scientific reports

OPEN

Hybrid assembly and comparative genomics unveil insights into the evolution and biology of the red-legged partridge

Abderrahmane Eleiwa[1], Jesus Nadal[2], Ester Vilaprinyo[1,2], Alberto Marin-Sanguino[1,2], Albert Sorribas[1,2], Oriol Basallo[1,2], Abel Lucido[1,2], Cristobal Richart[3], Ramona N. Pena[2,4], Roger Ros-Freixedes[2,4], Anabel Usie[2,5,6] & Rui Alves[1,2✉]

The red-legged partridge *Alectoris rufa* plays a crucial role in the ecosystem of southwestern Europe, and understanding its genetics is vital for conservation and management. Here we sequence, assemble, and annotate a highly contiguous and nearly complete version of its genome. This assembly encompasses 96.9% of the avian genes flagged as essential in the BUSCO aves_odb10 dataset. Moreover, we pinpointed RNA and protein-coding genes, 95% of which had functional annotations. Notably, we observed significant chromosome rearrangements in comparison to quail (*Coturnix japonica*) and chicken (*Gallus gallus*). In addition, a comparative phylogenetic analysis of these genomes suggests that *A. rufa* and *C. japonica* diverged roughly 20 million years ago and that their common ancestor diverged from *G. gallus* 35 million years ago. Our assembly represents a significant advancement towards a complete reference genome for *A. rufa*, facilitating comparative avian genomics, and providing a valuable resource for future research and conservation efforts for the red-legged partridge.

Alectoris rufa, also known as red-legged partridge, is a game bird that holds significant ecological and economic importance for rural areas in southwestern Europe[1]. Habitat degradation, captive breeding, and hunting management have led to the creation of a complex species situation, impacting both the ecosystems and society of the region. Across various hunting grounds, wild, farmed, and hybrid partridges coexist in varying proportions. While these partridges exhibit distinctions in behavior, physiology, morphology, anatomy, and genetics, the absence of a reference genome hinders our ability to molecularly differentiate these ecotypes, spanning from wild to domestic[2]. The haploid genome of *A. rufa* has 9 macro chromosomes and 30 micro chromosomes[3,4]. The advent of Next-Generation Sequencing (NGS) technologies, mainly based on short-read sequencing data, combined with decreasing DNA sequencing costs, led to an increase in the number of available genome sequences. However, those genomes were still highly fragmented due to the limitations inherent to short reads, where for example repetitive regions can lead to genome misassembly. The emergence of third-generation sequencing technologies partially overcame those limitations by generating long-read sequencing data. These long-reads helped to reduce assembly fragmentation and increase contiguity, greatly improving the quality of whole-genome assemblies[5]. Still, early long-read technologies had base-calling error rates of 10–14%, that are much higher than the less than 1% error rate found in short-read technologies[6]. In addition, the error profiles of both technologies are different. Errors in short-reads are mostly at the level of incorrect nucleotide substitutions, while errors in long-reads mostly involve incorrect insertions and deletions[7,8]. This difference makes long read errors more complex to resolve, requiring an error correction step prior to genome assembly. The error correction problem has been addressed either by self-correction, aligning long-reads against each other, or by a hybrid approach in which long-reads are corrected using short-reads. The latter approach is known to achieve more accurate genome assemblies than genomes assembled based only on short- or long-read technologies[9,10].

[1]Institut de Recerca Biomédica (IRBLleida), Lleida, Spain. [2]Universitat de Lleida (UdL), Lleida, Spain. [3]Universitat Rovira i Virgili (URV), Tarragona, Spain. [4]AGROTECNIO CERCA Center, Lleida, Spain. [5]Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo (CEBAL)/Instituto Politécnico de Beja (IPBeja), Beja, Portugal. [6]MED–Instituto Mediterrâneo para a Agricultura, Ambiente e Desenvolvimento & CHANGE–Global Change and Sustainability Institute, Évora, Portugal. ✉email: rui.alves@udl.cat

In this context, the quality of reference genome assemblies benefited from the combination of Illumina short-read sequencing with third-generation sequencing platforms such as Pacific Bioscience (PacBio)[11] or Oxford Nanopore Technologies (ONT)[12]. Application of these technologies improved contiguity, completeness, and accuracy compared to assemblies based on short-read sequencing alone[13,14]. In general, the number of contigs and scaffolds was significantly reduced, and N50 values increased, leading to better genome annotation and identification of more genes, including non-coding RNA genes, pseudogenes, and transposable elements[15,16]. Examples of genomes assembled using hybrid approaches in the avian clades include, for example, the Tibetan partridge[15], the Indian peafowl[17], the domestic turkey[14], or the common pheasant[18].

The first effort to sequence the red-legged partridge genome of a male individual, which was published in 2021 under the accession number GCA_019345075.1[19], was based on Illumina paired-end short reads sequence data resulting in a highly fragmented assembly, with 10 598 scaffolds, a contig/scaffold N50 of 11.57 Mb, and L90 equal to 131. A more recent version of *A. rufa*'s genome, based on ONT and short reads, was recently released at the NCBI under the accession number of GCA_947331505.1. That version has 426 scaffolds with N50 of 34 Mb and L90 of 32[20]. Both genomes lack detailed annotation. The contiguity of the GCA_947331505.1 assembly (~ 500 contigs) is approximately twenty five times better than that of assembly GCA_019345075.1 (~ 10,000 contigs). Finally, the BUSCO completeness assessment of the two assemblies reveals that assembly GCA_019345075.1 is missing approximately 500 single copy BUSCO orthologs with respect to GCA_019345075.1 and has approximately twenty times more duplicated gene copies. These discrepancies may lead to potential errors in gene order conservation (synteny) and contribute to large-scale assembly inaccuracies. In order to overcome some of the challenges and limitations found in the earlier genome assemblies of *A. rufa* and move towards a well annotated chromosome level assembly, we combined short- and long-read sequencing data in a hybrid approach. Here we report the resulting scaffold-level assembly and its annotation. We validated the assembly by comparing it to the reference genomes of chicken (*Gallus gallus*, NCBI reference GCF_016699485.2) and quail (*Coturnix japonica*, NCBI reference GCF_001577835.2), two closely related species. Overall, we provide a valuable resource for comparative and population genomics, improving our understanding of avian evolution, biogeography, and demography.

Results

Estimation of genome size and heterozygosity rate

We conducted genome profiling on sixty *A. rufa* individuals using *k*-mer analysis of short-read sequence data, and finding an estimated genome size between 1 and 1.06 Gb, and 0.1% ≤ heterozygosity ≤ 0.4% (Fig. 1).

A. rufa genome assembly, annotation and quality assessment

We tested and evaluated various pipelines to assemble the genome of the red-legged partridge. The NextDenovo pipeline produced a primary assembly with the best metrics. This assembly comprised 116 contigs, with an N50 length of 74 Mb and an N90 of 10 Mb (Supplementary Table S1). We further refined this assembly, recovering 96.8% (8078 out of 8332) of the single-copy genes found in the BUSCO dataset of avian single copy orthologous genes (aves_odb10, N = 8332 genes) (Supplementary Table S2). The contigs were then used as the basis for genome scaffolding, resulting in a final genome assembly of 115 scaffolds and 1.03 Gb. Table 1 summarizes the most relevant contiguity metrics of this assembly and its annotation.

The final assembly significantly improves the statistical metrics of contiguity of the earlier available assemblies (Table 2). Our L90 is 23, closer to the 9 macro-chromosomes present in the haploid genome of *A. rufa*, and at least five times smaller than that for assemblies GCA_947331505.1 (based on short-reads) and GCA_019345075.1 (based on long-reads). Our N50 (74 Mb) is twice that of the GCA_019345075.1 assembly and seven times that of the GCA_947331505.1 assembly. Our assembly contained 96.78% (n = 8053 genes) of complete and single-copy genes without duplications present in the BUSCO avian dataset, surpassing both the short-read (95.1%; n = 7933 genes) and the long read (96.58%; n = 7378) genome assemblies. Table 2 summarizes the main differences in terms of the genome contiguity and completeness metrics between those assemblies.

Additionally, we compared our genome assembly against that of eleven birds and one reptile, all of which possessed chromosome-level genome assemblies (Supplementary Table S3). Our assembly has the fifth highest scaffold N50 value for the bird genomes analyzed here (Fig. 2A; see also Supplementary Fig. S1 for the contig N50 statistics). Moreover, in terms of avian orthologs, our assembly also ranks within the top five of the highest number of both complete and single copy orthologs. The NCBI's Foreign Contamination Screening revealed no significant contamination in the assembly of those 115 scaffolds (Supplementary Table S4).

Annotation of transposable elements

RepeatMasker[21] annotated 13% of the *A. rufa* genome as repetitive sequences. Table 3 summarizes the analysis of transposable elements (TE), which revealed a higher percentage of repetitive elements, when compared to the previous draft genome based on short reads alone[19]. Long interspersed nuclear elements (LINE) are the most frequent transposable elements in the genome, representing 7.74% of the whole genome sequence. These elements have a greater divergence rate in comparison to other DNA transposable elements (Supplementary Fig. S2) identified in the genome. DNA transposons (2.33%) and long terminal repeat (LTR) elements (1.76%) are the second and third most abundant classes of transposable elements in the genome, respectively.

