (CB)
SP53	*Hybrid*	F1 Gyrfalcon-Peregrine	F	Black	Hybrid (CB)
CB5	*Hybrid (?)*	Saker-Gyrfalcon (?)	F	Light brown	Unknown (CB)
Lanner falcon					
Lanner	*F. biarmicus*	Lanner	F	Wildtype	Qatar

Table 3. A list of 45 falcons sequenced in this study. The table shows sample ID, species name, sex, plumage color as well as provenance. W refers to an individual of a wild origin, while CB refers to a captive-bred individual. (?) after the species name refers to taxonomically-ambiguous individuals, i.e. Altai-type falcons. The reference specimen is marked with *.

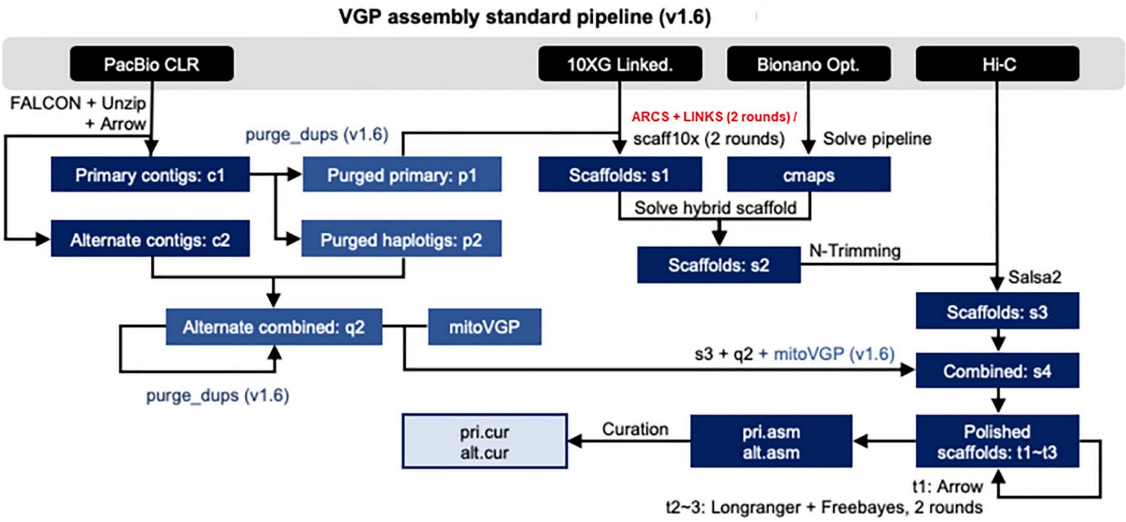

Fig. 6. VGP standard assembly pipeline 1.6, with the modified step of ARCS + LINKS in parallel with scaff10x, used in this project. Abbreviations: c = contigs; p = purged false duplications from primary contigs; q = purged alternate contigs; s = scaffolds; t = polished scaffolds. Further details on the pipeline and instructions are available at https://github.com/VGP/vgp-assembly. In brief, the raw data from PacBio sequencing CLR technologies are assembled to contigs using FALCON + Unzip, followed by polishing using Arrow to increase the accuracy of the base calling. The primary contigs refer to the full genome, whereas the alternate refers to detected haplotypes within the diploid genome. The primary contigs undergo purging to remove artificial duplications within the assemblies. This is followed up by scaffolding using 10 × Genomics linked-reads. The resultant scaffolded assembly is further upgraded to enhance the completeness and contiguity by integrating Bionano optical maps followed by Hi-C data to generate chromosome-level assembly. The resultant assembly is then polished, and gap-filled using the 10 × Genomics data. The assembly that is generated was submitted for manual curation and further contiguity enhancement based on the Hi-C data.

Fisher Scientific, USA). For each sample, ~ 1 µg of gDNA was used for preparing an indexed library for sequencing using the NEBNext® Ultra™ DNA Library Prep Kit from Illumina® (NEB, USA) following the manufacturer's protocol. The gDNA was randomly fragmented to a size of 350 bp by mechanical shearing. Subsequently, the DNA fragments were end-polished, A-tailed, and ligated with the NEBNext adapter for Illumina sequencing, and further PCR enriched by P5 and indexed P7 oligos. The PCR products were purified (AMPure XP system) and the resulting libraries were analyzed for size distribution on an Agilent 2200 TapeStation System (Agilent, USA). DNA resequencing was carried out for all the samples (Table 3) at the Sidra Medicine, Doha, Qatar, using Illumina HiSeq X platform at an 30 × average coverage with read length of 2 × 150 bp (PE). Raw sequencing data were filtered and assessed for quality using FastQC 0.11.8 and Trimmomatic 0.39[113]. The lanner DNA was sequenced using PacBio Sequel II system at a coverage of 25X at the Vertebrate Genome Laboratory (VGL), The Rockefeller University, New York, USA.

Reference genome sequencing

High-coverage DNA sequencing was done for a gyrfalcon (SP19) sample. The reference sample was also sequenced using Illumina HiSeq 4000 (150 bp paired-end) at a 101.2 × coverage at Novogene Genomics, Singapore. For long reads, PacBio Sequel System v5.0, using chemistry v2.1 was done at a 51.7 × coverage, in addition to Bionano Optical Mapping at a coverage of 246.43x (DLE-1 one enzyme non-nicking approach using the Bionano Saphyr instrument); and 10 × Genomics libraries at 125.8 × coverage , all at the Vertebrate Genome Laboratory (VGL), The Rockefeller University, New York, USA. Moreover, High-throughput chromosome conformation capture (Hi-C) was performed at Arima Genomics, USA at 98.9 × coverage. All raw data, and their statistics were uploaded to VGP's GenomeArk repository for raw data and assemblies (https://www.genomeark.org/genomeark-all/Falco_rusticolus.html).

