Fourth Report on Chicken Genes and Chromosomes 2022
Jacqueline Smith, James Alfieri, Nick Anthony, Peter Arensburger, Giridhar Athrey, Jennifer Balacco, Adam Balic, Philippe Bardou, Paul Barela, Yves Bigot, et al.

To cite this version:
Jacqueline Smith, James Alfieri, Nick Anthony, Peter Arensburger, Giridhar Athrey, et al.. Fourth Report on Chicken Genes and Chromosomes 2022. Cytogenetic and Genome Research, 2023, 162 (8-9), pp.405-528. 10.1159/000529376. hal-04217636

HAL Id: hal-04217636
https://hal.inrae.fr/hal-04217636
Submitted on 26 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives| 4.0 International License
Fourth Report on Chicken Genes and Chromosomes 2022

Prepared by

Jacqueline Smith, jacqueline.smith@roslin.ed.ac.uk

Correspondence to:

© 2023 The Author(s).

Published online: January 30, 2023

Received: January 16, 2023
Accepted: January 22, 2023

DOI: 10.1159/000529376

Cytogenet Genome Res 2022;162:405–527

This article is licensed under the Creative Commons Attribution 4.0 International License (http://www.karger.com/Services/OpenAccessLicense). Usage, derivative works and distribution are permitted provided that proper credit is given to the author and the original publisher.
Introduction
(Prepared by J. Smith)

The chicken continues to hold its position as a leading model organism within many areas of research, as well as being a major source of protein for human consumption. The First Report on Chicken Genes and Chromosomes [Schmid et al., 2000], which was published in 2000, was the brainchild of the late, and sadly missed, Prof. Michael Schmid of the University of Würzburg. It was a publication bringing together updates on the latest research and resources in chicken genomics and cytogenetics. The success of this first report led to the subsequent publication of the Second [Schmid et al., 2005] and Third Report on Chicken Genes and Chromosomes [Schmid et al., 2015] – each also proving popular references for the research community. It is now our pleasure to be able to introduce publication of the Fourth Report. Being 7 years since the last report, this publication captures the many advances that have taken place during that time. This includes pre-
sentation of the detailed genomic resources that are now available, largely due to increasing capabilities of sequencing technologies and which herald the pan genomic age, allowing for a much richer and more complete knowledge of the avian genome. Ongoing cytogenetic work also allows for examination of chromosomes, specific elements within chromosomes, and the evolutionary history and comparison of karyotypes. We also examine chicken research efforts with a much more “global” outlook with a greater impact on food security and the impact of climate change, and highlight the efforts of international consortia, such as the Chicken Diversity Consortium. We dedicate this Report to Michael.

**Multiple Chicken (Gallus gallus) Genome References to Advance Genetic Variation Studies**


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 noncoding 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 6 putative microchromosomes that were previously missing in GRCg6a. Using a pairwise genome comparison of the parental genomes, and 2 independent cohorts of sequenced chickens, we show small discernible 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 germplines 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 G 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 [Jarvis et al., 2014; Zhang G 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 2 functional instances of gained insight into embryonic germ and satellite muscle cell differentiation, respectively.

Since its first iteration in 2004 [International Chicken Genome Sequencing Consortium, 2004], we have worked to refine the assembly of the chicken genome as technology progressed over the past 2 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 [Koren et al., 2018; Rice et al., 2020; Bredemeyer et al., 2021; Rhie et al., 2021], 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 M 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 the trio binning approach to present 2 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 pseudoautosomal region (PAR), 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 (150 bp), and the F1 was sequenced with 80× PacBio reads (12 kb on average), and all reads were used as input to TrioCa-

<table>
<thead>
<tr>
<th>Common name</th>
<th>Assembly version</th>
<th>N50 contig, Mb</th>
<th>Total size, Mb</th>
<th>Total contigs</th>
<th>N50 scaffold length, Mb</th>
<th>Unplaced sequences, Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Jungle Fowl</td>
<td>GRCg6a</td>
<td>17.6</td>
<td>1,055</td>
<td>1,403</td>
<td>20</td>
<td>14.1</td>
</tr>
<tr>
<td>Broiler</td>
<td>GRCg7b</td>
<td>18.8</td>
<td>1,049</td>
<td>677</td>
<td>90</td>
<td>6.6</td>
</tr>
<tr>
<td>Layer</td>
<td>GRCg7w</td>
<td>17.7</td>
<td>1,046</td>
<td>685</td>
<td>90</td>
<td>7.1</td>
</tr>
</tbody>
</table>

Data based on NCBI assembly metrics.
ning 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 (online suppl. Material 1, 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 (online suppl. Material 2, 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 [International Chicken Genome Sequencing Consortium, 2004]. If the origin of the microchromosomes was initiated by a number of random fis-

---

**Fig. 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.
sion 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 microchromosomes. In both GRCg7b and GRCg7w, using long reads that get through GC-rich regions, as with zebra finch [Kim J et al., 1999], 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 2 methods: high resolution alignments to reveal major syntenic 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 (Fig. 2; online suppl. Material 1, Table 2, online suppl. Material 2, 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 (online suppl. Material 1, 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 (online suppl. Material 1, 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 noncoding genes (online suppl. Material 1, Table 4). Recent gene annotation of multiple chicken genomes revealed 1,335 more protein-coding genes relative to GRCg6a [Li M 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 M 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 (online suppl. Material 2, Fig. 5; online suppl. Material 1, Table 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 (online suppl. Material 1, 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 6 genetically diverse chicken samples (online suppl. Material 1, 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 3 references have comparable initial abilities to call SNVs or indels (online suppl. Material 1, 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 (online suppl. Material 1, 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/annotation_euk/Gallus_gallus/106/) and found very little difference between haplotypes (GRCg7b and

---

**Fig. 2.** Sequenced differences in the phased broiler and layer genomes for macro- (a) and micro-autosomes (b). From the inside out SNV density (red), window size of 500 kb, range of 0 to 2.5%, indels <50 bp (coral), 500 kb 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).
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 0.7–1.7% differences among 6 samples in the total percentage of reads uniquely mapping to each of the three references (online suppl. Material 1, Table 9). However, it is unclear why GRCg6a has a somewhat greater percentage of uniquely mapped reads across all samples (online suppl. Material 1, Table 9). We also highlight the two individual female and male layer muscle RNAseq samples with vastly different rates of unique mapping between assemblies, 70.6 and 68.5% to each GRCg7 haplotype, respectively, compared to 88% in GRCg6a (online suppl. Material 1, Table 9). In the female layer sample, after analyzing all secondary alignment counts (not read counts), the olfactory receptor 14C36-like gene (>10 M) is most abundant in GRC-g6a, whereas for GRCg7 haplotypes it is ribosomal RNAs (Fig. 4). Although the number of individual rRNAs 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 S 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 2 Avian Leukosis Virus subgroup E (ALVE) integrations: ALVE6 (ALVE-JFe-vA), 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 contain a total of 11 ALVEs: 5 from the maternal broiler and 6 from the paternal layer (summarized in Fig. 5b; detailed locations in online suppl. Material 1, Table 10). Leghorn layers typically have fewer than 6 ALVEs [Mason et al., 2020b, c], 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]. 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., 2008b], 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; online suppl. Material 1, Table 10), and 5 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* (TVBR; 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 in-

Fig. 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.
terpreted 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. [2021], 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. As a result, we are generating the requisite read-types to follow this same de novo assembly process in building multiple telomere–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. 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 h movie time to a total of ∼98 GB. 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

---

**Fig. 4.** The RNAseq alignment detection of multimapping events and rRNA number and size distributions by reference source.
steps. Separately, we assembled the mitochondrial (MT) genome with the mitoVGP pipeline (v2.2) [Formenti et al., 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-bp reads per individual.

Assembly Scaffolding and Curation

Various maps were constructed to facilitate scaffolding of the phased contigs [Rhee 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 10× library on an Illumina HiSeq X instrument with 150 bp read length to ~60× coverage. For optical mapping, the extracted
DNA (~750 μg) was labeled with a direct labeling enzyme (DLE-1) following the BioNano Prep Direct Label and Stain (DLS) Protocol (Document Number 30206). Labeled 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 ~70× coverage.

With 10× long-linked reads, BioNano, and Hi-C maps in hand, the earlier polished and purged haplotigs were scaffolded in 3 stages according to Rhie et al. [2021]: first, we used the 10× linked-reads in 2 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 2 rounds of short-read polishing using longranger (v2.2.2) [Bishara et al., 2015] and freebayes (v1.3.3) [Garrison and Marth, 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 10× 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 (online suppl. Material 1, 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, 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 [Margulis et al., 2006].

WGS and RNAseq Mapping
WGS short-read data of 6 chicken samples; a male and female layer, broiler and Ethiopian indigenous chicken breed were used to compare the mapping rate across the 3 genome assemblies (online suppl. Material 1, Table 6). WGS were first checked for quality using Fastqc [Andrews, 2010]. Trimmomatic [Bolger et al., 2014] was used to remove the remaining Illumina adapter sequences and low-quality bases with default parameters. Clean reads were mapped to the 3 reference assemblies (GRCg6a, GRCg7b, and GRCg7w) using bwa-mem with default parameters [Li and Durbin, 2009]. Picard (http://broadin-
stitute.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 8 chicken samples were used to compare the mapping rate across the 3 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 (online suppl. Material 1, 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 et al., 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 karyoplasteR (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, c]. ALVE susceptibility was assessed by identifying the TVB receptor (TNFRSF10B) genotype.

**Acknowledgments**

All figures were generated with the BioRender software.

**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. Moreover, all methods were performed in accordance with the ARRIVE guidelines.

**Conflict of Interest Statement**

All authors have no conflicts of interest.

**Funding Sources**

This work was supported by USDA NIFA 2020-67015-31574 to Y.D. at the Western University of Health Sciences, and USDA NIFA 2022-67015-36218 to W.C.W., at the University of Missouri. The work of T.D.M. and V.A.S. was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health.

**Data Availability Statement**

We have deposited the primary data underlying these analyses as follows: genome assemblies are deposited in the NCBI assembly archive (GRCg7b – PRJNA660757 and GRCg7w – PRJNA660758), 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 VGP database (https://genonemark.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.

**Karyotypes and Chromosome-Level Assemblies in Chickens and Other Dinosaurs**

(Prepared by D.K. Griffin, D.M. Larkin, and R.E. O’Connor)

Any article about chicken, or avian generally, genomics is inherently about dinosaur genomics. In the light of recent paleontological evidence, statements such as “birds
evolved from dinosaurs” and “birds are related to dinosaurs” require substantial revision. Rather, it is very clear that birds, including our humble chicken, are in fact extant dinosaurs. Omnipresent in literature, film, television, popular culture and the media since the original fossil discoveries, notions that dinosaurs were obliterated entirely by the latest mass extinction event have undergone radical revision. On the contrary, dinosaurs are the great survivors of mass extinction events and we suggest, this may be due, at least in part, to their unique genome organization (in other words, their karyotype). Studies of chromosome-level genome assemblies guided us to this conclusion.

**Chromosome-Level Assemblies**

Genomic approaches such as array comparative genomics hybridization (CGH) and next-generation sequencing (NGS) for cytogenetic application would not be possible for many clinical and veterinary uses were it not for “chromosome-level” assemblies (CLAs). In other words, one ultimate objective of any genome assembly is that all the sequences are correctly aligned and assigned to their place on the appropriate chromosome, chromosome arm, and chromosome band.

Similarly, genome assembly needs cytogenetic analysis. As Lewin et al. [2010] put it, “every genome needs a good map.” Thankfully, the outline of that map for any given species is provided in the shape of a karyotype. While karyotypes typically are used to detect chromosomal disease in humans, the karyotype can be considered the most basic low-resolution genomic map of an organism. When whole-genome assemblies are yet to reach the heights of a CLA (probably the case for most sequenced genomes), their applicability for evolutionary genomics is impeded. To give one of many examples, CLAs are crucial for furthering applied agricultural research in that an established order of sequences is an essential pre-requisite to establish genotype-phenotype associations, e.g., in genome-wide association studies (GWAS). CLAs have facilitated these in chicken, turkey, and duck as well as in mammals such as pig, cattle, and sheep. Other advantages of CLAs include discovery of mendelian traits, spotting cryptic chromosome translocations, as well as isolating quantitative trait nucleotides (QTNs), expression quantitative trait loci (eQTLs), and long-range regulatory interactions. The ultimate goal of these studies is increased food production efficiency as well as global food security. With the knowledge at our fingertips for a number of species’ CLAs, comparative genomics is much more practicable in silico, and intractable karyotyping of chromosome rearrangements can be detected by fluorescence in situ hybridization (FISH). Comparative genomics, thereafter, allows us to describe the genome structure of more obscure species (by comparison to a standard, such as chicken) and the identification of chromosome rearrangements that led to each species’ unique karyotype. CLAs also facilitate addressing fundamental biological hypotheses pertaining to genome evolution, e.g., the mechanisms of chromosome breakage and fusion, as well as the significance and genomic correlates of evolutionary breakpoint regions (EBRs) and homologous synteny blocks (HSBs). In the era of genomics, cytogenetics (or, more specifically, cytogenomics) is not just a descriptive science but it provides a framework for the conceptualization of the structure of any genome. It gives us a starting point through which we can understand genome-phenome correlations.

The particular challenge with establishing genome structure in birds is that while, in most species, correlating genome assembly with karyotype is a bit like identifying cities and towns in the major landmasses of Europe (some big, some small), doing it in birds is more like doing it for Polynesia where there are a lot of tiny islands, i.e., the microchromosomes. In our research, we have looked at CLAs in a multitude of bird species, providing novel insight into the genome organization of the avian forebears – the extinct dinosaurs.

**What Is a Dinosaur?**

The official definition of a dinosaur is “Triceratops, modern birds, their most recent common ancestor and all their descendants.” As biologists however, perhaps an easier way to visualise them is as reptiles (in this definition we class birds as reptiles) with hind limbs held erect beneath the body (like mammals). This distinguishes them from most other reptiles such as lizards and crocodilians where the legs are positioned to the side. If we then eliminate the sister group, the pterosaurs, which are easily distinguishable, we can usually spot a dinosaur with relative ease from another organism.

Accepting then that modern paleontological evidence is explicit that birds are dinosaurs, rather than considering them as a group of animals that the Chicxulub meteor wiped out, dinosaurs are, in point of fact, survivors of numerous extinction events including the latest – the K-Pg (Cretaceous-Paleogene) extinction event. This segment describes how groups mostly in Kent and the Royal Veterinary College combined molecular cytogenetics and bioinformatics to establish that this resilience and ability to recover from extinction event may be due, at least partly, to their karyotype.
The Dawn and Dusk and Dawn of the Dinosaurs

As illustrated in Figure 6, around 325 million years ago, the amniote lineage split into synapsids (ultimately becoming mammals, amongst other groups) and the reptile/bird lineage – diapsids. There are over 17,500 living diapsid species, the majority of which (around 11,000) are birds. Dinosaurs (including birds), pterosaurs, turtles, and crocodilians all share a common ancestor that lived 275 million years ago [Shedlock and Edwards, 2009; Hedges et al., 2015], with the turtles (Testudines) diverging first (around 255 million years ago), the crocodilians around 252 million years ago, the pterosaurs about 245 million years ago, and the true dinosaurs about 240 million years ago. For the next 30 million years, dinosaur species were few in number, but amid the Jurassic period, this number, their geographical spread, and their body size all dramatically increased [Benton et al., 2014]. The proceeding 135 million years of dinosaur evolution is spectacular for being a time when they had an incredible range of species diversity and thus became the dominant vertebrates on the planet. Once human beings evolved scientific investigation, popular culture and media however, then the dinosaurs’ legacy was complete. Incredibly, dinosaurs survived the Carnian-Norian and End-Triassic mass extinction events (228 and 201 million years, respectively) and now there are >1,000 discovered fossil species. About 30 more appear annually, not including birds, in the fossil record [Weishampel, 2004]. The devastation of the K-Pg extinction event 66 million years ago nearly wiped them out, but they bounced back again as modern birds, with more species than any other terrestrial vertebrate (Fig. 6). The cytogenomic study of birds is an independent line of enquiry to paleontology and circumvents some of the problems associated with fossil dating.

The remarkable diversity and species abundance that is observed in dinosaurs is often put down to the fact that competitor species were wiped out, thereby allowing the dinosaurs to thrive. It has nonetheless also been proposed that such impressive levels of abundance and diversity reflect genomic adaptations that are particular to dinosaurs, facilitating their survival over other species in harsh environments. Some examples include extraordinary bone growth rates and highly evolved respiration systems [Farmer and Sanders, 2010], including unidirectional respiration [O’Connor and Claessens, 2005]. These sorts of evolutionary adaptations could have led to the evolutionary success of avian species, clues of which may be found in their genome structure and organization.

Dinosaur (Bird) Evolution Just before, during, and after the Last Mass Extinction Event as Revealed by Multiple Genome Sequencing Efforts

As a result of a multiple genome sequencing effort in birds [Jarvis et al., 2014; Zhang G et al., 2014], a revised avian phylogeny based on genome assemblies (some of which were CLAs) corrected the timing of avian diversification. We can now consider the first avian evolutionary divergence at about 100 million years ago with the Palaeognathae (Ratites/Tinamous) branching from the Neog-
nathae (Galloanseres/Neoaves). The second is where Galloanseres (Galliformes and Anseriformes) and Neoaves then diverged 80 million years ago, with Galliformes (landfowl, e.g., chicken, turkey, quail, pheasant) and Anseriformes (waterfowl, e.g., geese, ducks, swans) diverging about 66 million years ago. Another major divergence of the Neoaves into Columbea (e.g., pigeons) and Passarea (e.g., songbirds) is dated slightly earlier (67–69 million years ago). Data from the Jarvis et al. [2014] analysis plus that of Prum et al. [2015] suggest that, following the K-Pg mass extinction event [Schulte et al., 2010], around the same time of these two divergences, there was a rapid period of diversification, with 36 lineages evolving in a very short evolutionary period of 10–15 million years [Jarvis et al., 2014]. Genomics has thus updated our understanding of dinosaurs through comparative studies, proving intriguing insights into relationships with diversity and phenotype [Jarvis et al., 2014; Zhang G et al., 2014]. The karyotype of dinosaurs was something therefore that warranted deeper investigation.

**Dinosaur Karyotype Evolution**

The extraction of intact DNA from Jurassic blood-sucking insects is an outstanding plot device for novelists and film-makers but, alas, not a feasible means of facilitating the making of metaphase preparations. We can, however, with enough avian CLAs, glean insight into extinct dinosaur karyotypes by inference. Analysis of (near) CLAs from 6 living birds plus an Anolis lizard outgroup allowed us to infer the most likely ancestral karyotype of all birds [Romanov et al., 2014]. Generally speaking, the common avian ancestor was probably a bipedal, terrestrial, chicken-sized small Jurassic dinosaur with some flying ability [Witmer, 2002], and our studies established that its karyotype was very similar to that of a chicken or a ratite bird. We then went on to reconstruct the most parsimonious sequence of events that led to the karyotypes we typically observe in avian species. The humble chicken (Gallus gallus) is in fact the closest, karyotypically, to the common avian ancestor (see Fig. 6). This underwent rapid change over ~20 million years and we established that most of the major features associated with a typical bird karyotype were already laid down in the archelosaur ancestor 255 million years ago. This is because most chicken (for this, read ancestral avian) chromosomes (numbered 1–28 + Z) are precisely syntenic to those of A. spinifera (2n = 66). Analogous studies using chicken chromosome painting on Chinese soft-shelled turtle (Pelodiscus sinensis) (2n = 66) [Matsuda et al., 2005], T. scripta [Kasai et al., 2012], and painted turtle (Chrysemys picta) chromosomes (both 2n = 50) [Badenhorst et al., 2015] further point to the notion that bird and turtle macrochromosomes are precise counterparts of one another. From this basic pattern that was laid down 255 million years ago in the archelosaur ancestor, only about 7 fissions would be needed to establish the familiar karyotype pattern directly observed in most of the major groups of birds including the Ratites, Galliformes, Anseriformes, Columbaea, Pas-...
seriformes, and others. Determining the precise timings that these “final” chromosome fissions occurred cannot be easily achieved with the available evidence. However, if the same fission rate that had occurred for the previous 20 million years continued for another 15 million years, a complete bird-like karyotype would have emerged before the appearance of the earliest dinosaurs and pterosaurs 240 million years ago [Baron et al., 2017].

To paraphrase, our evidence strongly suggests that the chromosomal pattern that we see in the majority of bird species that we choose to karyotype has remained mostly unchanged, not only in most birds, but with a reasonable degree of certainty, in many, if not most, extinct dinosaurs too [O’Connor et al., 2019 and unpublished results]. We thus go so far as suggesting that if we had the opportunity to make chromosome preparations “a la Jurassic Park/World” from the blood of extinct dinosaurs, then karyotype analysis and zoo-FISH results would differ very little from that of a modern chicken (Fig. 8). Those most closely related to modern birds such as the theropods (Spielberg favourites Tyrannosaurus rex and Velociraptor are representatives) are the most likely to have a very avian-like karyotype.

**Karyotype, the Reduction in Genome Size, and the Evolution of Flight**

Dave Burt [Burt, 2002] suggested that some avian microchromosomes were present in the avian ancestor >80 million years ago [Cracraft et al., 2015], purporting that it maybe had a karyotype of around 2n = 60. A very worthy summation of the evidence at the time of writing. As detailed above, however, on the basis of our new evidence over 15 years later, we challenged this notion. We also challenged the notion that this fragmented genome organization (i.e., a karyotype with 2n = 80 chromosomes) accompanied the genome size reduction in birds that has sometimes been associated with the evolution of flight [O’Connor et al., 2018b]. That is, it had previously been suggested that there was some evidence of a correlation between genomes with fewer chromosomes (and no microchromosomes) and larger genome sizes (2.5–3 Gb)

**Fig. 7.** Chromosome evolution from the diapsid ancestor, to the archosaur ancestor, via the theropod dinosaur lineage, to modern birds and finally to chicken. A recognisable avian pattern had evolved just before the dinosaurs emerged 240 million years ago. After this time, chromosome inversions were the principal mechanisms of change. a 275 million years ago. b 255 million years ago. c 240 million years ago at the dawn of the dinosaurs. d Snapshot 100–150 years ago. e Modern chicken.
such as is seen in mammals and crocodilians [St John et al., 2012; Kapusta et al., 2017]. We however suggest that the avian karyotype was in place first, and that this was followed by a reduction in genome size, which was then followed by the evolution of flight.

Why Not Change?

Why then is a near identical karyotype pattern present in most birds, and why has it been in place for around 255 million years? Conceptually there could be two reasons: that is, either there is little opportunity for change and/or the configuration is so evolutionarily successful that there is no pressure to change. In the case of the former, repetitive elements provide substrates for interchromosomal rearrangement, often seen in mammals but hardly even seen in avian species, suggesting that the bird karyotypes provide fewer opportunities for interchromosomal rearrangement because there are fewer recombination hotspots [Kawakami et al., 2014; Smeds et al., 2016], repeat structures [Mason et al., 2016; Gao et al., 2017; Warren et al., 2017], or endogenous retroviruses [Cui et al., 2014; Romanov et al., 2014; Farré et al., 2016]. We also provided evidence of purifying selection acting on some of the smallest microchromosomes [Damas et al., 2018].

A karyotype that hardly changes for 255 million years also however suggests that it is an evolutionarily successful one. The large chromosome number (especially microchromosomes) with high rates of recombination, could, we hypothesise, be the cause of the great variation we see in dinosaurs (including birds) mediated through random chromosome segregation and increased genetic recombination. Phenotypic variation is, of course, the driver of evolution and, although having many chromosomes is not the only mechanism through which variation can be generated, it may nonetheless explain this apparent paradox in dinosaurs of remarkable phenotypic diversity but very little karyotypic diversity. We should nonetheless recognise that it is possible (indeed likely) that some dinosaurs underwent a lot of interchromosomal change.

Kingfishers [Christidis, 1990] (many fissions), parrots [Nanda et al., 2007; O’Connor et al., 2018a], and falcons [Damas et al., 2017; Joseph et al., 2018] (many fusions) are modern examples of where this has occurred. Which specific extinct dinosaur groups did this, may however always remain a mystery.

Chromosome Inversion, the Role of Gene Ontology Analysis, and Dinosaur Phenotype

In the absence of interchromosomal change, the principal mechanism for chromosomal change in dinosaur genome evolution was probably chromosome inversion (also depicted in Fig. 7). Using the ancestral genome reconstruction tool Multiple Genome Rearrangement and Analysis (MGRA) [Avdeyev et al., 2016], we generated contiguous ancestral regions likely to represent the chromosomes of the diapsid ancestor. Compared to extant birds, we identified inversions along the path from the diapsid ancestor to the modern chicken, probably underestimating this number of inversions at 49 chromosome inversions. We believe that the rate of intrachromosomal change increased in modern times, even in the chicken [Romanov et al., 2014]. An even greater degree of change was seen however in some bird clades, particularly the songbirds [Skinner and Griffin, 2012; Zhang G et al., 2014; Farré et al., 2016], the group with the most species. It seems reasonable to hypothesise therefore that periods of faster speciation may have also been accompanied by increased chromosome inversion rates in other dinosaur groups [Skinner and Griffin, 2012; Romanov et al., 2014; O’Connor et al., 2018b].

We [O’Connor et al., 2018b], identified around 400 HSBs delineated by EBRs that characterize dinosaur genome evolution. Other genomic studies in other species (mostly mammals) established that EBRs commonly occur in gene-dense loci, with genes related to lineage-specific biology, transposable elements and other repetitive sequences [Pevzner and Tesler, 2003; International Chicken Genome Sequencing Consortium, 2004; Larkin et al., 2009; Rao et al., 2012]. HSBs on the other hand contain more developmental genes and regulatory elements

Fig. 8. Imagined dinosaur karyotype (deliberately manipulated image based on chicken and spiny soft-shelled turtle chromosomes).
Regions more likely to break, e.g., open chromatin areas or recombination hotspots, and chromosome breaks that do not disrupt key genes or provide a selective advantage, are more likely to be fixed in populations [Farré et al., 2016].

HSB analysis [O’Connor et al., 2018b] using gene ontology (GO) tools, established significant enrichments relevant to amino acid transmembrane transport and signalling plus synapse/neurotransmitter transport, nucleoside metabolism, cell morphogenesis, and cytoskeleton and sensory organ development. Former studies established that HSBs are enriched for GO terms related to evolutionary constant phenotypic features [Larkin et al., 2009], and our dinosaur results support this hypothesis. EBRs on the other hand, are often proposed to be where the “action” in genome evolution occurs [Sankoff, 2009].

We initially found GO terms in avian EBRs that were associated with specific adaptive features, e.g., enrichment for forebrain development in the budgerigar EBRs (consistent with vocal-learning) [Farré et al., 2016]. Later, we identified significant enrichments in genes and single GO terms pertaining to chromatin modification, chromosome organization, and proteasome/signalosome structure [O’Connor et al., 2018b].

The discovery that the avian karyotype likely dates back before the dawn of the dinosaurs complements palaeontological research that demonstrates that features such as feathers and pneumatised skeletons arose first among more ancient dinosaur or archosaurian ancestors [Zhou, 2004; Baron et al., 2017]. Dinosaurs were the dominant group of animals for around 200 million years, with significant radiations occurring in response to two mass extinction events and, despite being almost wiped out by a third (K-Pg), their resilience as a highly diverse and specialist clade (extant birds) [Barrowclough et al., 2016] is evident.

Conclusions
Far from being simply a curating exercise (and a speculative one at that), the study of likely dinosaur karyotype sheds new light into genome evolution, with distinct clues about phenotype and an alternative line of enquiry compared to more established methods. A highly fragmented genome that appears like it has been hit by a meteor is ironic since its pattern was most likely established around 200 million years before Chicxulub hit the earth. Our studies reveal an apparent paradox of an organization that is remarkably unchanging karyotypically during evolution, yet, quite possibly, the driver of such phenotypic evolutionary change. Did the fictional creators of Jurassic Park ever attempt karyotyping? Well, apparently, they did: Check out the last of the Jurassic World movies: at one point, an historical recording of the (now deceased) character Charlotte Lockwood appears, observed by her young daughter Maisie. In the clip, there is a screen with chromosomes on them in the background. There are however no microchromosomes! Unless Spielberg knows something that we don’t, these chromosomes were unlikely to be dinosaur in origin.

Conflict of Interest Statement
The authors have no conflicts of interest to declare.

Funding Sources
This research was funded by the Biotechnology and Biological Sciences Research Council (BB/K008226/1 and BB/J010170/1 to D.M.L, and BB/K008161/1 to D.K.G).

Chicken as a Model for Bird Genome Alignments
(Prepared by P.D. Waters, J.A.M. Graves, and H.R. Patel)

Chicken has long been a model for bird genomes [International Chicken Genome Sequencing Consortium, 2004], but it is only in recent years that a chromosome-level assembly has been available that includes the gene-dense microchromosomes.

The chicken karyotype is characterised by 9 macrochromosomes (defined as ≥35 Mb in the genome assembly), which include the Z, and 30 microchromosomes (defined as <35 Mb) that are characteristic of bird genomes [Mendonça et al., 2016]. Like all birds, chicken has a ZW female/ZZ male sex chromosome system, in which sex is determined by dosage of the gene DMRT1 on the Z chromosome [Smith et al., 2009].