Annotation of RNA and protein-coding genes

We validated 10,757 annotated protein genes through comparison of their intron–exon structure with *G. gallus* or *C. japonica* orthologs (Supplementary data file S1). To do so we BLASTed our annotated *A. rufa* proteome against that of *G. gallus*, to identify pairs of orthologs with conserved intro-exon structure. Then, we repeated the
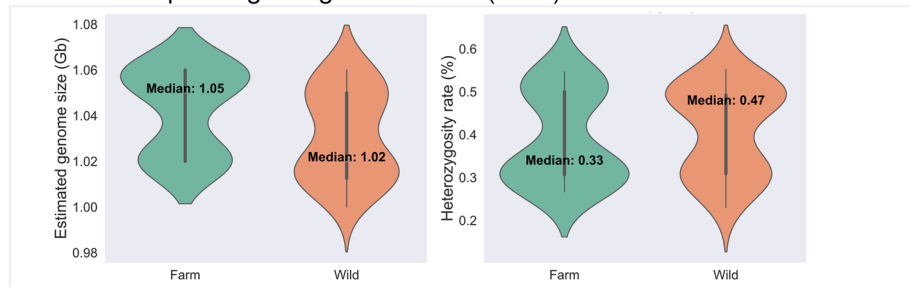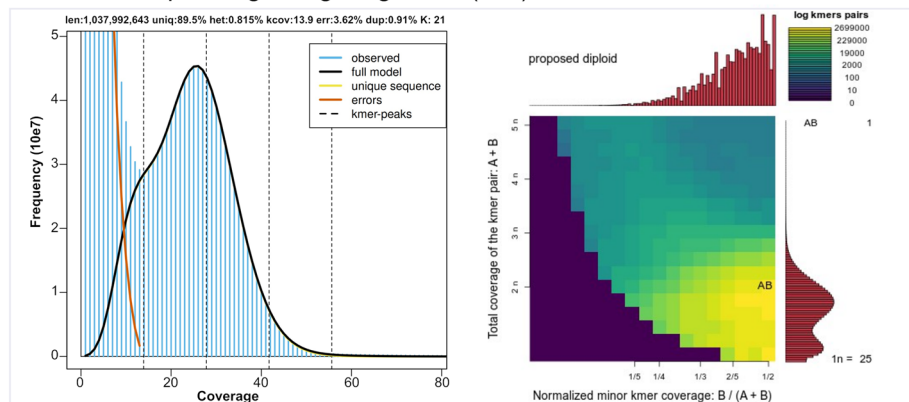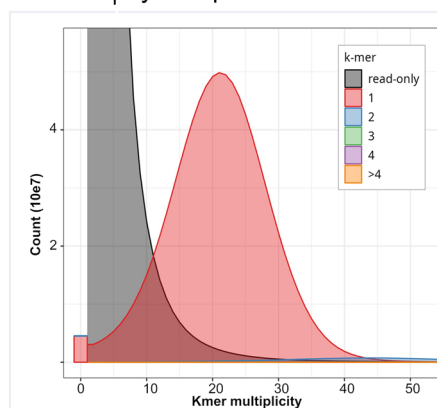
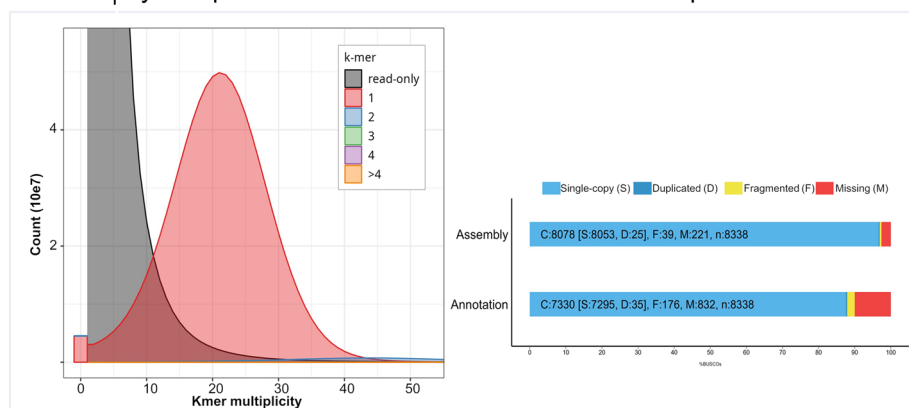**A**. Genome profiling using short-reads (N:60)

B. Genome profiling using long-reads(N:2)

C. Merqury completeness : 92%   **D**. BUSCOs completeness > 96 %

Figure 1. Assessment and profiling of *Alectoris rufa*'s genome assembly. **A** Distribution of estimated genome size and heterozygosity levels across sixty individuals of *A. rufa* sequenced with Illumina short-reads. B Genome size and ploidy level estimation using long reads from two individual *A. rufa* sequenced with ONT. Left: Genome size and heterozygosity plot. Right: genome ploidy inference. C Assessment of scaffolds completeness using a 21 k-mer spectrum approach. Scaffold completeness is estimated to be 92%. D Comparison the completeness between genome assembly and genome annotation, based on recovered core genes from the aves_odb10 dataset of BUSCO.

process between *A. rufa* and *C. japonica*. An additional 8,509 genes were also validated through mapping of the full transcript (Supplementary Table S5). This generated a high-confidence data set of 19,266 predicted protein genes. An additional 11,010 gene were annotated with lower confidence, making a total of 30,236 protein-coding genes in our assembly. We summarize the statistics for all these predicted genes in Table 4.

We identified known homologs for 95% (28,862) of the predicted protein genes in a non-redundant database merging the complete protein datasets downloaded from SwissProt, TrEMBL and NCBI. Of these, 18,865 (62.1%) proteins were simultaneous and consistently annotated among the three databases (Supplementary Fig. S3). We were able to assign InterProScan family and subfamily domains to 25,978 (85.9%) predicted genes, and GO biological functions to 13,371 (57.1%) genes (Supplementary data file S1).

A KEGG-based functional annotation mapped 12,377 of our protein-coding genes predicted with high confidence to their representative functional KEGG ortholog (KO) genes (Supplementary data file S1). The largest number of genes were mapped to genetic information processing (2968 genes), environmental information

Genome assembly	
Assembly size (Gb)	1.03
Number of scaffolds	115
Max. scaffold length (Mb)	149.6
N50 scaffold size (Mb)	74
N90	23
Mean scaffold length (Mb)	9.03
QV	36.3
Merqury completeness score	92%
Gene annotation	
Total Protein-coding genes	30236
Functional annotation (Total)	
Uniprot (SWISS_prot)	18878
Uniprot (TrEMBL)	28596
NCBI_NR	28799
Genomic features	
Repeat (%)	13.3
GC (%)	42.1
Protein coding genes (%)	46.7
Non protein-coding DNA (%)	54.3
Avg. protein length (aa)	441.3
Total microRNAs	246
Total tRNA	305
Total rRNA	130
Total snRNA	315

Table 1. Statistic for the *Alectoris rufa* genome assembly and annotation.

	This study	Chattopadhyay et al.[19]	GCA_947331505.1
Sequencing technology	Illumina NovaSeq 6000, ONT GridION	Illumina HiseqX	NovaSeq 6000, ONT GridION
Assembly size (bp)	1,027,480,606	1,039,068,021	1,142,486,555
Total scaffolds	115	10,598	426
Scaffold N50 (bp)	74,759,052	11,577,318	37,566,138
Scaffold N90 (bp)	10,103,736	1,025,037	6,380,696
Longest Scaffold length (Mb)	149	47	118
GC content (%)	42.1	41.4	42.3
Total predicted protein-coding genes	30,236	NA	NA
Complete recovered BUSCOs	8078	7970	8094
Complete and single-copy BUSCOs	8053	7933	7378
Complete and duplicated BUSCOs	25	37	716
Fragmented BUSCOs	39	122	37
Missing BUSCOs	221	246	207
Total BUSCO groups searched	8338	8338	8338

Table 2. Comparing *Alectoris rufa* genome assemblies. *ONT* Nanopore oxford technology.

processing (1785 genes), and molecular function-related signaling and cellular processes (1664 genes). The top five KEGG metabolic pathways were carbohydrate metabolism (342 genes), lipid metabolism (306 genes), glycan biosynthesis and metabolism (220 genes), amino acid metabolism (180 genes), and nucleotide metabolism (148 genes) (Supplementary data file S1, Supplementary Fig. S4).

