The gyrfalcon (*Falcon rusticolus*) genome

Andrea Zuccolo\(^1,2,^*\), Sara Mfarrej\(^3\), Mirko Celii\(^1\), Saule Mussurova\(^1\), Luis F. Rivera\(^1\), Victor Llaca\(^4\), Nahed Mohammed\(^1\), Arnab Pain\(^3\), Abdulmajeed Fahad Alrefaei\(^5\), Abdulwahed Fahad Alrefaei\(^6,^*\), Rod A. Wing\(^1,5,^*\)

\(^1\)-Center for Desert Agriculture (CDA), Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia

\(^2\)-Crop Science Research Center, Sant’Anna School of Advanced Studies, Piazza Martiri della Libertà 33, 56127 Pisa, Italy

\(^3\)-King Abdullah University of Science and Technology (KAUST), Pathogen Genomics Laboratory, Biological and Environmental Science and Engineering (BESE), Thuwal-Jeddah, 23955-6900, Saudi Arabia

\(^4\)-Research and Development, Corteva Agriscience, Johnston, Iowa 50131, USA

\(^5\)-Jamoum University College, Department of Biology, Umm Al-Qura University, Mecca 24382, Saudi Arabia

\(^6\)-Department of Zoology, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia

\(^7\)-Arizona Genomics Institute, School of Plant Sciences, University of Arizona, 24 Tucson, Arizona 85721, USA

* To whom correspondence should be addressed

Andrea Zuccolo, andrea.zuccolo@kaust.edu.sa, ORCID 0000-0001-7574-0714

Mirko Celii, mirko.celii@kaust.edu.sa, ORCID 0000-0001-9179-5758

Saule Mussurova, saule.mussurova@kaust.edu.sa

Luis Fernando Rivera, luis.riveraserna@kaust.edu.sa, ORCID 0000-0003-3978-7640

Sara Mfarrej, sara.mfarrej@kaust.edu.sa, ORCID 0000-0002-6460-8862

Victor Llaca, victor.llaca@corteva.com, ORCID 0000-0003-4822-2924

Nahed Mohammed, nahed.mohammed@kaust.edu.sa, ORCID 0000-0002-8857-3246

© The Author(s) 2023. Published by Oxford University Press on behalf of the Genetics Society of America. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Arnab Pain, arnab.pain@kaust.edu.sa, 0000-0002-1755-2819
Abdulmajeed Fahad Alrefaei, afrefaei@uqu.edu.sa, ORCID 0000-0003-0804-2339
Abdulwahed Alrefaei, afrefaei@ksu.edu.sa, ORCID 0000-0002-3761-6656
Rod A. Wing, rod.wing@kaust.edu.sa, ORCID 0000-0001-6633-6226

Keywords: gyrfalcon, *Falcon rusticolus*, conservation genomics, long reads, transposable elements, CR1, chromosome fusion

**ABSTRACT**

High-quality genome assemblies are characterized by high sequence contiguity, completeness, and a low error rate, thus providing the basis for a wide array of studies focusing on natural species ecology, conservation, evolutionary, and population genomics. To provide this valuable resource for conservation projects and comparative genomics studies on gyrfalcon (*Falco rusticolus*), we sequenced and assembled the genome of this species using third-generation sequencing strategies and optical maps. Here, we describe a highly contiguous and complete genome assembly comprising 20 scaffolds and 13 contigs with a total size of 1.193 Gbp, including 8,064 complete Benchmarking Universal Single-Copy Orthologs (BUSCOs) of the total 8,338 BUSCO groups present in the library aves_odb10. Of these BUSCO genes, 96.7% were complete, 96.1% were present as a single copy, and 0.6% were duplicated. Furthermore, 0.8% BUSCO genes were fragmented and 2.5% (210) were missing. A *de novo* search for transposable elements (TEs) identified 5,716 TEs that masked 7.61% of the *F. rusticolus* genome assembly when combined with publicly available TE collections. Long interspersed nuclear elements, in particular, the element Chicken-repeat 1 (CR1), were the most abundant TEs in the *F. rusticolus* genome. A *de novo* first-pass gene annotation was performed using 293,349 PacBio Iso-Seq transcripts and 496,195 transcripts derived from the assembly of 42,429,525 Illumina PE RNA-seq reads. In all, 19,602 putative genes, of which 59.31% were functionally characterized and associated with Gene Ontology terms, were annotated. A comparison of the gyrfalcon genome assembly with the publicly available assemblies of the domestic chicken (*Gallus gallus*), zebra finch (*Taeniopygia guttata*), and hummingbird (*Calypte anna*) revealed several genome rearrangements. In
particular, nine putative chromosome fusions were identified in the gyrfalcon genome assembly compared with those in the *G. gallus* genome assembly. This genome assembly, its annotation for TEs and genes and the comparative analyses presented, complement and strength the base of high-quality genome assemblies and associated resources available for comparative studies focusing on the evolution, ecology, and conservation of Aves.