Reference genome assembly and annotation

High-quality reference genome assembly was generated using VGP methodology[20]. For PacBio's long reads, Canu 1.8[114] and pb-assembly 0.0.6 (FALCON 1.3.0 and FALCON_unzip 1.2.0)[115] were used to generate a phased diploid and polished draft assembly. Thirty iterations of PacBio draft assemblies were generated to optimize the contiguity and phasing and decrease duplications, guided by the results of Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis (Fig. S1). The 10 × Genomics data was processed using the Longranger 2.2.2 software and subsequently employed to refine and scaffold the PacBio draft assembly using ARCS + LINKS Pipeline[116,117] (Fig. 6). To further optimize the scaffolding and resolve misassemblies, Bionano optical mapping data were integrated using Bionano Solve v3.2.1. Hi-C data were used to further scaffold the assembly to chromosome level using the Arima Genomics mapping pipeline (https://github.com/ArimaGenomics/mapping_pipeline) followed by Salsa2[118]. Finally, to improve the base quality of the assembly, two polishing methods were utilized: First, using PacBio long reads, polishing was conducted using Arrow (included in SMRT Link v.7). In Arrow, a range of coverage cutoff values were investigated to optimize the polishing efficiency, and the best was found to be—×5. Second, further polishing steps used 10 × Genomics short linked-reads; two rounds of short read polishing were performed using Longranger align 2.2.2, FreeBayes[119], and BCFtools consensus[120], according to VGP 1.6 pipeline[20] (Fig. 6). To confirm the quality and structure of the chromosome-level assembly, a Hi-C map was generated using PretextView 0.1.2. Manual curation was then implemented to detect any anomalies that could have resulted from the automated scaffolding process, as described by the VGP assembly protocol (Fig. 6)[20]. Briefly, manual curation involved mapping the Hi-C raw data to the draft primary assembly. This, along with Bionano maps, PacBio CLR long-read data, and gene alignments, guided the correction of any mis-joints, missed joins, and other misassemblies. The reference gyrfalcon specimen was selected to be a female to allow the capture of the W chromosome. Manual curation identified the Z and W chromosomes based on half sequencing coverage, homology alignments to sex chromosomes in other species, including the chicken, and the presence of avian sex chromosome-specific genes. At every step of the assembly pipeline quality assessment statistics were generated including contig N50 and scaffold N50, the quality value (QV), and genome completeness using BUSCO v3 software[121]. BUSCO provides quantitative measures for the assessment of genome assembly using a database of 2,586 near-universal single-copy vertebrate orthologs, included in the vertebrata_odb9 database. QV analysis was estimated using Illumina short reads and 10 × Genomics short reads using Merqury[122].

Draft annotation and gene prediction of the assembled genome was carried out using three ab initio pipelines: Augustus 3.3.2[123], using chicken (*Gallus gallus*) as the model avian species, after identifying and masking repetitive DNA using RepeatMasker 4.0.9[124]. Functional annotation was done using Blast2GO[125], using integrated InterProScan and non-redundant protein sequences NCBI Blast (nr v5) databases. The transcriptomic database from NCBI, which included data from the *Falco* genus was used to validate the gene annotation as it is more comprehensive than other available approaches. A functional annotation of the gyrfalcon annotation using Gene Ontology (GO) enrichment analysis was performed to evaluate the completeness of the draft assembly and the annotation. The genome was screened for TE using RepeatMasker (version 4.1.5) with default parameters and rmblastn (version 2.14.1 +), comparing the query sequences against the curated Passerellidae TE library to identify and classify transposable elements[42]. The reference-quality, manually curated, and annotated assembly for the gyrfalcon is available as a RefSeq at NCBI (GCF_015220135.1).

Genetic variation, admixture, and population structure

The paired-end short reads for the additional resequenced samples were mapped to the reference assembly of the gyrfalcon generated in this study using BWA-MEM alignment software[126]. Publicly available WGS datasets of a saker (PRJNA168071), two peregrines (PRJNA159791, PRJDB7811) and one lanner (PRJNA802108) were retrieved from NCBI and added to the analyses for comparison and validation. The mapped reads were then sorted using SAMtools 1.9. SNPs were then filtered (Phred quality score ≥ 20 and a minimum depth of 10 reads) and called using FreeBayes, VCFtools and BCFtools. The generated VCF file was split into three representing autosomal-, W- and Z-linked datasets. For admixture and population structure analyses, the filtered SNPs were then pruned to remove the SNPs that were in linkage or close proximity using an r^2 cut-off of 0.1. The QC process also involved filtering out SNPs with: a) very low minor allele frequencies (< 0.05); b) autosomal SNPs with frequencies that deviate significantly from Hardy–Weinberg Equilibrium (p-value < 10–6 threshold); c) excess missing genotypes (SNPs with more than 99.9% of genotypes missing), and d) low confidence call score (< Q30) score. The QC filtration was done using PLINK software v.1.9[127] using the following command lines:

```
plink –vcf $FILE –make-bed  –double-id –maf 0.05 –set-missing-var-ids @:# –hardy –hwe 1e-6 –vcf-min-qual 30 –indep-pairwise 50 10 0.1 –threads 32 –geno 0.999 –out pg.$FILE –allow-extra-chr.
```

```
plink –vcf $FILE –make-bed  –double-id –set-missing-var-ids @:# –extract pg.$FILE.prune.in –maf 0.05 –hardy –hwe 1e-6 –vcf-min-qual 30 –threads 32 –geno 0.99 –out pg.$FILE –allow-extra-chr.
```

Using the PLINK software v.1.9, a Principal component analysis (PCA) was conducted to represent the variation in the populations and explore the population substructure across the autosomes and sex chromosomes. The same dataset was also used to estimate the likelihood of individual ancestries using ADMIXTURE 1.3[128]. The results were visualized using ggplot2 package in RStudio Cloud software (https://rstudio.cloud/). PLINK v1.9 was also used to estimate the runs of homozygousity (ROH) with –homozyg command using the default parameters. PLINK v2 was used to calculate pairwise FST for GF, GLF, SF, and SLF populations using Hudson's and Weir and Cockerham's methods. (Fig. S2).