We reported the annotation profile of non-coding RNAs (ncRNA) in the assembled genome with respect to their Rfam families. We identified 305 transfer RNA (tRNA) through tRNAScan. Additionally, employing Infernal we were able to identify 246 micro-RNA (miRNA), 135 ribosomal RNA (rRNA) and 315 small nuclear RNA (snRNA) genes (Supplementary data file S1).
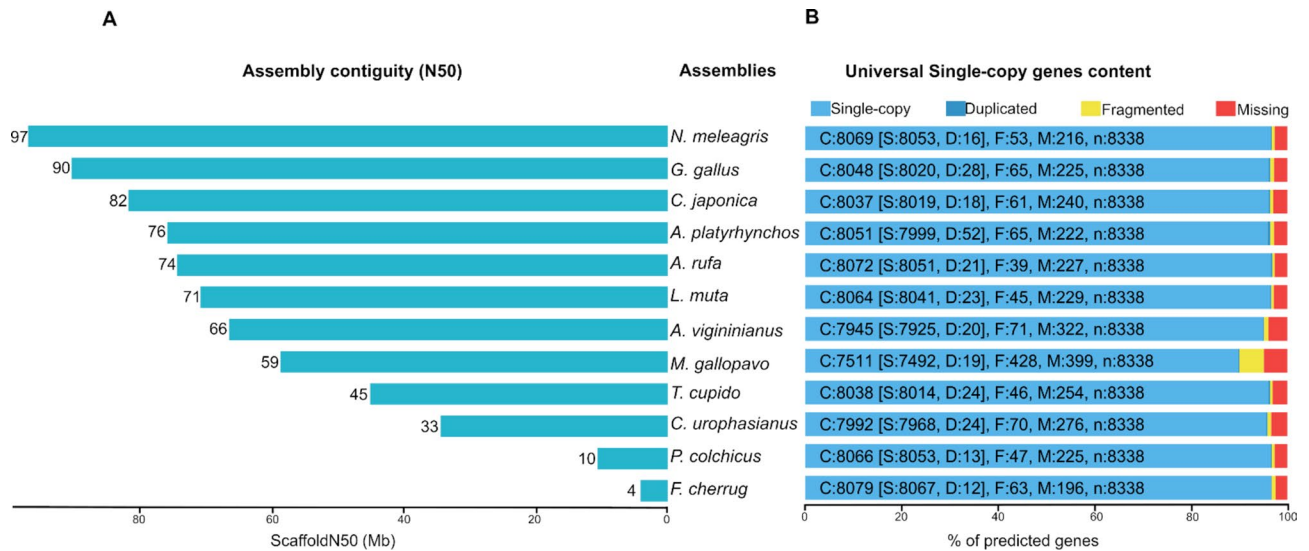
Figure 2. Assessing the completeness and correctness of the *A. rufa* assembly in comparison to closely related bird species. **A** Scaffold N50 statistic for each genome assembly. **B** Completeness of each assembly based on BUSCO results with the aves_odb10 dataset.

	This study		Chattopadhyay et al.[19]	
	Length (bp)	% of genome	Length occupied	% of genome
SINEs	723,913	0.07	179,454	0.02
LINEs	80,386,206	7.74	70,223,435	6.83
LTR elements	18,335,204	1.76	10,566,378	1.03
DNA transposon	24,247,058	2.33	7,824,132	0.76
Unclassified	14,562,721	1.40	10,691,987	1.04
Total	138,255,102	13.31	99,485,386	9.68

Table 3. Comparative statistics of repetitive elements between short read and hybrid *A. rufa* genome assemblies. *SINE* Short-interspersed element, *LINE* Long-interspersed elements, *LTR* Long terminal repeat.

Genomics features	Counts
Number of genes	30,236
Number of CDS	30,236
Number of exons	206,051
Number of intron in CDS	175,815
Number gene overlapping	0
Number of single exon gene	8625
Mean mRNAs per gene	10
Mean exons per CDS	68
Mean introns in CDSs per mRNA	58
Total gene length (bp)	485,558,167
Total CDS length (bp)	40,122,173
Total intron length per CDS	445,435,994
Mean gene length (bp)	16,058
Mean CDS length (bp)	1326
Mean CDS piece length (bp)	194
Mean intron in CDS length (bp)	2533

Table 4. Summary of features annotated in the genome of *A. rufa*.

Synteny analysis of the genome structures of *A. rufa, C. japonica* and *G. gallus*

A. rufa belongs to the Phasianidae (pheasants, partridges, chickens, turkeys, etc.) family of the Galliformes clade. While many relationships within Galliformes are well-supported, some uncertainties remain, particularly regarding the branching order within the species-rich Phasianidae family. One of the uncertainties in this family is the relationship between *A. rufa, C. japonica,* and *G. gallus*. The three birds are closely related and exhibit a shared karyotype of n = 39 chromosomes[22]. This karyotype similarity motivated us to compare the sequence of the largest 23 scaffolds of *A. rufa* (containing at least 90% of the assembled genome) across the three species. Figure 3 highlights significant syntenic regions across the three genomes. Scaffolds 2 and 5 of *A. rufa* align with chromosome 1 in the two other species. Similarly, scaffolds 1, 3 and 4 respectively align to chromosomes 2, 4 and 3 of both birds. Furthermore, *A. rufa* scaffolds 6 and 10 display near complete synteny with *C. japonica*'s sex chromosome Z, while scaffold 10 showing synteny with *G. gallus'* Z chromosome. Scaffolds 7 and 15 of *A. rufa* display considerable synteny with chromosome 5 of the other birds. The remaining 14 *A. rufa* scaffolds exhibit strong synteny with individual chromosomes of the other two bird species. Notably, twelve micro chromosomes from *C. japonica* and 20 micro chromosomes from *G. gallus* did not exhibit significant homology with any of the assembled *A. rufa* scaffolds.

Pairwise analysis of the chromosomal rearrangements between *A. rufa* and *C. japonica* or *G. gallus*

The scaffold-to-chromosome alignments revealed significant large-scale genomic rearrangements between *A. rufa* and both *C. japonica* and *G. gallus* genomes (Supplementary Fig. S5, Supplementary Table S6). Scaffold 2 exhibits a small 2.52 Mb inversion within the 105.87–108.07 Mb region of *C. japonica*'s chromosome 1. Scaffold 5 presents two similar-sized inversions, occurring at regions 19.06–20.95 Mb and 50.02–57.48 Mb of chromosome 1. Scaffold 1 displays a substantial inversion in its center relative to the centromeric region of *C. japonica*'s chromosome 2 (42.9–77.77 Mb). Scaffold 3 features two inversions near one of its ends compared to chromosome 3. Similarly, scaffolds 4 and 18 exhibit inversions when aligned to chromosomes 4 and 15, respectively.

Pairwise alignment of our scaffolds with *G. gallus* chromosomes unveiled repeated inversions, particularly at telomeric regions. Notably, scaffold 4 included two inversions totaling 4.37 Mb within regions 1.76–4.29 Mb and 0.02–1.77 Mb of chromosome 4. Similarly, scaffold 8 exhibited three inversions totaling 3.23 Mb between regions 7.3–8.46 Mb, 9.97–11.06 Mb, and 11.81–12.72 Mb, aligning with chromosome 6 of the *G. gallus* genome. Additionally, scaffold 11 featured a substantial 8.35 Mb inversion relative to the 0.06–8.07 Mb region of chromosome 8.

Overall, these results suggest that *A. rufa*'s genome is more similar to that of *C. japonica* than to that of *G. gallus,* indicating a closer evolutionary relationship between *A. rufa* and *C. japonica* when compared to the *G. gallus*. The similarities in genomic structures and rearrangements between *A. rufa* and *C. japonica* genomes imply a closer evolutionary proximity between the two birds with respect to *G. gallus*.

Comparative proteome of *A. rufa, C. japonica, G. gallus,* and *M. gallopavo*

Comparing the ortholog clusters of protein coding genes in the high confidence dataset between the four species reveals 10,111 shared orthologous gene families (Fig. 4A). We have also identified 113 gene families that are exclusive to *A. rufa*. Among these, 101 genes could be functionally associated to general biological processes using GO (Supplementary data file S1, summarized in Fig. 4B). Among the gene families linked to more specific GO components, 1 gene was associated with membranes, and 2 genes were associated with structural activities. The set of genes unique to *A. rufa* (Fig. 4C) is significantly enriched in genes related to viral processes (5 genes) regulation of immune response (8 genes) and microtubule depolymerization (16 genes).
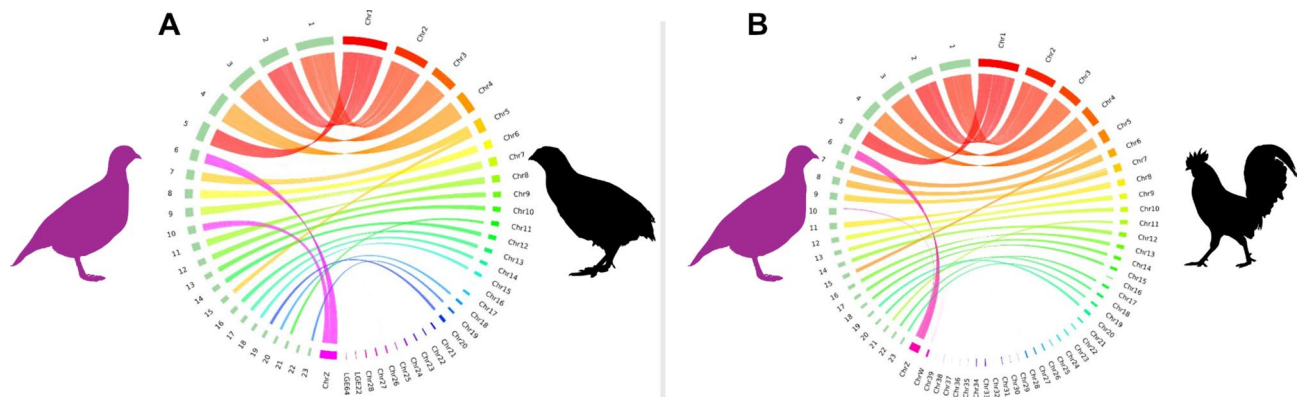