Both the macro- and microchromosomes are conserved across the most distantly related bird lineages [Waters et al., 2021], with the exception of some clades that have undergone rearrangement [Nanda et al., 2006; Huang et al., 2022] (Fig. 9). In fact, strong homologies are conserved with reptiles (lizards, snakes, turtles, and alligators), with differences mostly explained by chromosome fusions (less often fissions) so that large regions of synteny are retained [Waters et al., 2021].

Strikingly, it was demonstrated that almost all amphioxus chromosomes shared homology with one or two bird microchromosomes, in addition to two or three regions of macrochromosomes [Waters et al., 2021]. These multiple regions of amphioxus homology to the bird genome are likely to represent the four copies resulting...
from two rounds of whole-genome duplication in vertebrates [Simakov et al., 2020]. The fact that each amphioxus chromosome shared striking homology with one or two bird microchromosomes suggests that each bird microchromosome represents one copy of an ancestral chromosome.

Microchromosomes are CG-rich and gene-dense [Waters et al., 2021], as are amphioxus chromosomes. Additionally, microchromosomes cluster together in the nucleus and have many more interactions with each other than with the macrochromosomes [Liu J et al., 2021; Waters et al., 2021]. Therefore, they are physically isolated from macrochromosomes, so are less likely to fuse with them.

It is simplest to envisage that after genome duplication one complete copy was rapidly packaged up in the interior of the nucleus and isolated from the other genome duplicate, rather than a mix of chromosomes from duplications being isolated. Given their tight association in the nucleus, it is somewhat surprising that microchromosomes have not all fused with each other in birds, as they have in crocodiles and mammals. This may have resulted from the selective advantage of higher recombination due to obligate cross overs on each chromosome and their independent segregation at meiosis.

Here we report whole-genome alignments of the chicken genome with those of other bird representatives with chromosome-level assemblies. The chicken genome aligns very well with the emu genome, which is considered to represent the ancestral bird genome [Liu J et al., 2021]. The only whole-chromosome difference is the fusion of a microchromosome to chromosome 4, uniquely in chicken [Shetty et al., 1999; Burt, 2002]. There have been some small internal rearrangements at the termini of macrochromosomes, two that are chick-specific and three that are shared with other carinate clades so presumably occurred in the ancestor of all carinates. The Z chromosome has two large internal rearrangements with respect to the emu Z, but these are shared by other carinates.

As a gold standard genome, we propose that the chicken assembly is the best model for the bird genome. It has few autosomal rearrangements relative to the ancestral genome, and four of these are shared by other carinates.

**Fig. 9.** Bird whole-genome alignments. a Alignment of chicken to zebra finch [Rhie et al., 2021], Anna’s hummingbird [Rhie et al., 2021], superb fairywren [Penalba et al., 2020], and the ancestral emu genome [Liu J et al., 2021]. b Alignment of the chicken genome to species that have undergone significant chromosome rearrangement: saker falcon, California condor [Robinson et al., 2021], and golden eagle [Mead et al., 2021]. Macrochromosomes are green, microchromosomes purple, and Z chromosomes orange. Green ribbons show macro-to-macro homologies, purple show micro-to-micro homologies, blue show macro-to-micro homologies, and grey show Z-to-Z homologies.
All microchromosomes are represented in the most recent assembly. Therefore, whole-genome alignments to it can be used to assess assembly errors and/or real rearrangements in new bird assemblies for which karyotype information is available. This presents a more robust comparison than another popular bird model, the zebra finch, which has undergone macrochromosome fission and internal rearrangement (Fig. 9). It is also more useful for aligning the genomes of birds with very rearranged chromosomes, since the multiple rearrangements are independent in different clades (e.g., golden eagle and falcon/parrot, Fig. 9).

Conflict of Interest Statement
There are no conflicts of interest to declare.

Funding Sources
P.D.W. is supported by Australian Research Council Discovery Projects (DP180100931, DP210103512, and DP220101429). J.M.G. is supported by Australian Research Council Discovery Projects (DP210103512 and DP220101429). H.R.P. is supported by the Australian National University and NHMRC Synergy Grant (APP2011277). The authors acknowledge the provision of computing resources provided by the Australian BioCommons Leadership Share (ABLeS) program. This program is co-funded by Bioplatforms Australia (enabled by NCRIS), the National Computational Infrastructure and Pawsey Supercomputing Centre.

The Avian W Chromosome: Simply a Y with a Different Name?
(Prepared by P.D. Price, T.F. Rogers, and A.E. Wright)

Sex chromosomes have long fascinated biologists due to their unique gene content and evolutionary trajectories relative to the rest of the genome [Furman et al., 2020]. In particular, the halting of recombination between sex chromosome pairs has resulted in the evolution of highly degenerate sex-limited W and Y chromosomes in many species [Charlesworth, 1991]. Identifying the function of these chromosomes and understanding if and how they can resist the degenerative forces arising from reduced recombination has been the focus of numerous studies [Bachtrog et al., 2011].

We now know a considerable amount about Y chromosomes, despite the difficulties in sequencing highly heterochromatic and repetitive genomic regions [Tomaszkiewicz et al., 2017]. Their evolution is typically characterised by the accumulation of genes with male-specific functions, large-scale gene amplification, and rapid turnover of gene content across lineages [Bachtrog, 2013; Subrini and Turner, 2021]. In contrast, our understanding of the W chromosome has lagged. However, the last decade has seen an explosion in the number of W-linked genes sequenced across birds [Zhou Q et al., 2014; Bravo et al., 2021], ranging from songbirds [Smeds et al., 2015; Xu et al., 2019; Sigeman et al., 2021; Huang et al., 2022; Warmuth et al., 2022] to fowl [Moghadam et al., 2012; Ayers et al., 2013; Wright et al., 2014] to paleognaths [Xu and Zhou, 2020; Liu J et al., 2021], and a reference assembly of the chicken W chromosome [Warren et al., 2017]. In theory, W chromosomes are in many ways comparable to Y chromosomes, as both are sex-limited and often don’t recombine, and so they might be expected to share similar evolutionary fates. However, there are key differences, most notably that the W chromosome is limited to females whereas the Y chromosome is only present in males [Bachtrog et al., 2011; Mank, 2012]. Below, we outline new insights into avian W chromosomal evolution and ask whether W and Y chromosomes are really that different.

What Are the Evolutionary Dynamics of W Chromosomes across Birds?
It has been known for decades that the chicken W chromosome is a degenerated version of the Z, with the most recent build of the W reference (GRCg7b) identifying only ~80 protein-coding genes across ~9 Mb [Warren et al., 2017]. However, establishing whether the chicken W chromosome is representative of the avian W more generally has only recently been possible due to the plethora of W-linked sequences now available across the avian phylogeny [Wang Z et al., 2014; Zhou Q et al., 2014; Xu et al., 2019; Sigeman et al., 2020].

Sex chromosomes diverge as recombination is suppressed between them, typically assumed to occur in a stepwise process through sequential inversions [Charlesworth et al., 2005]. Consistent with this, ‘strata’ of different ages can be detected on avian Z and W chromosomes [Wang Z et al., 2014; Wright et al., 2014]. These strata are thought to reflect the halting of recombination through large-scale chromosomal rearrangements, such as inversions [Wright et al., 2016]. However, recent evidence suggests that recombination suppression at the earliest stages of avian sex chromosome divergence is a more mosaic and gradual process [Sigeman et al., 2021]. This cessation of recombination has occurred independently in different avian lineages [Zhou Q et al., 2014] and many species exhibit a heavily degraded W chromosome, similar to the chicken.

Despite degeneration proceeding independently across birds, the set of ancestral genes retained on the W chromosome is remarkably conserved, suggesting that...
The avian W chromosome lacks a candidate sex-determining gene [Schmid et al., 2015]. Instead, sex in birds is determined by dosage of the Z-linked gene, DMRT1 [Hirst et al., 2017]. However, there are several lines of evidence implicating the avian W chromosome in female fertility, although the precise functions of genes on the W have yet to be defined. First, W-linked genes are highly expressed in developing chicken ovaries [Moghadam et al., 2012; Ayers et al., 2013; Xu and Zhou, 2020]. This is consistent with feminization of the W [Mank et al., 2010], as key stages of oogenesis are restricted to embryogenesis unlike spermatogenesis, which is a continuous process throughout adult life. Second, W-linked genes expressed during late female development are convergently upregulated in the ovaries of chicken layer breeds subject to artificial selection for fecundity relative to their modern ancestor, the Red Jungle Fowl, and other chicken breeds [Moghadam et al., 2012].

However, unlike most Y-linked genes, which typically exhibit testes-specific expression [Subrini and Turner, 2021], expression of genes on the avian W chromosome is not limited to the ovary. Instead, studies from chicken and collared flycatcher show that many W genes are active in both somatic and reproductive tissue [Smeds et al., 2015; Bellott et al., 2017; Xu and Zhou, 2020]. While this does not preclude a specific role of the W chromosome in oogenesis, it has led to suggestions that this chromosome has instead been selected to maintain gene dosage and ancestral expression levels of essential genes. Consistent with this, many avian W-linked genes are subject to purifying selection [Wright et al., 2014; Sigeman et al., 2021], exhibit a high degree of sequence conservation as well as similar expression patterns to their Z-linked partner [Ayers et al., 2013; Smeds et al., 2015; Xu and Zhou, 2020], and have human orthologs that exhibit detrimental effects when haploid [Bellott et al., 2017; Xu et al., 2019; Bellott and Page, 2021; Sigeman et al., 2021].

It is plausible that apparent differences in the function of Y and W chromosomes could arise from their contrasting inheritance patterns. For instance, W-linked genes, which only pass through the female germ line, are not exposed to sperm competition and so might be subject to weaker sex-specific selection than genes on the Y chromosome. However, it is worth noting that our understanding of the function of the avian W is based on expression data from a limited number of species (chicken and flycatcher) taken across whole, heterogeneous, adult tissue. This precludes accurate contrasts of expression between Z and W orthologs [Price et al., 2022] and so could lead to false inferences of selection to maintain gene dos-
Several mechanisms have been proposed to drive the evolution of multi-copy gene families on the Y, including meiotic drive, sperm competition, genetic drift, and gene conversion [Skaletsky et al., 2003; Ellis et al., 2011; Cocquet et al., 2012; Larson et al., 2018; Bachtrog et al., 2019; Vegesna et al., 2020]. In theory, the strength of these processes might differ between the Y and W due to their contrasting inheritance patterns [Wright and Mank, 2013]. Notably, the Y chromosome is exposed to spermatogenesis, whereas the W is subject to oogenesis, and this likely leads to differences in the potential for antagonistic co-evolution between the X and Y versus the Z and W. Antagonistic co-evolution is predicted to drive the co-amplification of genes on sex chromosomes but should be weaker during oogenesis than spermatogenesis, potentially explaining the limited number of W-linked multi-copy gene families [Bachtrog, 2020]. Targeted avian gene knockouts [Ioannidis et al., 2021] provide an exciting opportunity to elucidate the functionality of HINTW copies, whether this varies across avian species, and the potential for antagonism between W and Z orthologs.

**How Do Multi-Copy Gene Families Evolve on the W Chromosome?**

Y chromosome degeneration is frequently characterised by massive gene amplification where many remaining Y-linked genes persist as members of multi-copy gene families [Skaletsky et al., 2003; Soh et al., 2014; Bachtrog et al., 2019; Vegesna et al., 2020]. However, until recently, gene amplification on the W chromosome has received comparatively less attention and it remained unclear whether large-scale gene amplification is a general feature of sex chromosome evolution or a peculiar quirk of the Y.

A handful of W-linked multi-copy gene families have been identified in a limited number of avian species [Hori et al., 2000; Backström et al., 2005; Davis et al., 2010; Smeds et al., 2015; Rogers et al., 2021]. The most comprehensively studied of these is histidine triad nucleotide-binding protein (HINTW), an ampicolic gene family that is hypothesized to play a role in female reproduction and oogenesis [O’Neill et al., 2000; Ceplitis and Ellegren, 2004]. At present, approximately 10 different copies of HINTW are annotated in the most recent chicken W chromosome assembly (GRCg7b), however, this is likely an underestimation with previous studies estimating over 40 copies [Hori et al., 2000; Backström et al., 2005]. Furthermore, large-scale amplification of HINTW is conserved across avian non-ratites [Hori et al., 2000]. Currently, evidence for the functionality of HINTW is lacking. However, it is known that HINT can form a heterodimer and the amino acid residues that form the dimer binding site are conserved in many HINTW copies, although many copies are nonfunctional [Hori et al., 2000; O’Neill et al., 2000]. Therefore, HINTW might act to disrupt the function of its Z-linked ortholog (HINTZ) by forming a heterodimer. Interestingly, the size of the HINTW gene family varies between chicken layer breeds subject to artificial selection for fecundity relative to other chicken breeds [Rogers et al., 2021], potentially suggesting a role of female-specific selection in driving gene amplification, although this relationship was absent across duck breeds.

The paucity of multi-copy gene families on the avian W chromosome is in stark contrast to the abundance of ampliconic genes often present on the Y chromosome. Several mechanisms have been proposed to drive the evolution of multi-copy gene families on the Y, including meiotic drive, sperm competition, genetic drift, and gene conversion [Skaletsky et al., 2003; Ellis et al., 2011; Cocquet et al., 2012; Larson et al., 2018; Bachtrog et al., 2019; Vegesna et al., 2020]. In theory, the strength of these processes might differ between the Y and W due to their contrasting inheritance patterns [Wright and Mank, 2013]. Notably, the Y chromosome is exposed to spermatogenesis, whereas the W is subject to oogenesis, and this likely leads to differences in the potential for antagonistic co-evolution between the X and Y versus the Z and W. Antagonistic co-evolution is predicted to drive the co-amplification of genes on sex chromosomes but should be weaker during oogenesis than spermatogenesis, potentially explaining the limited number of W-linked multi-copy gene families [Bachtrog, 2020]. Targeted avian gene knockouts [Ioannidis et al., 2021] provide an exciting opportunity to elucidate the functionality of HINTW copies, whether this varies across avian species, and the potential for antagonism between W and Z orthologs.

**Is There a “Toxic W” Effect?**

There appears to be a cost for males to carrying a degenerated Y chromosome [Brown et al., 2020; Xirocostas et al., 2020; Nguyen and Bachtrog, 2021; Connallon et al., 2022], where males in species with XY chromosomes tend to die earlier [Xirocostas et al., 2020]. Several hypotheses have been put forward to explain this phenomenon, including the presence of deleterious recessive mutations on the single X in males that would otherwise be shielded in females (“unguarded X”) or the accumulation of mutations and repetitive elements on the Y chromosome (“toxic Y”). There is also evidence that the Y chromosome acts as a heterochromatin sink, reducing the efficiency of heterochromatin maintenance across the rest of the male genome [Francisco and Lemos, 2014; Brown et al., 2020].

Similar processes may operate on the W chromosome, where females exhibit a shorter lifespan than males across a range of species [Xirocostas et al., 2020]. Consistent with a “toxic W” effect, the avian W chromosome is a haven for repetitive material and transposable elements in several species. For instance, females in species with a degenerate W carry between 20 and 90% more endogenous retroviruses than males [Peona et al., 2021]. Furthermore, transposable element suppression is less effective on the W chromosome than the rest of the genome, leading to higher expression of transposable elements in females [Warmuth et al., 2022]. Although transposable elements can facilitate adaptive evolution, they also have the potential to reduce fitness through the disruption of gene activ-
ity and the promotion of chromosomal rearrangements [McDonald, 1993]. In theory, they may also contribute to an increased chance of female sterility in hybrids, where mechanistic mismatches between transposable repressor mechanisms and the W chromosome lead to reduced female fertility. This would provide further support for Haldane’s rule where the heterogametic sex is more likely sterile in hybrids [Haldane, 1922].

**Final Remarks**

Recent studies have provided new insight into avian W chromosome evolution, challenging the traditional view that the avian W is genetically inert and highly conserved across species. There are clear parallels with Y chromosomal evolution but also key differences, primarily regarding the relative importance of the W in reproduction and fertility traits. Recent technological advances offer new potential to resolve uncertainty over the functionality of the W, for instance by using single-cell RNA-seq to establish fine-scale expression patterns of Z- and W-linked genes through development and across species [e.g., Estermann et al., 2020] and targeted gene knockouts to test gene function [e.g., Ioannidis et al., 2021]. Therefore, the next couple of years hold much promise for disentangling the function and evolution of the W chromosome in birds.

**Conflict of Interest Statement**

The authors declare no conflicts of interest.

**Funding Sources**

This work was funded by a NERC Independent Research Fellowship to A.E.W. (NE/N013948/1) and a NERC ACCE DTP to P.D.P.

**Recovering the “Missing” Avian Genes Using Multi-Omics Data**

(Prepared by Z.-T. Yin, J. Smith, and Z.-C. Hou)

Gene gain and loss are common events in the evolution of species, especially of birds, which have evolved many unique characteristics such as feathers, wings and flight capabilities, strong and lightweight skeletons, toothless beaks, high metabolic rates and heat absorption, sex, and unique respiratory and excretory systems [Kennedy and Veevers, 1976; Blomme et al., 2006]. The release of the first chicken genome provided the basis for systematic analysis of the similarities and differences between vertebrate and avian genomes [International Chicken Genome Sequencing Consortium, 2004]. In comparison with other amniotes, bird genomes are more compact, and this difference may be related to the overall smaller cell size [Hughes and Hughes, 1995; Hughes and Friedman, 2008]. The reductions in genome size may be the result of the loss of noncoding DNA sequences, with bird genomes having less repetitive DNA, fewer pseudogenes, and shorter introns than mammalian genomes [International Chicken Genome Sequencing Consortium, 2004; Hughes and Piontkivska, 2005]. Importantly, the evolution of avian genomes also appears to involve the loss of protein-coding genes, as the total number of uniquely identified avian coding genes is much smaller than in other tetrapods (i.e., 23,294 in humans, GRCh38.p14; 19,404 in lizards, AnoCar2.0; 17,007 in chickens, GRCg7b). Paralog analysis revealed a higher overall incidence of gene families with fewer members in birds compared to other vertebrates [Hughes and Friedman, 2008]. Likewise, birds have a high rate of chromosomal rearrangements compared to other organisms, all of which may result in the deletion of protein-coding genes [Backström et al., 2010]. In recent years, the genomes of a large number of birds and lizards have been assembled and annotated, including zebra finches [Warren et al., 2010], chickens [International Chicken Genome Sequencing Consortium, 2004], turkeys [Dalloul et al., 2010], and duck [Zhu et al., 2021]. Moreover, large-scale bird genome projects [Jarvis et al., 2014; Zhang G et al., 2014], and chicken pan-genomes [Wang K et al., 2021; Li M et al., 2022] have also generated considerable genomic data. These large comparative genomic datasets identified hundreds of lost genomic blocks in the bird genomes, and also suggested that hundreds of genes are missing in birds [Lovell et al., 2014; Zhang G et al., 2014].

The missing genes seem to be directly related to the unique physiological phenomena of birds. Several functionally important genes in mammals are supposed “missing” in chickens and have caused long-debated questions in bird biology. Spurious discovery of the missing/hidden genes in the bird genome has continued for decades. Previously, BGN [Blaschke et al., 1996], CORO1A [Xavier et al., 2008], MAPK3 [Lemoine et al., 2009], MMP14 [Simsa et al., 2007], TBX6 [Lardelli, 2003; Ahn et al., 2012], TSSK4 [Shang et al., 2013], and five adipokine genes [Dakovic et al., 2014] were reported to be missing in birds, however, several long-debated genes including TNF-α, and leptin have been cloned in birds [Prokop et al., 2014; Seroussi et al., 2016; Rohde et al., 2018]. This hide-and-seek game still continues, and does not appear to be ending anytime soon [Elleder and Kaspers, 2019].
Here we summarize recent efforts using multi-omics data to probe those genes missing/hidden in avian genomes.

**Reconstruction of Missing Genes in the Chicken Genome**

While the hypothesis of missing genes in birds has been proposed for decades, researchers have found that some of the missing genes were, in fact, present in chickens or other birds. In the presence of large gaps and imperfect gene annotation in the genome, de novo assembly of gene sequences using RNA-seq is considered to be an efficient way to identify unannotated genes in the genome. Attempts that only used a few tissues/organs have identified many missing genes in birds [Hron et al., 2015; Bornelov et al., 2017; Botero-Castro et al., 2017]. Recently, we used the raw data from 26 chicken tissues downloaded from the GenBank database to assemble and obtain 2,048,631 transcripts and identified 589 missing genes in birds [Yin et al., 2019b].

At the same time, the continuity and integrity of chicken genome assemblies have been rapidly improving. The chicken genome released in 2017 was assembled by third-generation sequencing technology, and the number of annotated genes increased significantly (2,768 noncoding and 1,911 protein-coding genes) [Warren et al., 2017]. In the Gallus_gallus-5.0 genome, 442 (77.41%, from a total of 571) genes thought to be missing in chickens [in Lovell et al., 2014, see Table S1, Table S6, plus select entries in Table S4 and Table S18] were annotated, indicating that there is no systematic deletion of genes in birds. With the development of sequencing and hybrid assembly technology, the genomes of different chicken breeds continue to be assembled and another 136 missing genes were further annotated in our recently assembled Silkie genome (unpublished). To date, it has now been shown that 528 (92.47%) genes that were thought to be missing, actually exist in chickens. This has been made possible by exploiting a large amount of multi-omics data available in chicken and has led to the revelation of genes with important functions such as TNF-α and leptin [Seroussi et al., 2016; Rohde et al., 2018]. Recent large-scale chicken pan-genome data have also identified thousands of genes that are not presented in the current chicken reference genome (Li M et al., 2022).

**Reconstruction of Missing Genes from Other Birds**

In addition to chicken, researchers have reconstructed many genes thought to be missing from other birds. We collected data from various important tissues from duck (24), pigeon (11), goose (8), and zebra finch (22) [Yin et al., 2019b], and an avian transcriptomic database containing a total of 9,296,247 transcripts was constructed by de novo transcriptome assembly. From this, we identified several genes in duck (583), pigeon (558), goose (537), and zebra finch (543) from 806 genes that were thought to be missing in birds [in Lovell et al., 2014, see Table S1; in Zhang G et al., 2014 see Table S10]. Only 135 genes were not found in this bird transcriptome database. The number of missing genes reconstructed in different birds by de novo assembly of large transcriptome data is similar, indicating that these genes thought to be missing exist across different bird species.

In recent years, duck functional genomics has developed rapidly. We have assembled the Mallard, Pekin duck, and Shaoxing laying duck genomes using a combination of third-generation sequencing, Bionano, and Hi-C sequencing technologies. These have proved to be a rich source of genetic information [Zhu et al., 2021]. In the Mallard duck the CAU_wild 1.0 genome has 1,872 more protein-coding genes annotated than the previous CAU 1.0 genome, including 89 genes previously thought to be missing in birds. Among these 89 genes, 5 genes have become pseudogenes, losing part of their gene function, 3 genes have been annotated as IncRNAs, and the remaining 81 genes remain as protein-coding genes. In addition, 240 genes were annotated as paralogous genes, and 108 genes had similar segments in the genome. Mining large multi-omics data assemblies and annotations now reveals that only 10 genes (from a total of 806 missing genes), to date, have not been reconstructed in birds, with the rest of the genes thought to be missing in birds having been shown to actually exist. The recovered gene list is shown in online supplementary Material 4.

**Development of New Methods to Identify More Missing Genes**

Summarizing the characteristics of these reconstructed missing genes in birds and the reasons why they are thought to be missing can provide insights and methods for us to identify more missing genes. First, these reconstructed gene sequences have high GC content and length in many birds. The GC content of most of these “missing” genes is more than 60%, and few genes even have over 80% (the median GC content of the chicken genome is 42.22% and the median GC content of the duck genome is 41.99%) [Hron et al., 2015; Bornelov et al., 2017; Botero-Castro et al., 2017; Yin et al., 2019b]. At the same time, the multi-tissue transcriptome expression profiles of birds
showed that most of the reconstructed genes usually have strong tissue-specific expression. These genes are generally expressed predominantly in one tissue and are rarely expressed in the other tissues [Yin ZT et al., 2019b]. High-throughput transcriptome-based assembly approaches have limitations for fully recovering missing genes due to technical factors such as the PCR amplification bias against GC-rich fragments [Beauclair et al., 2019]. Expression patterns, i.e., tissue-specific expression patterns, and low expression, also limit the ability for full transcriptome assembly. Now, the third-generation sequencing technologies, which have less GC bias, such as single-molecule real-time (SMRT) and nanopore sequencing technologies, can obtain full-length transcripts directly, without assembly [Yin et al., 2019a; Kuo et al., 2020]. The missing genes will continue to be discovered with the accumulation of full-length transcriptome data from more avian tissues from different physiological conditions.

Furthermore, the missing genes annotated in the chicken and duck genomes are mainly distributed on the microchromosomes, the ends of the chromosomes, and within regions showing a high content of tandem repeats clustering with non-canonical DNA structures [Zhu et al., 2021; Li M et al., 2022]. Long repetitive regions [Treangen and Salzberg, 2011], regions of high GC content [Chen et al., 2013], telomeric regions, fragmented microchromosomes [O’Connor et al., 2019], and adaptive assembly strategies have always proved problematic for enabling complete bird genome assembly. To fully resolve the whole chicken gene sets, a telomere-to-telomere (T2T) genome is necessary. The recently completed human T2T genome has now paved the way for the finished bird genome assembly [Miga et al., 2020; Hoyt et al., 2022; Mao and Zhang, 2022; Nurk et al., 2022]. Ultra-long ONT sequencing, high-precision HiFi sequencing data, multitype auxiliary assembly data, and hybrid assembly using multiple strategies will greatly promote the quality of bird genome assembly [Sohn and Nam, 2018]. For large presence/absence variations within species, we can enrich genomic information by constructing high-quality multi-breed pan-genomes [Vernikos et al., 2015]. The Bird 10,000 Genomes (B10K) Project [Zhang et al., 2015] has generated insightful results and the future bird T2T genome and pan-genome will undoubtedly reveal more genes. This complete gene map of birds will be critical for the further understanding of the biology and evolution of birds.

Finally, precise genome annotation will also provide the necessary sequence and structural information for mining more genes in birds. Annotation errors are unavoidable in genome annotation using automated processes, especially for some protein-coding genes that cannot be annotated in complex and high GC regions [Salzberg, 2019]. While applying full-length transcriptomic data for genome annotation [Nudelman et al., 2018; Wang X et al., 2019; Kuo et al., 2020], the use of novel annotation methods developed based on machine learning can further improve the accuracy of annotation [Mahood et al., 2020; Stiehler et al., 2020]. More accurate manual annotation of important genome regions is also necessary for novel gene identification [Dunn et al., 2019]. It can be seen that, with the continuous development of omics technology and analysis methods, the genome information will be more complete, the annotation will be more accurate, and the genes that were previously thought to be missing in birds will continue to be discovered.

Statement of Ethics
All experiments with birds were performed under the guidance of ethical regulations from the Animal Care and Use Committee of China Agricultural University, Beijing, China.

Conflict of Interest Statement
The authors declare no competing interests.

Funding Sources
The work was supported by the National Waterfowl-Industry Technology Research System (CARS-42), the National Nature Science Foundation of China (31972525, 31572388), Beijing Joint Research Program for Germplasm Innovation and New Variety Breeding (G20220628007), Beijing Municipal Science & Technology Commission (Z211100004621007).

Data Availability Statement
Data have been submitted to the public databases under the following accession numbers: The raw data for assembling the transcriptome database of chicken, duck, goose, pigeon, and zebra finch were deposited in Sequence Read Archive (SRA) database under the accession number SRP141084. The Mallard genome is stored in NCBI under accession number PRJNA554956. The raw data for the Silkie genome assembly can be found in the SRA database under the accession number PRJNA805080 (unpublished data).

Evolution of Protein-Coding and Long Noncoding Genes of the Chicken Genome through the Different Genome Assemblies and Their Associated Annotations
(Prepared by F. Degalez, K. Muret, and S. Lagarrigue)

The chicken genome was the first avian genome sequenced because of its importance in human food production, in fundamental biology like the study of devel-
opment or gene function conservation across evolution [International Chicken Genome Sequencing Consortium, 2004]. Since its first version in February 2004 (Galgal2/WUGSC1.0), five new genome assemblies have been released, each improving the genome sequences’ accuracy. Along with these genome assemblies, numerous genome annotations were released, providing at least models for gene loci and transcripts supporting them. Since the first annotated version (Ensembl v22 - May 25, 2004) associated with the Galgal2 assembly, the number of genes and the diversity of their biotypes have increased, especially in 2015 with the introduction of long noncoding RNAs (lncRNAs), which is concurrent with the first initiatives of lncRNA annotation [Chodroff et al., 2010; Necsulea et al., 2014; Li et al., 2015; Muret et al., 2017].