INTRODUCTION

Of the extant tetrapod vertebrates, birds (Aves) are the most diverse lineage (Prum et al., 2015) and include at least 40 orders that comprise over 10,000 living species (Brusatte et al., 2015). These represent the extant members of an adaptive radiation period that occurred approximately 150 million years ago (Chiappe and Dyke, 2002). Aves are characterized by a high diversity in morphology, ecology, and behavior (Gill, 1995). The genus to which falcons belong is a part of the Falconinae subfamily of the family Falconidae and comprises 38 species that are widely distributed throughout Asia, North America, and Europe (Wink, 2018) (Figure 1). Falcons can be roughly categorized into four groups: Kestrels, Hierofalcons, Peregrine falcons, and Hobbies (Wink, 2018).

The genus *Falco* underwent rapid and recent diversification, and evolutionary studies have investigated the ecological and geological factors driving it. The divergence within the subfamily Falconinae was inferred to date back to approximately 16 million years, and the most species-rich genus, *Falco*, which comprises about 60% of all Falconidae species, began to diverge approximately 7.5 million years ago (Fuchs et al., 2015). The timescale over which the falcons diverged and diversified is comparable to that of early hominids (Cade and Digby, 1982). Falcons underwent several radiations, which led to a higher diversity than that of most genera of Aves (Gill and Donkser, 2019). The unique evolutionary history of the genus *Falco* offers the possibility to study its speciation mechanisms at different evolutionary stages. Advances in genomics allow a better molecular-level understanding of the evolutionary mechanisms involved in generating the large diversity of falcons by investigating their genome sequence, structure, and function.
Most bird species have diploid karyotypes containing approximately 80 chromosomes. Generally, they comprise 7–10 pairs of large- and medium-sized chromosomes (macrochromosomes), several microchromosomes (30–33 pairs), often morphologically indistinguishable and the sex chromosomes (Masabanda et al., 2004). The karyotype of the Falconidae of the order Falconiformes is markedly different from this general pattern. The chromosome number per diploid genome is low and ranges from 40 chromosomes in merlin (Falcon columbarius) to 52 in common kestrel (F. tinnunculus). Additionally, the macrochromosomes show little size difference, and the number of microchromosomes is low. The fusion of microchromosomes with the macrochromosomes is the likely mechanism that leads to low chromosome counts in falcons. Indeed, tandem fusions of microchromosomes with the macrochromosomes and those between microchromosomes have frequently been observed (Nishida et al., 2008).

The gyrfalcon (F. rusticolus) is the largest and one of the fastest flying falcon. In both sexes, the average body length ranges from 41 to 56 cm, and the average body weight is between 800 and 2,100 g (Figure 1). The species is polymorphic; hence, its plumage varies greatly in color according to the environment in which it lives; it can have white, black, brown, or dark brown feathers (Del Hoyo, 1994).

Substantial advances in sequencing technology coupled with efficient assembly strategies and the availability of long-range sequencing technologies, such as optical mapping and HiC chromatin interaction-based analyses, have dramatically increased the overall quality of genome assemblies (Sedlazeck et al., 2018). Such high-quality genome assemblies constitute valuable resources for any through investigation of wild and domesticated species ecology, conservation, evolution and population genetics (Whibley et al., 2021). To provide genomic resources for comparative and conservation projects, we built an accurate and complete genome sequence of gyrfalcon using third-generation sequencing strategies with the support of optical maps. We obtained a genome assembly characterized by high contiguity and completeness, including 33 scaffolds and contigs of a total size of 1.193 Gbp. The Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis identified 8,064 complete BUSCOs of the 8,338 Aves BUSCO groups, corresponding to 96.7% complete and 0.8% fragmented genes. The genome assembly was searched de novo for transposable elements (TEs), resulting in the identification of 5,716 TEs. Together with
publicly available resources, these masked 7.61% of the entire genome assembly. Furthermore, 19,602 putative genes resulted from a first-pass annotation with the software Augustus (Stanke and Waack, 2003) using 293,349 PacBio Iso-Seq transcripts and 496,195 transcripts derived from the assembly of 42,429,525 Illumina paired-end RNA-seq reads as extrinsic support. Whole-genome comparisons with publicly available genomes of the domestic chicken (Gallus gallus, Warren et al., 2017), zebra finch (Taeniopygia guttata, Warren et al., 2010), and hummingbird (Calypte anna, Rhie et al., 2021) revealed several genome rearrangements encompassing all classes of structural variants that included nine chromosome fusions compared with the G. gallus genome. This genome assembly adds to those already available for birds (Bravo et al., 2021), specifically falcons (Zhan et al., 2013; Cho et al., 2019; Doyle et al., 2018; Wilcox et al., 2022) (Figure 1), complementing and strengthening the base of high-quality genome assemblies for comparative studies focusing on the evolution, ecology, and conservation of Aves.

