Multiple chicken (Gallus gallus) genome references to advance genetic variation studies.

Multiple chicken (*Gallus gallus*) genome references to advance genetic variation studies.

Wesley C. Warren[1], Olivier Fedrigo[2], Alan Tracey[3], Andrew S. Mason[4], Giulio Formenti[2], Francesco Perini[5], Zhou Wu[6], Terence Murphy[7], Valerie Schneider[7], Kyle Stiers[8], Edward S. Rice[1], Lyndon M. Coghill[8], Nick Anthony[9], Ron Okimoto[9], Rachel Carroll[1], Jacquelyn Mountcastle[2], Jennifer Balacco[2], Bettina Haase[2], Chentao Yang[10], Guojie Zhang[11], Jacqueline Smith[6], Yvonne Drechsler[12], Hans Cheng[13], Kerstin Howe[3], Erich D. Jarvis[2]

[1]Department of Animal Sciences, Data Science and Informatics Institute, University of Missouri, Columbia, MO 65201; [2]The Rockefeller University, Box 54, 1230 York Avenue, New York, New York 10065; [3]Tree of Life, Wellcome Sanger Institute, Cambridge CB10 1SA, UK; [4]Department of Biology, The University of York, Wentworth Way, York, YO10 5DD, UK; [5]Department of Agricultural, Food and Environmental Sciences, University of Perugia, Borgo XX Giugno, 74, 06121 Perugia (PG), Italy; [6]Department of Genetics and Genomics, The Roslin Institute, University of Edinburgh, Easter Bush Campus, Midlothian EH25 9RG, UK; [7]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA; [8]Department of Veterinary Pathology, University of Missouri, Columbia, MO, USA; [9]Cobb-Vantress, PO Box 1030, Siloam Springs, AR, USA; [10]BGI-Shenzhen, Shenzhen, 518083, China; [11]Center for Evolutionary and Organismal Biology, Zhejiang University School of Medicine, Hangzhou, 310058, China; [12]College of Veterinary Medicine, Western University of Health Sciences, 309 E. Second St, Pomona, CA 91766; [13]Avian Disease Biology Laboratory, Michigan State University, East Lansing, MI.

Corresponding author: warrenwc@missouri.edu
Running title: *Single haplotype chicken genome assemblies*.

Keywords: Genome assembly, variant detection, chicken

We present two phased chromosome-scale assemblies of chicken, a layer (GRCg7w) and broiler (GRCg7b), that better meet research demands to characterize segregating variation important for traits of interest. Annotation with existing long- and short-read RNAseq data improved contiguity, accuracy, and protein-coding and non-coding gene counts, when compared to the existing Red Jungle Fowl reference, GRCg6a. Most striking were the improvements in placed telomeres, corrections for erroneous microchromosome fusions, and gap reduction in these phased assemblies. We add six putative microchromosomes that were previously missing in GRCg6a. Using a pairwise genome comparison of the parental genomes, and two independent cohorts of sequenced chickens, we show small discernable differences in mapping rates of whole genome sequence (WGS) and RNAseq data, gene annotation, and called single nucleotide variants (SNVs) or indels. Structurally, some regional differences suggest future assembly curation will further improve variant ascertainment. These *Gallus* references also enabled a new genome-wide review of endogenous Avian Leukosis Virus (ALVE) integrations, exemplifying the improved representation of chicken genomic diversity by these phased genomes. Our genome references will collectively improve computational outcomes when testing multiple variant hypotheses that are at the core of understanding avian biology.

Today, the poultry industry faces many challenges, perhaps none more than the genetics underlying bird health. A constant balance must be maintained to select on several traits of immense economic impact, such as fast growth in broilers and reproductive success in layers, while not diminishing disease resistance. Genetic studies offer promising avenues to selectively maintain this trait balance with a new accounting of the most important contributing factors: genes and environment (Wolc et al. 2018). More complete and accurate genomic resources to support their continued discovery of these factors is paramount to generating the robust chicken germlines that can meet a growing demand for this protein food source.

The chicken also supports a vast model organism community that uses a collateral source of scientific data to comparatively inform developmental biology (see review (Cheng and Burt 2018)). The chicken genome is also one of the most frequently used resources for comparative genomic studies among vertebrates (Zhang et al. 2014). As the principal avian reference genome, it was used to transfer gene annotation evidence to over 50 bird genomes, which were in turn used for clade- and species-specific signals of genome evolution (Zhang et al. 2014; Jarvis et al 2014). Recent research attempts to determine the effect of structural variation (SV) on chicken phenotypic differences, although resolution beyond short-read mapping or hybridization methods must be considered (Rao et al. 2016). The site-directed gene knockouts of chicken *C2EIP* (Zuo et al. 2016) and *PAX* (Gandhi et al. 2017) genes are two functional instances of gained insight into embryonic germ and satellite muscle cell differentiation, respectively.

Since its first iteration in 2004 (Consortium 2004), we have worked to refine the assembly of the chicken genome as technology has progressed over the past two decades (Korlach et al. 2017). We recently

summarized these advances (Rhie et al. 2021), like the single haploid phasing of a diploid genome, i.e., trio binning (Koren et al 2018), which sorts and independently assembles divergent parental haplotypes from F1 hybrids (inter- and intraspecies crosses) as a highly efficient method for untangling complex sequence assembly graphs. This phasing strategy exploits the higher heterozygosity observed in some F1 hybrids in order to resolve diploid genomes more precisely and with fewer gaps. Several successful haplotype-resolved de novo assemblies for cats, cattle, zebra finches, and others have been created using this method (Bredemeyer et al. 2021) (Rhie et al. 2021) (Rice et al. 2020) (Koren et al 2018), highlighting the astounding improvements in contiguity, with some sequences spanning from telomere to telomere.

Recent characterization of different chicken genomes has demonstrated the necessity for pangenome resources in order to comprehend the comparative evolution of the *Gallus* genus (Li et al. 2022). To date, all chicken genetic studies have relied on the Red Jungle Fowl (RJF) genome, which portrays this diploid genome as a collapsed haploid genome containing a mixture of sequences from the two haplotypes. For the further investigation of trait selection indices, the adoption of additional high-quality chicken references, particularly those that better resemble commercial birds, is highly endorsed by the avian community and has broad applicability. In addition, these resources enable pangenome techniques that will provide higher resolution for discovering SVs that are exclusive to decades of artificial selection. Here, we used to trio binning approach to present two novel haploid *de novo* assemblies for chicken lines with extremely diverse genetic histories: one bred for muscle growth (broiler) and the other for egg production (layer). Large structural adjustments among microchromosomes, overall gap reduction, extension of the W chromosome by adding the pseudo-autosomal region, better chromosome 16 (MHC region) representation, and enhanced telomere sequence placements are notable. We demonstrate these new assemblies' application in determining the extent of alignment, SNV identification, ALVE integration, and structural expansions and contractions in a small sample of chickens.

Sequencing and Assembly. A parent-offspring trio composed of a paternal layer, a maternal broiler, and a female F1 offspring was sequenced to create these assemblies. Briefly, the parents were sequenced with low-coverage Illumina reads (150bp) and the F1 was sequenced with 80x PacBio reads (12kb on average), and all reads were used as input to TrioCanu (see review of methods: (Rhie et al. 2021)). Similar to cross-species trio assembly of cattle and yak, the amount of haplotyped long reads phased from each parental breed source was extremely similar (49.5 and 50.2%) with a low number of unknowns (0.16 percent) (Rice et al. 2020). Broiler (n=676) and layer (n=688) birds had half as many constructed contigs as RJF (n=1,403) birds, indicating that >53 percent of prior gaps have been bridged (currently 878 in RJF). Manual curation with orthogonal evidence, including chromatin proximity (Hi-C) and Bionano optical maps, delineated error locations that were fixed, e.g. 260 and 63 missed joins in GRCg7b and GRCg7w (Suppl. Table 1).