LncRNAs represent a large and heterogeneous class of genes defined by transcripts longer than 200 nucleotides without coding-potential capabilities [Derrien et al., 2012]. They represent a variety of regulatory elements implied in gene expression and can act at different levels by using diverse biological mechanisms based on DNA, RNA, or protein interactions [Guh et al., 2020]. As illustrated in Figure 10, lncRNAs can interact with DNA, RNA, and proteins and act at different molecular levels: nuclear organization (e.g., MALAT1 [Wang X et al., 2021b]/NEAT1 [Yamazaki et al., 2018]) (Fig. 10A), genome integrity (e.g., TERRA [Barral and Déjardin, 2020]) (Fig. 10B), histone marks modification for silencing (e.g., Fendrr [Grote et al., 2013]) or activating (e.g., GATA3-AS1 [Gibbons et al., 2018]) gene transcription (Fig. 10C), loop formation to connect enhancers to promoter regions (e.g., MYMLR [Kajino et al., 2019]) (Fig. 10D). LncRNAs can modulate RNA splicing (e.g., linc-HELLP [van Dijk et al., 2015]) (Fig. 10E), miRNA maturation (e.g., CCAT2 [Yu et al., 2017]/uc.372 [Guo et al., 2018]) (Fig. 10F), and protein translation (e.g., BC1 [Wang et al., 2002]/MC-M3AP-AS1 [Guo C et al., 2020]) (Fig. 10I) or their activity (e.g., NORAD [Munschauer et al., 2018]) (Fig. 10K). They can also control the stabilization or the degradation of molecules as miRNAs (e.g., ROR [Li C et al., 2017]/DSCR8 [Wang Y et al., 2018b]) (Fig. 10G), mRNAs (e.g., PTB-AS [Zhu L et al., 2019]/TINCR [Xu et al., 2015]) (Fig. 10H), and proteins (e.g., PiHL [Deng et al., 2020]/MALAT1 [Yang et al., 2016]) (Fig. 10J). LncRNAs can host small ORFs [Choi et al., 2019] which code for peptides (e.g., CASIMO1 [Polycarpou-Schwarz et al., 2018]/DWORF [Nelson et al., 2016]) (Fig. 10M) or host in their introns small RNAs [Sun Q et al., 2021] (e.g., MCM7 [Agranat-Tamir et al., 2014]/DLEU2 [Morenos et al., 2014]) (Fig. 10N). They can control protein transfers from cytoplasm to nucleus (e.g., NRON [Willingham et al., 2005]) or from nucleus to cytoplasm (e.g., Discn [Wang L et al., 2021]) (Fig. 10L). Finally, they can migrate to other cells with exosomes (e.g., ZFAS1 [Pan et al., 2017]//GAS5 [Chen et al., 2017]) (Fig. 10O).

Through their key roles in gene regulation, lncRNAs are consequently involved in diverse biological and pathophysiological processes [Ponting et al., 2009; Muret et al., 2019; Gil and Ulitsky, 2020; Statello et al., 2021]. Moreover, since most of the trait-associated variations identified by genome-wide association studies (GWAS) concerned noncoding intervals of the genome [Manolio et al., 2009; Bouwman et al., 2018], this reinforces the need to characterize the regulatory regions of domesticated species such as lncRNA genes. LncRNA genes have different characteristics compared to protein-coding genes (PCGs). They are less expressed [Derrien et al., 2012; Muret et al., 2017; Le Béguec et al., 2018; Jehl et al., 2020], explaining why they have been detected only recently – i.e., this last decade – by high-throughput transcriptome sequencing technologies (RNA-seq). Furthermore, lncRNA expression is more specific to tissues, life stages, and conditions than that of PCGs [Cabili et al., 2011; Derrien et al., 2012; Jehl et al., 2020]. The identification of these genic entities is therefore dependent on the variety of RNA-seq data available to detect them.

After presenting the different chicken genome assemblies developed over the last 2 decades, we discuss the associated genome annotations provided by NCBI’s RefSeq and EMBL-EBI’s Ensembl, the two reference annotation databases. We characterize them in terms of number of gene and transcript models, variety of biotypes, or in terms of models that are shared by the two reference databases. We show that lncRNA loci are even less well-known than PCG ones, although, for the latter, knowledge of their transcripts can be further improved. Finally, we discuss the impacts of these weaknesses and the value of gathering different genome annotation resources, in particular, for a better description of lncRNA loci, and then present two initiatives. The MANE project yet limited to the human genome aims to synergize the NCBI’s RefSeq and EMBL-EBI’s Ensembl “gene” databases to establish a consensus annotation. The second project, specifically realized for the chicken, is to provide a “gene” database built from various resources including the NCBI’s RefSeq and EMBL-EBI’s Ensembl databases and other resources such as FAANG multi-tissue resources and NONCODE database. This gene catalog is maintained at each significant update in chicken genome assembly and
genome annotation, the last version of June 2022, associated to the GRCg7b assembly, being composed of 23,926 PCGs and 44,428 lncRNA genes (available at http://www.fragencode.org).

Evolution of the Reference Sequence of the Chicken Genome

As illustrated in Figure 11a, while the overall coverage of Galgal2/WUGSC1.0 was 6.63× [International Chicken Genome Sequencing Consortium, 2004], this parameter is
doubled for the two successive assemblies, i.e., Galgal3/ WUGSC2.1 [NCBI RefSeq, 2006] and Galgal4 [NCBI RefSeq, 2011], which were released in November 2006 and 2011, respectively. Compared to the first version, these include 33 chromosomes (1–28; 32; W/Z; mitochondrial chromosome [MT]) but the number of scaffolds remained very high (~17,000), with 915 gaps between the scaffolds and a scaffold N50 which was quite low (~12 Mb) showing the incompleteness of the chicken genome sequence. Note that scaffold N50 is defined as the sequence length of the shortest scaffold at 50% of the total genome length. With the release of the Galgal5 assembly in December 2015 and with the improvement of sequencing technologies [NCBI RefSeq, 2015; Warren et al., 2017], the average coverage exploded and reached a global depth of 70×, leading to a better knowledge of the genome sequence. A few chromosomes were newly defined (addition of chromosomes 29–31). However, the quality of this genome sequence remained low with a high number of scaffolds (~24,000) and a lower scaffold N50 (~6.4 Mb) than before. These weak performances are likely due to long-read sequencing, which improved the detection of smaller scaffolds thus decreasing the N50 value [Warren et al., 2017].

In March 2018, a new assembly called GRCg6a was released by the Genome Reference Consortium, which has taken the lead concerning the chicken genome assembly previously managed by the International Chicken Genome Consortium [NCBI RefSeq, 2018]. Tremendous progress — due to the addition of long read sequences, improved de novo assembly algorithms, manual annotation of contigs, and integration of finished BAC clone sequences — was made regarding the genome accuracy, as indicated by the drop in the number of scaffolds (from ~24,000 to ~500) with only 68 gaps between the scaffolds and an increase in the scaffold N50 (from ~6.4 M to ~20 M).

Furthermore, with the latest GRCg7 genome assemblies [NCBI RefSeq, 2021a, b], the knowledge of the chicken genome sequence improved even more in two main ways. First, the accuracy of the genome sequence increased due to improvements in sequencing and especially assembly technologies. The chicken genome is now composed of 42 chromosomes (1–39; W/Z; MT), reaching the number observed in the chicken karyotype. This assembly includes more microchromosomes with ~250 scaffolds with no gap between scaffolds, and the scaffold N50 reaching 90 Mb. Second, whereas a Red Jungle Fowl breed (known as RJF #256) was always used in previous assembly versions, a trio of chickens from diverse breeds was used for GRCg7. The new reference sequence was

---

**Fig. 11.** Assembly versions associated with the chicken genome (a) and the number of publications associated with them (b). a RJF, Red Jungle fowl; W–L, White Leghorn; Cov., coverage; Scaf, scaffold; N50, scaffold N50. For more details, see online supplementary Material 5, Table 1. b Blue numbers, articles published using the corresponding assembly during the year 2020. Identification was made on PubMed Central by searching for the assembly name in different formats (e.g., GRCg6a or GRCg6 or Galgal6).
generated from a female offspring from a cross between a broiler female and a white leghorn laying male. The RJF breed, considered to be the descendant of domestic chickens, was used as a good representation of broiler and layer chicken breeds; however, such a choice has a significant impact on the detection of variants leading to the identification of false positives. Since actual breeds have diverged from the RJF, a variation (e.g., SNP) at one position may be detected according to the RJF genome sequence, whereas this position was fixed in the population of interest. As an illustration, we have previously shown, using RNA-seq data from the liver of 11 different breeds (~750 RNA-seq of ~400 birds) aligned on the Galgal5 assembly, that the SNP number with reliable genotypes was on average 549,634 per population, but this number dropped to 339,539 (~38.2%) with a minor allele frequency ≥10% [Jehl et al., 2021]. This drop is mainly due to fixed variants in the populations since the number decreased to 438,837 (~20.2%) after only excluding the fixed variations.

Consequently, two genome assemblies were released in January 2021: GRCg7b representing the broiler breed and considered as the new genome reference, and GRCg7w representing the laying breed and considered as an alternative.

Because of the quite frequent change of the genome assembly compared to the time needed to conduct and publish a scientific study, a lot of works are published in outdated versions which can lead to the publication of misleading results and in disagreement with more recent versions (Fig. 11b). For example, in 2020, two years after the release of GRCg6a, 79 studies using this genome reference were published against 46, 41, and 5 published with the Galgal5, Galgal4, and Galgal3 versions, respectively. Some tools such as Liftoff [Shumate and Salzberg, 2020] or LiftOver [Kuhn et al., 2013] can be used to convert coordinates from one version to another. The first tool is based on the alignments of the gene features from one annotation to another, whereas the second tool is based on alignments of the best/longest syntenic regions for each region of the genome between assemblies (chain files). However, the use of these tools must be done with caution, especially for remote versions, because of important changes in the genome sequence.

Evolution of the Two Reference Genome Annotations: A Breakthrough in 2015 with the Apparition of the IncRNA Gene Biotype

Genome annotation is not only evolving according to the version of the genome assembly but also to the evolution of annotation bioinformatics pipelines and data resources, mainly composed nowadays of RNA-seq data. In Figure 12a, genome annotations, from 2004 with the Galgal2/WUGSC1.0 assembly through 2022 with the GRC7b assembly, produced by the reference centers, NCBI’s RefSeq and EMBL-EBI’s Ensembl, have been analyzed. As illustrated, the gene number has increased, particularly due to the apparition in 2015 of IncRNAs, with 5,763 and 4,641 IncRNAs modeled by NCBI’s RefSeq (v103) and EMBL-EBI’s Ensembl (v94). This increase continues with the last genome annotation v107 (associated with the GRCg7b assembly) provided by EMBL-EBI’s Ensembl with 11,944 IncRNAs compared with 5,504 for v106 (associated with GRCg6a). The number of PCGs remains constant at around 17,000 (see further for more explanation regarding such evolution).

In parallel to the gene number, it is important to make some comments about the transcript models that support these genes. As observed in Figure 12b, the transcript models can still be improved, as illustrated by the numerous changes observed between the two versions v105 and v106 of NCBI’s RefSeq [NCBI RefSeq, 2022]. Only 6.4% of 93,980 transcripts identified in the 106 version are identical to those found in version v105. Such results can also be observed between genome annotations of different genome assemblies (not shown), or between genome annotations from the two reference centers, NCBI and EMBL-EBI, as illustrated in the next section.

Differences between the Latest NCBI RefSeq and EMBL-EBI Ensembl Genome Annotations

For the same genome assembly, the two genome annotation bioinformatics centers, EMBL-EBI and NCBI, do not provide the same annotations, as illustrated in Figure 13. First, EMBL-EBI’s Ensembl provides twice the number of IncRNA gene models compared to NCBI’s RefSeq (shown in Fig. 13a) resulting in a total of 30,108 gene models (associated to 72,689 transcripts) including 17,007 PCGs, 11,944 IncRNAs, and 674 miRNAs compared with 25,638 gene models (associated with 85,704 transcripts) including 18,024 PCGs, 5,791 miRNAs, and 799 miRNAs for RefSeq. These differences can be explained by the sample datasets and the annotation pipeline thresholds used specifically by the two bioinformatics centers. For example, NCBI’s RefSeq does not consider IncRNAs supported by a single mono-exonic transcript in contrast to EMBL-EBI’s Ensembl (with 1,157 IncRNA loci). Second, using the “GffCompare” software [Pertea and Pertea, 2020], we observed that most of the transcript models are different between the two ge-
nome annotations, as shown in Figure 13b. Among the 72,579 transcripts from EMBL-EBI’s Ensembl considered in the analysis, only 17.8% are strictly equal in the NCBI’s RefSeq annotation. More than half (55.9%) are identified as new isoforms of an existing locus and 26.1% (18,922) transcripts are associated with 9,958 new gene loci resulting in more than one-third of the 30,108 gene models from EMBL-EBI’s Ensembl not being known in NCBI’s RefSeq.

Important differences exist between PCG and lncRNA transcripts. For PCG, most of transcripts from EMBL-EBI’s Ensembl (70.8%) are new isoforms of the same gene loci existing in the two databases. These results show that the transcript isoforms are not well described with current RNA-seq resources. Indeed, most of RNA-seq data available in the public database are short-read RNA-seq; the long-read RNA-seq studies using the new technologies such as ONT or PacBio are still very limited [Kuo et al., 2017; Guan et al., 2022] due to the cost of these technologies and their low sequencing depth. For lncRNA, most of the IncRNA transcripts from EMBL-EBI’s Ensembl (77.4%) are considered as new loci compared to NCBI’s RefSeq. The main cause of this very low gene overlap between the two genome annotations is the difficulty in capturing and therefore modeling IncRNAs compared to PCGs, due to specific features of IncRNA. First, IncRNAs are characterized by a global low expression; around less than 10% of the total reads of a sample analyzed by common technologies support IncRNA transcripts [Lagarrigue et al., 2022]. Second, they are tissue-, developmental stage-, and condition-specific [Cabili et al., 2011; Derrien et al., 2012; Jehl et al., 2020], conditions which are not covered by the limited number of RNA-seq samples used by the reference genome annotation centers compared to the tens of thousands of short-read RNA-seq generated by the avian scientific community which are available in the public database.

Moreover, the transcript models from NCBI’s RefSeq are significantly longer than those of EMBL-EBI’s Ensembl, as shown in Figure 13c, particularly for IncRNAs (almost twice the length), with nearly two supplementary exons by transcript (resp. a median of 5 vs. 10 exons/trascript, \( p < 10^{-16} \)) for both PCG and IncRNA models, with median exon sizes which remain similar (\( \sim 250 \) bp). Moreover, NCBI’s RefSeq provides a higher extreme distribution of transcripts per gene for PCGs compared to
EMBL-EBI’s Ensembl (resp. 5 vs. 3 for the third quartile and resp. 9 and 5 for the last decile). Note that these numbers are far below what is described in human (resp. 6, 11, 18 transcripts per gene for the median, the third quartile, and the last decile (EMBL-EBI’s Ensembl v107 with the GRCh38.13 assembly). This discrepancy can be explained in part by the variety of samples each reference used. EMBL-EBI’s Ensembl combines a short-read RNA-seq dataset of 21 tissues from the Roslin Institute (1 stage-condition of a same breed per tissue, 21 samples of individual pool) with a short-read dataset of 7 tissues from the GENE-SWitCH project (3 stages of a same breed, 21 samples of individual pool) [EMBL EBI’s Ensembl, 2022]. NCBI’s RefSeq integrates data from various projects representing more than 20 tissues, different development stages and breeds for a total of 100 and 89 samples for short-read and long-read RNA-seq, respectively, in addition to “Cap Analysis Gene Expression” (CAGE) data including those from the FANTOM project [Lizio et al., 2017] for improving the annotation of transcription start sites [NCBI RefSeq, 2022].

For IncRNAs, the pattern in the distribution of the number of transcripts per gene is inverted between NCBI’s RefSeq and EMBL-EBI’s Ensembl with respectively 2 versus 3 for the third quartile and 3 and 5 for the last decile. Interestingly, these numbers are of the same order of magnitude in human (resp. 1, 2, and 5 transcripts per gene for the median, the third quartile, and the last decile), highlighting the general difficulty in capturing the transcript models associated with IncRNA genes.

**Interest in an Annotation Combining NCBI’s RefSeq and EMBL-EBI’s Ensembl**

In summary, different genome annotations coexist with important differences in transcript models for PCGs and gene models for IncRNAs. Initiatives like the MANE project [Morales et al., 2022] for the human genome aim to synergize the NCBI’s RefSeq and EMBL-EBI’s Ensembl reference genome annotations to establish a consensus, although, so far, these efforts have focused only on PCGs. Such initiatives have yet to exist for livestock species, especially chicken. So far, most RNA-seq studies have ana-

---

**Fig. 13.** Features of the current NCBI RefSeq (v106) and EMBL-EBI Ensembl (v107) genome annotations based on the latest GRCg7b genome assembly. **a** Number of genes and transcripts according to gene biotypes for the two genome annotations. **b** The transcript models were compared between the two annotations according to 4 main classes (Equal isoform, New isoform, New loci, and Artifacts) according to the software “GffCompare” (options: -S --no-merge) [Pertea and Pertea, 2020]. **c** #Tr/gn, number of transcripts per gene; #Ex/tr, number of exons per transcript; Tr. Size, transcript size considering only exonic regions; Ex. Size, exon size; In. Size, intron size. The median transcript sizes between RefSeq and Ensembl are 3,465 bp versus 2,317 bp, respectively, $p < 10^{-16}$ (Wilcoxon rank sum test); for PCG 3,634 bp versus 2,870 bp, $p < 10^{-16}$; for IncRNAs 2,952 bp versus 1,487 bp, $p < 10^{-16}$. 

**End of Document**
tissue studies have been performed in order to provide a trait or tissue of interest whereas the lncRNA catalogues few lncRNAs have been highlighted as associated with the trait in the species of interest. In most of these studies, animal groups contrasted for an economically important differential expression in a given tissue between breeds or strains. In addition to the gene atlas (i.e., gtf file), a functional annotation of the genes across 40 tissues using their high tissue specificity. We can point two studies, that are part of the Functional Annotation of ANimal Genome consortium (FAANG) [Andersson et al., 2015], which have provided a multispecies IncRNA annotation: the first, Foissac et al. [2019], used 3 tissues of 4 female and male biological replicates of 4 farm species including the chicken updated to 12 tissues (personal communication); the second one, Kern et al. [2018], used 8 tissues of 2 female and male biological replicates of chicken, pig and cattle and was recently updated to 19 tissues for the chicken species [Guan et al., 2022]. Nevertheless, for a given species such as chicken, these studies remain limited due to the range of tissues, stage of development, condition that may exist.

In this context, we proposed since 2020 to provide a comprehensive gene catalog for chicken by gathering different resources, including EMBL-EBI’s Ensembl, NCBI’s RefSeq, and other multi-tissue databases, that we update at each important change of chicken genome assembly and annotation. Since the release of the new GRCg7b chicken genome sequence, we have recently updated the gene catalogue of 52,075 genes published in 2020 [Jehl et al., 2020], considering the last NCBI’s RefSeq and EMBL-EBI’s Ensembl annotations available in June 2022. First, we gathered the two genome annotation references, i.e., the v106 of RefSeq and the v107 of Ensembl resources. In addition to these two references, we chose to gather the two updated FAANG multi-tissue resources described above [Foissac et al., 2019; Guan et al., 2022], in which IncRNAs have been modeled in parallel with the PCG loci. The NONCODE resource composed only of IncRNA loci has also been used, even if this resource has not been updated since 2014 for the chicken [Zhao et al., 2021]. As a result, the EMBL-EBI’s Ensembl and NCBI’s RefSeq references grew respectively from 17,007 to 18,024 to 23,926 PCGs and from 11,944 to 5,791 to 44,428 IncRNA genes. This atlas associated to GRCg7b assembly is publicly available at http://www.fragencode.org [Degalez et al., in preparation] as the previous ones published in 2020 and associated to the Galgal5 and GRCg6a genome assemblies. In addition to the gene atlas (i.e., gtf file), a functional annotation of the genes across 40 tissues using different public resources is also provided as well as the IncRNA gene naming according to the official HUGO gene nomenclature committee (HGNIC). Briefly, for the IncRNAs with an unknown function (frequent cases), the IncRNA adopts the symbol gene name of the gene harboring it, enriched by a suffix describing its geno-
Conclusion

This review provides an overview of the evolution of chicken genome assemblies from 2004 to June 2022 and their genome annotations provided by the two most widely used annotation databases, NCBI’s RefSeq and EMBL-EBI’s Ensembl. We show a great evolution of the genome assembly through 6 different versions due to various technical and technological advances, the latest GRCg7b offers a genome reference sequence composed of 42 chromosomes (1–39; W/Z; MT), reaching the number observed in the chicken karyotype with more microchromosomes than the previous versions and with no gap between the ~250 scaffolds. Moreover, we show that the annotation of the chicken genome is constantly evolving according to the version of the genome assembly, the evolution of bioinformatic annotation pipelines, and the RNA-seq data resources. We can highlight the recent emergence, in 2015, of IncRNA models in genome annotations associated with the Galgal5 genome assembly. Concerning the last GRCg7b genome assembly, the two reference genome annotations are quite different with 18,024 PCGs and 5,791 IncRNAs reported for NCBI’s RefSeq and 17,007 PCGs and 11,944 IncRNAs for EMBL-EBI’s Ensembl. The PCG entities mainly differ at the transcript model level whereas IncRNAs differ both at the transcript and gene loci levels. Gene loci display a very low overlap mainly explained by the specific features of IncRNAs (low expression, high tissue-, condition-specificity, …) and the limited number of RNA-seq samples used for generating these catalogs. To facilitate the reconstruction of full-length transcript models, and so accurate gene models, annotation centers will benefit in the near future from new technologies such as ONT or PacBio allowing long-read RNA sequencing. However, for properly catching IncRNAs, the low sequencing depths of these long-read technologies compared to short-read RNA-seq require preliminarily capture strategies used to boost the concentration of low-abundance transcripts in cDNA libraries. Such strategies have been applied to 4 human and mouse tissues by the GENCODE consortium [Lagarde et al., 2017]. However, the low sequencing depths and the high cost of these technologies limit for the moment their wider use. The main fuel of the genome annotation databases remains the short-read RNA-seq massively generated by the scientific community. In this context, to increase the completeness of the chicken genome annotation, especially IncRNAs, we highlight the interest to combine the two reference NCBI’s RefSeq and EMBL-EBI’s Ensembl genome annotation databases and even other databases and present two initiatives. One of them, applied to the chicken species and updated at each important change of the reference annotation, provides a catalogue of 23,926 PCG and 44,428 IncRNA gene models which includes all the gene loci of the last versions (June 2022) of NCBI’s RefSeq and EMBL-EBI’s Ensembl.

Conflict of Interest Statement
The authors have no conflicts of interest to declare.

Funding Sources
F. Degalez is a Ph.D. fellow supported by the Brittany region (France) and the INRAe’s Animal Genetics division.

Chicken Cell Atlas: Science and DeSci
(Prepared by M. Yamagata)

Chickens are not only widely consumed for their eggs and meat but are also used as a model species for biological and medical research [Stern, 2005; Burt, 2007; Haniissa et al., 2021]. To advance our understanding of this species, we need to chart the types and properties of all chicken cells across all organs and tissues, build a reference map of the mature and developing chicken bodies, and provide the resources for studying the biology of this species [Yamagata, 2022]. Similar to the mouse and human cell atlas enterprises in progress (see below), this project will generate single-cell transcriptome data for chickens, characterize each cell type, and provide foundational information integrating molecular, spatial, and temporal modalities. It will facilitate fundamental studies of chickens and other birds, including cell biology, molecular biology, developmental biology, neuroscience, physiology, oncology, virology, behavior, ecology, evolution, and animal husbandry.

The Current State of the Chicken Cell Atlas

Recent advances in single-cell RNA sequencing (scRNA-seq) have had significant effects on the study of complex tissues, leading to the discovery of novel cell types, cell states, and biomarkers [Stuart and Satija, 2019; Ruecken and Theis, 2019; Alfieri et al., 2022; Zeng, 2022]. The scRNA-seq technology has opened up a plethora of opportunities to perform novel studies using new and classic model animals, including chickens (Gallus gallus) [Yamagata, 2022] and other birds [Chen et al., 2021; Colquitt et al., 2021]. A chicken cell atlas project (aka,
Tabula Gallus) has been proposed to create a cell atlas of all tissues in the mature and developing chicken [Yamagata, 2022]. Like several ongoing cell atlas projects (see below), the chicken project will collect scRNA-seq data for chickens, characterize each cell type, and eventually make available information that integrates diverse modalities.

The chicken cell atlas is still in its infancy (Table 2) [Yamagata, 2022; Liu Y et al., 2022]. Nonetheless, a couple of pioneering studies have realized this endeavor, revealing various cell types and their states in the chick limb buds [Feregrino et al., 2019; De Lima et al., 2021], the early primitive streak stage [Vermillion et al., 2018; Guililot et al., 2021], the neural crests [Morrison et al., 2017; Gandhi et al., 2020; Williams et al., 2022], and the neural retina [Hoang et al., 2020; Tegla et al., 2020; Yamagata et al., 2021]. Most studies have used embryonic or juvenile tissues, likely due to accessibility. It is practicable because hatched chicks are highly active and generally considered mature. It takes 21 days, on average, for an egg to hatch.

### Table 2. Chicken cell atlas: scRNA-seq data

<table>
<thead>
<tr>
<th>Focused tissue</th>
<th>Methoda</th>
<th>Reference</th>
<th>doi</th>
<th>GEO accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>SMART-seq, dissociation, FACS, LCM</td>
<td>Morrison et al. [2015]</td>
<td>10.1002/dvdy.24274</td>
<td></td>
</tr>
<tr>
<td>Neural crest</td>
<td>SMART-seq, dissociation, FACS</td>
<td>Morrison et al. [2017]</td>
<td>10.7554/eLife.28415</td>
<td>GSE108230</td>
</tr>
<tr>
<td>Primitive streak</td>
<td>SMART-seq Fluidigm C1</td>
<td>Vermillion et al. [2018]</td>
<td>10.1016/j.ydbio.2018.04.007</td>
<td>GSE89910</td>
</tr>
<tr>
<td>Gonads</td>
<td>10x, dissociation</td>
<td>Estermann et al. [2020]</td>
<td>10.1016/j.celrep.2020.03.055</td>
<td>GSE143337</td>
</tr>
<tr>
<td>Neural crest (hindbrain)</td>
<td>10x, dissociation</td>
<td>Gandhi et al. [2020]</td>
<td>10.7554/eLife.57779</td>
<td>PRJNA624258</td>
</tr>
<tr>
<td>Skeletal muscle and fat (breast)</td>
<td>10x, dissociation</td>
<td>Li et al. [2020]</td>
<td>10.1186/s12864-020-07136-2</td>
<td>CRA002353</td>
</tr>
<tr>
<td>Developing retina</td>
<td>10x, dissociation</td>
<td>Tegla et al. [2020]</td>
<td>10.7554/eLife.54279</td>
<td>GSE142244</td>
</tr>
<tr>
<td>Limb muscle</td>
<td>10x, dissociation</td>
<td>De Lima et al. [2021]</td>
<td>10.1038/s41467-021-24157-x</td>
<td>GSE166981</td>
</tr>
<tr>
<td>Primitive streak</td>
<td>inDrops, dissociation</td>
<td>Guillot et al. [2021]</td>
<td>10.7554/eLife.64819</td>
<td>GSE161905</td>
</tr>
<tr>
<td>Developing heart</td>
<td>10x, dissociation</td>
<td>Mantri et al. [2021]</td>
<td>10.1038/s41467-021-21892-z</td>
<td>GSE149457</td>
</tr>
<tr>
<td>Retina</td>
<td>10x, dissociation</td>
<td>Yamagata et al. [2021]</td>
<td>10.7554/eLife.63907</td>
<td>GSE159107</td>
</tr>
<tr>
<td>Pituitary</td>
<td>10x, dissociation</td>
<td>Zhang et al. [2021]</td>
<td>10.3389/fphys.2021.562817</td>
<td>CRA003604</td>
</tr>
<tr>
<td>Developing hypothalamus</td>
<td>10x, methanol</td>
<td>Kim et al. [2022]</td>
<td>10.1016/j.celrep.2021.110251</td>
<td>GSE171649</td>
</tr>
<tr>
<td>Lymphocyte</td>
<td>10x, suspension</td>
<td>Qu et al. [2022]</td>
<td>10.3389/fmicb.2022.800618</td>
<td>PRJNA687808</td>
</tr>
<tr>
<td>Germ cell</td>
<td>10x, dissociation, FACS</td>
<td>Rengaraj et al. [2022]</td>
<td>10.1016/j.csbj.2022.03.040</td>
<td>PRJNA761874</td>
</tr>
<tr>
<td>Neural crest</td>
<td>10x, dissociation</td>
<td>Williams et al. [2022]</td>
<td>10.7554/eLife.74464</td>
<td>GSE131688, GSE181577</td>
</tr>
</tbody>
</table>

---

*a 10x, 10x chromium system from 10x Genomics (Pleasanton, CA, USA). Dissociation indicates enzymatic dissociation.*
Once incubation begins. However, embryos are often not consistently developed, thus staged according to the Hamburger and Hamilton series [Hamburger and Hamilton, 1951]. Interestingly, in single-cell analysis, one embryo consists of a series of cell types and their states covering different developmental stages. Therefore, making multiple developmental atlases at close time points is not essential. In contrast, auxiliary atlases must be generated reflecting different factors such as sex [Clinton et al., 2001] and variable genetic background [Núñez-León et al., 2019].