Figure 3. Circus plots comparing sequence homology between the largest 23 *A. rufa* scaffolds and the reference chromosomes of **A** *C. japonica*, and **B** *G. gallus*. Each line within the circle represents 10 Kb of sequence homology. Chromosomes are color coded to facilitate visualizing the synteny regions between *A. rufa* and the other two birds. There are 248386 regions of strong homology between *A. rufa* and *C. coturnix*, compared to 154686 regions of strong homology between the genome of *A. rufa* and that of *G. gallus*.
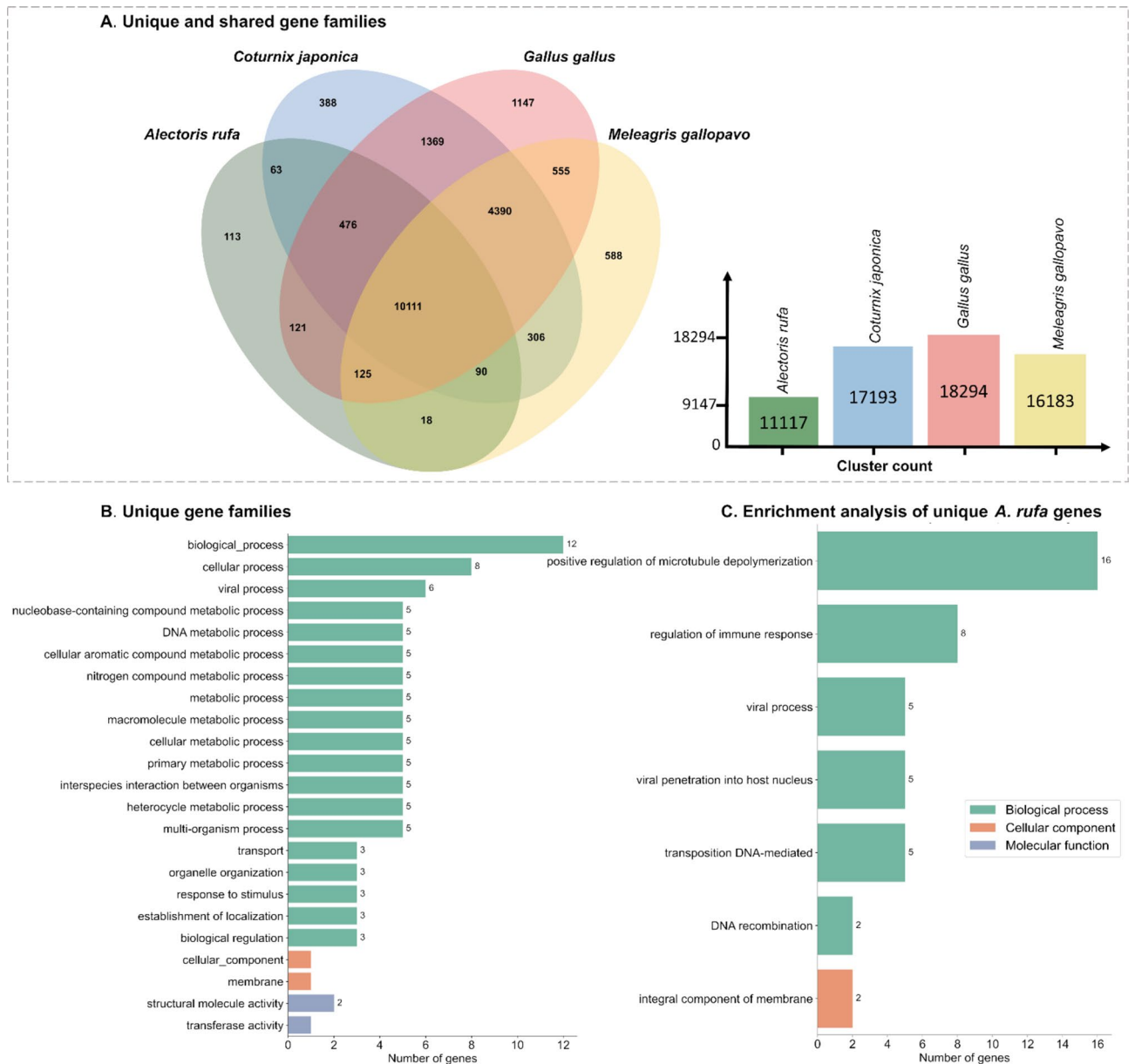
Figure 4. Functional comparison of the protein genes annotated with higher confidence in *A. rufa*'s proteome to the annotated NCBI proteomes of *C. japonica*, *G. gallus*, and *M. gallopavo*. **A** Comparison of orthologous gene families between *A. rufa, C. japonica, G. gallus* and *M. gallopavo*. **B** Generic GO enrichment terms for gene families that are unique to *A. rufa*. **C** Specific GO enrichment terms for gene families that are unique to *A. rufa*. Only GO categories that are associated to more than one gene were included in panels (**B**) and (**C**).

Phylogenetic analysis of *A. rufa* within the *Galliformes* clade

The phylogenetic tree (Fig. 5), constructed through the alignment of 8212 single-copy BUSCO genes found across thirteen genomes (Supplementary Table S3 and our *A. rufa* assembly), unveils pivotal points in evolutionary history measured in million years ago (Mya). The divergence between birds and reptiles occurred roughly 300 Mya. Anseriformes and Galliformes parted ways around 75 (95% credibility interval 46.14–106.85) Mya, with the Guinea fowl diverging from the main Galliformes lineage approximately 56 (95% credibility interval 33.41–78.34) Mya. The clade containing *G. gallus*, *C. japonica* and *A. rufa* separated from the rest of the Galliformes approximately 49 (95% credibility interval 27.88–69.01) Mya, with their last common ancestor estimated at roughly 35 (95% credibility interval 9.86–57.87) Mya. The divergence between *C. japonica* and *A. rufa* happened approximately 20 (95% credibility interval 0.0011–41.44) Mya ago. These predicted divergence timelines are consistent with the findings we report from the pairwise analysis of the chromosomal rearrangements between the three birds (Fig. 3). Because of the large confidence intervals for the divergence times we calculated the individual maximum likelihood gene trees for the single copy BUSCO orthologs identified in all genomes, using IQTree. We then used the multi-species coalescent model approach in ASTRAL to build the species tree from the individual gene-based trees (Supplementary Fig. S6). The speciation structure of the two trees is consistent.
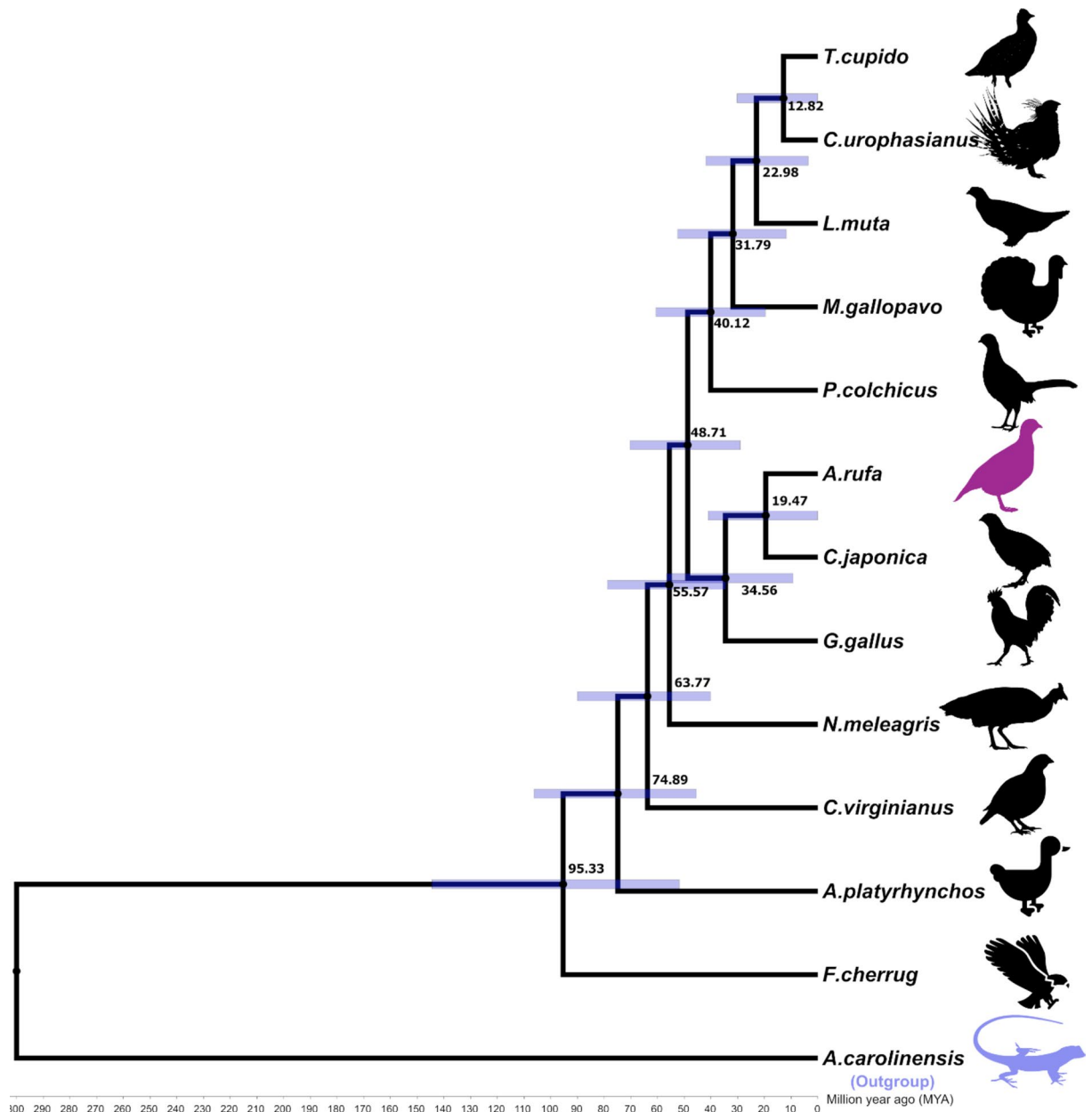
Figure 5. Phylogenetic analysis of *A. rufa*. The phylogenetic tree was reconstructed from concatenated single-orthologous genes of the complete genome of 11 birds plus *A. rufa* using IQTREE. The lizard *Anolis carolinensis* (lizard) was used as outgroup. *A. rufa* is closer to *C. japonica* than to *G. gallus*. Numbers at each node represent the estimated divergence time in million years. Blue lines indicate the 95% credibility interval for those estimates. Only branching points supported by 100% of bootstrapped trees are shown.