MATERIALS AND METHODS

-DNA extraction: High molecular weight DNA was extracted from the blood of an Arabian two-year-old female gyrfalcon weighing 1,400 grams using the Nanobind magnetic disc-based method from Circulomics (Circulomics, Baltimore, MD, USA). The optimization included efficient homogenization of the blood with Phosphate-buffered saline (PBS) to avoid a viscous lysate and ensure proper lysis, considering avian erythrocytes contain high amounts of nuclei. 50 μl avian blood provided more than 20 μg of DNA yield, quantified using broad range Qubit fluorometer (Invitrogen, USA). The DNA fragment size was 120 kb on the pulsed field gel electrophoresis.

-RNA extraction: RNA was extracted from falcon blood using Zymo-Direct Zol kit (Zymobiomics, Zymo Research Corporation, Irvine, CA, USA) with DNase I treatment. To get a clean aqueous RNA layer, the TRIzol (Invitrogen, USA) lysis incubation was extended, and chloroform was added during the initial lysis step. The quality control of purified RNA was performed using broad range Qubit kit (Invitrogen, USA) and RNA 6000 Nano LabChip kit (Agilent, Santa Clara, CA, USA) respectively, with an RNA integrity number of 9.

-Pac-Bio Libraries construction: PacBio DNA read and RNA Iso-Seq libraries were constructed by the KAUST Bioscience Core Lab using standard PacBio protocols.
The cDNA synthesis was done using NEBNext Single Cell/Low input cDNA synthesis & amplification module (NEB, Cat No: E6421S) with Iso-Seq Express Oligo Kit (Pacbio, 101-737-500) and Iso-seq library was built with SMRTbell Express Template Prep Kit 2.0 (PacBio, 100-938-900).

-DNA Sequencing: HMW gDNA was sequenced using the PacBio single-molecule real-time (SMRT) platform. Two SMRT cells were run for 15hr per cell with the continuous long read method (CLR) generating 13.93 million reads for a total of 345.3 Gbp of sequence data and providing an estimated ~287.8× coverage of the F. rusticolus genome, assuming a genome size of 1.2 Gbp (Wilcox et al., 2019).

-Assembly: Raw sequence data were assembled using two assemblers: MECAT2 (Xiao et al., 2017) and CANU v.1.8 (Koren et al., 2017). Both were run using default settings, assuming a genome size of 1.2 Gbp for the gyrfalcon genome.

-Polishing: Both the CANU and MECAT2 assemblies were subjected to two polishing rounds: First, PacBio reads were mapped onto each assembly using the software Blasr (Chaisson and Tesler, 2012). Subsequently, the software Arrow was used for polishing, as implemented in SMRTlink v.9.0 (Pacific Bioscience, Menlo Park, CA, USA) with default settings. Next, 412.7 million paired-end Illumina reads (151 bp/end) using DNA obtained from the same gyrfalcon blood sample were mapped onto both assemblies using bwa-mem (Li, 2013), followed by polishing with the software Pilon (Walker et al., 2014) under default settings.

-Quality evaluation: The initial assessment of the genome assembly metrics was conducted using the software Quast (Gurevich et al., 2013).

-Bionano library and hybrid scaffolding: bionano optical maps was generated by Corteva Agriscience (Johnston, IA, USA). The genome assembly and the corresponding Bionano maps were employed to build a hybrid scaffold using the software Bionano solve v.3.4 (https://bionanogenomics.com/) with the following settings: -B 2 -N 2 -f. Data visualization was performed using Bionano Access (v12.5.0) software package.