The raw data from and references to scRNA-seq studies are searchable at NCBI’s GEO database (https://www.ncbi.nlm.nih.gov/geo/) and several single-cell reference sites (e.g., https://panglaodb.se/papers.html, https://www.nxn.se/single-cell-studies/). To explore the original datasets, some interactive viewers for single-cell data are available (Single Cell Portal, https://singlecell.broadinstitute.org; UCSC Cell Browser, https://cells.ucsc.edu; EM- BL-EBI, https://www.ebi.ac.uk/gxa/sc/home). Thus, the first endeavor toward the chick cell atlas project will be to create an armamentarium, incorporate multimodal data, display those datasets and assist users. Tabula Muris, Tabula Sapiens/Human Cell Atlas (HCA)/HuBMAP, Tabula Drosophilae, and C. elegans atlas are examples of other species [Tabula Muris Consortium, 2018; HuBMAP Consortium, 2019; Hanifia et al., 2021; Taylor et al., 2021; Lindeboom et al., 2021; Eraslan et al., 2022; Li H et al., 2022; Tabula Sapiens Consortium, 2022]. Species-specific community websites like GEISHA (http://geisha.arizona.edu/geisha) and Chickspress (https://geneatlas.arizona.edu/) may provide vital starting points as a resource for chicken as in the cases for other model species. However, in the coming years, emerging crypto-technologies such as peer-to-peer data storage and smart contract protocols could transform data sharing methods (see below).

Next Steps: Multimodal, Spatial, and Temporal Atlases

In addition to scRNA-seq and relevant single-nucleus RNA sequencing, other multimodal single-cell technologies, which simultaneously profile multiple data types in the same cell, represent a frontier for discovering new cell types and characterizing cell states [Stuart and Satija, 2019]. The additional modalities include epigenome, proteome, glycome, metabolome, electrophysiology, morphology, and connectome (Guo S et al., 2021; Lee et al., 2021; Saunders et al., 2021; Sun YC et al., 2021; Mund et al., 2022).

Among those modalities, a series of single-cell sequencing methods for detecting heritable DNA methylation and altered chromatin configurations allow the description of epigenetic changes on a genome-wide scale. In particular, a single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) is the most commonly used method for studying epigenetic landscapes in single cells [Stuart and Satija, 2019; Armand et al., 2021]. Although scRNA-seq and scATAC-seq are different, both represent the activity of genes. Thus, it is vital to analyze proteomics and metabolomics to understand actual cell function.

The data gained from scRNA-seq and other dissociated protocols lead to the loss of spatial information. By contrast, spatial biology and spatial transcriptomics include histological, cellular, and subcellular information of transcripts in spatial context and coordinates [Rao et al., 2021; Zhuang, 2021; Palla et al., 2022; Moffitt et al., 2022; Chen A et al., 2022; Moses and Pachter, 2022]. The type and function of cells are further designated by cell morphology and cell interactions, including neuronal connectivity. eCHIKIN (electroporation- and CRISPR-mediated Homology-Instructed Knock-IN) is a technique for CRISPR-mediated genome editing in somatic cells to insert GFP or Cre cDNA into genes identified as cell-type specific in scRNA-seq data [Yamagata and Sanes, 2021]. This technique will reveal cell morphology and connectivity and potentially describe the molecular and spatial networks that organize the proteome [Cho et al., 2022]. These post-transcriptional modalities, together with spatial and temporal information, facilitate the resolution of cell types and states and provide more critical information on cell function.

Open Science and DeSci

In order to promote any scientific research, it is a prerequisite to have a supportive community, raise funds, and build facilities. Furthermore, all scientific data should be made openly accessible and maintained permanently. Nonetheless, most of the current “big science” projects have suffered from several drawbacks. For example, not all contributions and data submitted by individuals or institutions are satisfactorily credited. Instead, the management is often exceedingly centralized: only a handful of influential scientists are highly recognized as leaders of successful projects.

Decentralized science (DeSci) is an emerging movement that proposes to build a shared infrastructure for disseminating, assessing, funding, crediting, and storing data and knowledge using blockchain technologies.
more recently these groups coordinate their efforts to en-
2020; Tweedie et al., 2021; Bradford et al., 2022], and
[Burt et al., 2009; Kusumi et al., 2011; Fortriede et al.,
aliases). Gene nomenclature efforts were developed for
port literature indexing (referred to as synonyms or
is unique, the latter being the nomenclature most often
anced scientific communications. In addition, we anno-
sequences, the HLA region, or human blood groups). Close
omomic features or physiology not common to birds
(e.g., ovalbumin) or when human nomenclature refers to
ies about gene evolution in these animals. With support
22,315 genes automatically assigned nomenclature and
250
for vertebrate species. This consistency expedites comparative
studies amongst vertebrate species and enables discover-
ies between species.

The Chicken Gene Nomenclature Consortium (CGNC) was convened in 2009 to develop and promote
standardized gene nomenclature for chicken genes [Burt et al., 2009]. The CGNC provides updates following each
major annotation update of the chicken genome and has an ongoing manual annotation effort. Nomenclature
provided by CGNC is displayed and continually updated in NCBI’s Gene Database. CGNC cross-references both
NCBI Gene IDs (formerly Entrez Gene IDs) and Ensembl Gene IDs. Nomenclature may be assigned automatically
via multiple orthology and homology searches [Eyre et al., 2007] or by manually reviewing gene location, syn-
teny, and published papers. While we make every effort to assign standardized gene symbols and names that are
based upon human nomenclature when there is a clear orthology between genes, there are exceptions to this rule.
For example, exceptions are made in instances where the chicken gene names are well established in the literature
(e.g., ovalbumin) or when human nomenclature refers to
ies such as chick to other species in the same taxonomi-
class (i.e., birds), further promoting comparative studies and revealing genetic differences between species.

The last major update of CGNC was completed in May
2022, and this update coincides with updated gene annota-
tions for the GRCg7 assembly. This genome assembly
represents a distinct change from other chicken genome assemblies, as it is based upon modern chicken lines rath-
er than Red Jungle Fowl. Since both broiler and leghorn lines were sequenced, CGNC now provides information
about genes from both the maternal broiler assembly and the paternal white leghorn layer. There are currently
22,315 genes automatically assigned nomenclature and
3,602 manually approved genes. While initial manual cu-
ration efforts focused on assigning nomenclature for 1:1
orthologs between chicken and human, more recently we
are focused on assigning standardized names for genes

Conflict of Interest Statement
There are no conflicts of interest.

Standardized Chicken Gene Nomenclature to
Support Functional and Comparative Studies
(Prepared by F.M. McCarthy, S. Davey, M.C. Young,
K.C. Potter, A. Lanke, P. Barela, M.C. Casono, M. Chi-
dori, L. Cigan, N. Das, A. Goodell, S.M. Johnson, J.H.
Keroack, and D. Webb)

Standardized gene nomenclature facilitates unambigu-
ous communication about genes and allows accurate in-
dxing of associated scientific literature. Genes are as-
gigned a full-length name that succinctly describes what
is known about its function and a short gene symbol that
is unique, the latter being the nomenclature most often
used in scientific communications. In addition, we anno-
tate additional names used in published literature to sup-
port literature indexing (referred to as synonyms or
asiases). Gene nomenclature efforts were developed for
representative species of each of the vertebrate lineages
[Burt et al., 2009; Kusumi et al., 2011; Fortriede et al.,
2020; Tweedie et al., 2021; Bradford et al., 2022], and
more recently these groups coordinate their efforts to en-
sure that gene nomenclature is consistent across verte-
brate species. This consistency expedites comparative
studies amongst vertebrate species and enables discover-
ies about gene evolution in these animals. With support
from public resources such as NCBI, standardized no-
menclature can be propagated from a representative spe-
cies such as chick to other species in the same taxonomi-
class (i.e., birds), further promoting comparative studies and revealing genetic differences between species.

The Chicken Gene Nomenclature Consortium (CGNC) was convened in 2009 to develop and promote
standardized gene nomenclature for chicken genes [Burt et al., 2009]. The CGNC provides updates following each
major annotation update of the chicken genome and has an ongoing manual annotation effort. Nomenclature
provided by CGNC is displayed and continually updated in NCBI’s Gene Database. CGNC cross-references both
NCBI Gene IDs (formerly Entrez Gene IDs) and Ensembl Gene IDs. Nomenclature may be assigned automatically
via multiple orthology and homology searches [Eyre et al., 2007] or by manually reviewing gene location, syn-
teny, and published papers. While we make every effort to assign standardized gene symbols and names that are
based upon human nomenclature when there is a clear orthology between genes, there are exceptions to this rule.
For example, exceptions are made in instances where the chicken gene names are well established in the literature
(e.g., ovalbumin) or when human nomenclature refers to
genomic features or physiology not common to birds
e.g., when gene names reference the X or Y chromo-
somes, the HLA region, or human blood groups). Close
collaboration between vertebrate gene nomenclature
groups allows curators to come to a consensus and ensure
that gene nomenclature for orthologs conserved across a
diverse range of vertebrate species can be practically ap-
pied to all vertebrate lineages.

The last major update of CGNC was completed in May
2022, and this update coincides with updated gene annota-
tions for the GRCg7 assembly. This genome assembly
represents a distinct change from other chicken genome assemblies, as it is based upon modern chicken lines rath-
er than Red Jungle Fowl. Since both broiler and leghorn lines were sequenced, CGNC now provides information
about genes from both the maternal broiler assembly and the paternal white leghorn layer. There are currently
22,315 genes automatically assigned nomenclature and
3,602 manually approved genes. While initial manual cu-
ration efforts focused on assigning nomenclature for 1:1
orthologs between chicken and human, more recently we
are focused on assigning standardized names for genes

Conflict of Interest Statement
There are no conflicts of interest.

Standardized Chicken Gene Nomenclature to
Support Functional and Comparative Studies
(Prepared by F.M. McCarthy, S. Davey, M.C. Young,
K.C. Potter, A. Lanke, P. Barela, M.C. Casono, M. Chi-
dori, L. Cigan, N. Das, A. Goodell, S.M. Johnson, J.H.
Keroack, and D. Webb)

Standardized gene nomenclature facilitates unambigu-
ous communication about genes and allows accurate in-
dxing of associated scientific literature. Genes are as-
gigned a full-length name that succinctly describes what
is known about its function and a short gene symbol that
is unique, the latter being the nomenclature most often
used in scientific communications. In addition, we anno-
tate additional names used in published literature to sup-
port literature indexing (referred to as synonyms or
asiases). Gene nomenclature efforts were developed for
representative species of each of the vertebrate lineages
[Burt et al., 2009; Kusumi et al., 2011; Fortriede et al.,
2020; Tweedie et al., 2021; Bradford et al., 2022], and
more recently these groups coordinate their efforts to en-
that are not found in mammals. This latter gene set includes gene family expansions found in chickens or birds, or genes found in other vertebrate lineages but lost or significantly altered in mammals. Several of these projects have only been made possible with assistance from community experts, and some examples of these are described below.

**MHC-Y Region Genes**

The chicken major histocompatibility complex (MHC) contains two independent regions which segregate independently, MHC-B and MHC-Y [Briles et al., 1993; Miller et al., 1994]. The unique nature of MHC regions means that these genes need to be manually reviewed, both to identify the genes and to assign nomenclature in a standardized manner. With assistance from Dr. Miller, we reviewed and provided standardized gene nomenclature for genes associated with the MHC-Y locus. This includes a set of 107 genes and encompasses class I and II molecules, c-type lectin molecules, zinc finger proteins, and leukocyte receptor cluster members [Goto et al., 2022].

Class I genes are denoted with the “MHCY” prefix for gene symbols and the full gene name has the format “major histocompatibility complex Y, class I heavy chain”, with individual genes numbered sequentially and pseudogenes indicated with the designation “P” in the symbol and “pseudogene” for the full gene name. Likewise, class II genes are named using the “MHCY2B” gene symbol prefix and standardized gene names starting with “major histocompatibility complex Y, class II beta”. The MHC-Y region includes eight leukocyte receptor cluster genes that are most closely related to the human leukocyte receptor cluster member 9 family (LENG9; HGNC:16306). To clearly indicate this similarity while ensuring that these genes are not mistaken for direct homologs, the chicken leukocyte receptor genes are designated as “leukocyte receptor cluster member 9 like 3, MHCY region” and gene symbols have the prefix “LENG9Like.” Likewise, c-type lectin gene names include the information that these genes are located in the MHC-Y region. We note that 40 of these genes are not annotated on current reference assemblies and may be breed-specific; CGNC has reserved these gene names for future use. This systematic review of MHC-Y genes provides a clear and unambiguous naming system that can be extended to genes found in other chicken lines. We expect to extend this work to provide standard gene nomenclature for genes in the MHC-B regions and the nucleolus organizer region (NOR), although this effort may require assistance from other community experts.

**Chicken Immunoglobulin Receptor Genes**

Chicken microchromosome 31 contains more than 100 chicken Ig-like receptors, which function as activating (CHIRA), inhibitory (CHIRB), or bifunctional receptors (CHIRAB) [Nikolaides et al., 2005; Laun et al., 2006]. This nomenclature represents a challenge because, although it is well established in scientific literature (including in closely related species) [Windau et al. 2013], it is desirable to remove references to a specific species to promote transfer across multiple species. To resolve this dichotomy, we propose to retain the “CHIR” symbol for these genes but adjust the gene name to “cluster homolog of immunoglobulin like receptor” and to add appropriate “chicken Ig-like receptor” synonyms used in publications to support searching and indexing. We are currently working with Ms Brandi Sparling in Dr. Drechsler’s laboratory to ensure that these chicken genes are correctly identified as activating (CHIRA), inhibitory (CHIRB), or bifunctional (CHIRAB) receptors and named systematically.

**Heat Shock Proteins**

Heat shock proteins are an example of functionally related proteins that are expressed under stress conditions [Nakai and Ishikawa, 2001]. Multiple lines of independent research are actively studying stress responses in chickens [Olanrewaju et al., 2010; Zhao et al., 2011; Jastrebski et al. 2017; Sarsour and Persia, 2022], and the GRCg7 assemblies have 16 genes currently named with a temporary “LOC” designator that are assigned as “heat shock transcription factor, X-linked-like.” Using gene names and literature searches we identified a more comprehensive group of 55 chicken genes associated with heat shock responses. These genes were manually reviewed, and corrected gene nomenclature assigned where required. This process included assigning new names for the expanded gene set of heat shock transcription factors which had formerly been associated with mammalian X and Y chromosomes (HSFX1–HSFX4 and HSFY1). This project highlights the importance of manual review for gene nomenclature automatically assigned across taxonomic groups, and we are currently reviewing additional gene sets with names that reference X and Y chromosomes.

Manually assigning gene nomenclature requires that biocurators have a good working knowledge of how public databases store, integrate, and share information, as well as understanding key principles in genetics, molecular and cellular biology, evolutionary biology, and physiology. CGNC offers research experiences for undergraduate and high school students who wish to complete a...
research project. Students are introduced to the concept of standardized gene nomenclature and why it is important. They work collaboratively to review automatically assigned gene nomenclature, including reviewing associated publications and completing their own phylogenetic analyses. Students review each other’s work, and all nomenclature is quality checked before data entry to the CGNC database. This allows students the opportunity to see the results of their research reflected in databases which use CGNC information.

The current focus for manual curation efforts is to assign nomenclature for gene families and chicken genes that are currently designated with the “LOC” prefix in the NCBI gene set. Gene families frequently have contractions and expansions between species and homology searches often cannot identify orthologs within a family. Many of these LOC genes are similar to (or “-like”) genes found in other vertebrates but have no clear ortholog and are either new members of a gene family or represent lineage-specific genes. For example, avian gene families with notable changes compared to mammals include keratins, histones, ribosomal RNAs, cadherins, and olfactory receptors, and many of these gene families are included in the current LOC genes. Moreover, this process is complicated by the removal and merging of LOC genes during the process of assembly and annotation updates. However, the manual review of these genes necessarily includes closer inspection of annotation and assembly and can provide useful information about regions of the genome that can be improved.

Acknowledgments

D.W. and F.M.M.: We wish to acknowledge Dr. Marcia Miller (City of Hope), Ms Brandi Sparling (Western University of Health Sciences), and Dr. Carl Schmidt (University of Delaware) for their invaluable expertise in MHC-Y region, chicken immunoglobulin receptors, and heat shock genes, respectively. We also acknowledge and sincerely thank gene nomenclature biocurators at HGNC, the Vertebrate Gene Nomenclature Committee (VGNC), Xenbase and Zfin for their ongoing feedback and support.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

CGNC was developed and supported by National Institutes of Health (R24 GM079326) and US Department of Agriculture National Institute of Food and Agriculture (competitive grant no. 2011-67015-30332 to F.M.M.). F.M.M is partly supported by USDA NIFA Hatch funding from National Research Support Project-8 (NRARZT-1370570-R50-131) and Multistate Research Project NC1170 (ARZT-1370520-R50-130).

Distinct Hypothalamus and Breast Muscle Transcriptomic Response to Heat Stress under Newcastle Disease Virus Infection


Global warming has had adverse effects on mammals, birds, and plants [Bellard et al., 2012]. Physiological and behavioral disorders resulting from heat stress raise concerns for poultry production, health, and welfare. Decreased feed efficiency, growth rate, and egg production, as well as increased susceptibility to disease, are major challenges associated with heat stress in the poultry industry. Total economic losses due to heat stress amount to approximately 1.69–2.36 billion USD annually for the US livestock production industry, of which 128–165 million USD is lost from the poultry industry [St-Pierre et al., 2003]. Currently, in genetic selection and breeding programs, high-performance broilers and layers are superior in productivity; however, they cannot maintain their production in the presence of heat stress compared to non-selected birds [Kumar et al., 2013]. In addition, heat stress suppresses the chicken’s immune response, which increases the bird’s susceptibility to infectious disease threats [Monson et al., 2018]. Therefore, it is essential to better understand the deleterious effects and underlying mechanisms of heat stress on immunity and growth performance.

The hypothalamus in the host plays a critical role as a central regulator of temperature in chickens. It links the nervous system to the endocrine system and primarily mediates thermoregulation, food intake, and stress response [Chen et al., 2015]. It functions as a central hub to interact and coordinate downstream metabolism [Bohler et al., 2021]. Heat stress activates the hypothalamic-pituitary-adrenal (HPA) axis by initiating an appropriate response in the preoptic area of the hypothalamus [Tzschentke and Basta, 2002], which simulates the secretion of gamma-aminobutyric acid and then inhibits the dorsomedial nucleus (DMN) of the hypothalamus. As a result of this inhibition, circuits to muscle, fat, and the cardiovascular system are deactivated reducing heat production and increasing heat loss [Molina et al., 2019]. Activation of the HPA axis results in the release of both corticotrophin releasing hormone (CRH) and arginine vasopressin (AVP), and secretion of adrenocorticotropic hormone (ACTH) by the pituitary gland, which stimulates the release of glucocorticoids from the adrenal cortex. Circulating glucocorticoids interact with a variety of
cells to regulate both metabolism and immune function [Chen et al., 2015]. Disruption of the metabolic and immune system function in birds gives rise to lower productivity, such as low breast muscle yield, and an increased risk of disease and mortality.

Genetics plays an important role in the host response to heat in chickens [Wolc et al., 2019]. Therefore, genetic selection offers a feasible and sustainable option to improve heat stress resistance and immune response in chickens. Genetic selection of chickens with improved resilience to heat stress requires a better understanding of the genetic contribution and molecular mechanisms of the heat stress response. Previous studies conducted by our group have characterized transcriptome profiles of several chicken tissues under heat stress. Candidate genes differentially expressed in the comparisons of heat-stressed and control groups in chicken lungs, tracheas, Harderian gland, and livers were identified as a first step toward identifying targets for validation to serve as selection markers [Saelao et al., 2018, 2021; Wang Y et al., 2020]. A series of genome-wide association studies (GWAS) have identified several quantitative trait loci (QTL) associated with heat stress response such as body temperature, growth phenotypes such as body weight and breast muscle yield, and immune-related phenotypic traits such as mortality, viral replication, and antibody levels, highlighting the genetic contribution to heat stress tolerance in different chicken populations [Van Goor et al., 2015; Rowland et al., 2018; Saelao et al., 2019; Wolc et al., 2019]. However, underlying mechanisms of genetic resistance and susceptibility to heat stress are still not fully understood. Physiological parameters from two genetically distinct, highly inbred lines (Fayoumi and Leghorn) that were heat stressed under NDV infection, were reported [Wang Y et al., 2018a]. The relatively more heat tolerant Fayoumi birds showed distinct responses during acute and chronic heat stress compared to the relatively heat stress susceptible Leghorn birds [Wang Y et al., 2018a]. Better hemostasis was maintained in the Fayoumi birds under both abiotic and biotic stressors. Acid-base balance in the host was one of the key factors accounting for the genetic differences between these two lines [Wang Y et al., 2018a]. The liver transcriptome profile of these two inbred lines revealed that the relative heat tolerant and disease resistant Fayoumi line activated not only metabolic but also immune regulation with heat stress and viral infection, and the susceptible Leghorn birds were only able to maintain the basic metabolic responses [Wang Y et al., 2020].

Thermoregulation is mediated by both the nervous and endocrine systems with a key regulator of the nervous system being the hypothalamus. A major downstream organ, the breast muscle, deserves intensive study to help better understand different aspects of thermoregulation. Therefore, the current study was designed to survey the transcriptomic profiles of the hypothalamus and the breast muscle of these two inbred lines to characterize the global gene expression response to heat stress in distinct genetic backgrounds. Identifying candidate genes and pathways associated with heat tolerance in a variety of tissues will aid discovery of important molecular markers to target in genetic selection to improve heat resilience in poultry.

**Experimental Populations**

Two genetically distinct, highly inbred lines, Fayoumi (M 15.2) and Leghorn (GHs 6) having inbreeding coefficient of 99.95% [Zhou and Lamont, 1999], from the Iowa State University Poultry Teaching and Research Farm (Ames, IA) were used in the current study. Fifty-five Leghorns and 56 Fayoumi birds were housed in two temperature and humidity-controlled isolators. Birds were provided with ad libitum access to food and water. On day (d) 1 of age, 30 Leghorn and 31 Fayoumi birds were randomly chosen as the treatment groups and housed in one isolator and the rest of the birds were used as the non-treatment groups in another isolator. The two genetic lines were mixed in each isolator. From 14 days of age to the end of the experiment (41 days of age), the heat-treated groups were exposed to continuous heat stress of 38°C for the first 4 h and then decreased to 35°C, while the non-treatment groups were maintained at 29.4°C for the first week and then 25°C throughout the whole experiment. On d21, birds in the heat-treatment groups were inoculated with 10^7 EID₅₀ (one EID₅₀ unit is the amount of virus that will infect 50% of inoculated embryos) Newcastle disease virus (NDV) La Sota strain through both eyes and nares (50 μL per each eye and nostril). The non-treatment birds were given 200 μL phosphate-buffered saline (PBS) via the same routes. The animal experiment was performed according to the guidelines approved by the Institutional Animal Care and Use Committee, University of California, Davis (IACUC #17853).

**Blood Parameter Measurements**

Physiological blood parameters were measured at three stages: acute heat (AH) at 4 h, chronic heat (CH) at 7 days, and chronic heat combined with NDV infection.
(CH&NDV) at 10 days post-initiation of heat treatment. Thirteen parameters including four chemistry/electrolyte parameters (concentrations of sodium [Na⁺], potassium [K⁺], ionized calcium [iCa²⁺], and glucose [Glu]); seven blood gas parameters (blood pH, carbon dioxide partial pressure [PCO₂], oxygen partial pressure [PO₂], total carbon dioxide [TCO₂], bicarbonate [HCO₃⁻], base excess [BE], and oxygen situation [so₂]) were measured immediately by using an i-STAT Portable Blood Analyzer as we described in the previous study [Wang Y et al., 2018a].

**Tissue Sample Collection and RNA Isolation**

A total of 32 chickens (4 birds per line per treatment at d14 [4-h post-heat stress treatment, acute phase (AH)] and d23 [9 days post-treatment, chronic phase (CH)]) were randomly selected from treatment and non-treatment Leghorn and Fayoumi birds. The birds were euthanized, and the hypothalamus and the breast muscle samples were collected for RNA isolation. Total RNA was isolated using Trizol (Invitrogen, Carlsbad, CA) according to the manufacturer’s protocol. DNase I (Ambion, Austin, TX) digestion was carried out after RNA isolation, and the RNA quality and purity were determined by measuring absorbance at 260 nm and A260/A280 ratio using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE). The RNA samples were stored at −80°C until further use.

**RNA-Seq Library Preparation and Data Analysis**

For each sample, 500 ng of total RNA was used to construct a cDNA library by using the NEBNext Ultra II RNA library prep Kit for Illumina® (New England Biolabs, Ipswich, MA). In total, there were 64 RNA-seq libraries. The cDNA libraries were quantified by Qubit (ThermoFisher Scientific, Waltham, MA) and validated by Agilent Bioanalyzer High Sensitivity DNA Assay (Agilent, Santa Clara, CA). The RNA samples were stored at −80°C until further use.

Raw reads were trimmed to remove adapter sequences and low-quality bases were removed using the Trim Galore program (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The RNA-seq analysis was carried out by using the same method as described in the liver transcriptome study [Wang Y et al., 2020]. The statistical model included the effects of line and condition for each treatment in each tissue, along with the interactions between condition and line. DEGs were declared with a false discovery rate (FDR) <0.05. Gene expression profiles were compared between each pair of the 8 groups in each tissue: Leghorn non-treated (LC) and Leghorn treated (LT) with AH and CH&NDV; Fayoumi non-treated (FC) and Fayoumi treated (FT) with AH and CH&NDV. In each tissue, the comparison groups with the AH and CH&NDV treatment were FTFC (Fayoumi treatment vs. Fayoumi non-treatment), LTLC (Leghorn treatment vs. Leghorn non-treatment), FCLC (Fayoumi non-treatment vs. Leghorn non-treatment), and FTLT (Fayoumi treatment vs. Leghorn treatment).

**Gene Ontology**

Statistics related to the overrepresentation of functional categories were performed using DAVID [Huang et al., 2009; Sherman et al., 2022]. A fold enrichment >2 and FDR <20% was considered significant.

**Pathway Analysis**

Pathway analysis using the DEGs of within-line contrasts was performed by using the Ingenuity Pathway Analysis software (IPA; Qiagen, Redwood City, CA, USA) [Kramer et al., 2014]. Significant associations (p < 0.05) and a Z-score cutoff of |z| > 1.64 were used to identify significantly activated or inhibited pathways.

**Gene Co-Expression Network Analysis**

The Weighted Gene Co-expression Network Analysis (WGCNA) package in R was used for gene co-expression network analysis [Langfelder and Horvath, 2008, 2012]. A soft threshold of 13 was set for generating an adjacency matrix based on co-expression, and the minimum module size was arbitrarily set at 30. To evaluate associations of co-expressed gene clusters with line and treatment, the Leghorn and Fayoumi lines were given nominal values of 1 and 2 and non-treatment and treatment the nominal values of 0 and 1. For the continuous traits collected from the previous physiological study [Wang Y et al., 2018a], association of co-expressed gene clusters with the continuous traits pH, PCO₂, HCO₃⁻, BE, PO₂, so₂, Glu, Na⁺, K⁺, and iCa²⁺ were also evaluated. The driver genes were identified by high absolute values of gene significance (GS > 0.5) and module membership (MM > 0.5) with a threshold of p < 0.05 [Horvath and Dong, 2008].

**Summary of RNA-Seq Analysis**

Sixty-four chicken cDNA libraries were prepared from hypothalamus and breast muscle samples and sequenced...
by the Illumina HiSeq 4000, which included four treatment groups: treated and non-treated Leghorn (LT and LC); treated and non-treated Fayoumi (FT and FC), at the acute heat stress (AH) and the chronic heat stress & NDV inoculation (CH&NDV), respectively. Of the 24,357 annotated chicken genes in the chicken Galgal 6.0 database, more than 80% of the annotated genes were identified in both lines. The detailed mapping statistics are listed in Table 3. The two tissues showed distinct transcriptome profiles. Many genes (2,242 at the AH stage and 3,052 at the CH&NDV stage) were specifically identified in the hypothalamus. For the breast muscle, 627 genes at the AH stage and 384 genes at the CH&NDV stage were specifically expressed compared to the hypothalamus, many of which are related to muscle biogenesis (online suppl. Material 6).

**Differential Gene Expression in Different Comparisons**

Gene expression profiles were compared between each pair of the four groups with two treatments in the two tissues, respectively. The comparison groups were the same as previously reported, which were FC versus LC (FCLC), FT versus LT (FTLT), LT versus LC (L TLC) and FT versus FC (FTFC) at AH and CH&NDV for both hypothalamus and breast muscle [Wang Y et al., 2020]. The number of DEGs is shown in Figure 14 for the between-line comparisons and Figure 15 for the within-line comparisons of the two tissues. No genes were identified by the DEG analysis which demonstrated an interaction effect in each tissue. Detailed gene information and fold changes in each comparison are listed in online supplementary Material 7.

For DEGs between genetic lines, upregulated genes with higher expression in Fayoumi birds were Fayoumi-favored genes, while down-regulated genes with higher expression in Leghorn birds were Leghorn-favored genes. In the hypothalamus, more Leghorn-favored genes were identified than Fayoumi-favored genes at the AH stage, however, the pattern was switched by having more Fayoumi-favored DEGs than Leghorn’s at the CH stage (Fig. 14). Meanwhile, more DEGs were identified in the non-treated comparisons than in the treated comparisons. The numbers of DEGs identified between the treated Leghorn and Fayoumi line dramatically decreased at the CH&NDV stage with 320 Fayoumi-favored genes and 224 Leghorn-favored genes. In the breast muscle, there were always more Fayoumi-favored genes than Leghorn-favored genes across all comparisons (Fig. 14).