To account for genetic drift and gene flow, TreeMix 1.13 was used, which infers patterns of population split using allele frequency[129]. TreeMix was also used to calculate the F-statistics (F3 and F4) of all populations to estimate the mixing proportions of the admixture events. The D-statistic, also known as BABA-ABBA test, and its related f4-ratio ancestry proportion estimator was calculated for all populations using ADMIXTOOLS[130] through the *admixr* package version 0.7.1[131] in RStudio[132].

The distribution ranges of sakers and gyrfalcons were plotted on a global map in RStudio using data available at Map of Life[133,134].

Mitochondrial genome assembly

The mitochondrial reference genome for the gyrfalcon was assembled using NOVOPlasty[135], by using the chicken mitogenome (NCBI accession No. KM433666.1) as a reference and a seed, and a k-value of 71. In addition, whole-genome sequence datasets were also used to assemble the full mitogenomes of 34 falcons from five populations using NOVOPlasty. Five mitogenomes (SP50, SP17, SP14, SP37 and one lanner (PRJNA802108) failed to assemble fully, possibly due to the relatively low sequencing coverage and high computational requirements.

Phylogenetic analysis and demographic history

A time-calibrated phylogenetic tree was generated using the 35 falcon mitochondrial genomes assembled in this study, in addition to seven falcon mitogenomes retrieved from NCBI: common kestrel (NC_011307.1), American kestrel (NC_008547.1), merlin (KM264304.1), saker (NC_026715.1), peregrine 1 (NC_000878.1), peregrine 2 (JQ282801.1) and gyrfalcon (KT989235.1). The chicken mitogenome (NCBI Accession KT626858.1) was used as an outgroup. Following common practice, we excluded the control region (CR) due to its high variability, which could confound the phylogenetic signal[136–138]. After aligning the sequences with Clustal Omega v1.2[139], falcon species divergence times were estimated on BEAST v2.6.2[140] using StarBEAST2 module[141]. We assumed a strict molecular clock, integrated analytical population size, and an HKY (four categories) nucleotide substitution model. For the priors, we selected the Calibrated Yule speciation model, and calibrated the molecular clock by assuming the age of divergence between Falconiformes and chicken 88 ± 10 MYA[142]. Markov chain Monte Carlo chains (MCMC) were run for 60 million states, discarding the first 10% as burn-in. Convergence was inspected using Tracer v1.6, and a maximum clade credibility tree (chronogram) was built using TreeAnnotator v2.6.2[140].

A maximum-likelihood phylogenetic tree was constructed using an alignment of W-specific variants (3,007 SNPs) in 39 female falcons (38 of them were seqeunced in this study) using W-IQ-TREE[143,144]. The tree was constructed using K2P + R2, which was the best-fit model according to BIC using ModelFinder[145]. Branch labels represent SH-aLRT support (%), aBayes support, and ultrafast bootstrap support (%)[146], respectively.

To reconstruct the population histories of the falcon species, we used the Pairwise Sequentially Markovian Coalescent (PSMC) model, which uses diploid sequences to infer piecewise-constant population size histories as a function of time[147]. The raw sequencing reads were mapped to the reference, and heterozygous SNPs were called using SAMtools mpileup as in: samtools mpileup -C50 -uf ref.fa sample.bam | bcftools call -c | vcfutils.pl vcf2fq -d 10 -D 100 | gzip > sample.fq.gz. Autosomal SNPs were used to generate the PSMC files using the following command:

```
/utils/fq2psmcfa -q20 $name.fq.gz > $name.psmcfa.
psmc -N25 -t15 -r5 -p "4 + 25*2 + 4 + 6" -o $name.psmc $name.psmcfa.
```

Finally, the PSMC files were plotted using the saker's generation time of 6.6 years/generation according to previous estimates[18]. A genomic mutation rate of 4.6×10^{-9} per base per generation was used. This rate was obtained in a germline-based study for the collared flycatcher *Ficedula albicollis*[148] and has since been used as a reliable estimate for other birds as well[149,150].

Identification of candidate genes under selection and enrichment analysis

To identify candidate regions under selection, Hudson's *F*ST[151] method was used. Hudson's *F*ST was selected as it is independent of sample size, compared to other *F*ST calculation methods, including the classical Weir and Cockerham[152]. *F*ST and π were computed using the script popgenWindows.py (github.com/simonhmartin/genomics_general) with a sliding window of 20,000 bp and a minimum of 100 genotyped sites. Only windows corresponding to the upper 0.1% of the empirical genome-wide distribution of *F*ST were considered as high-*F*ST outliers and labeled as candidate regions. The aim was to capture the gradual and subtle differences between the falcon populations, and highlight the highly differentiated genomic regions that could probably explain the adaptive phenotypes involved in their respective habitats and the rate at which they are diverging. SnpEff was used to annotate the variants within these genes, especially W-linked, and predict their impact (HIGH, MODERATE, LOW, and MODIFIER). Variants with predicted high or moderate impact were then investigated in depth and validated manually by examining the VCF file.