Depending on the descriptive context, we use the assembled GenBank versions (GRCg6a, GRCg7b, and GRCg7w) and their common names (RJF, broiler, and layer) interchangeably throughout the remainder of this report.

While contig N50 length was comparable to GRCg6a (which also used Pacbio long read data to fill gaps), phasing and revised mapping data led to a 4.5-fold increase in N50 scaffold length and a 2-fold decrease in the number of unplaced sequences in broiler and layer assemblies (Table 1). The paternal layer contributes Z to the ZW sex chromosomes, while the maternal broiler was particularly chosen for her haplotype A mitochondrial genome and W. The female RJF reference is unique with a mitochondrial genome of haplogroup E. The layer Z chromosome is somewhat larger and contains more protein-coding genes than the BAC-curated GRCg6a version of Z (85.2 vs 80Mb; 1,492 vs 1,345 genes) (Bellott et al. 2010). Furthermore, the broiler W chromosome is more complete than the GRCg6a chromosome, which is 7.2Mb in size, due in part to the insertion of the pseudoautosomal region (PAR) that boosts its comparative utility (Suppl. Fig. 1). The initial about 500kb of the W chromosome show diploid coverage. We chose not to join this sequence to the beginning of Z since we lack precise coordinates, and this portion of Z is partially collapsed. In each phased reference, for the sake of completeness, Z and W were incorporated notwithstanding their parental origins. By searching the NCBI assembly archive for 'Gallus gallus', you can find all fully annotated assemblies (see data availability).

Assembly accuracy benchmarking. There are inherent assembly artifacts present in all reference genomes, including the human genome. With this knowledge, we wanted to estimate the detected errors in GRCg6a, given its extensive use in chicken genetic studies, and repair them in our new phased assemblies using a previously established iterative procedure (Howe et al. 2021). A greater number of GRCg7b and GRCg7w chromosomes exhibited telomere ends (24 and 13, respectively), than GRCg6a (just 3), demonstrating the much-improved completeness of the new assemblies. Among microchromosomes, we discovered many instances in which GRCg6a chromosomes were wrongly fused into a single chromosome instead of two distinct ones (Suppl. Table 1). The first 2 Mb of GRCg6a chr27 is not associated with chr27, but rather a variety of alignments to other chromosomes, including W and chr2, which, upon curation, accurately sizes this chromosome; 8 Mb as opposed to 5.2 Mb in GRCg7b (Fig. 1). Other errors include the fusion of chr31 and chr29 in GRCg6a, which is likely due to repeat sequences identified on the Hi-C heat map (Suppl. Fig. 2).

Avian microchromosomes show more frequent recombination, and thus the positive correlation between recombination and interspecies divergence observed in mammals is not seen in birds, at least at the resolution of whole chromosomes (Consortium 2004). If the origin of the microchromosomes was initiated by a number of random fission events that were channeled towards the present day

macro/microchromosome arrangement it was clear from earlier evaluations that there were not sufficiently long compositionally uniform regions in any of the sequenced avian genomes, that would satisfy some classifications, e.g., the classical isochore definition within microchromosomes (Waters et al. 2021). We now have corrected several assembly errors, mostly among the microchromosomes, to test more accurately these and other hypotheses regarding their evolution.

The chicken karyotype has a diploid number of 78 chromosomes, classified as a haploid autosome count of 10 macrochromosomes and 28 microchromosomes (Burt et al. 1999). In earlier chicken assemblies, microchromosomes 29 and 34-38 were absent, primarily due to the absence of linkage groups or physical maps that might assign missing scaffolds to any of these smaller chromosomes (Groenen et al. 2000), as well as difficulty in sequencing through high-GC rich microchromsomes. In both GRCg7b and GRCg7w, using long reads that get through GC-rich regions, as with zebra finch (Kim et al 2022), we identify these missing microchromosomes (Suppl. Fig. 3) and an additional microchromosome to a final total of 39 autosomes. Future cytogenetic evaluations or new combinatorial approaches that can yield telomere-to-telomere stepwise assembly of more complete chromosomes (Logsdon et al. 2021) will be necessary to rule out the possibility these nominated microchromosomes are not affiliated with other macro- or microchromosomes. Moreover, the availability of almost complete genome copies of these uniquely selected lines and others will drive reevaluations of all types of segregating variation in a pangenome-dependent manner (Siren et al. 2021).

Structural differences. To estimate the major structural differences among these phased references, we employed two methods: high resolution alignments to reveal major synteny differences using SyRi (Goel et al. 2019) and the predicted contractions and expansions of deletions, insertions, and repeat elements with different size distributions using Assemblytics (Nattestad and Schatz 2016). Across the chicken genome, differences in local chromosomal synteny were predominately one-to-one; however, when we discover discrepancies, they frequently occur towards chromosome ends, highlighting the difficult nature of placing sequences in these repetitive telomeric regions (Fig. 2). Regardless of their length distribution, Assemblytics alignment results show comparable total base size differences (Suppl. Table 2; Fig. 2; Suppl. Fig. 4). However, these differences vary by type, such as deletion versus insertion, which may be the result of numerous factors, including genetic diversity and assembly completeness and accuracy of each reference. When employing a phased assembly for pairwise broiler versus layer alignments, the total number and base sizes of discovered deletions and insertions drop relative to RJF, suggesting the more diversified origins of RJF and its mixed haplotype assembly architecture are the cause (Suppl. Table 2). A genome-wide perspective of SVs in RJF, layer, and broiler genomes, including overall alterations in repeat content, will require additional research to validate their patterns of segregation in larger populations of chickens and

their accuracy of ascertainment. Overall, we observe structural differences, despite the fact that the percent masked sequence, a measure of all repeat types, is comparable between these references (20.5, 20.2, and 20.3) using default WindowMasker output (Morgulis et al. 2006).

Gene annotation. First, protein-coding gene representation was evaluated with BUSCO, which demonstrated an average 54% reduction in the number of missing universal single-copy orthologs in both GRCg7 haplotypes compared to GRCg6a (Suppl. Table 3). Automated gene annotation of GRCg7b and GRCg7w using the NCBI workflow (Sayers et al. 2021) reveals an increase in the overall number of protein-coding and non-coding genes (Suppl. Table 4). Recent gene annotation of multiple chicken genomes revealed 1,335 more protein-coding genes relative to GRCg6a (Li et al. 2022). The addition of at most 546 genes to the GRCg7 phased assemblies is a modest increase, but not unexpected given the NCBI annotation methods are likely more conservative and do not rely on many varied genome annotations as in Li et al. (Li et al. 2022). We also analyzed the differences in gene set ontology between GRCg7b and GRCg7w given their distinct selection histories. As determined by enrichment analysis, we find 82.8% overlap between GRCg7b and GRCg7w, where uniqueness is most often in large gene families, e.g., immune genes (Suppl. Table 5; Suppl. Fig. 5). In the broiler annotated set, there are 154 genes not found in the layer set, but there were only 44 unique to the layer set (Suppl. Table 5). This disparity may reflect the slightly higher contiguity of the broiler reference (Table 1). Future study will be required to provide a precise accounting of the genes that are unique to breeds, commercial or research lines, and wild strains of *Gallus gallus*. In addition, solutions to the question of whether avian genes were genuinely lost during the ancient divergence of avian and mammalian lineages will begin with the availability of a full genome, as recently completed for a human (Nurk et al. 2022).