There were fewer DEGs within genetic lines than between lines, the DEG numbers were significantly decreased especially in the hypothalamus compared to the between-line comparisons (Fig. 15). More DEGs were upregulated at the AH stage for both lines, as compared to the CH&NDV stage in which more DEGs were downregulated. There were no DEGs shared by the four within-line comparisons in the hypothalamus (Fig. 16). Ten DEGs, including 3 heat shock protein family members (heat shock protein family B member 9 [HSPB9],

<table>
<thead>
<tr>
<th>Line</th>
<th>Treatment</th>
<th>Time point</th>
<th>Raw reads</th>
<th>Aligned reads</th>
<th>Alignment rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothalamus</td>
<td>Leghorn</td>
<td>Non-treated</td>
<td>Acute heat</td>
<td>90,131,605</td>
<td>82,596,745</td>
</tr>
<tr>
<td></td>
<td>Leghorn</td>
<td>Treated</td>
<td>Acute heat</td>
<td>99,594,536</td>
<td>92,251,909</td>
</tr>
<tr>
<td></td>
<td>Fayoumi</td>
<td>Non-treated</td>
<td>Acute heat</td>
<td>88,123,256</td>
<td>81,628,366</td>
</tr>
<tr>
<td></td>
<td>Fayoumi</td>
<td>Treated</td>
<td>Acute heat</td>
<td>88,021,678</td>
<td>81,552,876</td>
</tr>
<tr>
<td></td>
<td>Leghorn</td>
<td>Non-treated</td>
<td>CH&amp;NDV</td>
<td>85,586,252</td>
<td>78,499,724</td>
</tr>
<tr>
<td></td>
<td>Leghorn</td>
<td>Treated</td>
<td>CH&amp;NDV</td>
<td>82,230,451</td>
<td>75,662,675</td>
</tr>
<tr>
<td></td>
<td>Fayoumi</td>
<td>Non-treated</td>
<td>CH&amp;NDV</td>
<td>85,981,153</td>
<td>80,602,593</td>
</tr>
<tr>
<td></td>
<td>Fayoumi</td>
<td>Treated</td>
<td>CH&amp;NDV</td>
<td>84,433,550</td>
<td>78,218,756</td>
</tr>
<tr>
<td>Breast muscle</td>
<td>Leghorn</td>
<td>Non-treated</td>
<td>Acute heat</td>
<td>202,976,814</td>
<td>171,248,763</td>
</tr>
<tr>
<td></td>
<td>Leghorn</td>
<td>Treated</td>
<td>Acute heat</td>
<td>197,916,434</td>
<td>168,158,916</td>
</tr>
<tr>
<td></td>
<td>Fayoumi</td>
<td>Non-treated</td>
<td>Acute heat</td>
<td>199,002,995</td>
<td>172,003,441</td>
</tr>
<tr>
<td></td>
<td>Fayoumi</td>
<td>Treated</td>
<td>Acute heat</td>
<td>167,292,872</td>
<td>145,714,038</td>
</tr>
<tr>
<td></td>
<td>Leghorn</td>
<td>Non-treated</td>
<td>CH&amp;NDV</td>
<td>97,856,074</td>
<td>85,109,242</td>
</tr>
<tr>
<td></td>
<td>Leghorn</td>
<td>Treated</td>
<td>CH&amp;NDV</td>
<td>101,741,687</td>
<td>87,181,252</td>
</tr>
<tr>
<td></td>
<td>Fayoumi</td>
<td>Non-treated</td>
<td>CH&amp;NDV</td>
<td>100,416,625</td>
<td>86,497,655</td>
</tr>
<tr>
<td></td>
<td>Fayoumi</td>
<td>Treated</td>
<td>CH&amp;NDV</td>
<td>102,815,215</td>
<td>88,592,510</td>
</tr>
</tbody>
</table>
heat shock 70kDa protein 2 \([HSPA2]\), and heat shock protein 90 beta family member 1 \([HSP90B1]\), were shared by the Leghorn and Fayoumi birds at the AH stage (Table 4). Four DEGs were identified in both lines with the CH&NDV treatment, three of which were hemoglobin genes (Table 4). The list of DEGs specifically identified in each comparison provided more information about the line-specific gene regulatory response to heat stress under NDV infection. Six heat shock related genes were specifically identified in the LTLCAH contrast, 3 immune related genes (T cell leukemia homeobox 3 \([TLX3]\), interferon alpha inducible protein 6 \([IF16]\), and hemoglobin subunit epsilon 1 \([HBE1]\)) were specifically identified in the LTLCH group, 2 heat shock related genes (heat shock protein 30C-like \([HSP30C]\) and heat shock protein 90 alpha family class B member 1 \([HSP90AB1]\)), and 3 immune-related genes (Myxovirus resistance 1, interferon-inducible protein \([MX1]\), radical S-adenosyl methionine domain containing 2 \([RSAD2]\), and interferon induced protein with tetratricopeptide repeats 5 \([IFIT5]\)) were only identified in the FTFCAH contrast, and 7 metabolism-associated genes were only identified in the FTFCH contrast (Table 5).

The DEG numbers were higher in the breast muscle than hypothalamus, particularly during the CH&NDV stage, in which there were more upregulated DEGs than downregulated DEGs for all comparisons (Fig. 15). The heat shock protein family H member 1 \([HSPH1]\) gene was identified in all four groups (Fig. 17). Nine DEGs were identified in both lines at the AH stage and 200 DEGs were shared by the two lines at the CH&NDV stage. Of the line-specific DEGs, 53 DEGs for the LTLCAH, 33 DEGs for the FTFCAH, 316 DEGs for the LTLCH, and 677 DEGs for the FTFCH, the majority of DEGs were metabolism-related genes except a few immune-related genes such as interferon regulatory factor 2 \([IRF2]\), interferon alpha inducible protein 6 \([IF16]\) and interferon regulatory
factor 2 binding protein 2 (IRFBP2). Detailed gene information and fold changes for breast muscle line-specific DEGs are listed in online supplementary Material 8.

**Functional Categories of Differentially Expressed Genes**

Gene ontology (GO) was used to evaluate the function of DEGs from different comparisons. Because characterizing the line-specific responses to heat stress under NDV infection was one of our major objectives, the gene function analysis focused on the within-line comparisons in the two lines and two tissues. All DEGs in the within-line comparisons were performed by functional enrichment analysis through the DAVID program (DAVID Bioinformatics Resources 6.8). The biological process and KEGG pathways were presented as functional clusters. The significant enriched GO terms and KEGG pathways were presented if $p < 0.05$ and FDR $< 20\%$.

**Hypothalamus**

In the Leghorn line, even with fewer DEGs for the within-line comparisons identified in the hypothalamus, 7 GO terms including the protein folding, and one KEGG pathway (protein processing in endoplasmic reticulum) were enriched by the upregulated DEGs in the LTLCAH comparison. Most of them are metabolic-associated except 1 GO term, peptide antigen assembly with MHC class I protein complex, which was immune related (Fig. 18a). More GO terms (20) were significantly enriched by the downregulated DEGs in the Leghorn birds with the AH treatment, which included iron transport and cholesterol and cellular glucose homeostasis functions (Fig. 18b). No GO terms were enriched by the upregulated DEGs in the Leghorn line with the CH&NDV treatment. Three metabolic functional GO terms were enriched by the downregulated DEGs in the LTLCH comparison (Fig. 18c).
In the Fayoumi line, 2 metabolic functions were enriched by the upregulated DEGs with the AH treatment (Fig. 18d), same as the Leghorn line having the protein folding function enriched. Seven GO terms were enriched by the downregulated DEGs in the FTFCAH comparison, which included defense response to virus (Fig. 18e). With the CH&NDV treatment, 11 GO terms and 1 KEGG pathway (ABC transporters) were enriched by the upregulated DEGs, and 2 metabolic functional terms were enriched by the downregulated DEGs (Fig. 18f, g), and all of them were associated with metabolic and catabolic processes.

Breast Muscle
At the AH stage, upregulated DEGs in Leghorns enriched 5 GO terms including skeletal muscle cell differentiation and negative regulation of apoptotic signaling pathway (Fig. 19a). Downregulated Leghorn DEGs enriched 1 GO term, positive regulation of vasodilation, and 1 KEGG pathway, alanine, aspartate, and glutamate metabolism (Fig. 19b). Four GO terms were enriched by the upregulated DEGs in Fayoumi birds, which included both metabolic-related functions, such as responding to cold, and immune-related functions, such as positive regulation of T-cell activation (Fig. 19c). No GO terms were enriched by the downregulated Fayoumi DEGs.

At the CH&NDV stage, many GO terms and pathways were enriched by DEGs identified in the within-line comparisons due to the larger number of DEGs. Seventeen GO terms and 2 KEGG pathways were enriched by the upregulated Leghorn DEG genes (Fig. 19d). The innate immune response is one of the enriched functions. Downregulated Leghorn DEGs enriched 7 GO terms and 8 KEGG pathways, all of which are metabolic-associated functions, except the biosynthesis of antibiotics enriched by the downregulated Leghorn DEGs (Fig. 19e). In the Fayoumi line, both up- and downregulated DEGs are involved in metabolism functions such as skeletal muscle cell differentiation by upregulated DEGs and fatty acid beta-oxidation using acyl-CoA by the downregulated DEGs (Fig. 19f, g). Two immune related KEGG pathways were enriched by Fayoumi DEGs at this stage. Adipocytokine signaling pathway was enriched by upregulated Fayoumi DEGs and biosynthesis of antibiotics was enriched by downregulated Fayoumi DEGs (Fig. 19g).
### Table 4. DEGs identified in multiple contrasts in the hypothalamus

<table>
<thead>
<tr>
<th>DEG</th>
<th>Gene description</th>
<th>LTLC AH</th>
<th>FTFC AH</th>
<th>LTLC CH&amp;NDV</th>
<th>FTFC CH&amp;NDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSPB9</td>
<td>Heat shock protein family B (small) member 9</td>
<td>2.58∗</td>
<td>3.06∗</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSPA2</td>
<td>Heat shock 70kDa protein 2</td>
<td>1.23</td>
<td>1.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRELD2</td>
<td>Cysteine rich with EGF lik domains 2</td>
<td>1.07∗</td>
<td>1.04∗</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSP90B1</td>
<td>Heat shock protein 90 beta family member 1</td>
<td>0.96</td>
<td>0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPX1</td>
<td>Glutathione peroxidase</td>
<td>0.77</td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CALR</td>
<td>Calreticulin</td>
<td>0.64</td>
<td>0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDCL2L</td>
<td>Stromal cell derived factor 2 like 1</td>
<td>0.58</td>
<td>0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MANF</td>
<td>Mesencephalic astrocyte derived neurotrophic factor</td>
<td>0.57</td>
<td>0.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HYOU1</td>
<td>Hypoxia up regulated</td>
<td>0.37</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDIA3</td>
<td>Protein disulfide isomerase family A member 3</td>
<td>0.28</td>
<td>0.38∗</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBBA</td>
<td>Hemoglobin beta, subunit A</td>
<td></td>
<td></td>
<td>−1.18b</td>
<td>−1.54</td>
</tr>
<tr>
<td>HBA1</td>
<td>Hemoglobin subunit alpha 1</td>
<td></td>
<td></td>
<td>−1.26</td>
<td>−1.47</td>
</tr>
<tr>
<td>HBAD</td>
<td>Hemoglobin alpha, subunit D</td>
<td></td>
<td></td>
<td>−1.27</td>
<td>−1.59</td>
</tr>
<tr>
<td>EPB42</td>
<td>Erythrocyte membrane protein band 4.2</td>
<td></td>
<td></td>
<td>−2.73</td>
<td>−1.77</td>
</tr>
</tbody>
</table>

* Log2 fold change; bold values indicate upregulation. b Negative values (−) indicate downregulation.

### Table 5. Contrast specific DEGs identified in hypothalamus

<table>
<thead>
<tr>
<th>DEG</th>
<th>Gene description</th>
<th>LTLC AH</th>
<th>FTFC AH</th>
<th>LTLC CH&amp;NDV</th>
<th>FTFC CH&amp;NDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLC5</td>
<td>Chloride intracellular channel 5</td>
<td>1.03a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSPB8</td>
<td>Heat shock protein family B member 8</td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BAG3</td>
<td>BCL2 associated athanogene 3</td>
<td>0.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSPH1</td>
<td>Heat shock protein family H member 1</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSP5</td>
<td>Heat shock 70kDa protein 5</td>
<td>0.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNAJB6</td>
<td>DnaJ heat shock protein family (Hsp40) member B6</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDX3X</td>
<td>DEAD-box helicase 3, X-linked</td>
<td>0.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNAJC3</td>
<td>DnaJ heat shock protein family member C3</td>
<td>0.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLC38A2</td>
<td>Solute carrier family 38 member 2</td>
<td></td>
<td>−0.30b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NR1D1</td>
<td>Nuclear receptor subfamily 1 group D member 1</td>
<td></td>
<td>−0.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSP30C</td>
<td>Heat shock protein 30C-like</td>
<td></td>
<td>2.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSP90AB1</td>
<td>Heat shock protein 90 alpha family class B member 1</td>
<td></td>
<td>−0.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MX1</td>
<td>Myxovirus resistance 1, interferon-inducible protein</td>
<td></td>
<td>−0.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSAD2</td>
<td>Radical S-adenosyl methionine domain containing 2</td>
<td></td>
<td>−1.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFIT5</td>
<td>Interferon induced protein with tetratricopeptide repeats 5</td>
<td></td>
<td>−1.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TLX3</td>
<td>T cell leukemia homeobox 3</td>
<td></td>
<td>1.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFI6</td>
<td>Interferon alpha inducible protein 6</td>
<td></td>
<td>1.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBE1</td>
<td>Hemoglobin subunit epsilon 1</td>
<td></td>
<td>−2.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLC2A12</td>
<td>Solute carrier family 2 member 12</td>
<td></td>
<td>1.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRC1</td>
<td>macrophage mannose receptor 1-like</td>
<td></td>
<td>1.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLC13A5</td>
<td>Solute carrier family 13 member 5</td>
<td></td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGM2</td>
<td>Transglutaminase 2</td>
<td></td>
<td>0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRCSB</td>
<td>Cytokine receptor common subunit beta-like</td>
<td></td>
<td>0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS3ST1</td>
<td>Heparan sulfate-glucosamine 3-sulfotransferase 1</td>
<td></td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NKA1N4</td>
<td>Sodium/potassium transporting ATPase interacting 4</td>
<td></td>
<td>−0.22</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Log2 fold change; bold values indicate upregulation. b Negative values (−) indicate downregulation.
Fig. 18. Gene ontology (GO) biological processes and KEGG pathway overrepresentation (p < 0.05 and FDR <20%) for within-line comparisons in the hypothalamus. a GO terms and KEGG pathways significantly enriched by upregulated genes in the LTLCAH comparison. b GO terms and KEGG pathways significantly enriched by downregulated genes in the LTLCAH comparison. c GO terms and KEGG pathways significantly enriched by downregulated genes in the LTLCCH comparison. d GO terms and KEGG pathways significantly enriched by upregulated genes in the FTFCAH comparison. e GO terms and KEGG pathways significantly enriched by downregulated genes in the FTFCAH comparison. f GO terms and KEGG pathways significantly enriched by upregulated genes in the FTFCCH comparison. g GO terms and KEGG pathways significantly enriched by downregulated genes in the FTFCCH comparison.
Fig. 19. Gene ontology (GO) biological processes and KEGG pathway overrepresentation (p < 0.05 and FDR <20%) for within-line comparisons in the breast muscle. a GO terms and KEGG pathways significantly enriched by upregulated genes in the LTLCAH comparison. b GO terms and KEGG pathways significantly enriched by downregulated genes in the LTLCAH comparison. c GO terms and KEGG pathways significantly enriched by upregulated genes in the FTFCAH comparison. d GO terms and KEGG pathways significantly enriched by upregulated genes in the LTLCCH comparison. e GO terms and KEGG pathways significantly enriched by downregulated genes in the LTLCCH comparison. f GO terms and KEGG pathways significantly enriched by upregulated genes in the FTFCCH comparison. g GO terms and KEGG pathways significantly enriched by downregulated genes in the FTFCCH comparison.
Gene Network and Signaling Pathway Analysis

In addition to the GO analysis, the Ingenuity Pathway Analysis was used to identify gene networks and signaling pathways by using significant DEGs. Because of the limited DEG numbers in the hypothalamus, there was only one canonical pathway, unfolded protein response, identified in the Leghorn line with the AH treatment. In the breast muscle, no canonical pathways were identified in the AH stage. Twenty-five canonical pathways were identified in the two genetic lines with the CH&NDV treatment (Fig. 20). Thirteen out of 25 had similar activation and inhibition patterns. Five pathways were only activated in Leghorn birds, and 3 pathways were only identified in Fayoumi birds. Four pathways, phospholipase C signaling, calcium signaling, androgen signaling, and synaptic long-term potential, were activated in Leghorns and inhibited in the Fayoumi line with the CH&NDV treatment.

Weighted Gene Co-Expression Network Analysis

To overcome the limitation of DEG numbers specifically identified in the hypothalamus, a correlation-based WGCNA analysis was conducted. Seven gene modules were identified in total by using all the hypothalamus and breast muscle data (Fig. 21). Four out of 7 genes, black, turquoise, blue, and yellow modules, showed significant correlations with tissue, genetic lines, the CH&NDV treatment, and phenotypic traits. Four trait-correlated modules were further analyzed and the results are described in Table 6. The black module was the only one positively correlated with tissue, which means genes from the black module had higher expression levels in the hypothalamus. The turquoise module was only negatively correlated with tissue. Genes in the turquoise module had higher expression levels in the breast muscle. The blue module was positively correlated with line and Na⁺ and negatively correlated with pH, HCO₃, TCO₂, BE, PO₂, sO₂, and glucose levels in chicken blood. Genes in the blue module are highly expressed in Leghorn birds and correlated with higher sodium and lower blood gas levels. The yellow module was negatively correlated with the CH&NDV treatment and the blood PCO₂ level.

Furthermore, genes driving the biology within these trait-correlated modules were identified to understand the potential biological processes these two genetic lines underwent during treatments. Driver genes were not discovered in all correlated gene modules. Driver genes were only identified from tissue-correlated black and turquoise modules, the CH&NDV treatment-correlated yellow module, and the line, pH, PO₂, sO₂, and Glu correlated blue module. The top-5 driver genes with the highest absolute gene significance (GS) and module membership (MM) are listed in Table 7. All the detailed driver gene information is presented in online supplementary Material 9 and 10.

For example, driver genes, glutamate ionotropic receptor NMDA type subunit1 (GRIN1), tropomyosin 1 (TMP1), solute carrier family 17 member 6 (SLC17A6), and calcium voltage-gated channel auxiliary subunit...
gamma 1 (CACNG1) were identified in the turquoise module which was negatively correlated with tissue and had higher expression levels in the breast muscle. Some immune-related genes such as RAD52 motif containing 1 (RMD1), class I histocompatibility antigen, F10 alpha chain-like (HA1F), and MHC B-G antigen genes were identified as key driver genes in the blue module and highly expressed in the Leghorn line. The apoptosis inducing factor, mitochondria associated 2 (AIFM2) gene was identified as a driver gene for both pH and sO₂ levels. Only 1 driver gene, NADH-ubiquinone oxidoreductase chain 2 (MT-ND2) was identified associated with glucose levels. There were 245 genes in the blue module, in which 145 genes were identified as driver genes correlated with line, 11 driver genes with pH levels, 20 driver genes with PO₂, and 40 driver genes with sO₂.

GO terms from the biological process and KEGG pathways were further analyzed to understand the biological functions of driver genes in significantly correlated gene modules. Genes in the black module, which had higher expression levels in the hypothalamus, enriched 1 GO term: inter-kinetic nuclear migration. Seventy-seven GO terms and 15 KEGG pathways were enriched by the top-3,000 genes in the turquoise module that are highly expressed in the breast muscle. The top-5 enriched GO terms and KEGG pathways are listed in Figure 22. Biological functions enriched by most of these genes were related to muscle contraction and the glycolytic process in the metabolism function, which was not a surprise. Two GO terms, DNA-templated transcription, and initiation and nucleosome assembly were enriched by genes in the blue module and highly expressed in the Leghorn line. No GO terms and KEGG pathways were enriched by driver genes correlated with other traits which may be due to the lower number of genes.

The blue module was the most interesting one because it correlated with the genetic line as well as multiple phenotypic traits. A gene network of all genes in this module was generated (Fig. 23). More metabolic genes served as node genes in the network. Metabolic genes and immune-related genes were relatively clustered in the module, respectively.

The reported study is a part of the Innovation Lab for Genomics to Improve Poultry (GIP: http://gip.ucdavis.edu), which aims to genetically enhance resistance to heat stress and NDV infection in African poultry. Effects of the combination of both biotic (NDV inoculation) and abiotic (heat stress) stressors on the same two inbred lines (Leghorn: relatively susceptible and Fayoumi: relatively resistant used as experimental lines) have been investigated in previous studies [Deist et al., 2017, 2018; Deist and Lamont, 2018; Saelao et al., 2018; Wang Y et al., 2018a, 2020; Zhang J et al., 2018] as part of this program, with the Fayoumi line (originating in Egypt) as a representative of local African type chickens.

As demonstrated by the physiological responses of these two lines with the AH and CH&NDV treatments, the relatively heat tolerant Fayoumi birds were able to maintain electrolyte levels, respiratory alkalosis, and metabolic acidosis. Lower levels of TCO₂, HCO₃, and BE, higher levels of PO₂ and sO₂ in Fayoumi birds demonstrated heat resistance characteristics, while higher levels of iCa²⁺ and glucose and lower levels of sO₂ in Leghorn birds indicated heat stress susceptibility [Wang Y et al., 2018a]. Global transcriptome profile surveys of the host response to heat stress under NDV infection, or NDV infection alone, have been studied not only in organs where viral replication occurs: Harderian gland, lung, and trachea [Deist et al., 2017, 2018; Deist and Lamont, 2018; Saelao et al., 2018, 2021] but also of the liver, as a representative of a highly metabolic organ [Wang Y et al., 2020]. In the liver transcriptome study, Leghorn birds responded to heat stress and NDV infection mostly by regulating metabolism function. However, Fayoumi birds recruited many immune-related genes, which may regulate both metabolism and immune function to respond to heat stress and viral infection. Results from the liver transcriptome analysis provided insights into how Fayoumi birds are heat and viral infection resilient by activating immune functions even during acute heat stress without viral infection. The surveillance of the Fayoumi immune system was much more sensitive and active than in the

Table 6. Significantly correlated traits and gene modules

<table>
<thead>
<tr>
<th>Traits</th>
<th>Module</th>
<th>Turquoise</th>
<th>Blue</th>
<th>Black</th>
<th>Yellow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tissue</td>
<td>−1</td>
<td>0.15</td>
<td>0.60</td>
<td>−0.14</td>
<td>−0.45</td>
</tr>
<tr>
<td>Line</td>
<td>0.01</td>
<td>0.96</td>
<td>−0.03</td>
<td>−0.2</td>
<td>0.28</td>
</tr>
<tr>
<td>CH&amp;NDV</td>
<td>0.15</td>
<td>0.16</td>
<td>−0.17</td>
<td>−0.29</td>
<td>0.23</td>
</tr>
<tr>
<td>pH</td>
<td>−0.01</td>
<td>0.46</td>
<td>0.2</td>
<td>0.15</td>
<td>−0.06</td>
</tr>
<tr>
<td>PCO₂</td>
<td>−0.01</td>
<td>0.31</td>
<td>0.07</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>HCO₃</td>
<td>0.01</td>
<td>0.05</td>
<td>−0.08</td>
<td>0.13</td>
<td>−0.06</td>
</tr>
<tr>
<td>TCO₂</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.13</td>
<td>−0.06</td>
</tr>
<tr>
<td>BE</td>
<td>0.01</td>
<td>0.53</td>
<td>0.03</td>
<td>−0.06</td>
<td>−0.08</td>
</tr>
<tr>
<td>PO₂</td>
<td>0.01</td>
<td>0.25</td>
<td>0.01</td>
<td>0.13</td>
<td>−0.06</td>
</tr>
<tr>
<td>sO₂</td>
<td>0.01</td>
<td>−0.45</td>
<td>−0.17</td>
<td>−0.22</td>
<td>0.15</td>
</tr>
<tr>
<td>Na⁺</td>
<td>0</td>
<td>0.25</td>
<td>−0.06</td>
<td>−0.08</td>
<td>−0.08</td>
</tr>
<tr>
<td>Glu</td>
<td>0</td>
<td>0.45</td>
<td>−0.17</td>
<td>−0.22</td>
<td>0.23</td>
</tr>
</tbody>
</table>
Leghorns, which could be one of the reasons why Fay-oumi birds are more resilient to both heat stress and viral infection.

When birds are exposed to heat stress, the neuroendocrine system is the first responder, in which the hypothalamus is one of the key regulators for temperature regulation [Chen et al., 2015]. The hypothalamus belongs to the hypothalamic-pituitary-adrenal (HPA) axis, which is one of the two stress axes and can release glucocorticoids to circulate in the peripheral system. In circulation, glucocorticoids interact with a wide variety of cells to regulate both metabolic and immune functions [Nawaz et al., 2021]. Importantly in poultry, glucocorticoids promote proteolysis by damaging myofibrils in skeletal muscles through the ubiquitin-proteasome system which has negative effects on muscle metabolism and meat quality [Bell et al., 2016]. We observed gene expression regulation differences in the liver of heat resistant or susceptible lines. A more comprehensive understanding of the heat stress response from the upstream sensing by the hypothalamus to the downstream effectors such as the skeletal muscle can be obtained by profiling the transcriptomes of both hypothalamus and breast muscle from the same birds. The current study is the first to investigate transcriptome response in hypothalamus and breast muscle, two important organs related to heat stress, under both acute heat stress and chronic heat stress combined with NDV infection in the two inbred lines, which can further elucidate the specific molecular mechanism for heat stress resilience in poultry.

**Relatively Mild Response in the Hypothalamus**

In general, the hypothalamus had a mild response under both AH and CH&NDV treatments in the two genetic lines with very limited DEGs. Fewer DEGs (41 in LTLCAH, 35 in FTFCAH, 10 in LTLCCH, and 45 in FTFCCH) was what we expected based on other chicken hypothalamus transcriptome studies [Sun et al., 2015; Tu et al., 2016]. The two genetic lines showed minor differences in the hypothalamic gene regulation with the treatments, which suggests gene regulation in the hypothalamus during heat stress may undergo only fine-tuning or that the action of the hypothalamus is in a non-genetic role, with effects seen in other target tissues.

As a thermo-regulator, the hypothalamus regulates metabolic and immunological functions by secreting glucocorticoids through the HPA axis. Target genes would participate in the amino acid, glycerol, lipo-biogenesis, muscle biogenesis, and glucose metabolism in the metabolic category [Goel et al., 2021]. Meanwhile, they would also involve in inflammatory expression and homeostasis of T lymphocytes [Honda et al., 2015]. Even with small numbers of DEGs, many heat shock protein genes were identified under both treatments, and a few immune-re-

### Table 7. Top 5 driver genes in each significant trait-correlated gene module

<table>
<thead>
<tr>
<th>Factor</th>
<th>Correlation</th>
<th>Module color</th>
<th>Gene name (GSa, MMb)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tissue</strong></td>
<td>Positive</td>
<td>Black (hypothalamus)</td>
<td>Novel (0.82, 0.92), ELP1 (0.76, 0.91), HINT2 (0.75, 0.94), BRIX1 (0.72, 0.97), RPP25L (0.71, 0.93)</td>
</tr>
<tr>
<td>Line</td>
<td>Negative</td>
<td>Turquoise (breast muscle)</td>
<td>GRIN1 (−1, 1), TPM (−1, 1), VSTM2A (−1, 1), GABRE (−1, 1), GRIK2 (−1, 1)</td>
</tr>
<tr>
<td>Line</td>
<td>Positive</td>
<td>Blue (Leghorn)</td>
<td>KPFL (0.98, 0.94), Novel (0.98, 0.92), RMD1 (0.97, 0.89), KIFC1 (0.97, 0.92), HA1F (0.97, 0.91)</td>
</tr>
<tr>
<td>CH&amp;NDV Negative</td>
<td>Yellow</td>
<td>Blue</td>
<td>AUH (−0.55, 0.68), SKP2 (−0.53, 0.74)</td>
</tr>
<tr>
<td>pH Negative</td>
<td>Blue</td>
<td>Yellow</td>
<td>TNFRSF8 (−0.61, 0.69), MRERF2 (0.60, −0.71), AIFM2 (0.55, −0.85), MOG (0.51, −0.73), HA1F (−0.53, 0.91)</td>
</tr>
<tr>
<td>PO2 Negative</td>
<td>Blue</td>
<td>AIFM2 (−0.57, 0.78), BLEC3 (−0.59, 0.78), KIFC1 (−0.56, 0.92), PYROXD2 (0.55, −0.71), ENSGALG00000039964 (−0.54, 0.94), ENSGALG00000054582 (−0.54, 0.92)</td>
<td></td>
</tr>
<tr>
<td>sO2 Negative</td>
<td>Blue</td>
<td>BLEC3 (−0.59, 0.78), KIFC1 (−0.56, 0.92), PYROXD2 (0.55, −0.71), ENSGALG00000033498 (0.55, −0.92), Protein Manbal-like (0.54, −0.94)</td>
<td></td>
</tr>
<tr>
<td>Glu Negative</td>
<td>Blue</td>
<td>MT-ND2 (0.50, −0.82)</td>
<td></td>
</tr>
</tbody>
</table>

*Gene significance. Module membership.*

**Fig. 21.** Module-trait relationships from WGCNA. Each module (y axis) is correlated with each phenotype (x axis); the correlation and p values were reported for each comparison. Strong positive correlations are coloured in red, and strong negative correlations are coloured in green. TrtAH, acute heat stress; TrtCH, chronic heat stress and 2 dpi NDV infection. *Correlation coefficients. p values.*
**Fig. 22.** Top GO terms and KEGG pathways enriched by genes highly expressed in the turquoise module in the breast muscle.