-Iso-Seq data: RNA was sequenced using the PacBio single-molecule real-time (SMRT) platform using Sequel® II Binding Kit 2.1(Pacbio, 101-843-000), Sequel II Sequencing Kit 2.0 (101-820-200), and SMRT cell 8M Tray (101-389-001). Full-length isoforms were extracted using the tools included in SMRTlink v.9.0 (Pacific Bioscience) to perform the following analyses: CCS extraction from CLR transcriptome reads using the tool CCS (v.4.2.0);
barcode demultiplexing using the tool lima (v.1.11.0); removal of polyA and concatemers from
full-length reads using the software isoseq3 refine (v.3.3.0) and clustering sequences using
isoseq3 cluster software (v.3.3.0). Only the high-quality reads obtained through this pipeline
were used for subsequent analyses. Iso-Seq reads were mapped onto the genome assembly
using the tool minimap2 (Li, 2018) with the setting -ax splice:hq.

-RNA-seq: Paired-end RNA reads (42,429,525 151 bp) were assembled using the software
Trynity v.2.8.5 (Grabherr et al., 2011). Predicted transcripts were then clustered using the
software cd-hit (Li and Godzik, 2006) with the following settings: -c 0.98 -p 1 -d 0 -b 3.

-Comparison with the other bird genome-scale assemblies: The gyrfalcon genome assembly
was compared with those of the domestic chicken (GenBank assembly accession: 
GCA_000002315.5), hummingbird (GenBank assembly accession: GCA_003957575.1), and
zebra finch (GCA_000151805.2) using the software Mashmap (Jain et al., 2018) and D-
genies (Cabanettes and Flopp, 2018).

-BUSCO assessment: BUSCO evaluations were performed for the gyrfalcon genome
assembly using the BUSCO (v.5.1.2) software package (Simao et al., 2015) and the

-Circa plots: Circa plots were constructed using the Circa software available at
https://omgenomics.com/circa/.

-Gene prediction and functional annotation: Gene prediction was performed using the
software Augustus (Stanke and Waack, 2003), as implemented in the tool OmicsBox
(BioBam, 2019), with gene models calculated for the domestic chicken and the Iso-Seq reads
described above as extrinsic support. Functional annotation was performed, considering the
best match for each transcript calculated using Diamond BLASTp (v.0.9) searches of
predicted proteins against the nonredundant NCBI protein database. An e-value of 1e-6 was
used as the threshold for significant matches. Gene ontology analysis was performed using
InterProScan v.5.39 with default settings (Jones et al., 2014).

-TE identification and quantification: Using default settings, a TE library was obtained by
running the software EDTA (Ou et al., 2018) on the gyrfalcon genome assembly. The
RepBase (Bao et al., 2015) vertebrate TE library (July 2020) was combined with EDTA-
predicted TEs to generate the de novo FALCON_TE_LIBRARY_V3.fa. Quantification of TEs
in the F. rusticolus genome was performed with RepeatMasker software (Smit et al., 2015)
using the FALCON_TE_LIBRARY_V3.fa with default settings, except for the fact that the
option -qq was used to hasten the search.

RESULTS AND DISCUSSION

Genome assembly and statistics
The gyrfalcon genome was sequenced using two smart cells with PacBio Sequel sequencing
technology. A total of 14,329,526 reads were produced, amounting to 351.8 Gbp of sequence
data. Assuming an estimated genome size of ~1.2 Gbp (Wilcox et al., 2019), the sequence
data provided ~293× genome coverage. The PacBio reads were used as input for two
different genome assemblers, MECAT2 (Xiao et al., 2017) and CANU (Koren et al., 2017),
followed by two rounds of polishing using the Illumina data and the PacBio reads (see
Methods). The contigs’ N50 and L50 values were 48.17 Mbp and 9 for CANU and 39.78 Mbp
and 10 for MECAT2, respectively (Table 1). We used CANU for the baseline falcon genome
assembly based on these metrics.

Construction and quality assessment of the edited hybrid genome assembly
The baseline genome assembly was then used with Bionano optical maps to produce a
hybrid assembly using the Bionano solve software (www.bionanogenomics.com). The hybrid
assembly comprised 70 super-scaffolds manually inspected for possible inconsistencies with
the optical map. Conflicts were resolved by “breaking” the questionable super-scaffolds when
necessary. This process resulted in a dataset consisting of 20 super-scaffolds and 13 contigs.
Altogether, these 33 contigs and super-scaffolds represent the genome assembly of gyrfalcon
(Table 2). These were numbered according to length, from the longest (122.3 Mbp) to the
shortest (1.071 Mbp). The entire genome assembly has six gaps ranging from 183 bp to
113,856 bp. The total length of these gaps is 384,827 bp. The full length of the genome
assembly is 1.193 Gbp, which is highly consistent with the estimated genome size of 1.2 Gbp
for gyrfalcon (Wilcox et al., 2019).