W-linked genes in the gyrfalcon were analyzed for functional roles using The PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System version 17.0[153].

Data availability

All raw data of the reference genome assembly are available on VGP's GenomeArk repository for raw data and assemblies (https://www.genomeark.org/genomeark-all/Falco_rusticolus.html) and umbrella NCBI BioProject PRJNA562209. The sequence data is available under NCBI BioProject PRJNA1008133 . The primary assembly and the annotation data are available at NCBI BioProject PRJNA675116. The contig-level alternate assembly is available at NCBI BioProject PRJNA561989. The sequencing data of *Falco biarmicus* is available at VGP's GenomeArk (https://www.genomeark.org/genomeark-all/Falco_biarmicus.html) and NCBI BioProject PRJNA100 8134. The resequencing data is available from the authors upon reasonable request.

Received: 22 May 2024; Accepted: 24 January 2025
Published online: 04 February 2025

References

1. Pallas, P. S. & Tilesius von T., W. G. *Zoographia Rosso-Asiatica : Sistens Omnium Animalium in Extenso Imperio Rossico, et Adjacentibus Maribus Observatorum Recensionem, Domicilia, Mores et Descriptiones, Anatomen Atque Icones Plurimorum*. 1–588 (In officina Caes. Acadamiae Scientiarum Impress. MDCCCXI, Petropoli :, 1811). https://doi.org/10.5962/bhl.title.42222.
2. Moseikin, V. & Ellis, D. Ecological aspects of distribution for saker falcons Falco cherrug and Altai gyrfalcon F. altaicus in the Russian Altai. *Raptors Worldw.* 693–703 (2004).
3. Fox, N. & POTAPOV, E. Altai Falcon: subspecies, hybrid or colour morph. in *Proceedings of 4th Eurasian Congress on Raptors, Seville, Spain, 25–29 September 2001, Abstracts* 66–67 (2001).
4. Potapov, E. & Sale, R. *The Gyrfalcon* (Yale University Press, 2005).
5. Sundev, G. & Leahy, C. *Birds of Mongolia*. (Bloomsbury Publishing, 2019).
6. Cade, T. J. The Gyrfalcon. *Auk* **123**, 920–923 (2006).
7. Ellis, D. H. The Altay falcon: Origin, morphology and distribution. in *Proceedings of the Specialist Workshop, November 14–16, 1995* 143–168 Middle East Falcon Research Group, Abu Dhabi, United Arab Emirates, (1995).