**Fig. 23.** The blue module gene network. Red highlighted dots indicate immune-related genes, blue highlighted dots metabolic genes. Gray highlighted dots show all other genes in the blue module.
lated genes were identified in the two lines (Tables 4, 5). This is consistent with previous studies that heat shock family genes are mainly upregulated during heat stress [Hasan Siddiqui et al., 2020; Shehata et al., 2020]. For the genetic line-specific DEGs, immune-related genes were identified in both lines with the AH treatment. MX1, IFIT5, and RSAD2 genes were downregulated in the Fayoumi line and NR1D1 was downregulated in the Leghorn line. With the combination of heat stress and NDV infection, two immune-related genes, TLX3 and IFI6, were upregulated in Leghorn birds (Table 4). These findings contrast with our earlier liver transcriptome results for Fayoumi birds, which had an earlier immune response to heat stress than Leghorns [Wang Y et al., 2020]. That both Fayoumi and Leghorn birds’ hypothalamus regulates immune gene expression to deliver an immunological response is further supported by the GO analysis.

GO terms enriched by the upregulated DEGs in the two lines were still quite similar with acute heat stress, which contributed to protein folding and protein processing functions. Leghorn downregulated DEGs were involved in more aspects of metabolic functions, such as cholesterol, cellular glucose, circadian temperature homeostasis, glycogen biosynthesis process, and iron transportation, than Fayoumi downregulated DEGs. Interestingly, downregulated DEGs from both lines enriched immune-related functions such as the negative regulation of NF-kB signaling and negative regulation of toll-like receptor 4 signaling in the Leghorn line and defense response to the virus in the Fayoumi line. The NR1D1 gene, downregulated in the Leghorn birds, was a contributor to the two functions. NR1D1 is a ligand-sensitive transcription factor that negatively regulates gene expression in metabolic and inflammatory processes [Pivovarova et al., 2016]. Downregulation of NR1D1 may be one of the regulatory mechanisms used by the Leghorn birds to respond to heat stress. Three immune-related genes, MX1, IFIT5, and RSAD2, a part of biological function “defense response to virus,” were identified in Fayoumi birds when they were exposed to acute heat stress. The immune response triggered in the Fayoumi birds is more specific than in Leghorns.

Breast Muscle Responds to Heat Stress under NDV Infection Differently in the Two Genetic Lines

For differential gene regulation, the breast muscle had more dramatic responses to both treatments, especially with chronic heat under NDV infection. Many heat shock protein family genes were upregulated in breast muscle. In breast muscle, upregulation of heat shock protein genes can result in skeletal muscle remodeling to protect the muscle cell from damage [Abdelnour et al., 2019]. Functional GO terms, which were metabolically related and enriched by DEGs at the AH stage, were similar between the two lines, except for one immune-related GO term: positive regulation of T cell activation in the Fayoumi line. Two upregulated DEGs in Fayoumi with the AH treatment, Thy-1 cell surface antigen (THY1) and heat shock protein family D member 1 (HSPD1), contributed the most to this function. Therefore, Fayoumi birds had earlier and stronger immune responses than Leghorn birds, mediated by activating T cells.

More metabolic and immune functions were activated with the CH&NDV treatment in breast muscle of both genetic lines. Thermoregulation from the hypothalamus to downstream organs is modulated by glucocorticoids. Negative regulation of glucocorticoid receptor signaling pathway was enriched by DEGs in the Leghorn line. Two DEGs, cryptochrome circadian regulators 1 and 2 (CRY1 and -2), play important roles in the activation of this pathway and both of them were upregulated with the CH&NDV infection in Leghorns. We speculate that Leghorn birds may increase the gene expression levels of CRY1 and -2, subsequently reduce glucocorticoid receptor signaling [Lamia et al., 2011]. Meanwhile, the innate immune response was enriched by the upregulated Leghorn DEGs with both heat and viral stressors. Five DEGs, catelicidin-B1-like (CATHB1), tyrosine kinase non receptor 2 (TKNR2), joining chain of multimeric IgA and IgM (JCHAIN), tripartite motif containing 25 (TRIM25), and MX1 worked together to activate this innate immune response.

The canonical pathway analysis by IPA provided additional insight about the differential response of these two genetic lines. The phospholipase C signaling was activated in Leghorn birds and inhibited in the Fayoumi line with the CH&NDV treatment (Fig. 20). This pathway belongs to the intracellular and second messenger signaling and is associated with cell signaling, molecular transport, and vitamin and mineral metabolism [Putney and Tomita, 2012]. In Leghorn birds, upregulation of the protein kinase C genes (PKCs) activated many downstream genes such as nuclear factor of activated T cell family (NFAT), histon deacetylase 3 (HDAC), cAMP responsive element binding protein 1 (CREB) and nuclear factor kappa-B (NFkB) complex and then promoted downstream gene expression. With the inhibition of this pathway, Fayoumi birds decreased the expression levels of above genes and then suppressed gene expression (Fig. 24). Calcium is an important secondary messenger...
in the phospholipase C signaling pathway, and calcium signaling plays a critical role in muscle contraction [Zhu Y et al., 2019]. The calcium signaling pathway was also activated in Leghorns and inhibited in Fayoumi birds with the CH&NDV treatment (Fig. 25). Downregulation of Fayoumi DEGs may prevent Ca\(^{2+}\) transportation and then slow down cell growth and development. On the other hand, without the inhibition of Ca\(^{2+}\) transportation, Leghorn birds might be focusing on rapid cell growth, development, and inflammations. This is consistent with the GO term, potassium ion transport, enriched by the upregulated Fayoumi DEGs (Fig. 19f). Collectively, regulating mineral transportation might be one of the key differences between the two genetic lines in responding to chronic heat stress and NDV infection in the breast muscle.

**WGCNA Revealed Gene Modules and Driver Genes Important in the Response to Heat Stress**

WGCNA has been applied to many transcriptome studies, especially for complex traits such as diseases in farm animals [Kogelman et al., 2014; Deist et al., 2017; Monson et al., 2018; Fan et al., 2021; Farhadian et al., 2021]. Genes that share a similar function are clustered in gene modules and genes associated with interesting traits can be identified [Langfelder and Horvath, 2008]. In the current study, tissue, line, treatment, and physiological parameters were used to identify potential important driver genes.

Four gene modules showed significant correlation with traits, with the blue module having the most correlated traits (9 traits). Genes highly expressed in the blue module are Leghorn-favored genes. These genes were also correlated with higher levels of Na\(^+\) and lower levels of pH, HCO\(_3\), TCO\(_2\), BE, PO\(_2\), sO\(_2\), and the glucose level in the blood. Heat resilient Fayoumi birds had higher oxygen-related parameters and susceptible Leghorn birds had higher glucose, iron, and lower sO\(_2\) levels [Wang Y et al., 2018a]. Gene expression patterns in the blue module partially explained these physiological phenotypes with transcriptome data, in which genes highly expressed in Leghorn birds correlated with lower sO\(_2\), and higher Na\(^+\) levels.

Potentially important genes were selected from the top driver genes for each significantly correlated gene module with different traits. TNFR superfamily member 8 (TNFRSF8) gene was the top driver gene negatively correlated with the pH levels. As a member of the TNF-receptor superfamily, the TNFRSF8 gene is expressed by active T and B cells and leads to the activation of NFkB [Lee et al., 1996; Morais-Perdigao et al., 2022]. Subsequently, the lower pH level could be due to the activation of apoptosis by the TNFRSF8 gene [Wang M et al., 2008]. The pH level is an important parameter for heat stress treatments, which affects several other blood gas parameters. Further investigation of TNFRSF8’s effect on heat tolerance is desired. RDM1, identified by our previous liver transcriptome study, was also identified in this study in these two tissues that are highly expressed in Leghorn birds. It is potentially one of the Leghorn signature genes for response to heat stress. The AIFM2 gene was negatively correlated with both pH and sO\(_2\). This gene can be induced by cold stress and contributes to apoptosis in the presence of bacterial and viral DNA with oxidoreductase and NADH dehydrogenase activities [Nguyen et al., 2020]. It correlates with thermoregulation and needs further validation for its multiple effects on both antiviral and metabolic functions. Only one driver gene, the MT-ND2 gene, was correlated with glucose. MT-ND2 is the core subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase which can catalyze electron transfer from NADH through the respiratory chain and is essential for the catalytic activity [Rhooms et al., 2020]. The variant on this gene was reported to be associated with glucose metabolism in skeletal muscle in rats [Houstek et al., 2012]. This gene requires further investigation into the molecular mechanisms involved in glucose metabolism during heat stress in chickens. Most of these driver genes were also the node genes on the gene network generated by the blue module genes (Fig. 23). TNFRSF8 is on the boundary of the immune and metabolic gene clusters. MT-ND2 is distant from the immune gene cluster and close to the solute carrier family 9 member B2 (SLC9B2) gene which contributes to the regulation of intracellular pH and sodium homeostasis [Anderegg et al., 2022]. Gene network analysis here demonstrated the interactions of genes of interest and provided more hypotheses for potential future studies.

To understand the distinct physiological responses during acute heat stress or chronic heat stress combined with NDV infection, transcriptome profiles of two metabolically associated organs, the hypothalamus and the breast muscle, were surveyed in the current study to elucidate the molecular mechanisms of host responses in relatively heat stress-resistant Fayoumi and susceptible Leg-

![Fig. 24. Phospholipase C signaling pathway and gene heat map in within-line comparisons in the breast muscle. **a** Molecule activity prediction of the pathway in Leghorn birds with CH&NDV treatment. **b** Molecule activity prediction of the pathway in Fayoumi birds with AH. **c** DEG heatmap matching the pathway.](http://karger.com/cgr/article-pdf/162/8-9/405/3975130/000529376.pdf)
Fig. 25. Calcium signaling pathway and gene heat map in within-line comparisons in the breast muscle. **a** Molecule activity prediction of the pathway in Leghorn birds with CH&NDV treatment. **b** Molecule activity prediction of the pathway in Fayoumi birds with AH. **c** DEG heat map matching the pathway.
horn chicken lines. Both lines responded to heat stress and NDV infection by stimulating metabolic and immune functions. The heat and NDV-resistant Fayoumi line had earlier, more active, and specific immune regulation in the breast muscle than the Leghorn line with both treatments. Genes highly expressed in Leghorns correlated with heat-susceptible physiological phenotypes. Important driver genes, gene modules, and interactive networks identified in the current study provide valuable information for future validation of molecular mechanisms of resistance and for developing novel breeding programs to improve heat and disease resistance in chickens.

Annotation of Repeats in Avian Genomes

For most bird genomes repeats have been annotated automatically using a repeat database (most often Repeatfinder [Kojima, 2020]) and a library-based annotation tool (usually RepeatMasker). The resulting annotations include both full-length and fragmented copies of interspersed repeats. These consist mainly of TEs, tandemly repeated units of simple repeats ((RY)n and homonucleotideic motifs), low complexity repeats (minisatellite/variable number tandem repeat and microsatellite/short tandem repeat), and telomeric and centromeric macrosatellite DNA (large stretches, from thousand to millions of base pairs, of repeated units ranging in size from ~10 bp to several hundred bp). Some repeated genes encoding small non-coding RNAs such as tRNAs, 7SL RNAs, and ribosomal RNA are also annotated.

The main limitation of such library-based approaches is that their annotations depend heavily on the quality of the reference database used, including completeness and accuracy of consensus sequences. Furthermore, because these methods are primarily based on sequence similarity, they tend to overestimate the diversity of interspersed repeats by flagging very small non-overlapping hits that often display low sequence complexity. Finally, these methods are inappropriate for annotating satellite DNAs in newly sequenced species because those repeated units are absent from reference databases.

Some studies use signature-based methods instead, focusing on traits that are unique to certain TEs or repeats. For example, to detect retrotransposons with large terminal repeats (LTR) at both ends, programs such as LTR Finder, LocaTR, LtrHarvest, or ReroTector can detect specific DNA organization patterns and signatures (motifs) that are specific to retroviruses [Bolisetty et al., 2012; Mason et al., 2016; Ji and DeWoody, 2016]. Tandem repeat finder (TRF) [Benson, 1999], another signature method tool, is dedicated to detecting all types of uncom-
plicated tandem repeats such as simple repeats, microsatellites, minisatellites and satellite DNAs.

The final method for repeat annotation are DNA de novo consensus methods that combine a range of detection tools. The REPET pipeline [Pernal et al., 2012] includes the TEdenovo module which uses both de novo (RECON, GROUPER, and PILER) and signature-based tools such as TRF (it can also be set to include a library-based step). A second REPET module, TEannot, uses the output of the first module to annotate the genome. Such de novo consensus methods have historically been limited by the need for powerful calculation and storage resources, which has restricted their application to small eukaryotic genomes (~10 Mb to 500 Mb). However, advances in computing clusters and a recent REPET update have opened the way for the use of this software package with larger genomes such as those of vertebrates. A second de novo package, the RepeatModeler2 pipeline [Flyn et al., 2020], has an architecture that is similar, but simpler than that of REPET. It employs two discovery algorithms, RepeatScout and RECON, followed by consensus building and classification steps. In addition, RepeatModeler2 includes two signature-based tools LTRHarvest and LTR_retriever. The de novo repeat library produced by RepeatModeler2 is then used to annotate genomes using the RepeatMasker program. The main weakness of these de novo methods is that they are unable to detect repeats with very few copies, such as DNA transposable elements or endogenous viral elements (EVEs).

The method(s) used to estimate the amounts of repeats in a chicken genome model has a strong influence on the final result. Physico-chemical approaches, such as reassociation kinetics, indicate that the chicken genome is composed of ~30% repeats. However, annotations of galGal4 through 6 report repeat content of 8–21%, depending on the bioinformatic pipeline used [for review, see Guizard et al., 2016]. Compared to other vertebrate species, the TE diversity of chickens is lower (33 TE species) [Guizard et al., 2016].

Interspersed Repeats

Interspersed repeats are primarily composed of TEs. These are DNA segments with the potential to move and/or duplicate from one chromosomal location to another (i.e., transposition). The most common way to classify TEs is based on whether or not they use an RNA molecule as a transposition intermediate. Class I elements (a.k.a. retrotransposons) reverse transcribe into a DNA molecule in the process of integration, while class II elements, DNA transposons, do not [Kojima, 2020]. The advantage of classifying elements this way is that it is relatively simple to understand and widely used. However, a significant failing of this classification method is that it gathers many sequences into subclasses and families that are evolutionarily unrelated [for a more in-depth discussion, see Arensburger et al., 2016]. An alternative classification scheme divides TEs into at least 10 classes based on the enzymatic machinery used to transpose between loci [Arensburger et al., 2016]. There are 2 primary advantages to this alternative classification. First, it is easily amenable to the addition of new classes, as new TE transposition mechanisms are discovered. Second, its organization is compatible with the classification scheme adopted by the International Committee for the Taxonomy of Viruses (ICTV) with which it shares a number of retrotransposon, virus [Katzourakis and Gifford, 2010], and DNA transposon families [Koonin and Krupovic, 2017].

The chicken genome repeatome consists mostly of 4 kinds of interspersed repeats, making up 15.5% of the genome. Most abundant are non-LTR elements (a.k.a. long interspersed elements or LINEs) belonging to the CR1 group of retrotransposons (>410,000 annotated fragments, 11.8% in the galGal4 and galGal5 model chromosomes). CR1 elements contain 2 open reading frames (ORFs) coding for the ORF1 protein that binds to CR1 mRNA to assemble a ribonucleic particle, and a reverse transcriptase (RT) protein with an endonuclease domain (Fig. 26a). In bird genomes at least 22 CR1 subfamilies have been described [Liu et al., 2009], 8 of which are currently annotated in chicken genomes. These do not show signs of recent mobility, but full-length intact CR1s are present (fewer than 20) [Galbraith et al., 2021], suggesting that these might be mobilized in vitro under the right experimental conditions. In addition to the CR1s, 7 other LINE group elements have been described [Kojima, 2020], only two of which are found in bird genomes, R2 and RTE. Short interspersed elements (SINEs) are non-autonomous non-LTR retrotransposons derived from nuclear RNA (tRNA, rRNA, 7SL RNA, snRNA, etc.) that parasitize the enzymatic machinery of LINEs (Fig. 26b). There is some controversy regarding the presence of AmnSINE1 elements in the chicken genome [Nishihara et al., 2006]. This is an element that uses the transposition machinery of specific L1 LINE elements for its own transposition, but L1 elements are absent from all Sauropsida except Lepidopsauria. We have failed to find these elements in the chicken genome [Guizard et al., 2016]. Other SINEs, AviRTE [Suh et al., 2016] and TguSINE1 [Suh et al., 2017], that use the transposition machinery of spe-
specific CR1 and RTE elements, respectively, were found in other bird lineages.

The second most abundant interspersed repeats found in the chicken genome are LTR retrotransposons. These TEs have LTRs directly repeated at both ends of the element and contain ORFs encoding a group-specific antigen (Gag), a reverse transcriptase (RT), and in some cases a retroviral envelope protein (Env) (Fig. 26c). From an evolutionary and enzymatic standpoint, their mobility mechanism is strikingly different from that of LINEs and SINEs. At least 21 "species" of LTR retrotransposons were found in galGal4 and galGal5 [Guizard et al., 2016]. These have either 2 LTRs or are annotated as solo LTRs, likely resulting from the loss of the inner part of the LTR retrotransposon by recombination between the LTRs of each inserted element. We found no copies corresponding to complete, internally deleted, or partly truncated elements of 6 models of solo LTRs (Birddawg, putative_LTR_group 4, 9, 12, 22, 28, and 30). Among the 10 elements that could reliably be classified as LTR retrotransposons, 8 belonged to the endogenous retrovirus (ERV) superfamily (EAV, EAV-HP, ERV2, ERV7, ERv11, Kronos, Soprano, and RetroTux) and 2 were related to the Gypsy-Ty3 superfamily (retroCalimero, retroSaturnin). No element was found to belong to the Copia-Ty1 superfamily. A peculiarity of the LTR retrotransposons is that they are dramatically enriched on the chicken W chromosome.

DNA transposons, the third kind of interspersed repeat, cover about 1.8% of chromosome sequences in the chicken genome. These TEs typically contain a single ORF encoding a transposase enzyme that is generally sufficient to catalyze all the steps of its mobility. This ORF is flanked at both ends by terminal inverted repeats that are

---

**Fig. 26.** Sequence organization of transposable elements in avian genomes. 

**a** CR1 elements are composed of segments that resemble long retro-inserted messenger RNA (mRNA) with an A-rich tail at their 3′ end. Within a “species” of CR1s many copies are truncated at their 5′ ends. Full-length elements contain 2 open reading frames (ORFs) with ORF2 encoding a protein containing an apurinic endonuclease domain fused to a reverse transcriptase. 

**b** SINEs using the CR1 machinery are present in some bird species, but not in chickens. They consist of the fusion of a former tRNA to a 3′ CR1 end. 

**c** LTR retrotransposons have all the signatures of an endogenous retrovirus-like element, including long terminal repeats (LTR) at both ends and ORFs coding for a group antigen (Gag), a reverse transcriptase (RT), and in some case an envelope protein (Env). 

**d** DNA transposons that transpose directly from DNA to DNA have short terminal inverted repeats (arrows) at both ends. When these elements are intact, they may contain a gene encoding a transposase, an enzyme required for their own transposition. There are also internally deleted forms such as Galluhop and chimeric elements such as Charlie/Galluhop.

---

**Table 1.** List of transposable elements found in the chicken genome.
used as binding sites by the transposase to excise and reinsert the transposon. The diversity of DNA transposons in the chicken genome is extremely low, with only one member of the following superfamilies: IS630/Tc1/mariner, Mariner1_GG, hobo/Ac/Tam, and Charlie. Two other DNA transposons are amplified in the chicken genome, but one is an internally deleted derivative of Mariner1_GG (Galluphop) and one a chimeric element consisting of a Charlie copy with a Galluphop copy inserted (Fig. 26d). This DNA transposon profile is shared by most bird genome model species, even by the woodpecker genomes which display elevated TE abundances (17–30% of the genome) [Manthey et al., 2018]. Sequences related to DNA transposon superfamilies GingerI, Ginger2, hAT, IS630/Tc1/mariner, P, piggyBac, Polinton, Transib, Crypton, and Zisupton are also present as single or as a few repeated copies in various bird genomes. Most of these sequences are domesticated genes, derived from ORFs encoding transposases [for an inventory, see additional file 12 in Guizard et al., 2016].

Finally, there are also 7 interspersed repeat sequence types called Hitchcock, and undetermined_group_1 through 6 that do not appear to correspond to any existing classifications [Arensburger et al., 2016; Kojima, 2020]. Their annotation covers about 0.8% of chicken chromosomes. So far, their main outstanding feature is that they are enriched in the microchromosomes [Guizard et al., 2016].

**Tandem Repeats**

Noncoding tandem repeats are mainly composed of DNA minisatellites, microsatellites, and macrosatellites and account for approximately 4% of the chicken genome. In the case of DNA mini- and microsatellites, the percent coverage of the genome is unlikely to be severely biased by issues related to genome completion. These sequences were the subject of few studies in chickens because single-nucleotide polymorphisms (SNPs) were long used as genetic markers. Telomeric repeats, which contain large numbers of macro- and microsatellites, are actively maintained by the telomerase enzyme. These telomere ends display very different numbers of repeats, depending on which chromosome was examined, animal age, and whether tissues or cell lines were examined [Taylor and Delany, 2000; Rodrigue et al., 2005; O’Hare and Delany, 2009].

Unlike the mini- and microsatellites above, macrosatellites, which are located in centromeres, telomeres, and in non-coding regions of sex chromosomes, are likely underestimated. The most studied macrosatellites are those located in the Z chromosome (Z_rep in Repbase and Guizard et al. [2016]) and in the W chromosome [for review, see Komissarov et al., 2018], and between the MHC and rRNA genes repeated in tandem on chromosome 16 [Miller et al., 2014]. Those present in the inner regions juxtaposed to telomeres of autosomes remain unknown. In chickens, and other bird species, these regions are highly GC-rich (Fig. 27) [Federico et al., 2005], G-quadruplex (G4) rich, and contain genes with flanking regions and introns mainly composed of satellite DNA sequences. There are differences in the estimates of repeat content between those based on genome assemblies and those based on cytological and physico-chemical approaches. Resolving these discrepancies may require elucidating the sequence composition of centromeric and telomeric regions.

In addition to the sequences above, tandemly repeated genes represent between 1.5 and 2% of the chicken genome. They encode large ribosomal RNAs (18S, 5.2S, 28S) [Piégu et al., 2020], and account for a number of gene duplications in some bird genomes [Warren et al., 2010]. Unfortunately, when genome models are annotated, duplicated genes with sequences that are more than 95–97% identical are not reported as separate genes. In some cases, such as the PHF7 genes, annotations between genome models vary from 0 to 68 gene copies (likely paralogs plus segmental duplication). In the galGal6 release 105 model, these genes are distributed among 4 loci located on 2 different chromosomes [Fouchécourt et al., 2022]; while in the white leghorn breed, 39 PHF7-like gene copies are annotated at 10 different loci and distributed among 5 chromosomes. Therefore, much work remains to be done to fix the number of repeated genes present in wild fowls and domesticated lines.

**Balance between Genome Size and Repeats**

To our knowledge only 2 publications have examined the issue genome and repeats in birds. The first one [Piégu et al., 2020] showed that the genome of numerous domesticated chicken lines was smaller than that of the red jungle fowl. Both bioinformatics and molecular investigations showed that the genome coverage of various tandem repeat types found in the red jungle fowl genome (rDNA, telomeric repeats, macrosatellite DNA, and segmental duplications) were lower in domesticated lines, but that novel segmental duplications were present in these lines. This supports the hypothesis that domestic lines have been significantly reshaped during domestication and subsequently by human-mediated selection.
The second publication [Kapusta et al., 2017], focused on genome size variations on the scale of bird evolution. Prior to this publication, it was thought that bird genomes were small due to selection pressure related to metabolic constraints linked to flight, and to a dearth of active TEs able to expand their numbers. This view was supported by research showing that piwi RNA (piRNA) did not control TE activity in birds and which therefore would not be committed to an arm race with TEs [Lee et al., 2009]. However, Kapusta et al. [2017] showed that TE expansions in birds were counteracted by DNA losses, mainly through large segmental deletions (>10 kb). This new view, where TEs have remained active in bird genomes, has been strengthened by reports that (1) the chicken genome contains full-length and intact CR1 elements that are putatively active in transposition [Nishihara et al., 2006], (2) it contains recently active LTR retrotransposons [Wang Z et al., 2013], and (3) that W retrotransposons may act as refugia for active ERVs [Peona et al., 2021]. This new view was further supported by reports that the chicken genome uses the piRNA system to control CR1 elements and at least some ERVs [Sun et al., 2017; Chang et al., 2018]. Together, these findings support the “accordion” model of genome size evolution [Kapusta et al., 2017] and have changed our understanding of genome plasticity in birds, from a static [Wicker et al., 2005; Ellegren, 2010; Gao et al., 2017] to a dynamic view where all forms of recombination are involved. However, it is important to note that the analyses above all used genome assemblies made from somatic DNA. We currently do not know the organization of the germinal genome and therefore are unable to determine if variations in genome organization observed in somatic cells are either due to differences in recombination events occurring in germ cells over generations or to differences in programmed DNA rearrangements or aging during development at each generation, or both.

**Future Challenges and Technical Perspectives about Repeats in Avian Genomes**

Thanks to the latest versions of Pacific Biosciences and Oxford Nanopore MinION sequencing technologies, the Telomere-to-Telomere (T2T) consortium was able to resolve telomeric and centromeric regions of the human genome, adding another 8% of new sequences (∼200 million bp) including 1,956 new genes, 99 of which are predicted to be protein-coding [Nurk et al., 2022]. Such a T2T project for avian genomes would be very helpful for resolving highly GC-rich genomes such as those of ratites, but also the GC-rich telomeric regions of macro- and microchromosomes of numerous bird species. However, one should not be too optimistic about the ability of these new reads to resolve the telomeric and subtelomeric regions of birds. Indeed, studies of Pacific Biosciences sequences of mRNA and genomic DNA have demonstrated that these sequences are high in G4 motifs that may lead to formation of G-quadruplexes [Beauclair et al., 2019]. These structures interfere with sequence library fabrication for both Pacific Biosciences and Oxford Nanopore MinION technologies, as well as with the sequencing of these libraries. HiFi reads from Pacific Biosciences do not solve this issue because G4 motifs block...
DNA polymerization and no sequences are produced [Zhu et al., 2016].

In theory there are technical solutions to circumvent these problems. Indeed, the formation of G-quadruplexes requires the presence of Na+ or K+ cations [Guiblet et al., 2021]. While these cations are normally used during library preparation by reverse transcriptases and by DNA polymerase enzymes, it is possible to substitute the Na+ or K+ cations with Li+ or Cs+ cations using alkali chloride salts in the buffers [Ramos-Alemán et al., 2018]. However, as Pacific Biosciences and Oxford Nanopore do not make the content of their solutions accessible to the public, users cannot modify them to replace the salts in the buffers [Flores-Juárez et al., 2016].

The elucidation of these enigmatic GC-rich telomeric, subtelomeric, and centromeric regions of bird genomes is an exciting challenge. Indeed, these regions contain genes where all noncoding segments are filled with GC-DNA satellites, they are G4 motifs repeated in tandem [Beauclair et al., 2019]. In mammalian genomes G-quadruplexes are detected in vivo [Zheng et al., 2020] and have been shown to be important as transcription factor binding hubs [Bochman et al., 2012; Hall et al., 2017; Spiegel et al., 2021] and as epigenetic modulators of chromatin [Guilbaud et al., 2017]. Birds have significantly expanded these motifs in telomeric, subtelomeric, and centromeric regions likely to control the expression. From an evolutionary standpoint, it would be important to determine if this characteristic is specific to birds or evolved early in the Sauropsida and expanded in bird genomes.

Conflict of Interest Statement
The authors report no competing interests.

Ribosomal DNA Repeats in Chicken and Guinea Fowl Genomes
(Prepared by A. Dyomin, S. Galkina, A. Davidian, and E. Gaginskaya)

The ribosomal DNA (rDNA) is one of the key elements in the cellular genome. It consists of multiple tandemly arranged repeats that bear coding sequences for ribosomal RNAs (rRNAs). rRNA molecules play an essential role in the ribosome functioning. They define the size and shape of ribosomal subunits forming a structural scaffold for the specific placement of proteins inside the ribosome. rRNAs are involved in all events of translation, including mRNA initial binding [Martin et al., 2016], codon recognition by aminoacyl-tRNAs [Ogle et al., 2001; Demeshkina et al., 2012], peptide bond formation [Nissen et al., 2000], and tRNA/mRNA translocation [Mohan and Noller, 2017; Noller et al., 2017; Djumagulov et al., 2021]. These processes are based on conformational rearrangements of the rRNA molecules which constitute 2 ribosomal subunits. In the cytoplasmic ribosome of a eukaryotic cell, 28S, 5.8S, and 5S rRNA make up the large subunit, while 18S rRNA is the core molecule of the small one. Due to the role of rRNA in the mechanism of protein synthesis in the cell, rDNA is sometimes referred to as a separate subgenome (rDNAome) [Symonová [2019].