WGS Mapping and SNV analysis. It is probable that the choice of broiler, layer or RJF as a reference for alignment of various resequenced chicken populations could contribute to SNV ascertainment bias as has been shown in human (Schneider et al. 2017). We examined the mapping rates of WGS short-read data of six genetically diverse chicken samples (Suppl. Table 6): a male and female for layer and broiler chicken, as well as an Ethiopian indigenous chicken breed. For various mapping metrics, regardless of the reference, we find no large differences, indicating that for measures of genetic diversity, all three references have comparable initial abilities to call SNVs or indels (Suppl. Table 7).

Despite similar WGS mapping rates across references, the optimal SNVs set for the experimental purpose intended is not certain. Next, we mapped WGS data from a separate cohort of solely broilers (n=10) to all three genome assemblies and called SNVs using GATK version 4.2.0. Although we found differences in total SNVs, these were not large, suggesting that SNV detection will be comparable when beginning

with any of the three assemblies (Suppl. Table 8). However, regional variations may be encountered and must be addressed if certain loci are of great experimental relevance (Fig. 3).

RNAseq mapping. The mapping of RNAseq data to estimate transcriptome changes between samples for biological interpretation is a crucial reference usage. To address this application, we first analyzed a large number of diverse tissues where total percent mapping is available in the NCBI gene annotation report (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Gallus_gallus/106/) and found very little difference between haplotypes (GRCg7b and GRCg7w). However, since RNAseq alignment in this application is optimized for verifying gene model predictions, we also tested the STAR aligner, which is typically considered best practice for bulk RNAseq studies (Dobin et al. 2013). Using a small number (n=8) of RNAseq samples from diverse tissue origins, including ileum, bone derived macrophages, and uterus, we observe a small average range of 0.7 to 1.7% differences among six samples in the total percentage of reads uniquely mapping to each of the three references (Suppl. Table 9). However, it is unclear why GRCg6a has a somewhat greater percentage of uniquely mapped reads across all samples (Suppl. Table 9). We also highlight the two individual female and male layer muscle RNAseq samples with vastly different average rates of unique mapping between assemblies, 70.6 and 68.5% to each GRCg7 haplotype, respectively, compared to 88% in GRCg6a (Suppl. Table 9). In the female layer sample, after analyzing all secondary alignment counts (not read counts), the olfactory receptor 14C36-like gene (>10M) is most abundant in GRCg6a, whereas for GRCg7 haplotypes it is ribosomal RNAs (Fig. 4). Although the number of individual rRNAs in each reference is comparable (~90), the total base size is notably different. The total assembled lengths of rRNA in each genome are GRCg6a (20,568), GRCg7b (113,817), and GRCg7w (55,916) (Fig. 4), suggesting that when evaluating unique versus multi-mapping events for any of these references the type of library sequenced, i.e., ribosome depleted or polyA selected should be considered (Zhao et al. 2018). Overall, for the majority of RNAseq samples studied, we observe minimal differences in unique mapping rates across all references; nonetheless, prior to conducting RNAseq mapping experiments, the reference choice should be considered.

Contrasting ALVE diversity in varied genome assemblies. To show an additional advantage of these phased references, we revisit the question of the RJF reference being unrepresentative of ancestral *Gallus gallus* ALVE diversity. ALVEs are species-specific retroviral integrations which retain the potential for retrotransposition and retroviral expression (Fig. 5A). The previous RJF reference assembly contained two Avian Leukosis Virus subgroup E (ALVE) integrations: ALVE6 (ALVE-JFevA), a truncated ALVE widespread across many breeds; and ALVE-JFevB, an intact integration found in no other chicken to date (Mason et al. 2020a). The new GRCg7 phased assemblies contains a total of eleven ALVEs: five from the

maternal broiler and six from the paternal layer (summarized in Fig. 5B; detailed locations in Suppl. Table 10). Leghorn layers typically have fewer than six ALVEs (Mason et al. 2020b; Mason et al. 2020c), but the identified ALVE1, ALVE3, ALVE15, ALVE_ros034 and the slow feathering-associated ALVE21 are representative of this Leghorn layer breed. ALVE_ros005, however, has previously only been identified in brown-egg layers and Ethiopian indigenous birds (Mason, et al. 2020b; 2020c). The ALVEs of the broiler haplotype are widely found across brown-egg commercial layers and broiler lines, representing their recent shared ancestry (Muir et al. 2008), and the presence of ALVE-TYR supports the observed recessive white phenotype (Fox and Smyth 1985) (Chang et al. 2006).

Seven of the ALVEs are full-length (Fig. 5B; Suppl. Table 10), and five have completely intact retroviral ORFs, accounting for the -1 ribosomal frameshift between gag and pol (Nikolic et al. 2012). Despite this, ALVE transmission between cells is unlikely, as both parental haplotypes exhibit ALVE-resistance at the TVB receptor (*TNFRSF10B*): maternal Q58* (TVB^R; rs736008824) and paternal P61L (rs318006572). This perhaps represents the effects of selection against P27 expression in commercial birds. Additionally, the ALVE_ros034 gag ORF truncates within P27, and similar mutations have been observed in ALVEs in other commercial backgrounds (Fig. 5B; (Mason et al. 2020b). Additional high quality chicken genome references of diverse genetic backgrounds interpreted in pangenome visualization modes will continue to resolve the evolution of ALVEs and their role in trait presentation.

Future chicken references. Pangenomic starting points as opposed to single linear representations have been proposed in humans (Siren et al. 2021) to overcome reference bias in genotyping. In Siren et al., the utility of a human pangenome reference in variant ascertainment demonstrates that this is the optimal course of action for future chicken genetic studies, particularly structural analyses (Siren et al. 2021). As a result, we are generating the requisite read-types to follow this same *de novo* assembly process in building multiple telomere-to-telomere single haplotype reference sources. Using these individual linear genome graphs to construct pangenome references will ensure the availability of the next generation of computational resources for optimally estimating segregating variation for significant genotype to phenotype connections in poultry production. The phased assemblies of the broiler and layer genomes as well as the RJF reference provide new insights into their general structure. In addition, we believe a new era in the use of avian genome references has already begun due to the rapid development of methods to build full genome copies.

Bird husbandry. The parent-offspring trio of this study is composed of a male White Leghorn and female broiler, the parents, each raised at the University of Arkansas avian housing facilities. A female F1 offspring was chosen from this cross for sequencing, to GRCg7b obtain both Z and W sex chromosomes. DNA for each parent and the F1 was extracted from white blood cells using standard practices for each intended use.