Discussion and conclusions

We achieved a highly contiguous genome assembly for *A. rufa* by integrating accurate short reads from Illumina sequencing with lower accuracy ultra-long reads from the Oxford Nanopore Technology (ONT). The resulting assembly is scaffold-level, comprising 115 DNA scaffolds, with a L90 of 23. Our approach demonstrates superior contiguity and scaffolding accuracy compared to previous assemblies relying solely on either short-read[19] or long-read data (accession number GCA_947331505.1 at the NCBI), further validating the efficacy of the combined sequencing approach for de novo genome assembly in non-model organisms. Additionally, the sequences from sixty *A. rufa* individuals provides a valuable reference for future genetic studies characterizing genome size, ploidy, and heterozygosity rates in different *A. rufa* populations. Our assembly contributes to the collection of avian genomes and highlights the effectiveness of integrating long-read and high-quality short-read data from Illumina[10,16,23].

Notably, the contiguity statistics for the *A. rufa* genome is above average with respect to the other eleven fully sequenced *Galliformes* genomes analyzed (Fig. 2A). Still, we note that the Bird10K genome sequencing initiative

is having tremendous success in generating highly contiguous genomes and these have better contiguity statistics than ours[24–27]. We expect to further improve our contiguity by generating and using HiC data to improve the assembly in the future. Assessment our assembly's completeness using BUSCO[28] shows that it has the highest number of single copy orthologous genes identified with respect to the other analyzed genomes (Fig. 2B). We note that the BUSCO assessment of the gene annotation using BUSCOs is lower (87.9% completion rate, Fig. 1D) that that for the assembled genome. This discrepancy between the recovered BUSCO genes and the annotated gene set is consistently observed in similar cases[29]. The highly contiguous assembly facilitated a comprehensive genome annotation by leveraging diverse functionally annotated sequence databases and pre-existing transcriptomic data. As a result, we could use sequence homology to assign biological function for over 95% of all genes identified in our assembly. Overall, our assessment of the annotation quality using RNA sequencing data showed a complete alignment with gene models, with no missed single exons (Supplementary Table S5). Supplementary data file S1 contains all details of the annotation.

We identified 19,226 protein genes with high confidence. Of these, 10,757 protein genes were verified to maintain their intron–exon structure when compared to *G. gallus* or *C. japonica* orthologs (Supplementary data file S1), suggesting these genes are also correctly annotated. The remaining 8,509 genes were verified through mapping the full transcript. Notably, transcriptomic data is currently limited to the spleen and skin tissues, yet it aligns well with the annotated gene models. Despite varying parameters during the process of masking DNA transposable elements, we observed a minimal impact of those changes on the number of annotated protein-coding genes. Given these findings, we anticipate that incorporating transcriptomic data from additional tissues will refine the gene models specific to *A. rufa* and potentially reduce the overall count of annotated genes, mirroring observations in other model organisms[30]. Additional ab initio annotation identifies 11,010 genes with lower confidence.

The genomic annotation of TEs in the *A. rufa* genome shows a high abundance of LINE (7.74% of the genome) and LTR repeat elements (1.76% of the genome). These numbers are higher than those found in the genomes of *G. gallus*[31] (~3% LINE and ~0.5% LTR) and *C. japonica*[32] (~5.60% LINE, and ~0.60% LTR, Supplementary Table S7). The genomes of *C. californica*[33] and *C. virginianus* both have a percentage of LINE (6.9% and 5.6% respectively) and LTR (5.6% and 1.73% respectively) more similar to that found in *A. rufa*. Given that transposable elements were found to influence color[34,35] in insects, mammals, birds and other vertebrates, a future analysis of the genome should reveal if any genes involved in color determination are found within regions containing TEs.

Notably, 13% of the annotated genes are associated with metabolic functions, while 11% are involved in processing environmental information, including 9% dedicated to signal transduction tasks. The distribution of tRNA genes in the *A. rufa* genome indicates that 11% code for alanine-tRNA and 9% for serine-tRNA (Supplementary data file S1, Supplementary Table S8). Our gene enrichment analysis suggests that *A. rufa* evolved a distinct set of regulatory genes and viral response proteins, likely shaped by species-specific infections and pressures. These findings align with previous transcriptomic analyses that highlighted heightened immune responses in the *A. rufa*[36].

A. rufa, *C. japonica*, and *G. gallus* (*Phasianidae* family) have a diploid genome with 78 chromosomes while *C. virginianus* or *C. californica* (*Odontophoridae* family) have 82 and 84, respectively. A structural genomic comparison between the three *Phasianidae* birds using chromosome-mapping approaches shows that chromosomal coverage and synteny is stronger between *A. rufa* and *C. japonica* than between *A. rufa* and *G. gallus*. Still, several chromosomal inversions (Supplementary Fig. S5, Supplementary Table S6) highlight that the divergence between *A. rufa* and *C. japonica* is not recent. For example, aligning scaffold 1 of *A. rufa* to chromosome 2 of *C. japonica* reveals an inversion that contains the centromeric region of the chromosome. Our sequence comparison between scaffold 4 of *A. rufa* and chromosome 4 of *G. gallus* reveals another centromeric inversion. This inversion had been previously reported by[4] based on cytogenetic analysis. Still, we note that inversions detected close to centromeres and telomeres may result from mis-assemblies, due to the higher DNA repeat content in those genomic regions. However, a more detailed analysis of those inverted regions shows that they were all associated to DNA mobile elements, rather than with tandem repeats (Supplementary Table S6). In addition we realigned the raw long reads to our assembly and this alignment is consistent with the assembly direction. As such, we strongly believe that those inversions are not an artefact of assembly. In fact, they are also consistent with similar massive inversions observed within independent populations of *C. coturnix*, another quail species, and associated to an expansion of phenotypic diversity between populations[37]. These genomic rearrangements were reported to associate with adaptive divergence in other species of animals[38]. These and other observations in our analysis emphasize the potential interest of future research focusing on *A. rufa*'s evolutionary chromosomal rearrangements. We are currently developing efforts to generate HiC data that would facilitate obtaining a map of physical interactions that would allow us to generate a chromosome level assembly. This would contribute to fully discard the possibility of the chromosomal rearrangements being assembly artifacts.

The genome assembly provided here is also of interest for phylogenetic studies. Phylogeny proposes an evolutionary tree that aids our comprehension of species divergence over time, drawing upon evidence from paleontology, biogeography, and genetics[39–41]. The integration of both mitochondrial and nuclear markers significantly advanced the accuracy of those studies[42,43]. Still, phylogenetic trees based on individual genes may be biased[44–46], due to factors such as incomplete lineage sorting, gene flow dynamics, and horizontal gene transfer. Coalescent-based methods are often helpful in reducing that bias[47]. Still, combining genome-based trees with estimates of divergence time gleaned from fossil records and genetic clocks has produced robust phylogenies that can be used to generate strong hypotheses about speciation events[48–50]. We also combined fossil record-based divergence times, concatenated gene-based trees, and coalescent-based trees to reconstruct the phylogeny of *A. rufa* in the Galliformes order. We found the phylogenetic tree topologies to be robust for the alternative approaches.

Our tree suggests a divergence time around 75 Mya for the Galliformes clade, consistent with previous estimates[42,48,51]. Notably, within the bird group and the order Galliformes, a variety of studies using a limited set of genetic markers proposed multiple and often coinciding clade formation hypotheses[51,52]. Our results are consistent with those hypotheses. However, divergence times are slightly different because earlier efforts used a smaller number of genetic markers and a larger number of species. Still, those earlier estimates are well within the 95% credibility intervals for our divergence times. Leveraging future assemblies of *A. rufa*'s and other bird genomes to create genome wide alignments from which to create phylogenetic trees will likely enable a more accurate understanding of the evolutionary history of life.

Overall, our assembly and annotation provide a significant contribution towards a reference genome of the red-legged partridge, which will aid in developing genetics applied to phylogeny, zoology, demography, and ecology of the species. This near-chromosome assembly provides a foundation upon which to anchor future comparative genomics research between different *A. rufa* populations and across *Phasianidae* species. It is a valuable resource, potentially enabling the development of more effective strategies for management and conservation of *A. rufa* and wildlife.

Methods
Genome sequencing data
Total DNA was obtained from the muscle of sixty frozen *A. rufa* individuals (muscle from 30 wild birds obtained from hunter's bags and 30 farm birds obtained from slaughtered partridges from a farm in Ciudad Real) for whole-genome sequencing on the NovaSeq6000 Illumina platform producing short paired-end reads with a read length of 151 bp as described in[2]. Additionally, blood was collected from the brachial vein in the wing of two live individuals (one male and one female, no anesthetics were used) using a sterile syringe with a 20 G needle. We then extracted high molecular weight (HMW) DNA from that blood for library preparation with the genomic DNA sequencing kit of Oxford Nanopore technology (ONT) and then sequenced the libraries using a GridION platform. This had the purpose of facilitating an assembly of both sex chromosomes when HiC data becomes available. The study was conducted in full compliance with Spanish laws and regulations, including the licence of "Las Ensanchas" for sampling shot partridges. The protocol was approved by the Committee on the Ethics of Animal Experiments of the University of Lleida (Ref. 1998–2012/05). The ten essential ARIVE guidelines were followed in designing and reporting this study.