Most eukaryotes feature 2 main types of rDNA loci. The first type includes the 5S rRNA genes. The other type, nucleolus organiser regions (NOR), is specific for encoding the 18S, 5.8S, and 28S rRNAs. rDNA repeating units are organised the same way in both types of loci: a transcriptional unit (gene) is followed by a spacer sequence. The 5S rRNA gene (~120 bp) is followed by the so-called non-transcribed spacer (NTS), which can be of varying length. 5S rDNA is transcribed by RNA-pol III, regulator sites being located within the coding sequence [Pieler et al., 1987; Paule and White, 2000; Hall, 2005]. In the eukaryotic NOR, rDNA repeats are organised in a more complex way [Singer and Berg, 1991; Shaw and Brown, 2012; Hori et al., 2021]. Each of them consists of a cluster of 18S, 5.8S, and 28S rRNA genes followed by an intergenic spacer (IGS) sequence. The 18S, 5.8S, and 28S rRNAs coding sequences are separated by 2 internal transcribed spacers (ITS1 and ITS2) and flanked by 2 external transcribed spacers (5’ETS and 3’ETS). The rDNA cluster is transcribed by RNA-pol I into the primary 45/47S rRNA molecule (pre-rRNA) as a single transcriptional unit. The coding sequences for each rRNA are highly conserved in length and nucleotide composition across taxa. However, rRNA inter- and intra-individual genomic polymorphisms have been described in several species [Pieler et al., 2012; Locati et al., 2017; Kim JH et al., 2018, 2021; Parks et al., 2018; Ding et al., 2022]. The spacer regions containing splicing cleavage sites evolve more rapidly. Their high variability can be detected even within the same genome [Kim JH et al., 2018, 2021].

The IGS, which plays important regulatory roles in the cell function, is one of the most interesting regions in the rDNA repeat. In many animals, IGS have been found to contain such functional elements as pre-rRNA promoters [Haltiner et al., 1986; Caudy and Pikaard, 2002; Massin et al., 2005; Agrawal and Ganley, 2018], several transcription termination sites (Sal box) [Pfeiderer et al., 1990; Agrawal and Ganley, 2018], noncoding RNA binding sites involved in cell stress response and regulation of
rDNA transcription [Audas et al., 2012; Agrawal and Ganley, 2018], cdc27 pseudogene [Grandori et al., 2005; Agrawal and Ganley, 2018], and putative c-Myc and p53 binding sites [Gonzalez et al., 1993; Zentner et al. 2011; Agrawal and Ganley, 2018]. Moreover, the formation of R loops (RNA-DNA duplex) at certain IGS loci prevents RNA-polymerase I from reading sense ncRNAs, which can disrupt RNA expression in the human nucleolus [Abraham et al., 2020]. Recently, we have shown a localization of the functional 5S gene within the IGS in turtles and crocodiles that is unique for the vertebrates [Davidian et al., 2022]. This NOR-5S rRNA gene is only active in oocytes and apparently plays a role in producing a maternal pool of extra ribosomes during NOR amplification in oogenesis. All this supports the importance of studying the IGS structure. However, the number of complete ribosomal repeats annotated with their IGS constituents remains extremely low among the available genome assemblies. So far few publications on the identification of structural and functional blocks within IGS in Xenopus [Caudy and Pikaard, 2002], mice [Grozdanov et al., 2003], human and other Apes [Gonzalez and Sylvester, 1995; Agrawal and Ganley, 2018], chicken [Dyomin et al., 2019], and some reptiles [Davidian et al., 2022] exist. This under-investigation of spacer regions is a significant obstacle to understanding their function and evolution.

**NOR rDNA Repeat Sequences**

Information on the genomic location and cytogenetic features of NOR rRNA genes, the rDNA copy number and repeat size variability in chicken has long been available [Delany and Krupkin, 1999; Schmid et al., 2000, 2005]. However, complete sequence details and organisation of the rDNA repeat unit remained unknown until very recently. The 45S rDNA cluster sequence of Gallus gallus was completely assembled in 2016 [Dyomin et al., 2016]. A complete sequence of the rDNA repeat including IGS was annotated in 2019 [Dyomin et al., 2019]. We also introduce new data on the rDNA repeat sequence of the guinea fowl Numida meleagris so that we can compare complete rDNA repeats in representatives of 2 galliform families, Phasianidae and Numidae.

Integrated data on the chicken demonstrate the single NOR location on microchromosome 16 [Bloom and Bacon, 1985; Delany et al., 2009; Solinhac et al., 2010], the rDNA repeat copy number intra- and inter-individual variability to be 150–250 per haploid genome, and average rDNA array sizes in various chicken populations having a range from 5 to 7 Mb [Delany and Krupkin, 1999; Delany, 2000; Schmid et al., 2005]. The latter authors also showed that rDNA repetitive units in the chicken NORs vary from 11 to 50 kb. This difference depends on the IGS size, which is significantly larger in broiler breeds [Delany and Krupkin, 1999; Schmid et al., 2005]. Four chicken rDNA repeats identified to date are of 34,497 bp, 26,999 bp, 27,055 bp, and 25,865 bp, the latter detected in the red jungle fowl [Dyomin et al., 2019]. A sample of the complete chicken rRNA gene cluster sequence was first assembled from raw reads and annotated as being 11,863 bp long (NCBI accession number KT445934) [Dyomin et al., 2016]. This was later verified by data from sequencing BAC clone WAG137G04 containing 3 complete rDNA clusters using PacBio RSII [Dyomin et al., 2019]. The results obtained were very similar to the previously assembled sequence, both in sequence homology and cluster size (11,871 bp, 11,830 bp, and 11,855 bp).

The recently released N. meleagris genome assembly [Vignal et al., 2019] lacks rDNA data. We identified 2 complete rDNA clusters and 3 IGSs of the guinea fowl in NCBI JABXER010000123 contig (online suppl. Material 11, Table S1). The sequence lengths and GC content in the elements that make up the rRNA gene clusters have been determined for both chicken and guinea fowl (Fig. 28; Table 8; online suppl. Material 11, Table S1). Their 18S, 5.8S, and 28S rRNA gene sequences are typical of vertebrates but ITS1 and ITS2 sequences are more extended in size, and show higher GC content compared to the majority of other Deuterostomia [Dyomin et al., 2017]. The secondary structure of rDNA cluster sequences should therefore be complicated and infusible while containing multiple hairpins, as was shown for the chicken ITS1 [Dyomin et al., 2016]. The ITSs may be the source of species-specific microRNAs, as was shown for human ITS1 (miRNA-663) [Chak et al., 2015] and for mouse ITS2 (miRNA-712) [Son et al., 2013].

**Sequence Complexity and GC Content in Intergenic Spacers**

When deciphering the guinea fowl NOR rDNA repeats, we also found that the repeat length difference between chicken (∼27–34 kb) and guinea fowl (∼18–20 kb) is due to the great difference in the IGS lengths (Fig. 28; Table 9). Even within a single NOR, IGS sequences may differ in length (Table 9) and, as a consequence, in their nucleotide composition. The difference in IGS lengths correlates with the difference in the amount of internal repeats (Fig. 29), which may be caused by unequal crossing-over [Erickson and Schmickel, 1985; Smirnov et al., 2016]. It is noteworthy that the red jungle fowl IGS is the shortest, which suggests the possibility of the repeat num-
ber increasing in the course of domestication [Dyomin et al., 2019]. We should remember the splendid data by M. Delany and coworkers [Delany and Krupkin, 1999; Delany, 2000; Schmid et al., 2005] of broiler chicken breeds having the longest rDNA repeats (up to 50 kb) due to the longer IGS sequences. Interestingly, despite the strongly differing lengths of IGS, the GC content in IGS remains particularly high in chicken and guinea fowl (Table 9). Compared to mammals [Agrawal and Ganley, 2018], the IGS of Galliformes representatives feature a strict hierarchy and order: they contain internal tandem repeats conserved in the unit lengths and are arranged in long blocks and oriented in the same direction. According to Dyomin et al. [2019], the chicken IGS contains 3 internal repeat blocks: 5′ SV-AL block, central EL block, and 3′ VAL block (Fig. 29a). The 5′ block consists of GC-rich AL repeats (~250 bp) alternating with AT-rich SV repeats (~150 bp). The central block is the longest (9,297–14,414 bp) and consists of short GC-enriched tandem EL repeats (~93 bp). The 3′ block is separated from the central block by a poly-A motif and consists of VAL repeats (~85 bp). The guinea fowl IGS also contains internal repeat blocks of 3 types: the conserved SV-AL and VAL blocks and apparently a species-specific Nme block (Fig. 29b). Both chicken and guinea fowl IGS have a poly-T motif (~10–49 bp) at the 5′ end of the IGS (Fig. 29), which is usually considered the transcription termination motif [Mason et al., 1997]. However, according to the results of transcriptome analysis, the SV-AL repeat block following the poly-T motif is transcribed in all analysed chicken tissues [Dyomin et al., 2019]. Chicken IGS contains one long unique highly conserved sequence (~1,940 bp) between SV-AL and EL repeat blocks and one short unique region (~191 bp) after the 3′ VAL repeat block. At the same time, the guinea fowl IGS has at least 3 unique regions: between Nme and VAL repeat block (~1,350 bp), between VAL blocks (~735 bp), and after the 3′ VAL block (169 bp) (Fig. 29; online suppl. Material 11, Data S1). Each unique region contains a single sequence of interspersed repeat (IR, 93 bp), which is enriched with adenine (Fig. 29). (CT)_n repeats (32–36 bp) have also been found in both

### Table 8. Length and GC content in the rDNA cluster and cluster elements in chicken and guinea fowl

<table>
<thead>
<tr>
<th>Species</th>
<th>Cluster</th>
<th>Cluster elements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>length, bp</td>
<td>5′ETS</td>
</tr>
<tr>
<td>Gallus g. domesticus (KT445934)</td>
<td>11,863</td>
<td>1,836</td>
</tr>
<tr>
<td>Numida meleagris (JAXER010000123)</td>
<td>12,133</td>
<td>1,792</td>
</tr>
</tbody>
</table>
chicken and guinea fowl unique sequence regions (Fig. 29). It is noteworthy that the size of the third guinea fowl IGS is enlarged due to an insertion into the VAL block between VAL_C and VAL_B units (Fig. 29b, dotted rectangle). This insertion duplicates the IGS fragment containing the unique area and a part of the VAL block. Thus, the longest IGS in the contig contains 3 VAL repeat blocks and 4 unique regions, with 3 of these regions carrying a copy of the IR.

Neither within chicken IGS [Dyomin et al., 2019] nor within the IGS of guinea fowl (this research) were functionally significant sites detected.

As it turned out, the variants of EL and VAL repeats are non-randomly scattered over the corresponding blocks. They are strictly organised into groups and form high-order repeats (HORs) (Fig. 30). This pattern is common to domestic and red jungle chicken in case of EL repeats (Fig. 29a, 30) and to guinea fowl in case of VAL repeats (Fig. 29b, 30). A comparison of different IGS copies demonstrates that internal repeats play the key role in the variability of IGS lengths at the individual and species levels (Fig. 29). In both chicken and guinea fowl, the IGSs are very rich in GC and CpG. This brings them closer to the turtle IGS [Dyomin et al., 2019; Davidian et al., 2022]. However, it is still unclear whether the IGS internal repeats perform any function. The IGS internal repeats and unique sequence regions described herein are not homologous to mammalian and amphibian IGS elements. This may reflect a separate evolutionary pathway for avian rDNA regulatory sequences, which are still to be discovered.

Table 9. Intraindividual variability of IGS lengths in chicken and guinea fowl

<table>
<thead>
<tr>
<th>Species</th>
<th>IGS_I Length, bp</th>
<th>IGS_I GC, %</th>
<th>IGS_II Length, bp</th>
<th>IGS_II GC, %</th>
<th>IGS_III Length, bp</th>
<th>IGS_III GC, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gallus g. domesticus (WAG137G4)</td>
<td>22,627</td>
<td>68.1</td>
<td>15,169</td>
<td>69.2</td>
<td>15,241</td>
<td>68.1</td>
</tr>
<tr>
<td>Numida meleagris (JABXER010000123)</td>
<td>6,530</td>
<td>67.6</td>
<td>6,562</td>
<td>68.1</td>
<td>8,223</td>
<td>68.1</td>
</tr>
</tbody>
</table>

Fig. 29. Chicken and guinea fowl IGS structure. a Four aligned chicken IGS sequences. IGS_I, IGS_II, and IGS_III are from the Gallus g. domesticus BAC-clone containing rDNA (WAG137G04). IGS_IV belongs to a red jungle fowl (AADN04001305.1). All four IGS have different sizes caused by the difference in repeat blocks of each type (SV-AL, EL, VAL). The unique regions are almost of the same length in all analysed IGS (see also online suppl. Material 11, Table S1). b Three aligned IGS sequences from the guinea fowl JABXER010000123 contig. Two of them are completely identical, IGS_III has an insertion at 5,500 bp (dotted rectangle). Each of the IGS contains no EL repeats, one SV-AL repeat block and at least two VAL repeat blocks, differentiated into 6 repeat variants (see also online suppl. Material 11, Fig. S1, S2, Data S1). A species-specific Nme repeat block following SV-AL block is marked in black. Sequence gaps are designated with fine black lines.
5S rDNA in Galliformes

The previous report on chicken genes [Schmid et al., 2005] has estimated the copy number of 5S rRNA genes in chicken as 35–41 copies with a predominant repeat (5Sα) of 2.2 kb [Daniels and Delany, 2003; Schmid et al., 2005]. Genetic linkage analysis and cytogenetic localization assigned the 5S rDNA to chromosome 9 [Daniels and Delany, 2003]. This is fully consistent with what can be found in the latest chicken genome assembly GRCg7w: one 5S rDNA locus (partly annotated) is situated at ∼0.5 Mb on chromosome 9 (NC_052540.1). It comprises 37 full and 1 incomplete tandemly repeated unit made up of a 5S rRNA gene followed by a spacer sequence (NTS) that can be of 2,038–2,268 bp. Such a polymorphism in the length is mainly due to the (G)n microsatellite present in the middle of each NTS. Traces of endogenous retroviruses (ERV-LTR) are also present in the NTSs. Two 5S rRNA gene copies seem to be nonfunctional; they comprise 117 and 118 bp due to nucleotide deletions. Some copies contain 1 or 2 nucleotide substitutions, allowing insights into the intradividual sequence variability (online suppl. Material 11, Fig. S3a). A variant 5Sβ rRNA gene repeat of 0.6 kb [Daniels and Delany, 2003] is not present in the Red Jungle fowl genome, but can be found in many WGS contigs of the Rhode Island, Liyang, Houdan, White Leghorn, Cornish, and Naked Neck chicken genomes. The GRCg7w assembly contains 3 single 5S rRNA gene copies not followed by the NTS (online suppl. Material 11, Fig. S3b). Two of them at chr2:44,384,176 bp (LOC112531980 at NC_052533.1) and at chr6:9,273,471 bp (NC_052537.1) are slightly degenerated (7 and 8 SNPs correspondingly), whereas the gene situated at chr9:3,007,209 bp is completely identical to the reference 5S rRNA gene sequence (NCBI X01,309.1). We could assume the existence of a single functional 5S rDNA site in the chicken genome on chromosome 9. However, this assumption would conflict with the earlier descriptions of 2 [Krol et al., 1981; Lazar et al., 1983] and even 3 [Keith et al., 1986] 5S rRNA types from various somatic tissues of the chicken. The alignment of 5S rDNA sequences with the available 5S rRNA sequences of the two types [Lazar et al., 1983] shows that 5S rRNA genes located on chromosome 9 correspond to 5S rRNA type I (NCBI M13920, online suppl. Material 11, Fig. S3b). A variant gene on chromosome 2 is identical to 5S rRNA type II (NCBI M13919, online suppl. Material 11, Fig. S3b). Both types of 5S rRNAs were found in the cytoplasm polysome fractions of chicken liver and brain [Lazar et al., 1983]. The 5S rRNA type II coding sequence is not associated with a common NTS.

In the N. meleagris genome assembly no 5S rRNA genes have been annotated so far. The corresponding array consisting of 18 full repeats is present in NCBI WGS contig JAXRER010000004. The predicted NTSs are much shorter (937–961 bp) than in chicken. In general, the guinea fowl 5S rDNA repeats are highly homogeneous as only 1 nucleotide substitution in the genic regions (online suppl. Material 11, Fig. S3b) and a dozen mutations in the NTS were found. Neither guinea fowl NTSs nor chicken repeats show the complex organisation as described above for the IGS, but they both feature a high GC content (~70%). When aligned, NTSs of both species share no related significant features. Only simple oligonucleotide motifs can be found (e.g., CCCGC, GGGTCG, GCGTG, GGAGCAG etc.). The biological meaning of these features is still to be elucidated.
Conclusion

Structural features and variability of chicken rDNA arrays are well studied [Delany and Krupkin, 1999; Schmid et al., 2000, 2005; Daniels and Delany, 2003; Dyomin et al., 2016, 2017, 2019]. This report additionally introduces the first decoding and analysis of the guinea fowl rDNA repeat. The SS rRNA genes are variable in their structure, copy number, and location in the chicken genome but appear to be homogeneous in the guinea fowl. Both studied species, and presumably all representatives of Galliformes, do not have a special oocyte type of SS rRNA genes, in contrast to fish, amphibians, and archelosaur reptiles. This perfectly coincides with the absence of mRNA gene amplification in avian oogenesis.

The complete sequences of the ribosomal repeat from the NORs in chicken and guinea fowl, representatives of 2 closely related galliform families, Phasianidae and Numididae, are compared. On the whole, the rDNA clusters were found to be quite similar in both species. The 18S, 5.8S, and 28S rRNA coding sequences are typical of higher eukaryotes. The ITSs are longer and more enriched in GC than in other Deuterostomia [Dyomin et al., 2017]. The similarity on high GC and CpG content in the ITS1 and ITS2 sequences of both species may indicate the existence of a general evolutionary mechanism that maintains the same proportions of nucleotides in avian rDNA spacers. Avian IGS seem to differ essentially from the IGS in Mammalia. In both studied species of birds, the IGS sequences separating rRNA gene clusters contain several blocks of internal tandem repeats of the same type. The internal tandem repeats of chicken IGS are HORs enriched in GC and CpG. This brings the IGS closer to turtles and distinguishes them from fish, amphibians, and mammals [Dyomin et al., 2019]. The same is true for the guinea fowl IGS. Some of the internal repeats, such as SV-AL, (CT)$_n$, VAL, were found to be common to the IGS of both species and perhaps to IGS of all Galliformes. At the same time, specific internal IGS repeats exist, such as the Nme repeats for the guinea fowl IGS and the EL repeats for the chicken IGS. The existence of specific repeat types within relatively close taxa of birds, the high degree of degeneracy of SV-AL and VAL, and the low interspecies homology of unique sequences indicate a high rate of evolution of both the entire IGS structure and of its elements. No functionally significant sites associated with RNA-polymerase activity were found in the IGS sequences of the chicken and guinea fowl using conventional transcription databases. In the identification of internal regions of IGS with high functional significance, it could be quite useful to study the IGS structure in representatives of all major taxa of birds. Such a study could reveal the main evolutionarily conserved sequences in the genomes of Galliformes. It could also clarify the mechanisms that determine the structural diversity and rapid evolution of IGS sequences compared to other spacer sequences of the avian rDNA arrays, even at the intraspecific level.

Acknowledgments

We are sincerely grateful to Alissa Gousseva for her kind help in English. The analytical facilities were provided by “Chromas” Resource Center of the Saint Petersburg State University Scientific Park. The study was supported by Russian Science Foundation grant #22-24-00538.

Conflict of Interest Statement

The authors declare that they have no competing interests.

Funding Sources

This work was supported by the grant from the Russian Science Foundation #22-24-00538.

Data Availability Statement

One full rDNA repeat annotation of *N. meleagris* in the JABX-ER010000123 contig, the alignments of the SS rRNA sequences found in the chicken genome assembly bGalGal1.mat.broiler. GRCg7b and the guinea fowl JABXER010000004 NCBI WGS contig are available in the Supplementary material.

Endogenous Avian Leukosis Viruses: The Chicken-Specific Endogenous Retrovirus

(Prepared by A.S. Mason)

As with all birds, the chicken genome is repeat sparse, extending across all repetitive element classes (comprehensively described by Bigot and Arensburger in this report). In mammals, approximately 10% of the genome consists of endogenous retroviruses (ERVs), a subclass of long terminal repeat (LTR) retrotransposons which reflect ancestral retroviral infections and integrations into the germline [Bromham, 2002]. However, and consistent with other Neognathae birds, ERVs only constitute approximately 3% of the chicken genome, though this may change with greater assembly contiguity [Mason et al., 2016; Kapusta and Suh, 2017; Warren et al., 2017]. Chicken ERVs include full-length and degraded examples of spumaviruses, beta- and gamma-retroviruses, but the endogenous alpha-retroviruses are of particular interest as this retroviral group is endemic to birds. Chicken genomes consist of 2 major groups of endogenous alpha-retroviruses: endogenous avian viruses (EAVs) and avian leukosis virus subgroup E (ALV-E) integrations, previously known as “ev” loci [Payne and Nair, 2012; Sacco and Nair, 2014].
EAVs are ancestral, found broadly across Galliformes with element divergence matching host co-speciation patterns. EAVs are further subdivided into 3 phylogenetically distinct clades (EAV-0, EAV-HP, and EAV-E51/E33), although inter-clade recombination has been described, resulting in ART-CH elements (avian retrotransposon in chicken; EAV-HP/EAV-51 recombinant) [Gudkov et al., 1992; Sacco and Nair, 2014]. Most chicken EAVs are replication-incompetent, but many are polymorphic with some retaining transcriptional or regulatory potential, associated with phenotypic traits such as blue-green eggshell colour [Wang Z et al., 2013; Wragg et al., 2013]. Most prominently, recombination between an intact EAV-HP envelope and exogenous ALV-A produced the emergent ALV-J, with the recombinant virion enabling altered haemopoietic tropism, inducing myeloma cells rather than the typical B cell lymphoma [Payne et al., 1991; Benson et al., 1998; Sacco et al., 2004]. Even these limited examples highlight the potential impact of ERVs when they retain high structural integrity.

ALV-E integrations are evolutionarily recent and recurrent additions to the chicken genome, as these ERVs are endemic only to *Gallus gallus* and exemplify an evolutionarily brief period in retroviral life history where exogenous and endogenous forms co-exist [Payne and Nair, 2012; Kanda et al., 2013]. Consequently, ALV-Es are highly polymorphic and are present at low copy number, but typically retain high structural integrity. Replication-competent and transcriptionally active ALV-Es are common, even in commercial flocks, where gag expression has been associated with reductions in muscle mass and egg number, size, and shell thickness [Crittenden et al., 1984; Fox and Smyth, 1985; Kuhnlein et al., 1989; Gavora et al., 1991]. Furthermore, ALV-E expression has complex impacts on the immunological status of flocks. ALV-Es can be shed and transmitted horizontally through a flock, high antigen titre can lead to persistent viremia and immune exhaustion, and yet expression of endogenous envelope can provide protection against exogenous ALV by receptor interference [Robinson et al., 1981; Smith et al., 1990]; a form of ERV-derived immunity (EDI) [Aswad and Katzourakis, 2012; Hurst and Magiorkinis, 2015]. Modulated infection dynamics have also been observed with other viruses [Mays et al., 2019]. Efforts to eradicate ALV-Es in commercial flocks have been hindered by their close association with desirable traits such as recessive white [Chang et al., 2006], henny feathering [Li J et al., 2019], and sex-linked slow feathering [Bacon et al., 1988; Elferink et al., 2008].

**ALV-E Structure, Retrotransposition, and Expression**

Intact ALV-E integrations have a typical length of 7,524 bp. This includes LTRs of 274 bp which are identical upon integration, terminally flanked by 6-bp target site duplications. Due to their recent integration, most ALV-E LTRs remain identical, and many ALV-Es retain full structural integrity (e.g., ALVE21, ALVE-TYR). Degraded ALV-Es, including terminal (e.g., ALVE6, ALVE9) or internal (e.g., ALVE3) truncations, and solo LTRs (e.g., ALVE15, ALVE_ros005) are common, but less frequent than in more ancient ERV groups [Stoye, 2001].

ALV-E LTRs are shorter than exogenous ALV, limiting endogenous promoter activity and transformation capability [Ruddell, 1995; Benachenhou et al., 2013]. The phenotypic impact of ALV-E integrations is therefore dependent on the specific integration site. Like all alpha-retroviruses and lentiviruses, the presence of a nuclear localisation signal in the retroviral integrase leads to an enrichment of ALV-E integrations within open chromatin encompassing protein-coding genes [Narezkina et al., 2004; Justice and Beemon, 2013]. This distribution holds across all identified ALV-E loci, even after the impact of selection [Mason et al., 2020c].

ALV-Es exhibit the canonical retroviral structure without accessory genes, expressed by host RNA polymerase II from 2 frame-separated open reading frames: gag-pol and env. Like other retroviruses, translation of the gag-pol transcript is regulated by ribosomal frameshifting. Frameshifting is successful 5–10% of the time, resulting in far lower abundance of pol proteins compared to those encoded by gag, including alpha-retrovirus-specific inclusion of the protease domain [Arad et al., 1995]. env translation is inhibited by host miR-155 binding within the surface receptor gp85 domain [Hu et al., 2016], although some integrations, such as ALVE6 [Mason et al., 2020a], have escaped regulation by mutation, resulting in high envelope titres [Robinson et al., 1981].

Novel horizontal transmission of ALV-Es within flocks, and between cells of the host, is dependent on the cell entry receptor genotype [Hunt et al., 2008]. ALV-E, as well as exogenous ALV subgroups B and D, use tumour virus cell entry receptor B (TVB), encoded by *tumour necrosis factor receptor superfamily member 10b* (*TNFRSF-F10B*). Wildtype TVB (*TVB*S1) is susceptible to infection, but alleles which result in either truncation before the transmembrane domain (e.g., Q58*, Q100*), or direct or indirect disruption of disulfide bridges in cysteine-rich domains (e.g., P61L, C62S, C101R, C125S), have been observed to provide resistance to ALV-E infection in commercial flocks [Adkins et al., 2000, 2001; Klucking and...
Early work characterising ALV-E diversity focused on White Leghorn lines (typically 1–3 loci) due to the well-described detrimental effects on egg-laying success [Gavara et al., 1991]. Expansion into the more genetically diverse brown-egg layers (typically 5–10 loci) and broilers (genotypes rarely published from commercially relevant lines) was inhibited by available technologies [Iraqi et al., 1991; Sabour et al., 1992; Grunder et al., 1995; Muir et al., 2008b]. Restriction fragment length polymorphisms (RFLPs) became harder to interpret with a greater number of ALV-Es, especially following the discovery that some RFLPs varied between breeds for the same loci [Aarts et al., 1991; Boulliou et al., 1991]. Taken together, these studies showed that whilst some ALV-Es were shared between white- and brown-egg layers and broilers (notably ALVE3 and ALVE6), many loci were novel, and there was no clear indication of an ancestral ALV-E complement.

By the time of the publication of the draft red junglefowl (RJF) reference genome [International Chicken Genome Sequencing Consortium, 2004], almost 50 different ALVE loci had been identified, many with diagnostic and commercially-utilised PCR genotyping assays [Benkel, 1998]. It was therefore surprising that the RJF individual used for the reference contained only 2 ALVEs: ALVE6 (ALVE-JFevA), widespread but polymorphic in commercial layers and broilers yet not found in other RJF, and the intact ALVE-JFevB, which is, so far, unique to the reference genome individual [International Chicken Genome Sequencing Consortium, 2004; Benkel and Rutherford, 2014; Mason et al., 2020a]. Even the study of just these 2 ALV-E elements has not been straightforward, as ALVE6 is near the chromosome 1 p-arm telomere, and was only observed in brown-egg layers and broilers, including ALVE-TYR, responsible for the recessive white phenotype of the maternal bird. Both haplotypes contain TVB resistance alleles.

The direct benefit of long read-scaffolded assemblies was the full characterisation of ALV-E integrity in these haplotype references. This remains an outstanding issue with ALV-E identification from short-read technologies alone, which can optimally identify integration sites, but are unable to uniquely resolve internal integrity; crucial for predicting expression and retrotransposition potential.

Not only does the reference RJF assembly poorly represent the diversity of ALV-Es in commercially relevant chickens, it also poorly represents RJF ALV-E diversity [Mason, 2021]. This corresponds with the well-described White Leghorn introgression into this RJF individual, as well as the general issues studying polymorphic repetitive elements [Ulfah et al., 2016]. Unfortunately, despite the literature on ALV-E polymorphic diversity, and the clear absence of a conserved, ancestral ALV-E complement, the RJF genome is still used to represent the pre-domesticated state [Hu et al., 2017; Sun YH et al., 2017; Chen S et al., 2019], leading to ambiguous or overreaching results, both with ALV-Es and more broadly.

Pangenomes will more comprehensively document ALV-E diversity. The recently derived Chinese and South-East Asian chicken pangenome consisting of 664 individuals [Wang K et al., 2021], many of which were previously analysed for their ALV-E content [Mason et al., 2020b], was certainly a good start, however even these resources can never be “complete.” Researchers studying ALV-Es need to be sure which specific integrations are present in their study system. As high-throughput sequencing is the only unambiguous method for novel integration detection, the reality of having a highly contiguous but nearly ALV-E-blank genome, such as GRCg6a, is actually quite appealing, as the first step of many detection algorithms is to mask homologous repeats.