To assess the completeness of our final assembly, we conducted a BUSCO analysis
identifying 8,064 complete genes out of the 8,338 included in aves_odb10, which corresponds
to 96.7% of the complete genes. Of these 96.1% were single copies, and 0.6% were
duplicated. There were 0.8% fragmented and 210 (2.5%) missing genes.
The hybrid assembly included another 42 small super-scaffolds shorter than 1.8 Mbp, totaling 11.5 Mbp, which did not show any significant similarity with the genome of *G. gallus*. These short super-scaffolds could include unassigned microchromosomes and unique and highly divergent regions of the gyrfalcon genome or could be artifactual products of the hybrid scaffolding process generated mainly because of highly repetitive simple sequences.

Comparative genomics with domestic chicken, zebra finch, and hummingbird

Comparisons with the domestic chicken genome: identification of chromosomal rearrangements

Comparisons between the *F. rusticolus* genome and the other three high-quality bird reference genomes showed a remarkable level of conservation in terms of sequence synteny and overall sequence similarity. Long stretches of contiguous sequences with an overall sequence similarity of >85% were observed (Figures S1, S2, and S3). This high overall level of conservation among the four species is remarkable, considering the fact that they diverged >60 million years ago (Prum et al., 2015; Cho et al., 2019).
However, major chromosomal rearrangements, including structural variants such as inversions and translocations, were identified, along with evidence of chromosomal fusions and breakages. Particularly, 9 of the 33 falcon genome assembly contigs and super-scaffolds appeared to result from chromosome fusions compared with the chicken genome (Table 2). Seven contigs demonstrated homology to two chicken chromosomes each, one to three chicken chromosomes, and one (2_sc, whose total length is 104.5 Mbp) included regions homologous to tracts from four different chicken chromosomes (Figure 2). Tracts of the super-scaffold 2_sc from 16.3 Mbp to 35.9 Mbp and from 35.9 Mbp to 42.2 Mbp showed similarity to the entire length of chicken chromosomes 14 and 28, respectively. Two other regions were homologous to extensive stretches of chicken chromosome 12 (from the beginning of 2_sc to 16.3 Mbp) and chromosome 2 (from 42.2 Mbp to the end of 2_sc). All genome assembly regions in which potential rearrangements were detected were independently validated with optical maps, thereby eliminating any possible misassembled artifacts (Figures S4 to S11).

The large number of genome rearrangements in the gyrfalcon genome was consistent with the evidence collected in other species of the Falcon genus, such as *F. peregrinus* (Nishida et al., 2008; O’Connor et al., 2018; Penalba et al., 2019), indicating that the Falconiformes karyotype is an exception to the avian karyotypes otherwise characterized by a limited number of interchromosomal changes (Damas et al., 2018).

Of note, five contigs (24_co, 25_co, 27_co, 28_co, 30_co) which are not part of sex chromosomes showed features typical of microchromosomes such as high GC content, high gene density and TE depletion (McQueen et al., 1998; Warren et al., 2017; Waters et al., 2021). Specifically, they had a length ranging from 2.47 Mbp to 6.06 Mbp, a GC content (50.27%-57.69%) significantly higher than the average one (42.83%), a TE content (3.05% to 6.40%) lower than average (7.61%) and a gene density (40.76-83.30 genes/Mbp) strikingly higher than the average one *i.e.* 16.4 genes/Mbp (Table 2).

Interestingly, the comparative analysis of 28_co (Figure S5) showed a convincing similarity of this sequence to the microchromosomes 25 and 33 of *G. gallus* providing a suggestive example of microchromosomes fusion leading to chromosome reduction number as described in Joseph et al., 2018. These kinds of rearrangements involving microchromosomes are not frequent in birds with the notable exceptions of Falconiformes and Psittaciformes (Joseph et al., 2018; O’Connor et al., 2018; Kretscher et al., 2021).
**TE identification, abundance, and distribution**

Using the EDTA pipeline (Ou et al., 2019), a library of TEs was created for *F. rusticolus*. This library included 5,716 entries (Supplementary file 2) and was used together with the RepBase (Bao et al., 2015) vertebrate TE library (5,038 entries) to mask the falcon genome assembly. Altogether, the TE library included 10,759 entries (Table S1) and repeat-masked 7.61% of the genome assembly (Table 3).