Sequencing and primary genome assembly. We followed the workflow established by the Vertebrate Genomes Project (VGP) to create the haplotype-phased chicken assembly (Rhie et al. 2021). Libraries were sequenced on the PacBio Sequel II instrument with the sequencing kit 2.1 (#101-310-500) and 10 hours movie time to a total of ~98GB. Because sequence coverage is lowered when phasing a diploid genome, we targeted a high read coverage of ~80x, to attempt the accurate assembly of repetitive microchromosomal regions and ZW sex chromosomes. TrioCanu (v1.8+287) was used to bin Consensus Long Reads (PacBio) of the F1 female into maternal and paternal haplotypes using haplotype-specific 21-mer markers derived from the Illumina short reads of the mother and father. Following binning, TrioCanu independently generated contigs for each haplotype (haplotigs). From this point, the maternal and paternal haplotigs independently underwent the same steps. Separately, we assembled the mitochondrial (MT) genome with the mitoVGP pipeline (v2.2) (Formenti 2020) and added it to the haplotigs to keep any raw MT reads from being mapped to nuclear sequences preventing conversion of possible mitochondrial nuclear integrations into MT sequence during the polishing steps. We used Arrow from smrtlink (v6.0.0.47841) to improve base calling accuracy and purge_dups (v1.0.0) (Guan et al. 2020) in an adapted trio mode to remove erroneous duplications.

The median insert sizes of WGS libraries were approximately 400 bp and individual libraries were tagged with unique dual index DNA barcodes to allow pooling and minimize the impact of barcode hopping. Libraries were pooled for sequencing on the NovaSeq 6000 (Illumina) to obtain at least 750 million 151-base pair reads per individual.

Assembly scaffolding and curation. Various maps were constructed to facilitate scaffolding of the phased contigs (Rhie et al. 2021). Briefly, long linked read libraries were generated from unfragmented high molecular weight DNA on the 10X Genomics Chromium instrument (Genome Library Kit & Gel Bead Kit v2 PN-120258, Genome Chip Kit v2 PN-120257, i7 Multiplex Kit PN-120262). We sequenced this 10X library on an Illumina HiSeq X instrument with 150bp read length to ~60X coverage. For optical mapping, the extracted DNA (~750ug) was labeled with a direct labeling enzyme (DLE-1) following the BioNano Prep Direct Label and Stain (DLS) Protocol (Document Number 30206). Labelled samples were imaged on the Bionano Saphyr instrument. Finally, Hi-C crosslinks were generated by Arima Genomics (https://arimagenomics.com/) using the Arima-Hi-C kit (P/N: A510008). From size selected fragments, Illumina-compatible libraries were generated using the KAPA Hyper Prep kit (P/N: KK8504). The resulting libraries were sequenced on an Illumina HiSeq X instrument to ~70x coverage.

With 10X long-linked reads, BioNano, and Hi-C maps in hand, the earlier polished and purged haplotigs were scaffolded in three stages according to Rhie et al. (Rhie et al. 2021): first, we used the 10x

linked-reads in two rounds of scaff10x (v4.1.0) (https://github.com/wtsi-hpag/Scaff10X) to generate the primary scaffolds. Second, we generated BioNano cmaps and used BioNano Solve (v3.2.1_04122018) (Lam et al. 2012) for hybrid scaffolding and to break mis-assemblies. Third, we used Salsa2 (v2.2) (Ghurye et al. 2019) to generate chromosomal-level scaffolds using the molecular contact information from Hi-C linked reads. Finally, we performed a second round of Arrow polishing on the maternal and paternal scaffolds with the binned long reads. During this round of polishing, gaps between contigs were closed by the gap-filling function of Arrow. The two haplotypes were then combined in a single assembly and underwent two rounds of short read polishing using longranger (v2.2.2) (Bishara et al. 2015) and freebayes (v1.3.3) (Garrison 2017). After separating the scaffolds back into their respective haplotypes and removing the MT genome from each assembly, the two phased assemblies underwent manual curation using gEVAL as described previously (Chow et al. 2016) (Howe et al. 2021), particularly to correct structural assembly errors.

Assembly statistics and evaluation. Following each stage of the assembly, we calculated various metrics of assembly quality, for example, N50 contig length, number of contigs, and quality value (QV) scores for each base call to assess progress. We used Merqury (v1.0) for overall assembly evaluations (including k-mer completeness and spectra copy number analysis) as well as phasing assessment with hap-mers. We first generated 21-mer databases (dbs) from the raw F1 10x data and the parental Illumina data using meryl. We then built inherited hap-mer dbs by taking the difference between the maternal and paternal k-mer dbs, filtering according to the filter level used by TrioCanu for binning, intersecting both with the F1 dbs, and filtering again, as below (steps 1-4). For evaluation of genome completeness and protein-coding gene representation, we ran BUSCO v4.0.2 (Manni et al. 2021) on our phased assemblies to determine the representation of near-universal single-copy orthologs in the vertebrate avian lineage (n=8,338); aves_odb10 (Suppl. Table 3).

Genome synteny and structural variation. To estimate sequence structural changes between assemblies for synteny, structural variation, and repeat expansion and contractions we used SyRi (Goel et al. 2019) with default parameters or Assemblytics v1.2.1 (Nattestad and Schatz 2016) with a unique sequence length requirement of 10,000 on nucmer alignments between GRCg6a, GRCg7b, and GRCg7w assemblies.

Gene Annotation. Both assemblies, GRCg7b, and GRCg7w, were gene annotated using the standard NCBI pipeline (Pruitt et al. 2014), including masking of repeats prior to ab initio gene predictions, for evidence-supported gene-model building. All annotation processes used publicly available RNA-seq and Iso-Seq data from diverse tissue sources. We relied on the NCBI gene annotation report release 106 to compare the

outcomes for each assembly. GRCg6a gene annotation data were reported earlier in NCBI release 104 using the same process as above.

Interspersed repeat estimation. Two independent assessments were made to estimate the percentage of repeats to confirm their similarities between assemblies. RepeatMasker v4.0.9 (Smit A 2013) with *-excln* and *-species chicken* was used to identify and annotate repetitive regions of each genome while ignoring gap sequence then WindowMasker analysis was carried out using default parameters (Morgulis et al. 2006).

WGS and RNAseq mapping. WGS short-read data of six chicken samples; a male and female layer, broiler and Ethiopian indigenous chicken breed were used to compare the mapping rate across the three genome assemblies (Suppl. Table 6). WGS were first checked for quality using Fastqc (Andrews 2010). Trimmomatic (Bolger, Lohse, and Usadel 2014) was used to remove the remaining Illumina adapter sequences and low-quality bases with default parameters. Clean reads were mapped to the three reference assemblies (GRCg6a, GRCg7b, and GRCg7w) using bwa-mem with default parameters (Li and Durbin 2009). Picard (http://broadinstitute.github.io/picard/) was used to sort the mapped files and merge files from multiple sequencing runs and to mark duplicate reads. Finally, SAMtools (Li et al. 2009) was used to assess mapping quality.

RNAseq data from eight chicken samples were used to compare the mapping rate across the three genome assemblies. We retrieved all sequence data from the NCBI Sequence Read Archive that included the diverse tissue sources of ileum, bone-derived macrophages, uterus and muscle from a male and female layer (Suppl. Table 9). Sequencing quality was checked by FastQC software (v 0.11.7), qualifying reads were mapped using STAR software (v 2.5.3a) with default parameters to all assemblies (GRCg6a, GRCg7b, and GRCg7w), and the percentage of uniquely mapped reads, multiple mapped reads, and reads mapped to too many loci were taken for the comparison. Moreover, the resulting bam files were used for assessing the mapping rate for each sample with the Samtools (v 1.9) 'flagstat' command (Li et al. 2009). Percentages of correctly paired reads were used for comparison.