Processing sequence data
The Illumina sequencing yielded an average of 218 million raw reads per individual, with an average depth sequencing of 32X per sample. We assessed the quality of those reads using FastQC[53]. The per-base quality scores were consistently high across all samples, and no adaptor content within the reads was found. Thus, it was determined that additional cleaning and adapter removal procedures were unnecessary.

We generated 2 million raw ultra-long reads of the Oxford Nanopore Technology (ONT), yielding 48 Gb with an average read length of 20.68 kb (Supplementary Table S9). We used Porechop V.0.2.4[54] with default parameters in order to scan for known Nanopore adapters and to trim them out of the long reads, ensuring a high-quality dataset, free of adaptor contamination. We assessed the quality of this dataset using NanoPlot v1.40.2 (part of the NanoPack software suite)[55]. We then used Filtlong[56] to split the reads into two subsets applying different criteria. For the first subset, we prioritized read length over average read quality, selecting a coverage depth of 40X(–min_length 15 kb -t 40 Gb). For the second subset, we prioritized average read quality over read length, generating a coverage depth of 20X (–min_mean_q 12 -t 20 Gb). By using these two different subsets, we aimed at improving genome contiguity while also correcting structural errors, ensuring a more reliable and accurate analysis of the sequencing data.

Genome size estimation
We used a 21-mer-based approach in Jellyfish v.2.2.10[57] to estimate *k*-mer histogram frequencies from the Illumina paired-end sequencing data of each of the sixty individual birds. The output of Jellyfish was then used in GenomeScope2[58] to estimate genome size and heterozygosity level for the genome of each bird. In addition to the genome profiling with genomescope2 on those short reads, Smudgeplot[58] was used to estimate the ploidy level using Nanopore long reads sequencing data.

Hybrid genome assembly
Supplementary Fig. S7 summarizes the pipeline we employed to create a de novo assembly for the genome of *A. rufa*, using a hybrid approach. The raw ONT long-reads were assembled de novo with Flye[59], Canu[60], Wtdbg2[61], and NextDenovo v2.2.4[62]. In order to select the best primary assembly for further procedures we compared the performance of the four assemblers. We used QUAST v5.2.0[63] to calculate the contiguity statistics of each assembly statistics and the aves_odb10 dataset of Benchmarking Universal Single-Copy Orthologs (BUSCO) v.5.4.3[28] to assessed their completeness. Based on these numbers, we chose the NextDenovo contig-level assembly for further improvement.

We combined long and short read information to improve the contig-level assembly. This hybrid approach comprised two main steps to enhance the assembly quality. First, we mapped the subset of long-read ONT with 40X and min length size of 15 kb, to the contig-level assembly using minimap2[64]. This alignment was then input into RACON v.1.5.1[65] for one polishing iteration, improving the contiguity of the contig-level assembly by correcting several structural assembly errors. Then, we aligned the short reads from the sixty *A. rufa* individuals to the RACON-improved assembly using BWA-mem2 v2.2.1[66]. This alignment was the input for Polypolish v0.5.0[67], which we used to polish the RACON-improved draft and fix small SNPs and indels, leveraging the high

coverage of short-reads to generate a high-quality consensus assembly that represents the genetic diversity of *A. rufa*'s genome. We completed the scaffolding of the assembly using the REDUNDANS pipeline[68]. We ran this pipeline using the "–non-reduction –nogapclosing" parameters to enhance genome scaffolding, using a subset of both long and short reads in combination from the original raw reads. We combined the subset of accurate long-read ONT with 20X sequencing depth and the short reads of the two animals with the highest genome coverage, aiming at improving scaffold accuracy. The final scaffold-level assembly served as the foundation for downstream genome annotation and comparative analysis.

Genome screening for contamination sequences

Before annotating the assembled genome, we conducted a thorough screening process to identify and eliminate any sequences that might be contaminants related to the assembled genome of *A. rufa*. To do this, we employed NCBI's Foreign Contamination Screening (FCS) tools[69] FCS-adapter and FCS-GX. We used FCS-adapter to detect adaptors and vectors. We used FCS-GX to identify foreign DNA contamination sequences by aligning our assembly against the NCBI database of genomes. We ran each of these tools independently using default settings, except for the taxonomic identifier, which was set to be that of *A. rufa* (NCBI: txid 9079). This rigorous screening process helped ensure the integrity of our assembled genome data before proceeding with annotation.

Genome annotation

Annotation of transposable elements
We used EDTA v2.1.1[70] to annotate the DNA transposable elements (TEs) in our assembled genome. EDTA integrates a set of open-source programs for TE annotation based on homology and/or ab initio search methods. We used two independent data sets to increase the accuracy of EDTA annotation. First, we downloaded a curated library from the gold-standard database of repetitive sequences msRepDB[71]. This library contained DNA transposable sequences for six closely related bird species (*Alectoris barbara*, *Alectoris philbyi*, *Alectoris melanocephala*, *Coturnix japonica*, *Meleagris gallopavo*, and *Gallus gallus*; Supplementary Table S10). Then, the CDS sequences of *G. gallus* were downloaded from ENSEMBL release 109[72], to remove gene-related sequences. In parallel, we used RepeatModeler V2.0.3[73] with default parameters for additional ab initio annotation of repetitive elements. Finally, we combined the results from EDTA and RepeatModeler to build a non-redundant library of repetitive elements using our in-house scripts. This custom TEs library was used as input to the RepeatMasker v4.1.4[21,74] for soft masking of the *A. rufa* genome. We ran RepeatMasker using the following parameters: "-e ncbi -gff -s -a -inv -no_is -norna -xsmall -nolow -div 40", against the Dfam[75] and RepBase update 18. We then used the soft-masked genome for further annotation.

Divergence distribution of transposable element
We analyzed RepeatMasker's alignment output file using the parseRM.pl script v5.8.2 available at[74]. We determined the percentage of divergence from the consensus for each TE fragment, considering the elevated mutation rate at CpG sites and employing the Kimura 2-Parameter divergence metric. This divergence percentage serves as a measure of the age of the TE fragments, as older TE invasions accumulate more mutations. We further categorized TE fragments by age, organizing them into bins of 1 million years, based on the substitution rate calculated by parseRM.pl. We then plotted the distribution landscape of TE using a custom R script.

Gene structure annotation

We combined three strategies to annotate the protein coding genes in the soft-masked genome: homology-based, transcriptome-based, and ab initio predictions:

1- We ran Miniprot v.0.10-r225[76] for homology-based gene prediction. A dataset comprising 3,044,546 protein sequences was generated. These sequences were obtained from the NCBI reference sequence of proteins (accessed on April 15, 2023). Specifically, we focused on the Aves NCBI:txid8782 lineage to ensure retrieval of only avian proteins. Additional details about this dataset can be found in Supplementary Table S11.
2- We ran PASApipeline v.2.5.3[77] to perform gene prediction based on the transcriptional evidence provided by the transcriptome assembly of *A. rufa* published in 2017[36].
3- For ab initio gene prediction, we ran BRAKER2 v.2.1.6[78], training it with the same dataset we used for Miniprot.

The annotation results of the three approaches were then combined using EVidenceModeler v.2.1.0[77] to produce a consensus gene set model of the assembled *A.rufa* genome. The pipeline is summarized in Supplementary Fig. S8.

We then took the annotated proteome of *A. rufa* and BLASTed it against the annotated proteome of *G. gallus*, to identify all pairs of orthologs, filtering by e-values $\leq 10^{-30}$, mutual best BLAST result, and mutual alignments over more than 80% of query and target proteins. Finally, we mapped each ortholog to its corresponding genome, to compare intron structures between orthologous genes. We repeated this comparison between *A. rufa* and *C. japonica*.

Non-coding RNA gene annotation

We also annotated non-coding RNA genes (ncRNAs) in our genome assembly. We used tRNAscan-SE2 v.2.0.11[79] to identify transfer RNAs (tRNAs). Infernal v.1.1.4[80] was run to identify microRNAs (miRNAs), ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs), based on the Rfam database (release 14.0)[81].

Functional annotation

We assigned functions to the predicted gene models combining various approaches. A standard e-value cutoff of 1e-6 was applied for sequence comparisons, unless otherwise specified. Initially, we utilized eggnog-mapper v.2.1.10[82] against the eggNOG database[83] to assign Gene Ontology terms. Subsequently, Blastp v2.12.0 + was employed against SwissProt, TrEMBL[84], and NCBI NR[85] databases for homology-based functional annotation (all the public protein databases mentioned above were downloaded on April 15, 2023). Priority was given to matches with over 95% identity from SwissProt and TrEMBL, as annotation of proteins in these databases is more reliable because of manual curation. The resulting functional annotations were combined with InterProScan v5.64.-96.0[86]. InterProScan identified protein domains, families, and superfamilies in annotated protein-coding genes using specified parameters "-m diamond –sensmode fast –go_evidence non-electronic". KofamKOALA v1.3.0[87] assigned KEGG orthologs (KO)[88] and pathways with an e-value cutoff of 1e-9. Functional annotations from the Uniprot database (minimum Blastp homologue identity match of 95%) were integrated into the final genome annotation file using the GAG tool v2.0.1[89].

Quality assessment of genome assembly and annotation

We used QUAST v5.2.0 to calculate correctness and contiguity metrics for the genome assembly. We used BUSCO against the aves_odb10 v2019-11-20 database to assess both the completeness of the assembly and of the annotation of structurally predicted protein-coding genes.