8. Ayé, R., Schweizer, M. & Roth, T. *Birds of Central Asia*. Bloomsbury Publishing, (2020).
9. Sushkin, P. P. *Birds of the Soviet Altai and adjacent parts of north-western Mongolia* (Academy of Science of USSR Press, 1938).
10. Ferguson-Lees, J. & Christie, D. A. *Raptors of the World: A Field Guide*. (Bloomsbury Publishing, 2020).
11. Eastham, C. P. Morphological studies of taxonomy of the saker (Falco cherrug - Gray 1833) and closely allied species. (University of Kent at Canterbury, 2000).
12. Ellis, D. H. What is falco altaicus menzbier?. *J. Raptor Res.* **29**, 11 (1995).
13. Yan, F. et al. The Chinese giant salamander exemplifies the hidden extinction of cryptic species. *Curr. Biol.* **28**, R590–R592 (2018).
14. Gippoliti, S., Cotterill, F. P. D., Groves, C. P. & Zinner, D. Poor taxonomy and genetic rescue are possible co-agents of silent extinction and biogeographic homogenization among ungulate mammals. *Biogeogr. J. Integr. Biogeogr.* https://doi.org/10.21426/B633039045 (2018).
15. Huang, S.-L. & Karczmarski, L. 2014 Indo-Pacific Humpback Dolphins: A Demographic Perspective of a Threatened Species. *Primates and Cetaceans: Field Research and Conservation of Complex Mammalian Societies* (eds. Yamagiwa, J. & Karczmarski, L.) Springer Japan, Tokyo 249–272 https://doi.org/10.1007/978-4-431-54523-1_13
16. Delić, T., Trontelj, P., Rendoš, M. & Fišer, C. The importance of naming cryptic species and the conservation of endemic subterranean amphipods. *Sci. Rep.* **7**, 3391 (2017).
17. Nittinger, F., Gamauf, A., Pinsker, W., Wink, M. & Haring, E. Phylogeography and population structure of the saker falcon (Falco cherrug) and the influence of hybridization: mitochondrial and microsatellite data. *Mol. Ecol.* **16**, 1497–1517 (2007).
18. Zhan, X. et al. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nat. Genet.* **45**, 563–566 (2013).
19. Doyle, J. M. et al. New insights into the phylogenetics and population structure of the prairie falcon (Falco mexicanus). *BMC Genom.* **19**, 233 (2018).
20. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
21. Lischer, H. E. L. & Shimizu, K. K. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinform.* https://doi.org/10.1186/s12859-017-1911-6 (2017).
22. Weissensteiner, M. H. et al. Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Res.* **27**, 697–708 (2017).
23. Christmas, M. J. et al. Chromosomal inversions associated with environmental adaptation in honeybees. *Mol. Ecol.* **28**, 1358–1374 (2019).
24. Wellenreuther, M., Mérot, C., Berdan, E. & Bernatchez, L. Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Mol. Ecol.* **28**, 1203–1209 (2019).
25. Weissensteiner, M. H. et al. Discovery and population genomics of structural variation in a songbird genus. *Nat. Commun.* **11**, 3403 (2020).
26. Trimble, W. L. et al. Short-read reading-frame predictors are not created equal: Sequence error causes loss of signal. *BMC Bioinform.* **13**, 183 (2012).
27. Li, M. et al. Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome Res.* **27**, 865–874 (2017).
28. Kim, J. et al. False gene and chromosome losses in genome assemblies caused by GC content variation and repeats. *Genome Biol.* **23**, 204 (2022).
29. Ko, B. J. et al. Widespread false gene gains caused by duplication errors in genome assemblies. *Genome Biol.* **23**, 205 (2022).
30. Smeds, L. et al. Evolutionary analysis of the female-specific avian W chromosome. *Nat. Commun.* **6**, 7330 (2015).
31. Bellott, D. W. et al. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nat. Genet.* **49**, 387–394 (2017).
32. Tomaszkiewicz, M., Medvedev, P. & Makova, K. D. Y and W chromosome assemblies: Approaches and discoveries. *Trends Genet. TIG* **33**, 266–282 (2017).
33. Deakin, J. E. et al. Chromosomics: Bridging the gap between genomes and chromosomes. *Genes* https://doi.org/10.3390/genes10080627 (2019).
34. Zuccolo, A. et al. The gyrfalcon (Falco rusticolus) genome. *G3 GenesGenomesGenetics*. https://doi.org/10.1093/g3journal/jkad001 (2023).
35. Howe, K. et al. Significantly improving the quality of genome assemblies through curation. *GigaScience* https://doi.org/10.1093/gigascience/giaa153 (2021).
36. Murigneux, V. et al. Comparison of long-read methods for sequencing and assembly of a plant genome. *GigaScience* https://doi.org/10.1093/gigascience/giaa146 (2020).
37. Wang, J. et al. Benchmarking multi-platform sequencing technologies for human genome assembly. *Bioinform. Brief.* **24**, bbad300 (2023).
38. Meng, Y. et al. Genome sequence assembly algorithms and misassembly identification methods. *Mol. Biol. Rep.* **49**, 11133–11148 (2022).
39. Wang, K. et al. African lungfish genome sheds light on the vertebrate water-to-land transition. *Cell* **184**, 1362-1376.e18 (2021).
40. Rutkowska, J., Lagisz, M. & Nakagawa, S. The long and the short of avian W chromosomes: no evidence for gradual W shortening. *Biol. Lett.* **8**, 636–638 (2012).
41. Graves, J. A. M. Avian sex, sex chromosomes, and dosage compensation in the age of genomics. *Chromosome. Res.* **22**, 45–57 (2014).
42. Benham, P. M. et al. Remarkably high repeat content in the genomes of sparrows: The importance of genome assembly completeness for transposable element discovery. *Genome Biol. Evol.* **16**, evae067 (2024).
43. Eastham, C. P. & Nicholls, M. K. Morphometric analysis of large Falco species and their hybrids with implications for conservation. *J. Raptor Res.* **39**, 386–393 (2005).
44. Zheng, Y. & Janke, A. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. *BMC Bioinform.* https://doi.org/10.1186/s40657-017-0088-z (2018).
45. Ottenburghs, J. et al. Avian introgression in the genomic era. *Avian Res.* **8**, 30 (2017).
46. Roquet, C., Lavergne, S. & Thuiller, W. One tree to link them all: a phylogenetic dataset for the European tetrapoda. *PLoS Curr* https://doi.org/10.1371/currents.tol.5102670fff8aa5c918e78f5592790e48 (2014).
47. Johnson, J. A., Brown, J. W., Fuchs, J. & Mindell, D. P. Multi-locus phylogenetic inference among New World Vultures (Aves: Cathartidae). *Mol. Phylogenet. Evol.* **105**, 193–199 (2016).
48. Nielsen, Ó. K. & Pétursson, G. Population fluctuations of gyrfalcon and rock ptarmigan: analysis of export figures from Iceland. *Wildl. Biol.* **1**, 65–71 (1995).
49. Robinson, B. W., Booms, T. L., Bechard, M. J. & Anderson, D. L. Dietary Plasticity in a Specialist Predator, the Gyrfalcon (Falco rusticolus): New Insights into Diet During Brood Rearing. *J. Raptor Res.* **53**, 115–126 (2019).
50. Cahill, J. A., Soares, A. E. R., Green, R. E. & Shapiro, B. Inferring species divergence times using pairwise sequential Markovian coalescent modelling and low-coverage genomic data. *Philos. Trans. R. Soc. B Biol. Sci.* https://doi.org/10.1098/rstb.2015.0138 (2016).
51. Mather, N., Traves, S. M. & Ho, S. Y. W. A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. *Ecol. Evol.* **10**, 579–589 (2020).