SNV analysis. To estimate SNV differences in the starting reference alignments we used short-read sequences from a small cohort of broilers (n=10) representing commercial birds generated by Cobb-Vantress (available upon request). All samples attained genome coverage depth greater than 20x and individual reads were aligned to each reference with the Nvidia Clara Parabricks (version 3.6) implementation of the BWA algorithm. Variants were called in GVCF mode with Nvidia Parabricks HaplotypeCaller and GVCF files were loaded into GenomicsDB using GATK 4.2.0 (Poplin 2017),

GenomicsDBImport, and joint-genotyped with GATK's GenotypeGVCFs. Hard-filtering was performed on the resulting raw VCF using GATK's current best-practices for filtering.

BCFTools 1.12 was used to extract statistics on SNVs and insertions and deletions (indels) per chromosome. Variants that did not pass the filtering criteria were removed and mapping data for chromosomes were compared between all assemblies using command-line tools and then plotted in R. Ideograms were generated using karyoploteR (v1.16.0) in R. Colored regions of the chromosome denote annotated feature regions for that chromosome. Rainfall plots of variants depict where variants were found in the analysis along each chromosome. Each unique color indicates a different type of substitution. We only include variants that passed all filters and were heterozygous in the reference source.

ALVE annotation. Assembled Avian Leukosis Virus subgroup E (ALVE) integrations were identified by BLAST v2.10.0 (Altschul et al. 1990) using the ALVE1 reference sequence (GenBank: AY013303.1) and annotated for ORFs and miR-155 recognition sites (Hu et al. 2016). Analogous GRCg6a locations were identified using flanking sequences, then compared with known ALVE integration sites and target site duplications (TSD) (Mason et al. 2020b; Mason et al. 2020c). ALVE susceptibility was assessed by identifying the TVB receptor (*TNFRSF10B*) genotype.

Statement of Ethics

For these experiments we used CO2 gas euthanasia, following the current standards for poultry euthanasia provided in AVMA Guidelines for the Euthanasia of Animals (2020 Edition). All experiments presented herein were carried out in accordance with the approval of the Institutional Animal Care and Use Committee, University of Arkansas, Fayetteville, AR (protocol approval number). Moreover, all methods were performed in accordance with the ARRIVE guidelines.

Conflict of Interest Statement

All authors have no conflicts of interest.

Funding sources and acknowledgements

This work was supported by USDA NIFA 2020-67015-31574 to YD at the Western University of Health Sciences, and USDA NIFA 2022-67015-36218 to WCW, at the University of Missouri. All figures were generated with the BioRender software.

Author Contributions

Overall project coordination WCW; Genome assembly GF, OF; Genome data processing BH, JM, VS; Sample planning and acquisition RO, NA, JM, HC; Assembly curation AT, KH, EDJ; ALVE analysis AM; Gene annotation curation TM; RNAseq analysis FP, ZW, JS; Variant calling analysis WCW, KS, LC; Genome synteny analysis CT, GZ; Structural variation analysis RC, WCW; Manuscript writing WCW, EDJ, OF, JS, AM, HC, and YD.

Data Availability

We have deposited the primary data underlying these analyses as follows: genome assemblies are deposited in the NCBI assembly archive (GRCg7b – Bioproject PRJNA660757 and genome GCA_016699485.1; and GRCg7w – PRJNA660758 and GCA_016700215.2), PacBio SMRT reads associated with each reference are found in the SRA under the Bioproject number PRJNA673216. In addition, all sequence types and files are available in the GenomeArk database (https://genomeark.s3.amazonaws.com/index.html?prefix=species/Gallus_gallus/bGalGal1/). The mitochondrial genome is available under NCBI accession number NC_053523.1. All Illumina data used in evaluating mapping rates, WGS or RNAseq, are described in various supplemental tables.

Table 1. Phased assembly comparisons of broiler and layer genomes to RJF for Gallus gallus. Each assembly contains the Z and W sex chromosomes despite there being only one copy of each from the parents.

Common name	Assembly version	N50 contig (Mbp)	Total size (Mbp)	Total contigs	N50 Scaffold Length (Mbp)	Unplaced Sequences (Mbp)
Red Jungle Fowl	GRCg6a	17.6	1,055	1,403	20	14.1
Broiler	GRCg7b	18.8	1,049	677	90	6.6
Layer	GRCg7w	17.7	1,046	685	90	7.1

[1] NCBI assembly metrics.

Figure legends.

Figure 1. Assembled structural errors detected in RJF compared to broiler for chromosome 27 using Hi-C mapped data to the scaffolds. Genetic linkage map markers (n=125) displayed as green tic marks below the x-axis for the chromosome 27 heat map were mapped to each assembly to validate sequence order and orientation.

Figure 2. Sequenced differences in the phased broiler and layer genomes for A) macro and B) micro autosomes. From the inside out SNV density (red), window size of 500kb, range of 0 to 2.5%, indels <50bp (coral), 500kb window size and 0-0.8%; large indels (blue) per Mb, range of 0 to 60; CNV count per Mb (green); highlighted inversions (black dashes); chicken karyotype (varied color); ideograms of GRCg7b and GRCg7w chromosomes (varied colors).

Figure 3. The distribution of called heterozygous SNVs across chicken macrochromosome 7 (A) and microchromosome 20 (B) in the three assemblies. Rainfall plots of heterozygous variants depict their location, and each unique color indicates a different type of base substitution. We only include variants that passed all filters and were heterozygous in either reference source.

Figure 4. The RNAseq alignment detection of multimapping events and rRNA number and size distributions by reference source.

Figure 5. ALVE integration, propagation, and degradation within the chicken genome. A) Shows the retroviral genomic lifecycle. Retroviral positive sense, single stranded RNA is reverse transcribed into cDNA and associates with the retroviral integrase integration complex, which primes the cDNA 3' ends and initiates strand transfer with genomic DNA. Integration creates overhangs which are repaired by host machinery, creating target site duplications (TSDs; grey). Following integration, retroviral expression and retrotransposition is possible. Over evolutionary timescales integrated ERVs degrade, either by non-homologous recombination events (I, II) or internal LTR recombination leaving solo LTRs (III). B) Schematic indicates an intact ALVE with putative transcripts, with the ribosomal -1 frame slip and recognition site for miR-155 indicated. Phased chicken genome ALVE content and integrity is shown, with likely transcript and regulatory implications. CA: capsid; INT: integrase; LTR: long terminal repeat; MA: matrix; NC: nucleocapsid; PR: protease; RH: RNaseH; RT: reverse transcriptase; SU: surface; TM: transmembrane.

Supplemental files

Supplemental Tables

Supplemental Table 1. A select grouping of major structural errors corrected in the GRCg6b reference. All alignment outcomes were summarized using a combination of proximity map and whole genome sequence alignments to GRCg6a.

Supplemental Table 2. A summary of total counts and bases for each pairwise reference alignment using the Assemblytics software.

Supplemental Table 3. A summary of gene completeness using BUSCO v4.1.4 output for each assembly.

Supplemental Table 4. A summary of protein-coding and non-coding gene counts for all chicken assemblies.

Supplemental Table 5. Analysis of protein-coding genes annotated and unique to GRCg7w or GRCg7b.

Supplemental Table 6. A summary of WGS of chickens with known genetic diversity used to assess mapping efficiency.

Supplemental Table 7. A summary of WGS mapping rates to each reference of a cohort of diverse chickens.

Supplemental Table 8. A summary of SNVs and indels detected depending on the reference use for a small cohort of whole genome sequenced broilers.

Supplemental Table 9. Percent of unique mapped RNAseq reads by reference source for different tissue sources.