Furthermore, to assess the accuracy of the genome, Merqury v1.3[90] was used. This involved comparing the original ONT raw reads to the final version of the genome assembly, which had been polished using high-quality data from 60 partridge samples. This analysis provided insights into the QV metric and the accuracy of the consensus sequence.

Quality assessment of the genome annotation using RNA sequencing

As part of evaluating the accuracy of gene model annotations, we downloaded a set of RNA sequencing transcriptome from the spleen and the skin of the red-legged partridge (*A. rufa*) that are deposited in the NCBI SRA database (Supplementary Table S12). We employed the STAR aligner tool v2.7.10b[91] for mapping these reads to the soft-masked assembly version. Subsequently, each sample underwent transcript-assembly guided using Stringtie v2.2.1 reference-guided assembler of transcripts. The spliced transcripts from all samples were combined using Stringtie into a one master list of transcripts, the output of Stringtie was retrieved in the GFF file format for suitable downstream analysis. Next, we used the GffCompare v2.12.6 tool[92] to compare this list of annotated transcripts with respect to the final annotated gene set model. This comparison helped determine the number of new spliced transcripts that were not previously identified in our gene set, contributing to our assessment of gene annotation quality.

Comparison to the reference genomes of *C. japonica* and *G. gallus*

We used MUMMER v.4[93] to perform whole-genome alignment between our assembly and the fully sequenced genomes of *C. japonica* and *G. gallus*. The genome pairwise alignment results and synteny blocks of 10 kb were visualized with DOT-PLOT viewer[94] and Circos v.0.69-8[95].

Gene family analysis

We used the OrthoVenn3 pipeline[96] to compare gene families between *A. rufa*, *C. japonica*, *G. gallus*, and *G. pavo*. In brief, Orthofinder[97] was used to compute the orthologs between the species of interest and to cluster gene families based on GO functional annotation categories. Additionally, we also used the pipeline to automatically conduct GO terms enrichment analysis by considering the evolutionary relationship between the four species.

Phylogenomic analysis and divergence time tree building

We performed phylogenetic analysis to infer the divergence time of *A. rufa* with respect to other birds with fully sequenced genomes within the *Galliformes* order, in a way that is similar to previous reports[14,98,99]. We included all *Galliformes* reference genomes available at the NCBI RefSeq database at the time of submission. In addition to *A. rufa*, we included 8 genome protein sequences of Galliformes species, of which 7 species belong to Phasianidae family and one to Numididae family (*Numida meleagris*). We also included one genome from the *Anseriformes* order (*Anas platyrhynchos*), and another bird species for the Falconiformes order (*Falco cherrug*). As an outgroup we used *Anolis carolinensis*[100] from the Reptilia class.

Genome assemblies for the birds and outgroup (*Anolis carolinensis*) from Reptilia were downloaded from the NCBI. Detailed information about those species can be found in Supplementary Table S3. We started by using the aves_odb10 database of the BUSCO tool[28] to identify shared single-copy genes in the twelve analyzed genomes. The aves_odb10 database contains 8338 genes. We used the custom Python script available at https://github.com/jamiemcg/BUSCO_phylogenomics.git to extract the 8212 shared single-copy orthologs common to all species. We independently created multiple alignments for each of the orthologs common to all species, using MUSCLE[101]. We concatenated the resulting multiple alignments to create a supermatrix alignment. To ensure alignment quality, we applied trimAI[102] and removed unreliable aligned sites and gaps.

Subsequently, a phylogenetic tree was constructed using IQTREE v[103], incorporating 1000 bootstrap replicates. The best model for tree construction was determined using the ModelFinder package[99] from the IQTREE suite. Then we used ASTRAL v5.7.3[104] to handle the possibility of incomplete lineage sorting that might impact gene-based trees. Finally, to estimate the divergence time of *A. rufa* in relation to the other birds, we used the MCMCtree tool from the PAML package[105]. MCMCtree used the phylogenetic tree generated by IQTREE and the

alignment file to achieve reliable divergence time estimation, minimizing potential outliers. Three fossil calibration times from the TimeTree5[106] were employed for divergence estimation: *G. gallus*–*C. japonica* ($\approx$ 32.9–46.1 Mya), Numida-Mallards ($\approx$ 72.5–85.4 Mya), and the divergence time between birds and reptiles ($\approx$300–250 Mya)[107]. We ran MCMCTREE on protein-coding sequences, sampling 20,000 times with a sampling frequency of 10, following a burn-in of 2000 iterations. We used default parameter for the other settings.

Supplementary Table S13 summarizes all bioinformatics pipelines, tools versions, and settings used during the genome assembly and annotation process and other related analysis used in this work.

Ethics approval and consent to participate

The study was conducted in full compliance with Spanish laws and regulations, including the licence of "Las Ensanchas" for sampling shot partridges. The protocol was approved by the Committee on the Ethics of Animal Experiments of the University of Lleida (Ref. 1998–2012/05). The ten essential ARIVE guidelines were followed in designing and reporting this study.

Data availability

Data Availability This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JBCGZB000000000. The version described in this paper is version JBCGZB010000000. The Nanopore raw read data are available via ENA (Bioproject accession PRJEB67643, Biosample: ERS16499794, ERS16499793, ERS16499792, ERS16499791, ERS16499790). The sixty Illumina sample of the partridge sequencing raw reads used for polishing have been deposited in the NCBI database under BioProject PRJNA824288. The high confidence gene annotation dataset is available at https://figshare.com/articles/dataset/Filtered_gene_set_model_of_the_i_Alectoris_rufa_i_genome/25982689. The complete annotation dataset is available as supplementary material. The source code and relevant data files used to generate each figure in this manuscript are available on the GitHub repository page of the Systems Biology and Statistical Methods Group at https://github.com/BioModelLab/A.rufa_genome.git This work was performed under the scope of the Catalan Biogenome Project (CBP).

Code availability

The source code and relevant data files used to generate each figure in this manuscript are available on the GitHub repository page of the Systems Biology and Statistical Methods Group at https://github.com/BioModelLab/A.rufa_genome.git

References

1. Farfán, M. Á. *et al.* The red-legged partridge: A historical overview on distribution, status, research and hunting. In *The Future of the Red-legged Partridge: Science, Hunting and Conservation* (eds Casas, F. & García, J. T.) 1–19 (Springer International Publishing, Cham, 2022). https://doi.org/10.1007/978-3-030-96341-5_1.
2. Ros-Freixedes, R., Pena, R. N., Richart, C. & Nadal, J. Genomic diversity and signals of selection processes in wild and farm-reared red-legged partridges (Alectoris rufa). *Genomics* **115**, 110591 (2023).
3. Arruga, V. *et al.* Estudios genéticos en 'alectoris rufa' y 'a. graeca' en España. *Arch. Zootec.* **45**(170–171), 339–344 (1996).
4. Kasai, F., Garcia, C., Arruga, M. V. & Ferguson-Smith, M. A. Chromosome homology between chicken (*Gallus gallus domesticus*) and the red-legged partridge (*Alectoris rufa*); evidence of the occurrence of a neocentromere during evolution. *Cytogenet. Genome Res.* **102**, 326–330 (2003).
5. Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T. & Sandhu, M. S. Long reads: Their purpose and place. *Hum. Mol. Genet.* **27**, R234–R241 (2018).
6. Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nat. Rev. Genet.* **14**, 157–167 (2013).
7. Ma, X. *et al.* Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.* **20**, 50 (2019).
8. Pfeiffer, F. *et al.* Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **8**, 10950 (2018).
9. 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).
10. De Maio, N. *et al.* Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genomics* https://doi.org/10.1099/mgen.0.000294 (2019).
11. Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics Proteomics Bioinform.* **1**, 3. https://doi.org/10.1016/j.gpb.2015.08.002 (2015).
12. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol.* https://doi.org/10.1186/s13059-016-1103-0 (2016).
13. Jaiswal, S. K. *et al.* Genome sequence of peacock reveals the peculiar case of a glittering bird. *Front. Genet.* **9**, 392 (2018).
14. Dalloul, R. A. *et al.* Multi-platform next-generation sequencing of the domestic Turkey (*Meleagris gallopavo*): GENzOME assembly and analysis. *PLoS Biol.* **8**, e1000475 (2010).
15. Xuejuan, L. I. *et al.* A de novo assembled genome of the Tibetan Partridge (*Perdix hodgsoniae*) and its high-altitude adaptation. *Integr. Zool.* **18**, 225–236 (2023).
16. Dhar, R. *et al.* De novo assembly of the Indian blue peacock (*Pavo cristatus*) genome using Oxford Nanopore technology and Illumina sequencing. *GigaScience* **8**, 1–13 (2019).
17. Liu, S. *et al.* A high-quality assembly reveals genomic characteristics, phylogenetic status, and causal genes for leucism plumage of Indian peafowl. *GigaScience* **11**, 1–16 (2022).
18. Liu, Y. *et al.* Genome assembly of the common pheasant phasianus colchicus: A model for speciation and ecological genomics. *Genome Biol. Evol.* **11**, 3326–3331 (2019).
19. Chattopadhyay, B. *et al.* Novel genome reveals susceptibility of popular gamebird, the red-legged partridge (Alectoris rufa, Phasianidae), to climate change. *Genomics* **113**, 3430–3438 (2021).
20. González-Prendes, R., Pena, R. N., Richart, C., Nadal, J. & Ros-Freixedes, R. Long-read de novo assembly of the red-legged partridge (Alectoris rufa) genome. 2024.01.23.576805 Preprint at https://doi.org/10.1101/2024.01.23.576805 (2024).
21. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013).