The overall amount of TEs in the falcon genome was similar to that observed in the genomes of other birds and confirmed the underrepresentation of TEs in these genomes compared with those in the other metazoans. Notably, the overall number of TEs is higher than that estimated in previous studies on falcon species. Zhan et al., in 2013, estimated the amount of the repetitive fraction (not limited to TEs) in both saker and peregrine falcons to be 6.80%. Zhang et al., in 2014, estimated the TE content of *F. peregrinus* to be 5.50%. Recently, Watson et al. (2022) annotated an average of 6.5% of the genome of eight Falcon species as TE-related. However, the amount of TEs was highly variable across the different scaffolds and contigs and ranged from 3.05% to 59.98% (Table 2 and Figure 3). For example, three contigs/scaffolds tentatively assigned to chromosome W had a TE content of >50%. The abundance of different TE types was consistent with that observed in other birds; all the main TE classes were represented, with LINEs being the most abundant (2.66%).

Most LINEs in the gyrfalcon assembly exhibited high similarity with the element “Chicken repeat 1” (CR1), an element highly abundant in most published bird genomes. CR1 retrotransposons are the most common family of TEs found in the genomes of birds, crocodilians, turtles, and snakes (Suh et al., 2017). The length of the complete CR1 element is approximately 4.5 kbp; however, most copies are incomplete. Indeed, of all the CR1 paralogs in the chicken, only 0.6% are complete (International Chicken Genome Sequencing Consortium, 2004). Complete CR1 elements encode for two proteins: an RNA-binding protein (ORF1p) and a multifunctional protein with endonuclease and reverse transcriptase enzymatic activities (ORF2p). A TblastN search of the gyrfalcon genome, using a 100-amino acid tract of the CR1 ORF2 as a query, showed 13,124 significant (e-value 1e-5) hits. Theoretically, if all these hits corresponded to complete full-length elements, it would translate to 59.1 Mbp of sequence, amounting to ~5% of the genome assembly. This is not the case in
our assembly because the entire LINE complement totaled 2.66% of the assembly, thereby indicating the presence of many incomplete CR1 elements in the gyrfalcon genome.

Gene annotation

De novo gene identification analysis was performed on the falcon genome assembly masked for TEs. The search used the gene predictor Augustus (Stanke and Waack, 2003) in the OmicsBox (BioBam, 2019) software package. As extrinsic support data, 293,349 RNA high-quality PacBio Iso-Seq sequences and 496,195 transcripts derived from the assembly of 42,429,525 Illumina PE RNA-seq reads were used. The Hidden Markov Model (HMM) used for the ab initio search was that computed for G. gallus. The search identified 19,602 putative coding regions (Figure 3), 2,069 of which appeared to be monoexonic. This gene count is on par with the values obtained for other falcons, such as F. peregrines (16,263 genes) and F. cherrug (16,204) (Zhan et al., 2013) and is consistent with the number of genes in many bird genomes (Zhang et al., 2014).

The average gyrfalcon gene length was 20.8 kbp, with a median of six exons per gene and an average of eight exons per gene. The overall number of introns was 146,180, with an average length of 2,606 bp. Of the predicted genes, 17,084 (87.15%) had InterProScan positive hits and 10,625 (59.31%) could be functionally characterized and associated with Gene Ontology terms. A description of the locations and structures of the predicted genes is available in the Supplementary file 1. As further support to the genome assembly completeness, it is worth noting that 291,029 Iso-Seq reads of the 293,249 (99.25%) reads were mapped onto the gyrfalcon assembly.

DATA AVAILABILITY

The genomics data presented in this manuscript were submitted to the National Center for Biotechnology Information with the BioProject accession number BioProject ID PRJNA872351. The genome assembly project has been deposited at DDBJ/ENA/GenBank under the accession JAPSEQ000000000. The version described in this paper is version JAPSEQ010000000.
Supplementary file 1 and Supplementary file 2 are available in figshare: https://doi.org/10.25387/g3.21769628.

**FUNDING**

This project was funded in part by KAUST baseline funding to R.A.W., and Green Sustainability Company for Environmental Services (GSCES).

**Acknowledgments**

The authors extend their appreciation to Alnadir for Falcons Center in Jeddah for assistance with samples provision, and Green Sustainability Company for Environmental Services (GSCES) for their support.