52. Fuchs, J., Johnson, J. A. & Mindell, D. P. Rapid diversification of falcons (Aves: Falconidae) due to expansion of open habitats in the Late Miocene. *Mol. Phylogenet. Evol.* **82**, 166–182 (2015).
53. Zheng, W. et al. Large-scale genome sequencing redefines the genetic footprints of high-altitude adaptation in Tibetans. *Genome Biol.* **24**, 73 (2023).
54. Chen, B., Li, D., Ran, B., Zhang, P. & Wang, T. Key miRNAs and genes in the high-altitude adaptation of tibetan chickens. *Front. Vet. Sci* https://doi.org/10.3389/fvets.2022.911685 (2022).
55. Terefe, E., Belay, G., Han, J., Hanotte, O. & Tijjani, A. Genomic adaptation of Ethiopian indigenous cattle to high altitude. *Front. Genet.* https://doi.org/10.3389/fgene.2022.960234 (2022).
56. Edea, Z., Dadi, H., Dessie, T. & Kim, K.-S. Genomic signatures of high-altitude adaptation in Ethiopian sheep populations. *Genes Genomics* **41**, 973–981 (2019).
57. Buroker, N. et al. SNPs, linkage disequilibrium and transcriptional factor binding sites associated with acute mountain sickness among Han Chinese at the Qinghai-Tibetan Plateau. *Int. J. Genom. Med.* **3**, 2332–0672 (2015).
58. Rubin, C.-J. et al. Rapid adaptive radiation of Darwin's finches depends on ancestral genetic modules. *Sci. Adv.* **8**, eabm5982 (2022).
59. Lamichhaney, S. et al. A beak size locus in Darwin's finches facilitated character displacement during a drought. *Science* **352**, 470–474 (2016).
60. Chen, S.-Y., Luo, Z., Jia, X., Zhou, J. & Lai, S.-J. Evaluating genomic inbreeding of two Chinese yak (Bos grunniens) populations. *BMC Genomics* **25**, 712 (2024).
61. Hou, H. et al. Genome-wide association study of growth traits and validation of key mutations (MSTN c.C861T) associated with the muscle mass of meat pigeons. *Anim. Genet.* **55**, 110–122 (2024).
62. Yu, S. et al. Resequencing of a pekin duck breeding population provides insights into the genomic response to short-term artificial selection. *GigaScience* https://doi.org/10.1093/gigascience/giad016 (2023).
63. Recuerda, M. et al. Adaptive phenotypic and genomic divergence in the common chaffinch (Fringilla coelebs) following niche expansion within a small oceanic island. *J. Evol. Biol.* **36**, 1226–1241 (2023).
64. Shi, J. et al. MiRNA sequencing of Embryonic Myogenesis in Chengkou Mountain Chicken. *BMC Genom.* **23**, 571 (2022).
65. Gong, S., Ge, Y., Wei, Y. & Gao, Y. Genomic insights into the genetic basis of eagle-beak jaw, large head, and long tail in the big-headed turtle. *Ecol. Evol.* **13**, e10361 (2023).
66. Resnyk, C. W. et al. Transcriptional analysis of abdominal fat in genetically fat and lean chickens reveals adipokines, lipogenic genes and a link between hemostasis and leanness. *BMC Genom.* **14**, 557 (2013).
67. Tarsani, E. et al. Discovery and characterization of functional modules associated with body weight in broilers. *Sci. Rep.* **9**, 9125 (2019).
68. Sun, Y. et al. GATA Binding Protein 3 Is a Direct Target of Kruppel-Like Transcription Factor 7 and Inhibits Chicken Adipogenesis. *Front. Physiol.* **11**, 610 (2020).
69. Zhao, M. et al. OTUD7A Regulates Inflammation- and Immune-Related Gene Expression in Goose Fatty Liver. *Agriculture* **12**, 105 (2022).
70. Wang, D. et al. TBK1 Mediates Innate Antiviral Immune Response against Duck Enteritis Virus. *Viruses* **14**, 1008 (2022).
71. Seki, R. et al. Functional roles of Aves class-specific cis-regulatory elements on macroevolution of bird-specific features. *Nat. Commun.* **8**, 14229 (2017).
72. Hu, F. et al. Different expression patterns of sperm motility-related genes in testis of diploid and tetraploid cyprinid fish†. *Biol. Reprod.* **96**, 907–920 (2017).
73. Castillo, J., Jodar, M. & Oliva, R. The contribution of human sperm proteins to the development and epigenome of the preimplantation embryo. *Hum. Reprod. Update* **24**, 535–555 (2018).
74. Shen, X. et al. Quantitative proteomic analysis of chicken serum reveals key proteins affecting follicle development during reproductive phase transitions. *Poult. Sci.* **100**, 325–333 (2021).
75. Renaud, S. J. et al. OVO-like 1 regulates progenitor cell fate in human trophoblast development. *Proc. Natl. Acad. Sci.* **112**, E6175–E6184 (2015).
76. Peona, V. et al. The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities. *Philos. Trans. R. Soc. B Biol Sci.* **376**, 20200186 (2021).
77. Sakashita, A. et al. Endogenous retroviruses drive species-specific germline transcriptomes in mammals. *Nat. Struct. Mol. Biol.* **27**, 967–977 (2020).
78. Ballan, M. et al. Genomic diversity and signatures of selection in meat and fancy rabbit breeds based on high-density marker data. *Genet. Sel. Evol.* **54**, 3 (2022).
79. Wang, N., Wang, R., Wang, R. & Chen, S. Transcriptomics analysis revealing candidate networks and genes for the body size sexual dimorphism of Chinese tongue sole (Cynoglossus semilaevis). *Funct. Integr. Genomics* **18**, 327–339 (2018).
80. Jiang, J. et al. Spin1z induces the male pathway in the chicken by down-regulating Tcf4. *Gene* **780**, 145521 (2021).
81. Scholz, B. et al. Sex-dependent gene expression in early brain development of chicken embryos. *BMC Neurosci.* **7**, 12 (2006).
82. Zhang, S. O., Mathur, S., Hattem, G., Tassy, O. & Pourquié, O. Sex-dimorphic gene expression and ineffective dosage compensation of Z-linked genes in gastrulating chicken embryos. *BMC Genomics* **11**, 13 (2010).
83. Widelitz, R. B. et al. Morpho-regulation in diverse chicken feather formation: Integrating branching modules and sex hormone-dependent morpho-regulatory modules. *Dev. Growth Differ.* **61**, 124–138 (2019).
84. Hirst, C. E. et al. Sex Reversal and Comparative Data Undermine the W Chromosome and Support Z-linked DMRT1 as the Regulator of Gonadal Sex Differentiation in Birds. *Endocrinology* **158**, 2970–2987 (2017).
85. Weiss, K. et al. Haploinsufficiency of ZNF462 is associated with craniofacial anomalies, corpus callosum dysgenesis, ptosis, and developmental delay. *Eur. J. Hum. Genet.* **25**, 946–951 (2017).
86. Wang, B. et al. Zfp462 deficiency causes anxiety-like behaviors with excessive self-grooming in mice. *Genes Brain Behav.* **16**, 296–307 (2017).
87. Pilarowski, G. O. et al. Missense variants in the chromatin remodeler CHD1 are associated with neurodevelopmental disability. *J. Med. Genet.* **55**, 561–566 (2018).
88. Cade, T. Biological Traits of the Gyrfalcon (Falco rusticolus) in Relation to Climate Change. in *Gyrfalcons and Ptarmigan in a Changing World* eds. Watson, R. T., Cade, T. J., Fuller, M., Hunt, G. & Potapov, E. The Peregrine Fund., Idaho, USA, (2011).
89. Schluter, D. Evidence for Ecological Speciation and Its Alternative. *Science* **323**, 737–741 (2009).
90. vonHoldt, B. M. et al. Growth factor gene IGF1 is associated with bill size in the black-bellied seedcracker Pyrenestes ostrinus. *Nat. Commun.* **9**, 4855 (2018).
91. Kerr, B., Garcia-Rudaz, C., Dorfman, M., Paredes, A. & Ojeda, S. R. TrkA and TrkB receptors facilitate follicle assembly and early follicular development in the mouse ovary. *Reprod. Camb. Engl.* **138**, 131–140 (2009).
92. Esmaeili-Fard, S. M., Gholizadeh, M., Hafezian, S. H. & Abdollahi-Arpanahi, R. Genome-wide association study and pathway analysis identify NTRK2 as a novel candidate gene for litter size in sheep. *PLOS ONE* **16**, e0244408 (2021).
93. Wu, Y., Liu, X., Hou, S., Xiao, H. & Zhang, H. Identification of adipose differentiation-related protein gene in Peking duck and its expression profile in various duck tissues. *Mol. Biol. Rep.* **38**, 2479–2484 (2011).
94. Zhao, X., Zhu, Q., Wang, Y., Yang, Z. & Liu, Y. Tissue-specific expression of the chicken adipose differentiation-related protein (ADP) gene. *Mol. Biol. Rep.* **37**, 2839–2845 (2010).