13

22. Schmid, M. *et al.* Second report on chicken genes and chromosomes 2005. *Cytogenet. Genome Res.* https://doi.org/10.1159/000084205 (2005).
23. O'Connor, R. E. *et al.* Chromosome-level assembly reveals extensive rearrangement in saker falcon and budgerigar, but not ostrich, genomes. *Genome Biol.* https://doi.org/10.1186/s13059-018-1550-x (2018).
24. Palacios, C. *et al.* Genomic variation, population history, and long-term genetic adaptation to high altitudes in tibetan partridge (Perdix hodgsoniae). *Mol. Biol. Evol.* **40**, msad214 (2023).
25. Luo, H. *et al.* A high-quality genome assembly highlights the evolutionary history of the great bustard (Otis tarda, Otidiformes). *Commun. Biol.* **6**, 1–11 (2023).
26. Mirarab, S. *et al.* A region of suppressed recombination misleads neoavian phylogenomics. *Proc. Natl. Acad. Sci.* **121**, e2319506121 (2024).
27. Stiller, J. *et al.* Complexity of avian evolution revealed by family-level genomes. *Nature* https://doi.org/10.1038/s41586-024-07323-1 (2024).
28. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing genomic data quality and beyond. *Curr. Protoc.* **1**, e323 (2021).
29. Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
30. Zerbino, D. R., Frankish, A. & Flicek, P. Progress, challenges, and surprises in annotating the human genome. *Annu. Rev. Genomics Hum. Genet.* **21**, 55–79 (2020).
31. Wicker, T. *et al.* The repetitive landscape of the chicken genome. *Genome Res.* **15**, 126–136 (2005).
32. Kulak, M. *et al.* Genome organization of major tandem repeats and their specificity for heterochromatin of macro- and microchromosomes in Japanese quail. *Genome* **65**, 391–403 (2022).
33. Benham, P. M. *et al.* A highly contiguous genome assembly for the California quail (*Callipepla californica*). *J. Hered.* https://doi.org/10.1093/jhered/esad008 (2023).
34. Galbraith, J. D. & Hayward, A. The influence of transposable elements on animal colouration. *Trends Genet.* **39**, 624–638 (2023).
35. Mundy, N. I., Kelly, J., Theron, E. & Hawkins, K. Evolutionary genetics of the melanocortin-1 receptor in vertebrates. *Ann. N. Y. Acad. Sci.* **994**, 307–312 (2003).
36. Sevane, N., Cañon, J., Gil, I. & Dunner, S. Transcriptomic characterization of innate and acquired immune responses in red-legged partridges (*Alectoris rufa*): A resource for immunoecology and robustness selection. *PLoS One* https://doi.org/10.1371/journal.pone.0136776 (2015).
37. Sanchez-Donoso, I. *et al.* Massive genome inversion drives coexistence of divergent morphs in common quails. *Curr. Biol.* https://doi.org/10.1016/j.cub.2021.11.019 (2022).
38. Jackson, C. E. *et al.* Chromosomal rearrangements preserve adaptive divergence in ecological speciation. 2021.08.20.457158 Preprint at https://doi.org/10.1101/2021.08.20.457158 (2021).
39. Pimm, S. L. *et al.* The biodiversity of species and their rates of extinction, distribution, and protection. *Science* **344**, 1246752–1246752 (2014).
40. Prum, R. O. *et al.* A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **526**, 569–573 (2015).
41. Meseguer, A. S., Lobo, J. M., Ree, R., Beerling, D. J. & Sanmartín, I. Integrating fossils, phylogenies, and niche models into biogeography to reveal ancient evolutionary history: The case of hypericum (Hypericaceae). *Syst. Biol.* **64**, 215–232 (2015).
42. Kan, X.-Z. *et al.* Phylogeny of major lineages of galliform birds (Aves: Galliformes) based on complete mitochondrial genomes. *Genet. Mol. Res. GMR* **9**, 1625–1633 (2010).
43. Martínez-Fresno, M., Henriques-Gil, N. & Arana, P. Mitochondrial DNA sequence variability in red-legged partridge, Alectoris rufa, Spanish populations and the origins of genetic contamination from A. chukar. *Conserv. Genet.* **9**, 1223–1231 (2008).
44. Huerta-Cepas, J., Dopazo, H., Dopazo, J. & Gabaldón, T. The human phylome. *Genome Biol.* **8**, R109 (2007).
45. Castresana, J. Topological variation in single-gene phylogenetic trees. *Genome Biol.* **8**, 216 (2007).
46. Hahn, M. W. Bias in phylogenetic tree reconciliation methods: Implications for vertebrate genome evolution. *Genome Biol.* **8**, R141 (2007).
47. Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* **3**, 380–390 (2002).
48. Wang, N., Kimball, R. T., Braun, E. L., Liang, B. & Zhang, Z. Ancestral range reconstruction of Galliformes: The effects of topology and taxon sampling. *J. Biogeogr.* **44**, 122–135 (2017).
49. Chen, D. *et al.* Divergence time estimation of Galliformes based on the best gene shopping scheme of ultraconserved elements. *BMC Ecol. Evol.* **21**, 209 (2021).
50. Kimball, R. T., Hosner, P. A. & Braun, E. L. A phylogenomic supermatrix of Galliformes (Landfowl) reveals biased branch lengths. *Mol. Phylogenet. Evol.* **158**, 107091 (2021).
51. Wang, N., Kimball, R. T., Braun, E. L., Liang, B. & Zhang, Z. Assessing phylogenetic relationships among galliformes: A multigene phylogeny with expanded taxon sampling in Phasianidae. *PLoS One* **8**, e64312 (2013).
52. Brown, J. W., Wang, N. & Smith, S. A. The development of scientific consensus: Analyzing conflict and concordance among avian phylogenies. *Mol. Phylogenet. Evol.* **116**, 69–77 (2017).
53. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.
54. rrwick/Porechop: adapter trimmer for Oxford Nanopore reads. https://github.com/rrwick/Porechop.
55. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
56. rrwick/Filtlong: quality filtering tool for long reads. https://github.com/rrwick/Filtlong.
57. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
58. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* https://doi.org/10.1038/s41467-020-14998-3 (2020).
59. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
60. 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).
61. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2019).
62. Hu, J. *et al.* An efficient error correction and accurate assembly tool for noisy long reads. *bioRxiv* 2023.03.09.531669 (2023). https://doi.org/10.1101/2023.03.09.531669.
63. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).
64. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
65. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* https://doi.org/10.1101/gr.214270.116 (2017).
66. Jung, Y. & Han, D. BWA-MEME: BWA-MEM emulated with a machine learning approach. *Bioinformatics* **38**, 2404–2413 (2022).

67. Wick, R. R. & Holt, K. E. Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLoS Comput. Biol.* **18**, e1009802 (2022).
68. Pryszcz, L. P. & Gabaldón, T. Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* **44**, e113 (2016).
69. Astashyn, A. *et al.* Rapid and sensitive detection of genome contamination at scale with FCS-GX. 2023.06.02.543519 Preprint at https://doi.org/10.1101/2023.06.02.543519 (2023).
70. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 1–18 (2019).
71. Liao, X. *et al.* msRepDB: A comprehensive repetitive sequence database of over 80 000 species. *Nucleic Acids Res.* **50**, D236–D245 (2022).
72. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
73. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).
74. GitHub - 4ureliek/Parsing-RepeatMasker-Outputs: Few scripts facilitating the extraction of info from Repeat Masker .out files. https://github.com/4ureliek/Parsing-RepeatMasker-Outputs.
75. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 1–14 (2021).
76. Li, H. Protein-to-genome alignment with miniprot. *Bioinform. Oxf. Engl.* **3**, 9. https://doi.org/10.1093/bioinformatics/btad014 (2022).
77. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
78. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinform.* **3**, lqaa108 (2021).
79. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
80. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
81. Kalvari, I. *et al.* Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkaa1047 (2021).
82. Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
83. Huerta-Cepas, J. *et al.* eggNOG 50: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
84. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* https://doi.org/10.1093/nar/28.1.45 (2000).
85. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkv1189 (2016).
86. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
87. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
88. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
89. Geib, S. M. *et al.* Genome annotation generator: A simple tool for generating and correcting WGS annotation tables for NCBI submission. *GigaScience* **7**, giy018 (2018).
90. 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).
91. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
92. Pertea, G. & Pertea, M. GFF utilities: GffRead and GffCompare. *F1000research.* https://doi.org/10.12688/f1000research.23297.2 (2020).
93. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
94. MariaNattestad/dot: Dot: An interactive dot plot viewer for comparative genomics. https://github.com/marianattestad/dot.
95. Krzywinski, M. I. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* https://doi.org/10.1101/gr.092759.109 (2009).
96. Sun, J. *et al.* OrthoVenn3: An integrated platform for exploring and visualizing orthologous data across genomes. *Nucleic Acids Res.* **51**, W397–W403 (2023).
97. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).
98. He, C. *et al.* Chromosome level assembly reveals a unique immune gene organization and signatures of evolution in the common pheasant. *Mol. Ecol. Resour.* **21**, 897–911 (2021).
99. 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).
100. Alföldi, J. *et al.* The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477**, 587–591 (2011).
101. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkh340 (2004).
102. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btp348 (2009).
103. 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).
104. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* **19**, 153 (2018).
105. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msm088 (2007).
106. Kumar, S. *et al.* TimeTree 5: An expanded resource for species divergence times. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msac174 (2022).
107. Blair Hedges, S. & Poling, L. L. A molecular phylogeny of reptiles. *Science* **283**, 998–1001 (1999).

Acknowledgements

Author contributions

Funding

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-70018-0.

Correspondence and requests for materials should be addressed to R.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature portfolio 16