**REFERENCES**


Gill F & Donsker D. 2019. IOC world bird list (v 9.1).


Joseph S, O’Connor RE, Al Mutery AF, Watson M, Larkin DM, Griffin DK. 2018. Chromosome Level Genome Assembly and Comparative Genomics between Three Falcon Species Reveals an Unusual Pattern of Genome Organisation. Diversity. 10:113


rearrangement in saker falcon and budgerigar, but not ostrich, genomes. Genome Biol. 19(1),171.


### Table 1. Gyrfalcon genome assembly metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>MECAT2</th>
<th>CANU</th>
</tr>
</thead>
<tbody>
<tr>
<td># contigs</td>
<td>1,161</td>
<td>1,219</td>
</tr>
<tr>
<td># contigs &gt; 50,000 bp</td>
<td>539</td>
<td>938</td>
</tr>
<tr>
<td>Total length (bp)</td>
<td>1,250,464,509</td>
<td>1,308,989,646</td>
</tr>
<tr>
<td>Largest contig (bp)</td>
<td>104,620,499</td>
<td>122,222,403</td>
</tr>
<tr>
<td>GC (%)</td>
<td>42.55</td>
<td>42.83</td>
</tr>
<tr>
<td>N50 (bp)</td>
<td>39,771,069</td>
<td>48,172,526</td>
</tr>
<tr>
<td>N75 (bp)</td>
<td>29,809,375</td>
<td>29,473,197</td>
</tr>
<tr>
<td>L50</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>L75</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>Ns per 100 kbp</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Falcon assembly</td>
<td>Number of predicted genes</td>
<td>Genes/ Mbp</td>
</tr>
<tr>
<td>----------------</td>
<td>---------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>1_sc</td>
<td>122,326,283</td>
<td>1,254</td>
</tr>
<tr>
<td>2_sc</td>
<td>104,580,349</td>
<td>1,945</td>
</tr>
<tr>
<td>3_sc</td>
<td>104,319,969</td>
<td>974</td>
</tr>
<tr>
<td>4_sc</td>
<td>95,117,127</td>
<td>1,302</td>
</tr>
<tr>
<td>5_sc</td>
<td>91,389,740</td>
<td>1,013</td>
</tr>
<tr>
<td>6_co</td>
<td>77,507,961</td>
<td>1,298</td>
</tr>
<tr>
<td>7_sc</td>
<td>73,427,893</td>
<td>1,271</td>
</tr>
<tr>
<td>8_sc</td>
<td>66,236,224</td>
<td>1,095</td>
</tr>
<tr>
<td>9_sc</td>
<td>58,459,340</td>
<td>690</td>
</tr>
<tr>
<td>10_sc</td>
<td>55,126,675</td>
<td>1,031</td>
</tr>
<tr>
<td>11_sc</td>
<td>37,966,174</td>
<td>925</td>
</tr>
<tr>
<td>12_co</td>
<td>35,775,413</td>
<td>671</td>
</tr>
<tr>
<td>13_sc</td>
<td>34,096,887</td>
<td>511</td>
</tr>
<tr>
<td>14_co</td>
<td>31,933,560</td>
<td>1,050</td>
</tr>
<tr>
<td>15_co</td>
<td>29,744,985</td>
<td>559</td>
</tr>
<tr>
<td>16_sc</td>
<td>27,799,139</td>
<td>302</td>
</tr>
<tr>
<td>17_sc</td>
<td>24,209,053</td>
<td>456</td>
</tr>
<tr>
<td>18_sc</td>
<td>23,548,544</td>
<td>485</td>
</tr>
<tr>
<td>19_sc</td>
<td>17,715,289</td>
<td>635</td>
</tr>
<tr>
<td>20_co</td>
<td>14,691,556</td>
<td>148</td>
</tr>
<tr>
<td>21_sc</td>
<td>12,341,734</td>
<td>106</td>
</tr>
<tr>
<td>22_co</td>
<td>8,268,704</td>
<td>221</td>
</tr>
<tr>
<td>23_sc</td>
<td>8,146,813</td>
<td>192</td>
</tr>
<tr>
<td>24_co</td>
<td>6,660,032</td>
<td>343</td>
</tr>
<tr>
<td>25_sc</td>
<td>6,398,898</td>
<td>251</td>
</tr>
<tr>
<td>26_sc</td>
<td>5,459,845</td>
<td>104</td>
</tr>
<tr>
<td>27_co</td>
<td>5,078,416</td>
<td>207</td>
</tr>
<tr>
<td>28_co</td>
<td>4,212,995</td>
<td>350</td>
</tr>
<tr>
<td>29_sc</td>
<td>3,107,971</td>
<td>22</td>
</tr>
<tr>
<td>30_co</td>
<td>2,469,339</td>
<td>130</td>
</tr>
<tr>
<td>31_co</td>
<td>2,304,850</td>
<td>29</td>
</tr>
<tr>
<td>32_co</td>
<td>2,015,512</td>
<td>21</td>
</tr>
<tr>
<td>33_co</td>
<td>1,071,326</td>
<td>11</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1,193,708,382</td>
<td>19,602</td>
</tr>
</tbody>
</table>
### Table 3. Gyrfalcon transposable element content