Supplemental Table 10. ALVEs of GRCg7b and GRCg7w. Identified ALVEs are shown with haplotype location, putative TSD location in each assembly, orientation, total length, TSD sequence, overlap with protein-coding genes, ORF integrity, and presence of the miR-155 recognition site AGCATTA within the env ORF. Abbreviated domains: INT: integrase; LTR: long terminal repeat; RH: RNaseH; RT: reverse transcriptase. P27 is a gag subunit conserved between ALV subtypes, hence its commercial use for exogenous ALV detection by ELISA. Notes: identical 16bp deletion identified in both LTRs (AY013303.1:11_26del; AY013303.1:7262_7277del) not affecting transcription factor binding sites, TATA box motif or transcription start site; b ALVE1 has a +1 frameshift in RT truncating the rest of pol, but the gag and env domains are intact; ALVE3 is missing RT but the gag-pol ORF is otherwise intact; miR-155 recognition site was found to be mutated ([A>G]GCATTA) as previously described with ALVE6.

Supplemental Figures

Supplementary Figure 1. Whole genome alignments and HiC heat maps support the allocation of a pseudoautosomal region to W. Some regions beyond PAR show evidence of probability of order and orientation. The red dotted square denotes the PAR location on the W chromosome in the assembly alignment and proximity map. rearrangements with W associated with GRCg6a. A. whole genome alignments using nucmer and B. chromatin proximity mapping with red showing the weighted scores for probability of order and orientation.

Supplementary Figure 2. Improper fusion of chromosome 29 to chromosome 31 in GRCg6a. The unplaced sequences are an unplaced scaffold in GRCg7 assemblies.

Supplementary Figure 3. The identification of four possible microchromosomes based on chromosome proximity heat map evidence. It is possible to nominate additional microchromosomes but karyotyping at this resolution is unproven. The putative microsomes are named M1 to M4 with black circles denoting their map location.

Supplementary Figure 4. A summary of structural variation for each pairwise assembly comparison. We used aligned contigs of each assembled genome as input for Assemblytics. Only the upper detected range of 500 to 10,000 bp is reported.

Supplementary Figure 5. Gene annotation overlap and uniqueness between GRCg7b and GRCg7w.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. 'Basic local alignment search tool', *J Mol Biol*, 215: 403-10.

Andrews, S. . 2010. 'FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].'.

Bellott, D. W., H. Skaletsky, T. Pyntikova, E. R. Mardis, T. Graves, C. Kremitzki, L. G. Brown, S. Rozen, W. C. Warren, R. K. Wilson, and D. C. Page. 2010. 'Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition', *Nature*, 466: 612-6.

Bishara, A., Y. Liu, Z. Weng, D. Kashef-Haghighi, D. E. Newburger, R. West, A. Sidow, and S. Batzoglou. 2015. 'Read clouds uncover variation in complex regions of the human genome', *Genome Res*, 25: 1570-80.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30: 2114-20.

Bredemeyer, K. R., A. J. Harris, G. Li, L. Zhao, N. M. Foley, M. Roelke-Parker, S. J. O'Brien, L. A. Lyons, W. C. Warren, and W. J. Murphy. 2021. 'Ultracontinuous Single Haplotype Genome Assemblies for the Domestic Cat (Felis catus) and Asian Leopard Cat (Prionailurus bengalensis)', *J Hered*, 112: 165-73.

Burt, D. W., C. Bruley, I. C. Dunn, C. T. Jones, A. Ramage, A. S. Law, D. R. Morrice, I. R. Paton, J. Smith, D. Windsor, A. Sazanov, R. Fries, and D. Waddington. 1999. 'The dynamics of chromosome evolution in birds and mammals', *Nature*, 402: 411-3.

Chang, C. M., J. L. Coville, G. Coquerelle, D. Gourichon, A. Oulmouden, and M. Tixier-Boichard. 2006. 'Complete association between a retroviral insertion in the tyrosinase gene and the recessive white mutation in chickens', *BMC Genomics*, 7: 19.

Cheng, Y., and D. W. Burt. 2018. 'Chicken genomics', *Int J Dev Biol*, 62: 265-71.

Chow, W., K. Brugger, M. Caccamo, I. Sealy, J. Torrance, and K. Howe. 2016. 'gEVAL - a web-based browser for evaluating genome assemblies', *Bioinformatics*, 32: 2508-10.

Consortium, International Chicken Genome Sequencing. 2004. 'Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution', *Nature*, 432: 695-716.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29: 15-21.

Formenti, G.; Rhie, A.; Balacco, J.; Haase, B.; Mountcastle, J.; Fedrigo, O.; Samara Brown, S.; et al. . 2020. 'Complete Vertebrate Mitogenomes Reveal Widespread Gene Duplications and Repeats', *bioRxiv*.

Fox, W., and J. R. Smyth, Jr. 1985. 'The effects of recessive white and dominant white genotypes on early growth rate', *Poult Sci*, 64: 429-33.

Gandhi, S., M. L. Piacentino, F. M. Vieceli, and M. E. Bronner. 2017. 'Optimization of CRISPR/Cas9 genome editing for loss-of-function in the early chick embryo', *Dev Biol*, 432: 86-97.

Garrison, E.; Marh, G. 2017. 'Haplotype-based variant detection from short-read sequencing', *arXiv*, 1207:3907v2.

Ghurye, J., A. Rhie, B. P. Walenz, A. Schmitt, S. Selvaraj, M. Pop, A. M. Phillippy, and S. Koren. 2019. 'Integrating Hi-C links with assembly graphs for chromosome-scale assembly', *PLoS Comput Biol*, 15: e1007273.

Goel, M., H. Sun, W. B. Jiao, and K. Schneeberger. 2019. 'SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies', *Genome Biol*, 20: 277.

Groenen, M. A., H. H. Cheng, N. Bumstead, B. F. Benkel, W. E. Briles, T. Burke, D. W. Burt, L. B. Crittenden, J. Dodgson, J. Hillel, S. Lamont, A. P. de Leon, M. Soller, H. Takahashi, and A. Vignal. 2000. 'A consensus linkage map of the chicken genome', *Genome Res*, 10: 137-47.

Guan, D., S. A. McCarthy, J. Wood, K. Howe, Y. Wang, and R. Durbin. 2020. 'Identifying and removing haplotypic duplication in primary genome assemblies', *Bioinformatics*, 36: 2896-98.

Howe, K., W. Chow, J. Collins, S. Pelan, D. L. Pointon, Y. Sims, J. Torrance, A. Tracey, and J. Wood. 2021. 'Significantly improving the quality of genome assemblies through curation', *Gigascience*, 10.

Hu, X., W. Zhu, S. Chen, Y. Liu, Z. Sun, T. Geng, X. Wang, B. Gao, C. Song, A. Qin, and H. Cui. 2016. 'Expression of the env gene from the avian endogenous retrovirus ALVE and regulation by miR-155', *Arch Virol*, 161: 1623-32.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P., Prosdocimi F., Samaniego J.A., Vargas Velazquez A.M., Alfaro-Núñez A., Campos P.F., Petersen B, Sicheritz-Ponten T, Pas A, Bailey T, Scofield P, Bunce M, Lambert DM, Zhou Q, Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jønsson K.A., Johnson W., Koepfli K.P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T., Zhang G. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 2014 Dec 12;346(6215):1320-31. doi: 10.1126/science.1253451.