95. Zhao, X., Liu, Y., Jiang, X., Du, H. & Zhu, Q. Association of Polymorphisms of Chicken Adipose Differentiation-related Protein Gene with Carcass Traits. *J. Poult. Sci.* **46**, 87–94 (2009).
96. Wright, K., Nip, K. M., Kim, J. E., Cheng, K. M. & Birol, I. Seasonal and sex-dependent gene expression in emu (Dromaius novaehollandiae) fat tissues. *Sci. Rep.* **12**, 9419 (2022).
97. DeMoranville, K. J. et al. PPAR expression, muscle size and metabolic rates across the gray catbird's annual cycle are greatest in preparation for fall migration. *J. Exp. Biol.* **222**, jeb198028 (2019).
98. Jiao, Y. et al. Mex3c Mutation Reduces Adiposity and Increases Energy Expenditure. *Mol. Cell. Biol.* **32**, 4350–4362 (2012).
99. Sassu, E. D. et al. Mio/dChREBP coordinately increases fat mass by regulating lipid synthesis and feeding behavior in Drosophila. *Biochem. Biophys. Res. Commun.* **426**, 43–48 (2012).
100. Lü, Z. et al. Large-scale sequencing of flatfish genomes provides insights into the polyphyletic origin of their specialized body plan. *Nat. Genet.* **53**, 742–751 (2021).
101. Zu, S. et al. Duck PIAS2 negatively regulates RIG-I mediated IFN-β production by interacting with IRF7. *Dev. Comp. Immunol.* **108**, 103664 (2020).
102. Garcia-Morales, C. et al. Cell-Autonomous Sex Differences in Gene Expression in Chicken Bone Marrow-Derived Macrophages. *J. Immunol.* **194**, 2338–2344 (2015).
103. Valdebenito, J. O. et al. Seasonal variation in sex-specific immunity in wild birds. *Sci. Rep.* **11**, 1349 (2021).
104. Vincze, O. et al. Sexual dimorphism in immune function and oxidative physiology across birds: The role of sexual selection. *Ecol. Lett.* **25**, 958–970 (2022).
105. Schluter, D. & Conte, G. L. Genetics and ecological speciation. *Proc. Natl. Acad. Sci.* **106**, 9955–9962 (2009).
106. Ferguson-Lees, J. & Christie, D. A. *Raptors of the World*. Houghton Mifflin Harcourt, (2001).
107. Sarà, M. et al. First evidence by satellite telemetry of Lanner falcon's Falco biarmicus feldeggii natal dispersal outside Sicily, and a review of existing data. *Avocetta* **43**, 75–80 (2019).
108. Alabdulhafith, B. et al. Predicting the potential distribution of a near-extinct avian predator on the Arabian Peninsula: implications for its conservation management. *Environ. Monit. Assess.* **194**, 535 (2022).
109. Porter, R. & Aspinall, S. *Birds of the Middle East* (Bloomsbury Publishing, 2013).
110. Olave, M., Nater, A., Kautt, A. F. & Meyer, A. Early stages of sympatric homoploid hybrid speciation in crater lake cichlid fishes. *Nat. Commun.* **13**, 5893 (2022).
111. Elgvin, T. O. et al. The genomic mosaicism of hybrid speciation. *Sci. Adv.* **3**, e1602996 (2017).
112. Paez, S. et al. Reference genomes for conservation. *Science* **377**, 364–366 (2022).
113. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
114. 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).
115. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
116. Warren, R. L. et al. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience* **4**, 35 (2015).
117. Yeo, S., Coombe, L., Warren, R. L., Chu, J. & Birol, I. ARCS: scaffolding genome drafts with linked reads. *Bioinformatics* **34**, 725–731 (2018).
118. Ghurye, J. et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Comput. Biol.* **15**, e1007273 (2019).
119. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *ArXiv12073907 Q-Bio* (2012).
120. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
121. Rm, W. et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msx319 (2018).
122. 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).
123. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465-467 (2005).
124. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* https://doi.org/10.1002/0471250953.bi0410s25 (2009).
125. Götz, S. et al. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).
126. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).
127. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
128. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
129. Pickrell, J. K. & Pritchard, J. K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genet.* **8**, e1002967 (2012).
130. Patterson, N. et al. Ancient Admixture in Human History. *Genetics* **192**, 1065–1093 (2012).
131. Petr, M., Vernot, B. & Kelso, J. admixr—R package for reproducible analyses using ADMIXTOOLS. *Bioinformatics* **35**, 3194–3195 (2019).
132. RStudio Team. RStudio: Integrated Development for R. RStudio, PBC. (2020).
133. Jetz, W., McPherson, J. M. & Guralnick, R. P. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol.* **27**, 151–159 (2012).
134. 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).
135. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **45**, e18–e18 (2017).
136. Murtskhvaladze, M., Tarkhnishvili, D., Anderson, C. L. & Kotorashvili, A. Phylogeny of caucasian rock lizards (Darevskia) and other true lizards based on mitogenome analysis: Optimisation of the algorithms and gene selection. *PLOS ONE* **15**, e0233680 (2020).
137. Leslie, M. S., Archer, F. I. & Morin, P. A. Mitogenomic differentiation in spinner (Stenella longirostris) and pantropical spotted dolphins (S. attenuata) from the eastern tropical Pacific Ocean. *Mar. Mammal Sci.* **35**, 522–551 (2019).
138. Zhang, D. et al. "Ghost Introgression" As a Cause of Deep Mitochondrial Divergence in a Bird Species Complex. *Mol. Biol. Evol.* **36**, 2375–2386 (2019).
139. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
140. Bouckaert, R. et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).
141. Ogilvie, H. A., Bouckaert, R. R. & Drummond, A. J. StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates. *Mol. Biol. Evol.* **34**, 2101–2114 (2017).