<table>
<thead>
<tr>
<th>TE class</th>
<th>Count</th>
<th>bp Masked</th>
<th>%masked</th>
</tr>
</thead>
<tbody>
<tr>
<td>hAT</td>
<td>4,405</td>
<td>906,713</td>
<td>0.07%</td>
</tr>
<tr>
<td>CACTA</td>
<td>72,610</td>
<td>16,829,073</td>
<td>1.41%</td>
</tr>
<tr>
<td>Harbinger</td>
<td>7,017</td>
<td>1,206,649</td>
<td>0.10%</td>
</tr>
<tr>
<td>Mutator</td>
<td>29,652</td>
<td>6,480,296</td>
<td>0.54%</td>
</tr>
<tr>
<td>Mariner</td>
<td>8,693</td>
<td>1,510,257</td>
<td>0.13%</td>
</tr>
<tr>
<td>Helitron</td>
<td>5,461</td>
<td>3,135,059</td>
<td>0.26%</td>
</tr>
<tr>
<td>Other DNA TE</td>
<td>9,200</td>
<td>814,246</td>
<td>0.07%</td>
</tr>
<tr>
<td>LINE</td>
<td>57,441</td>
<td>29,573,603</td>
<td>2.66%</td>
</tr>
<tr>
<td>LTR-RT Copia</td>
<td>146</td>
<td>58,429</td>
<td>0.00%</td>
</tr>
<tr>
<td>LTR-RT-Gypsy</td>
<td>15,529</td>
<td>8,696,215</td>
<td>0.73%</td>
</tr>
<tr>
<td>LTR-RT</td>
<td>19,187</td>
<td>9,510,118</td>
<td>0.80%</td>
</tr>
<tr>
<td>SINE</td>
<td>3,986</td>
<td>524,148</td>
<td>0.04%</td>
</tr>
<tr>
<td>Others</td>
<td>37,396</td>
<td>11,572,395</td>
<td>0.80%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>270,723</td>
<td>90,817,201</td>
<td>7.61%</td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

Figure 1. A) Picture of a gyrfalcon B) Placement of falcons in the avian tree of life (modified and simplified from Prum et al., 2015 and Wink, 2018) C) Phylogenetic analysis of falcons modified and simplified from Wink, 2018. For the species that have been sequenced (indicated by a double-helix) we provided along to the genome assembly accession number, the genome assembly details in parentheses as follow: PA: number of primary assemblies; S: number of assemblies at scaffold level; C: number of assemblies at chromosome level; I: Illumina technology; P: PacBio technology; H: Hi-C chromatin interaction data; B: bionano optical maps:

For *F. pelegrinoides* which was sequenced but is not included in the phylogenetic tree the genome assembly information are: PA (1), S(1), I

Figure 2. Details of predicted chromosomal rearrangements in the gyrfalcon super-scaffold 2_sc. A) Circa plot comparing 2_sc with the entire set of domestic chicken chromosomes (specified as “gg_chromosome number”). Regions showing significant similarity are connected by violet lines. B) Bionano optical map validation for 2_sc. NGS: assembled sequence; BNG: Bionano map. C) Dot plot of 2_sc versus four chicken chromosomes showing homology. 2_sc is on the x-axis, and chicken chromosomes are on the y-axis. The chicken chromosomes are coded using the color assigned to them in A.

Figure 3. Circa plot of the *Falcon rusticolus* genome assembly showing (outer circle inward) GC content distribution, gene density, and TE content.
Figure 1
406x356 mm (3.4 x DPI)
Figure 2
406x203 mm (3.4 x DPI)
Figure 3

133x133 mm (3.4 x DPI)