Kim J., Lee C., Ko B.J., Yoo D.A., Won S., Phillippy A.M., Fedrigo O., Zhang G., Howe K., Wood J., Durbin R., Formenti G., Brown S., Cantin L., Mello C.V., Cho S., Rhie A., Kim H., Jarvis E.D. False gene and chromosome losses in genome assemblies caused by GC content variation and repeats. Genome Biol. 2022 Sep 27;23(1):204. doi: 10.1186/s13059-022-02765-0.

Koren S., Rhie A., Walenz B.P., Dilthey A.T., Bickhart D.M., Kingan S.B., Hiendleder S., Williams J.L., Smith T.P..L, Phillippy A.M. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018 Oct 22:10.1038/nbt.4277. doi: 10.1038/nbt.4277.

Korlach, J., G. Gedman, S. B. Kingan, C. S. Chin, J. T. Howard, J. N. Audet, L. Cantin, and E. D. Jarvis. 2017. 'De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads', *Gigascience*, 6: 1-16.

Lam, E. T., A. Hastie, C. Lin, D. Ehrlich, S. K. Das, M. D. Austin, P. Deshpande, H. Cao, N. Nagarajan, M. Xiao, and P. Y. Kwok. 2012. 'Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly', *Nat Biotechnol*, 30: 771-6.

Li, H., and R. Durbin. 2009. 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25: 1754-60.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Subgroup Genome Project Data Processing. 2009. 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25: 2078-9.

Li, M., C. Sun, N. Xu, P. Bian, X. Tian, X. Wang, Y. Wang, X. Jia, R. Heller, M. Wang, F. Wang, X. Dai, R. Luo, Y. Guo, X. Wang, P. Yang, D. Hu, Z. Liu, W. Fu, S. Zhang, X. Li, C. Wen, F. Lan, A. Z. Siddiki, C. Suwannapoom, X. Zhao, Q. Nie, X. Hu, Y. Jiang, and N. Yang. 2022. 'De Novo Assembly of 20 Chicken Genomes Reveals the Undetectable Phenomenon for Thousands of Core Genes on Microchromosomes and Subtelomeric Regions', *Mol Biol Evol*, 39.

Logsdon, G. A., M. R. Vollger, P. Hsieh, Y. Mao, M. A. Liskovykh, S. Koren, S. Nurk, L. Mercuri, P. C. Dishuck, A. Rhie, L. G. de Lima, T. Dvorkina, D. Porubsky, W. T. Harvey, A. Mikheenko, A. V. Bzikadze, M. Kremitzki, T. A. Graves-Lindsay, C. Jain, K. Hoekzema, S. C. Murali, K. M. Munson, C. Baker, M. Sorensen, A. M. Lewis, U. Surti, J. L. Gerton, V. Larionov, M. Ventura, K. H. Miga, A. M. Phillippy, and E. E. Eichler. 2021. 'The structure, function and evolution of a complete human chromosome 8', *Nature*, 593: 101-07.

Manni, M., M. R. Berkeley, M. Seppey, and E. M. Zdobnov. 2021. 'BUSCO: Assessing Genomic Data Quality and Beyond', *Curr Protoc*, 1: e323.
Mason, A. S., J. E. Fulton, and J. Smith. 2020a. 'Endogenous avian leukosis virus subgroup E elements of the chicken reference genome', *Poult Sci*, 99: 2911-15.
Mason, A. S., A. R. Lund, P. M. Hocking, J. E. Fulton, and D. W. Burt. 2020b. 'Identification and characterisation of endogenous Avian Leukosis Virus subgroup E (ALVE) insertions in chicken whole genome sequencing data', *Mob DNA*, 11: 22.
Mason, A. S., K. Miedzinska, A. Kebede, O. Bamidele, A. S. Al-Jumaili, T. Dessie, O. Hanotte, and J. Smith. 2020c. 'Diversity of endogenous avian leukosis virus subgroup E (ALVE) insertions in indigenous chickens', *Genet Sel Evol*, 52: 29.
Morgulis, A., E. M. Gertz, A. A. Schaffer, and R. Agarwala. 2006. 'WindowMasker: window-based masker for sequenced genomes', *Bioinformatics*, 22: 134-41.
Muir, W. M., G. K. Wong, Y. Zhang, J. Wang, M. A. Groenen, R. P. Crooijmans, H. J. Megens, H. Zhang, R. Okimoto, A. Vereijken, A. Jungerius, G. A. Albers, C. T. Lawley, M. E. Delany, S. MacEachern, and H. H. Cheng. 2008. 'Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds', *Proc Natl Acad Sci U S A*, 105: 17312-7.
Nattestad, M., and M. C. Schatz. 2016. 'Assemblytics: a web analytics tool for the detection of variants from an assembly', *Bioinformatics*, 32: 3021-3.
Nikolic, E. I., L. M. King, M. Vidakovic, N. Irigoyen, and I. Brierley. 2012. 'Modulation of ribosomal frameshifting frequency and its effect on the replication of Rous sarcoma virus', *J Virol*, 86: 11581-94.
Nurk, S., S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, N. C. Chen, H. Cheng, C. S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin, T. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Fungtammasan, E. Garrison, P. G. S. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V. V. Maduro, T. Marschall, A. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B. Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sovic, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W. Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, and A. M. Phillippy. 2022. 'The complete sequence of a human genome', *Science*, 376: 44-53.
Poplin, R.; Ruano-Rubio, V.; DePristo, M.A.; Fennell, T.J.; Carneiro, M.O.; Van der Auwera, G.A.; Kling DE, Gauthier, L.D.; Levy-Moonshine, A.; Roazen, D.; Shakir, K.; Thibault, J.; Chandran, S.; Whelan, C.; Lek, M.; Gabriel, S.; Daly, M.J.; Neale, B.; MacArthur, D.G.; Banks, E. . 2017. 'Scaling accurate genetic variant discovery to tens of thousands of samples', *bioRxiv*, 201178.
Pruitt, K. D., G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, C. M. Farrell, J. Hart, M. J. Landrum, K. M. McGarvey, M. R. Murphy, N. A. O'Leary, S. Pujar, B. Rajput, S. H. Rangwala, L. D. Riddick, A. Shkeda, H. Sun, P. Tamez, R. E. Tully, C. Wallin, D. Webb, J. Weber, W. Wu, M. DiCuccio, P. Kitts, D. R. Maglott, T. D. Murphy, and J. M. Ostell. 2014. 'RefSeq: an update on mammalian reference sequences', *Nucleic Acids Res*, 42: D756-63.
Rao, Y. S., J. Li, R. Zhang, X. R. Lin, J. G. Xu, L. Xie, Z. Q. Xu, L. Wang, J. K. Gan, X. J. Xie, J. He, and X. Q. Zhang. 2016. 'Copy number variation identification and analysis of the chicken genome using a 60K SNP BeadChip', *Poult Sci*, 95: 1750-6.
Rhie, A., S. A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, M. Uliano-Silva, W. Chow, A. Fungtammasan, J. Kim, C. Lee, B. J. Ko, M. Chaisson, G. L. Gedman, L. J. Cantin, F. Thibaud-