142. Jarvis, E. D. et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014).
143. Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A. & Minh, B. Q. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **44**, W232-235 (2016).
144. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
145. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
146. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
147. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
148. Smeds, L., Qvarnström, A. & Ellegren, H. Direct estimate of the rate of germline mutation in a bird. *Genome Res.* **26**, 1211–1218 (2016).
149. Ericson, P. G. P., Qu, Y., Blom, M. P. K., Johansson, U. S. & Irestedt, M. A genomic perspective of the pink-headed duck Rhodonessa caryophyllacea suggests a long history of low effective population size. *Sci. Rep.* https://doi.org/10.1038/s41598-017-16975-1 (2017).
150. Hanna, Z. R. et al. Northern spotted owl (strix occidentalis caurina) genome: Divergence with the barred owl (strix varia) and characterization of light-associated genes. *Genome Biol. Evol.* **9**, 2522–2545 (2017).
151. Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).
152. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
153. Mi, H. et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **49**, D394–D403 (2021).

Acknowledgments

The authors would like to thank Al-Gannas Qatari Society, The Cultural Village Foundation-Katara, Doha, Qatar, for their support and funding since the inception of the project. We also thank Souq Waqif Falcon Hospital, Qatar University, Sidra Medicine, Um Haish Falcon and Houbara Conservation Center, Monash University Malaysia Genomics Facility and The Vertebrate Genomes Laboratory, Rockefeller University, for their contributions to this work. We also would like to express our gratitude to the Animal Resources Department, and Environmental Protection Reserves & Wildlife Department, Ministry of Municipality and Environment in Qatar for their support and assistance for this project. QA is supported by strategic funding from Monash University (STG-000114).

Author contributions

Conceptualization: FA Sampling: FA, IK Methodology: FA, SR, QA, EJ, OF, GF, AT, KH, YS Investigation: FA Visualization: FA Supervision: SR, QA, AAA Writing—original draft: FA Writing—review & editing: SR, QA, EJ, AAA

Funding

Al-Gannas Qatari Society, State of Qatar. The Cultural Village Foundation-Katara, Doha, State of Qatar. Monash University Malaysia, Malaysia (STG-000114).

Declarations

Competing interests

The authors declare that they have no competing interests.

Ethics approval

Sample collection was carried out under the official approval of the Ministry of Municipality and Environment in Qatar (Reference No. 2017/283748) and by trained veterinarians at Souq Waqif Falcon Hospital, Doha, Qatar. The blood samples were only taken during routine diagnostics from falcons being checked for purposes unrelated to the research activity and the researcher had no influence on the timing or location of the medical check-up. These blood samples would have normally been discarded if the researcher had not asked for them. The Monash University Animal Ethics Committee (AEC) advised that such activity, considered as scavenging, did not require animal ethics approval.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-88216-9.

Correspondence and requests for materials should be addressed to F.O.A.-A. or Q.A.