Nissen, L. Haggerty, I. Bista, M. Smith, B. Haase, J. Mountcastle, S. Winkler, S. Paez, J. Howard, S. C. Vernes, T. M. Lama, F. Grutzner, W. C. Warren, C. N. Balakrishnan, D. Burt, J. M. George, M. T. Biegler, D. Iorns, A. Digby, D. Eason, B. Robertson, T. Edwards, M. Wilkinson, G. Turner, A. Meyer, A. F. Kautt, P. Franchini, H. W. Detrich, 3rd, H. Svardal, M. Wagner, G. J. P. Naylor, M. Pippel, M. Malinsky, M. Mooney, M. Simbirsky, B. T. Hannigan, T. Pesout, M. Houck, A. Misuraca, S. B. Kingan, R. Hall, Z. Kronenberg, I. Sovic, C. Dunn, Z. Ning, A. Hastie, J. Lee, S. Selvaraj, R. E. Green, N. H. Putnam, I. Gut, J. Ghurye, E. Garrison, Y. Sims, J. Collins, S. Pelan, J. Torrance, A. Tracey, J. Wood, R. E. Dagnew, D. Guan, S. E. London, D. F. Clayton, C. V. Mello, S. R. Friedrich, P. V. Lovell, E. Osipova, F. O. Al-Ajli, S. Secomandi, H. Kim, C. Theofanopoulou, M. Hiller, Y. Zhou, R. S. Harris, K. D. Makova, P. Medvedev, J. Hoffman, P. Masterson, K. Clark, F. Martin, K. Howe, P. Flicek, B. P. Walenz, W. Kwak, H. Clawson, M. Diekhans, L. Nassar, B. Paten, R. H. S. Kraus, A. J. Crawford, M. T. P. Gilbert, G. Zhang, B. Venkatesh, R. W. Murphy, K. P. Koepfli, B. Shapiro, W. E. Johnson, F. Di Palma, T. Marques-Bonet, E. C. Teeling, T. Warnow, J. M. Graves, O. A. Ryder, D. Haussler, S. J. O'Brien, J. Korlach, H. A. Lewin, K. Howe, E. W. Myers, R. Durbin, A. M. Phillippy, and E. D. Jarvis. 2021. 'Towards complete and error-free genome assemblies of all vertebrate species', *Nature*, 592: 737-46.

Rice, E. S., S. Koren, A. Rhie, M. P. Heaton, T. S. Kalbfleisch, T. Hardy, P. H. Hackett, D. M. Bickhart, B. D. Rosen, B. V. Ley, N. W. Maurer, R. E. Green, A. M. Phillippy, J. L. Petersen, and T. P. L. Smith. 2020. 'Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle', *Gigascience*, 9.

Sayers, E. W., J. Beck, E. E. Bolton, D. Bourexis, J. R. Brister, K. Canese, D. C. Comeau, K. Funk, S. Kim, W. Klimke, A. Marchler-Bauer, M. Landrum, S. Lathrop, Z. Lu, T. L. Madden, N. O'Leary, L. Phan, S. H. Rangwala, V. A. Schneider, Y. Skripchenko, J. Wang, J. Ye, B. W. Trawick, K. D. Pruitt, and S. T. Sherry. 2021. 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Res*, 49: D10-D17.

Schneider, V. A., T. Graves-Lindsay, K. Howe, N. Bouk, H. C. Chen, P. A. Kitts, T. D. Murphy, K. D. Pruitt, F. Thibaud-Nissen, D. Albracht, R. S. Fulton, M. Kremitzki, V. Magrini, C. Markovic, S. McGrath, K. M. Steinberg, K. Auger, W. Chow, J. Collins, G. Harden, T. Hubbard, S. Pelan, J. T. Simpson, G. Threadgold, J. Torrance, J. M. Wood, L. Clarke, S. Koren, M. Boitano, P. Peluso, H. Li, C. S. Chin, A. M. Phillippy, R. Durbin, R. K. Wilson, P. Flicek, E. E. Eichler, and D. M. Church. 2017. 'Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly', *Genome Res*, 27: 849-64.

Siren, J., J. Monlong, X. Chang, A. M. Novak, J. M. Eizenga, C. Markello, J. A. Sibbesen, G. Hickey, P. C. Chang, A. Carroll, N. Gupta, S. Gabriel, T. W. Blackwell, A. Ratan, K. D. Taylor, S. S. Rich, J. I. Rotter, D. Haussler, E. Garrison, and B. Paten. 2021. 'Pangenomics enables genotyping of known structural variants in 5202 diverse genomes', *Science*, 374: abg8871.

Smit A, Hubley R, Green P. 2013. "RepeatMasker " In. http://repeatmasker.org.

Waters, P. D., H. R. Patel, A. Ruiz-Herrera, L. Alvarez-Gonzalez, N. C. Lister, O. Simakov, T. Ezaz, P. Kaur, C. Frere, F. Grutzner, A. Georges, and J. A. M. Graves. 2021. 'Microchromosomes are building blocks of bird, reptile, and mammal chromosomes', *Proc Natl Acad Sci U S A*, 118.

Wolc, A., W. Drobik-Czwarno, J. E. Fulton, J. Arango, T. Jankowski, and J. C. M. Dekkers. 2018. 'Genomic prediction of avian influenza infection outcome in layer chickens', *Genet Sel Evol*, 50: 21.

Zhang, G., C. Li, Q. Li, B. Li, D. M. Larkin, C. Lee, J. F. Storz, A. Antunes, M. J. Greenwold, R. W. Meredith, A. Odeen, J. Cui, Q. Zhou, L. Xu, H. Pan, Z. Wang, L. Jin, P. Zhang, H. Hu, W. Yang, J. Hu, J. Xiao, Z. Yang, Y. Liu, Q. Xie, H. Yu, J. Lian, P. Wen, F. Zhang, H. Li, Y. Zeng, Z. Xiong, S. Liu, L. Zhou, Z. Huang, N. An, J. Wang, Q. Zheng, Y. Xiong, G. Wang, B. Wang, J. Wang, Y. Fan, R. R. da Fonseca, A. Alfaro-Nunez, M. Schubert, L. Orlando, T. Mourier, J. T. Howard, G. Ganapathy, A. Pfenning, O. Whitney, M. V. Rivas, E. Hara, J. Smith, M. Farre, J. Narayan, G. Slavov, M. N. Romanov, R. Borges, J. P. Machado, I. Khan, M. S. Springer, J.

Gatesy, F. G. Hoffmann, J. C. Opazo, O. Hastad, R. H. Sawyer, H. Kim, K. W. Kim, H. J. Kim, S. Cho, N. Li, Y. Huang, M. W. Bruford, X. Zhan, A. Dixon, M. F. Bertelsen, E. Derryberry, W. Warren, R. K. Wilson, S. Li, D. A. Ray, R. E. Green, S. J. O'Brien, D. Griffin, W. E. Johnson, D. Haussler, O. A. Ryder, E. Willerslev, G. R. Graves, P. Alstrom, J. Fjeldsa, D. P. Mindell, S. V. Edwards, E. L. Braun, C. Rahbek, D. W. Burt, P. Houde, Y. Zhang, H. Yang, J. Wang, Consortium Avian Genome, E. D. Jarvis, M. T. Gilbert, and J. Wang. 2014. 'Comparative genomics reveals insights into avian genome evolution and adaptation', *Science*, 346: 1311-20.

Zhao, S., Y. Zhang, R. Gamini, B. Zhang, and D. von Schack. 2018. 'Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion', *Sci Rep*, 8: 4781.

Zuo, Q., Y. Wang, S. Cheng, C. Lian, B. Tang, F. Wang, Z. Lu, Y. Ji, R. Zhao, W. Zhang, K. Jin, J. Song, Y. Zhang, and B. Li. 2016. 'Site-Directed Genome Knockout in Chicken Cell Line and Embryos Can Use CRISPR/Cas Gene Editing Technology', *G3 (Bethesda)*, 6: 1787-92.