Whilst no individual- or pangenome can fully represent ALV-E diversity, the newly derived, haplotype-phased layer (GRCg7w; paternal) and broiler (GRCg7b; maternal) references are good representatives of their breeds, and highly informative for the study of Western commercial stock [Warren et al., this report]. GRCg7w contains 6 ALVEs which, while more than usual for a White Leghorn, contain the common ALVE1, ALVE3, ALVE15, and ALVE21. GRCg7b has 5 ALVEs commonly observed in brown-egg layers and broilers, including ALVE-TYR, responsible for the recessive white phenotype of the maternal bird. Both haplotypes contain TVB resistance alleles.

The direct benefit of long read-scaffolded assemblies was the full characterisation of ALV-E integrity in these haplotype references. This remains an outstanding issue with ALV-E identification from short-read technologies alone, which can optimally identify integration sites, but are unable to uniquely resolve internal integrity; crucial for predicting expression and retrotransposition potential.
Future Relevance of ALV-Es in Disease

Much of the impetus for studying ALV-Es was based on the detrimental impact of these ERVs on productivity traits [Crittenden et al., 1984; Fox and Smyth, 1985; Kuhnlein et al., 1989; Gavora et al., 1991]. However, direct and indirect selection against these elements, particularly in commercial layers, has largely eradicated these effects. A recent association analysis suggested only linkage disequilibrium in ALV-E/trait associations, even for integrations such as ALVE3 [Fulton et al., 2021]. The full picture is less simple, as ALV-Es fixed in a population cannot be measured in such studies. ALVE-TYR, for example, is structurally intact, common in layers and broilers, and its impact on growth rate has been appreciated since the 1980s [Fox and Smyth, 1985]. Further work is needed to characterise the phenotypic effects of individual ALV-Es to prioritise their eradication from flocks, particularly given the poorly understood role of ALV-E loci in spontaneous lymphoid leukosis [Cao et al., 2015; Mays et al., 2019].

Outside intensive selection, ALV-E diversity and abundance appears to support a more natural role through ERV-derived immunity [Mason et al., 2020c]. Whilst this may improve the host response to novel ALV infections, the deleterious productivity associations impact the food and economic security of subsistence and small-holder poultry farmers, and will introduce unwanted deleterious loci in the generation of "localised" commercial birds in areas such as sub-Saharan Africa.

Unregulated flocks also enable the more nebulous possibility of novel, emergent recombinant retroviruses. This has been documented extensively in China with exogenous ALV subgroups A, J, and K [Chesters et al., 2001; Liu et al., 2011; Dong et al., 2015; Přikryl et al., 2019], but coinfection could facilitate recombination with other retroviruses, leading to novel tropisms or host expansion.

Conflict of Interest Statement
The author has no conflicts of interest to declare.

Funding Sources
A:S.M. is supported by a Research Fellowship in Cancer Informatics by York Against Cancer. This work was supported by a Biotechnology and Biological Sciences Research Council pump priming award (BB/W510737/1) and a researcher mobility award from The Houghton Trust (HT/RMG/22/01).

Recombination Frequency and Distribution along the Chicken Chromosomes
(Prepared by P.M. Borodin, L.P. Malinovskaya, and A.A. Torgasheva)

What Is the Recombination Rate and Why Is It Important?

Recombination is essential to orderly chromosome segregation and generation of new allele combinations. The efficiency of artificial selection is critically dependent on the recombination rate; that is, the number of recombination events per whole genome, chromosome, and chromosome region. Populations with higher recombination rates demonstrate a higher response to selection [Martin et al., 2006; Dapper and Payseur, 2017; Gonen et al., 2017]. The distribution of recombination events along chromosomes is another important variable affecting the efficiency of selection. A position of 2 crossing overs too close to each other does not affect the linkage phase [Gorlov and Gorlova, 2001; Berchwitz and Copenhagen, 2010]. Similarly, crossing overs located too close to the centromere of an acrocentric chromosome or to the telomere do not produce new allele combinations. Thus, estimates of genome-wide and chromosome- and region-specific recombination rates in livestock are important for breeding programs.

Recombination is a stochastic but tightly controlled process occurring in the prophase of the first meiotic division [Zickler and Kleckner, 2015; Gray and Cohen, 2016]. It starts with chromatin remodeling and the scheduled generation of multiple DNA double-strand breaks (DSBs), followed by a RAD51-mediated search for homologous DNA sequences and formation of heteroduplexes involving DNA strands of homologous chromosomes. Polymerization of the synaptonemal complex, a meiosis-specific proteinaceous structure, stabilizes homologous chromosome synopsis. A small percentage of DSBs are repaired in a crossover manner, while the majority are repaired in a non-crossover manner. Most of the crossover sites are located in recombination hotspots, regions 1–2 kb long, usually flanked by longer cold regions with lower than average recombination frequency [Paul et al., 2016]. The sites of crossing over can be visualized at the pachytene stage as recombination nodules containing MLH1 (mismatch repair protein), and at the diploteine-diakinesis stage as chiasmata. Sister chromatid cohesion beyond the chiasmata holds homologues together at metaphase I, ensuring proper orientation and orderly segregation. Crossover and non-crossover chromatids segregate at the second meiotic division.
How to Measure the Recombination Rate

There are genetic and cytological methods of assessing the recombination rate. Genetic assessment is based on linkage analysis. It requires large sets of well-controlled crosses or well-characterized pedigree records. This approach provides precise estimates of recombination rate even between closely linked markers. However, it is expensive and time- and labor-consuming. Its efficiency is critically dependent on the number and distribution of the markers. Chiasma count at diplotene and diakinesis provides an unbiased estimate of the genome-wide and chromosome-specific rate of recombination. However, the efficiency of this approach is restricted by difficulties in obtaining the cells at the particular stages of meiosis and by the accuracy of ascertaining chiasma position. The cytological methods of recombination mapping based on the electron microscopic visualization of the recombination nodules or immunolocalization of MLH1 protein in pachytene spermatocytes and oocytes provide highly reliable estimates of the total recombination rate, as well as the frequency and distribution of recombination events in individual chromosomes [Anderson et al., 1999]. The material for this analysis is easily available in the testis of adult males during the breeding season and in the ovaries of juvenile females during the first week after hatching [Pigozzi, 2016]. Figure 31 shows examples of the application of this method to chicken spermatocytes and oocytes.

Genetic and Cytological Estimates of the Chicken Recombination Rate

The first genetic map of the chicken chromosomes was constructed by Serebrovsky and Petrov [1930]. It contained 12 markers at 4 linkage groups and 4 more unlinked markers. The total length of the map was 252 cM. The most recent high-density consensus linkage map is based on the analysis of the segregation of 9,268 SNPs and other markers in 3 different mapping populations. The total length of the linkage map is 3,098 cM for female meiosis and 3,145 cM for male meiosis [Groenen et al., 2009].

The first cytological estimate of the recombination rate has been carried out by chiasmata count in the cockerel spermatocytes at the diakinesis-metaphase I stage [Pollock and Fechheimer, 1978]. There was a significant interindividual variation: from 56 to 66 chiasmata per spermatocyte. Analysis of the number and distribution of chiasmata along the lampbrush chromosomes in chicken diplotene oocytes gave a similar interval of variation (59–64) [Rodionov et al., 1992]. The average numbers of recombination nodules estimated by electron microscopic visualization [Rahn and Solari, 1986] and immunolocalization of MLH1 protein in pachytene oocytes [Pigozzi, 2001] were rather close to each other (57.5 and 65.0, correspondingly) and to the chiasma count in the spermatocytes. Because each chiasma or recombination nodule represents one crossingover (50 cM), the total length of the sex averaged chicken genetic map is estimated to be 2,800–3,300 cM. There is a reasonably good correspondence between the genetic and cytological estimates of the recombination rate in chicken. It has, for a long time, been considered the highest among birds. It was suggested that such a high recombination rate could have resulted from domestication and strong artificial selection [Groenen et al., 2009; Backström et al., 2010]. Further studies clarified the intermediate position of the chicken recombination rate between those of the white wagtail (3,805 cM) and the black tern (2,155 cM) [Semenov et al., 2018].
The Distribution of the Recombination Nodules along the Macrochromosomes

Mapping of recombination nodules at the synaptonemal complexes made it possible to visualize the recombination landscapes of chicken macrochromosomes. They showed a highly positive correlation between the length of synaptonemal complex and the number of recombination nodules [Pigozzi, 2001; Rahn and Solari, 1986], which is typical for vertebrate chromosomes. The macrochromosomes of most birds examined show a highly polarized distribution of the recombination nodules with steep peaks near the telomeres and deep valleys near the centromeres [Pigozzi and del Priore, 2016]. The recombination landscapes of the chicken macrochromosomes are slightly more flatter than those of other Galloanserae [Calderón and Pigozzi, 2006; del Priore and Pigozzi, 2015; Pigozzi and del Priore, 2016].

The recombination landscape at ZW bivalent is of special interest. In all Neaves examined, it contains a single crossover located in a very small pseudoautosomal region (PAR). Using the limits of MLH1 foci distribution, Torgasheva et al. [2021] estimated the size of PAR in domestic chicken as 5% of the completely paired ZW bivalent. Interestingly, in one oocyte they detected the second MLH1 focus located at 21% of the ZW length from the telomere, far beyond the most distant MLH1 foci found in other chicken oocytes. Although this exceptional MLH1 focus could be an artifact, it might indicate that rare recombination events are possible in the dispersed regions of residual homology between Z and W detected by Zhou Q et al. [2014].

Interbreed Variation in the Recombination Rate

The recombination rate shows substantial interbreed and individual variation. Malinovskaya et al. [2019] analyzed the number and distribution of MLH1 foci in spermatocytes of the roosters of 6 chicken breeds. They revealed significant effects of breed (\( R^2 = 0.17; \ p < 0.001 \)) and individual (\( R^2 = 0.28; \ p < 0.001 \)) on variation in this trait, determined mainly by variation in recombination density on macrochromosomes. There was an interesting correspondence between the age of the breed and its recombination rate. Those with high recombination rates were the breeds created during the last century by crossing several local breeds. The breeds with a low recombination rate were the ancient local breeds.

The linkage experiment also revealed significant interpopulation differences in recombination rates [Groenen et al., 2009]. F1 male and female hybrids between Red Jungle fowl and White Leghorn showed significantly lower recombination rates than purebred broiler populations. The authors suggested that a higher recombination rate in purebred domestic animals was the result of strong artificial selection. Alternatively, a lower recombination rate in the F1 hybrids might be the result of negative heterosis. Malinovskaya et al. [2021] found that F1 hybrids between 2 purebred breeds had a significantly lower recombination rate (2,950 cM) than the cockerels of both parental breeds, Russian Crested and Pervomai (3,150 cM and 3,350 cM, respectively). The authors explained the negative heterosis for the recombination rate by difficulties in homology matching between the DNA sequences of genetically divergent breeds.

Future Prospects

Controlling recombination frequency and, more importantly, its distribution, is a necessary future step in order to increase the efficiency of selection and overcome its limitations. Identification of the key molecular regulators of crossing-over in plants leads to new technologies for increasing recombination rate, which could potentially benefit plant breeding [Blary and Jenczewski, 2019].

The first group of methods relies on suppressing the genes limiting meiotic recombination, such as topoisomerase TOP3α, DNA-helicases FANCM and RECQ4, and AAA-ATPase FIGL1 [Girard et al., 2015; Séguela-Arnaud et al., 2015, 2016; Mieulet et al., 2018]. Combinations of mutations in these genes resulted in up to an 8-fold increase in recombination frequency in Arabidopsis thaliana hybrids [Fernandes et al., 2018].

This promising result, however, leaves doubts about the feasibility of developing such methods and their applicability for livestock. Although simulation studies show that an increase in genome-wide recombination would indeed result in an increased response to selection, a significant effect can only be achieved with a 10–20-fold increase in recombination rate [Battagin et al., 2016; Gonen et al., 2017]. Such a huge increase in genome-wide recombination rate, if ever possible, will break up existing beneficial allele combinations and end up in decreased genomic selection accuracy [Battagin et al., 2016]. This might outweigh the benefits of reshuffling genetic material and a potential increase of genetic gain [Blary and Jenczewski, 2019]. In addition, due to the relatively high stability of recombination hotspots in birds [Singhal et al., 2015], a genome-wide increase in recombination may not lead to the appearance of new crossover sites. Indeed, a 19% interbreed variation in recombination rate did not affect pattern of crossover localization along the chicken macrochromosomes [Malinovskaya et al., 2019].
By manipulating the distribution of crossover events (for example, by stimulating recombination in cold regions), the negative aspects of increased recombination might be avoided. Another approach for controlling crossing-over tested in Saccharomyces cerevisiae operates by targeting SPO11, a protein which induces DSBs, to specific regions [Peciña et al., 2002; Sarno et al., 2017]. Simulation study predicts that inducing crossing-over in non-recombining regions decreases the loss of genetic variability and increases genetic gain, especially in cases when polymorphisms associated with the trait are clustered [Gonen et al., 2017]. Thus, the development of a technology for local modifying recombination patterns, taking into account the genetic architecture of a trait, might have the potential to improve breeding programs in the future. Considering the relatively high stability of recombination hotspots in birds [Singhal et al., 2015], their redistribution could have potential for chicken breeding.

**Conclusion**

Being an important trait for artificial selection, recombination rate is actively studied in the domestic chicken by various cytological and molecular genetics methods. Recent estimates are consistent and indicate that chicken is characterized by a high level of recombination rate, which is typical for birds. It also shows substantial inter-breed variation and a relatively stable pattern of distribution along chromosomes. Methods for controlling the position of crossovers could potentially be useful for chicken breeding in the future.

**Conflict of Interest Statement**

The authors have no conflicts of interest to declare.

**Funding Sources**

The work was supported by the Ministry of Science and Higher Education of the Russian Federation, grant numbers FWNR-2022-2015 and 2019-0546 (FSUS-2020-0040).

**The Broad-Scale Recombination Landscape of Chicken Macrochromosomes Inferred by in situ Mapping of Crossovers**

(Prepared by M.I. Pigozzi)

Meiotic crossovers (CO) serve to maintain and generate genetic variability in genomes by breaking associations between alleles at linked loci, resulting in new haplotypes. This has implications for the effect of selection on molecular evolution, as non-recombining genomic regions of sexually reproducing organisms accumulate deleterious mutations and deteriorate [Reeve et al., 2016]. Also, the correct number and placement of meiotic CO have a vital role to ensure faithful segregation, with failures often resulting in aneuploidy and infertility. Recombination rates, or the number of recombination events per Mb every generation, may be explained by both differences in CO positions within a genome (recombination landscape) and genome-wide recombination (how many CO events occur per meiosis). A direct method to record both features in individual meiosis is the immunodetection of the MLH1 protein, a component of mature recombination nodules in pachytene (Fig. 32a). The length of the recombination map in centimorgans (cM) can be obtained multiplying the average number of MLH1 foci per cell by 50 map units, as one recombination event is equivalent to 50 cM. MLH1 focus data from chicken oocytes established that the average female genetic map is between 3,150 and 3,250 cM, in white layer and broiler lines, respectively [Pigozzi, 2001; del Priore and Pigozzi, 2020]. The use of the same methodology in spermatocytes from 5 different breeds established similar genetic lengths in males with interbreed variations of up to 19% in overall genomic recombination rates [Malinovskaya et al., 2019]. These independent studies also demonstrated a close match to the existing chicken linkage maps [Groenen et al., 2009].

From the MLH1 count in a number of domestic and wild bird species, global recombination rates and recombination patterns for macrochromosomes have been determined, offering a source for comparative analysis [Semenov et al., 2018; Pigozzi, 2022]. The usage of meiotic chromosomes at pachytene is particularly favorable to investigate the distribution of CO along chromosome arms, because at this stage of meiosis genomic distances are proportional to physical distances. This relationship is based on the relatively uniform size of the DNA loops attached to the synaptonemal complexes (SCs) of pachytene bivalents [Veller et al., 2019]. The theory of the regular spacing of DNA loops along the meiotic chromosomal axis can also be extended to the chicken where SC/DNA ratios exhibit little fluctuation [Pigozzi, 2007; del Priore and Pigozzi, 2021]. Thus, physical distances measured in micrometers or percentages along pachytene bivalents can be converted into genomic positions (in base pairs) along chromosomes. The resolution of the method has the limitation of the fluorescence microscopy. Figure 32b, c shows that the FISH signal of a BAC located at 1.2 Mb from the sequence start of the assembly of GGA1 can be clearly separated from the end of the synaptonemal com-
plex, indicating that accurate measurements could be made for sequences located in pachytene oocyte spreads that were separated by 2 Mb [Pigozzi, 2001; del Priore and Pigozzi, 2021].

Taking this resolution into account, the recombination rates along the 8 largest autosomal bivalents were calculated here for 2.5-Mb intervals based on MLH1 focus data in 138 chicken oocytes from previous work of our laboratory [del Priore and Pigozzi, 2020]. The fluctuation of the recombination rates along the macrobivalents is shown in the form of a heatmap where the colour of the cells represents the range of recombination rates (Fig. 33). The mean recombination rate in these intervals was 2 cM/Mb, with 54% of the intervals showing rates between 1.5 and 2.5 cM/Mb. Even though crossovers can be found anywhere along chromosome arms, recombination rates are higher near chromosome ends and lower in the center. Regions without foci were rare (5 in 283 intervals) and they were limited to the 0.5-μm intervals closest to the centromeres of metacentric/submetacentric chromosomes.

An MLH1 recombination map can be used when only the genetic (cM) positions of markers are known to predict their physical positions on a chromosome [Anderson et al., 2004]. One of the initial scans for potential quantitative trait loci (QTL) for production-related traits in the chicken identified a region on GGA1 spanning the area of 263 and 285 cM [Hansen et al., 2005]. The cumulative cM frequency along GGA1 calculated from the MLH1 recombination map predicts that this region comprises about 20 Mb between 145 and 165 Mb from the end of the short arm (Fig. 34a). This estimate could have been done even with the approximate size of GGA1 obtained by image cytometry [Smith and Burt, 1998; Mendonça et al., 2016]. Although this particular prediction is now outdated, it is an example of how cM maps based on MLH1 foci can help integrate physical and genetic maps at their initial stages. Identification of genomic loci governing complex traits involves time-consuming and expensive procedures such as QTL mapping or genome-wide association studies (GWAS). Linkage disequilibrium (LD) between QTL or markers plays a central role in gene lo-
calization [Pritchard and Przeworski, 2001]. As recombination is known to shuffle genetic material, leading to decay of LD, prior knowledge of the broad recombination landscape in regions of interest could be valuable before embarking on more complex genetic analyses. The MLH1 recombination maps can be overlaid onto assembled chromosomes to obtain cM distances between markers of interest. For instance, a region of GGA5 contains the majority of QTL for woody breast and white striping myopathies which are located in an 8-Mb-long area on GGA5, between 8 and 16.5 Mb of the chromosome sequence [Lake et al., 2021]. The MLH1-cM map shows that this segment is located in an area with recombination rates higher than the average for that chromosome and is bordered by regions with low recombination rates (Fig. 34b).

Fig. 33. Recombination rates along GGA macrochromosomes 1 to 8. Each chromosome is divided in cells of 2.5 Mb. The recombination rates are represented by the different colours as indicated in the legend. The arrowheads point at the centromere positions calculated from the centromeric indexes in synaptonemal complex spreads.

The recombination rates obtained by MLH1 mapping have been applied directly or indirectly to analyze macro-evolutionary processes, the variation in crossing over between populations or species, the existence of sex-specific recombination landscapes or the evolution of genome-wide recombination rates [Segura et al., 2013; Semenov et al., 2018; Guo X et al., 2020; Peterson and Payseur, 2021]. The wide range of biological issues analyzed by the immunocytological location of crossovers highlights the importance of this approach to studying meiotic recombination as well as its versatility to comparing data from independent studies.

Statement of Ethics
Handling and euthanasia of birds were conducted in accordance with Argentinian and international guidelines for the use of farm and laboratory animals and were approved by the Animal Care and Use Committee of the University of Buenos Aires School of Medicine (EXP-UBA 0047533/16, Res 2116/16).

Conflict of Interest Statement
The author declares no conflict of interest.

Funding Sources
Current research at the author’s laboratory is supported by ANPCyT, FONCYT BID-PICT 2016 #2302.

Data Availability Statement
The data presented in this study are openly available at Mendeley Data, V2, 10.17632/w3n9xp5dnp.2.

The BF-BL and BG Regions of the Chicken B Locus Differ in Genome Dynamics
(Prepared by J. Kaufman)

Genomes do not evolve homogeneously, and different regions may evince divergent properties. A classic example is the presence of isochores that separate by Giemsa staining and by density centrifugation. In mammals, GC-rich isochores are considered to replicate earlier in the cell cycle than AT-rich isochores, and also have a greater density of genes which are more compact [Constantini and Musto, 2017; Bernardi, 2021]. Analyses of the chicken B locus illustrate regions that differ in another way: the BF-BL region (which corresponds to the major histocompatibility complex, MHC) is relatively compact and stable with little obvious recombination, whereas the BG region undergoes frequent expansion and contraction leading to copy number variation (CNV). These two regions are organised differently, with the structure of the BF-BL region differing significantly from the MHC of mammals and having a profound effect on function, while the pheno-
typic effects of recombination and deletion on the BG genes are at present still a mystery [Kaufman et al., 1999a; Salomonsen et al., 2014].

The long history of the B locus has been extensively reviewed [with many more citations: Miller and Taylor, 2016; Afrache et al., 2020; Kaufman, 2021]. It was discovered as a polymorphic alloantigen system on chicken red blood cells, then found to co-segregate with a variety of functions associated with the MHC in mammals, and later found to co-segregate with the nucleolar organiser region (NOR) on a microchromosome, now numbered chromosome 16 (Fig. 35). Two regions became apparent by immunoprecipitation of the antigens recognised by the alloantisera: the BG region encoding large erythrocyte membrane proteins without obvious glycans, and the BF-BL region encoding membrane glycoproteins corresponding to MHC classical class I molecules (BF antigens with wide tissue distribution) and class II molecules (lymphocyte BL antigens). Later, it became clear that both BG and BL molecules are specifically expressed on various other cell types. Different B locus haplotypes were found to determine decisive resistance and susceptibility to various economically important pathogens, later narrowed down to the BF-BL region [Briles et al., 1983; Miller and Taylor, 2016; reviewed in Kaufman, 2021].

More recently (Fig. 35), a variety of approaches have shown that the BF-BL and BG regions flank a region primarily encoding tripartite motif-containing (TRIM) genes, and are separated by high levels of recombination from non-classical MHC-like genes in the Rfp-Y region, the ribosomal RNA genes of the NOR, olfactory receptors, and scavenger receptors [Ruby et al., 2005; Shiina et al., 2007; Delany et al., 2009; Miller et al., 2014]. Moreover, the BF-BL region was discovered to extend towards the telomere with a few genes found in the MHC class III region and with some non-polymorphic non-classical class I-like genes encoding CD1 molecules, which in mammals are found on other chromosomes in so-called MHC-paralogous regions [Maruoka et al., 2005; Miller et al., 2005; Salomonsen et al., 2005; reviewed in Kaufman, 2021]. The non-classical class I-like YF genes in the Rfp-Y region are apparently polymorphic and likely bind hydrophobic ligands, but they are not found in mammals nor do they determine rapid graft rejection or other classical MHC functions [Thoraval et al., 2003; Goto et al., 2022].

An early description of the BF-BL region was a “minimal essential MHC” [Kaufman et al., 1999a, b], based on the fact that most genes expected from the MHC of typical mammals are missing (including genes encoding the complement components C2 and factor B, and the class
Fig. 35. Organisation of regions on chicken chromosome 16, as currently published. 

**a** Depiction of chromosome 16, based on analysis by FISH, radiation hybrids, genetics, southern blotting, and sequencing. B, B locus; GC, G+C-rich region of PO1 repeats; Y, Rfp-Y region; NOR, nucleolar organiser region; BLA, class II A gene; fB, factor B gene; ORs, olfactory receptor genes; SRCRs, scavenger receptor with cysteine repeat genes. Double-headed arrows indicate recombination frequencies between B and BLA, fB and Rfp-Y, and B and Rfp-Y. 

**b** Region of the B locus currently sequenced, including the BF-BL region, the TRIM region and the BG region. Genes represented by boxes. Rising and falling stripes indicate genes of the classical class I and class II presentation system, respectively; stippled indicate class II region genes; black indicates lectin-like genes and pseudogenes; horizontal stripes indicate TRIM family genes; vertical stripes indicate BG genes. Names of genes above indicate transcription from left to right, below indicate transcription from right to left; note the homologous genes in opposite transcriptional orientation in the BF-BL region but in the same transcriptional orientation in the BG region, strongly affecting the dynamics of evolution based on recombination. Figure from Kaufman [2021].
II A chain BLA), with those few genes present being critical to classical MHC function, including the genes encoding the TAP1 and TAP2 chains forming the peptide transporter, the class I-dedicated chaperone tapasin, and the class II-dedicated chaperones formed by DMA, DMB1, and DMB2. Also present is the BRD2 (or RING3) gene encoding a serine-threonine kinase, somewhat mysteriously present in the MHC of all jawed vertebrates examined down to cartilaginous fish. However, the organisation of the BF-BL region is significantly different from the MHC of typical mammals, particularly with the class III region on the outside of the class I and class II regions, and with the two class I (BF) genes flanking the TAP genes. Moreover, at one end of the BF-BL region there is an unexpected pair of genes encoding lectin-like membrane glycoproteins, likely a natural killer (NK) receptor and ligand which in mammals are found on a different chromosome in the natural killer complex (NKC) [Rogers et al., 2005; Rogers and Kaufman, 2008], and a single gene resembling the many genes in the BG region, now called BG1 [Goto et al., 2009; Salomonsen et al., 2014].

Compared to mammalian MHCs, the BF-BL region is notably compact and overall there has been little evidence for ongoing recombination, so that the BF-BL region exists as relatively stable haplotypes with potential co-evolution between the closely linked polymorphic genes. Unlike typical mammals, the peptide-loading genes (TAP1, TAP2, tapasin, DMA, DMB1, and DMB2) are polymorphic and, as an example, the peptide-translocation specificity of the TAP protein allele from a particular BF-BL haplotype correlates with the peptide-binding specificity of the dominantly expressed class I molecule encoded by the BF2 gene. A few examples of apparent recombination within the BF-BL region have been described, including one between B15 and B12 haplotypes giving rise to the famous B19 haplotype. In fact, out of 242 haplotypes recently identified by sequence (PCR-NGS) typing of commercial, fancy, and African chickens [Tregaskes et al., this volume], 53% could be related by potential recombination between the BF and BLB genes, suggesting that even low levels of recombination could lead to new combinations given sufficient time. Thus far, deletions have only been identified due to short direct repeats in the BF1 gene, leading to disabling of the promoter or to loss of the gene locus.

The stability of the BF-BL region may also depend on the presence of many pairs of sequence-related genes in opposite transcriptional orientation: BNK and Blc, BLB1 and BLB2, BF1 and BF2, and TAP1 and TAP2 [Kaufman et al., 1999b]. Homologous recombination between the members of such pairs should lead to inversion rather than deletion or unequal crossing-over, thus preserving the essential genes in a putative “minimal essential MHC.” The presence of BLB1-like sequences in the BLB2 locus adjacent to the BRD2 gene might be a result of such inversions, although gene conversion or some other form of segmental exchange might also be responsible [Afrache et al., 2020]. The orientations of TAP and tapasin genes in closely related galliform birds are consistent with historic inversions [He et al., 2021].

In contrast to the several different kinds of genes in the BF region that exist in relatively stable haplotypes, the BG region is relatively homogeneous but very dynamic. Almost all the genes in the BG region have similar structures, with exons encoding a signal peptide, a single extracellular immunoglobulin-like (Ig) domain, a hydrophobic transmembrane (TM) region, and a long cytoplasmic tail composed of heptad repeats [Miller et al., 1991; Salomonsen et al., 2014]. BG molecules are dimers, so the two chains interact through the Ig and TM regions, and the cytoplasmic tails wind around each other in a coiled coil. All these genes are in the same transcriptional orientation, which is well-suited for unequal crossing-over and deletion, leading to CNV which has been observed by sequence comparisons and by fibre fluorescent in situ hybridisation (FISH) [Salomonsen et al., 2014].

Further evidence for the dynamic recombinational landscape of the BG region comes from sequence comparisons [Salomonsen et al., 2014]. The BG genes in the BF region are expressed either in haemopoietic cells (such as erythrocytes, lymphocytes, and myeloid cells) or in non-haemopoietic cells (generally epithelial cells), and this expression correlates with 2 kinds of promoter and 5′UTR sequences. The exons encoding the cytoplasmic tail and the 3′UTR also fall into 2 classes. Overall, the BG genes represent the 4 combinations of the 2 classes of promoter/5′UTR and the 2 classes of cytoplasmic tail/3′UTR, with the exons for the signal peptide and the Ig-like region not following any particular pattern, as through recombination in the middle of the genes creates chimeric genes.

This kind of recombination leads to a conundrum. All the exons of BG genes are polymorphic, but determining whether the variation is selected has depended on being able to identify alleles, which is difficult in an expanding and contracting multigene family. A comparison of BG transcripts from different BG haplotypes that are expressed in the same cell type (so-called “functional alleles”) showed that selection is apparent only in the cytoplasmic tails, despite the fact that the B locus was originally identified by serological reactions with the extracellular Ig-like domain [Chattaway et al., 2016; Chen et al., 2018]. The conundrum is that recombination will
lead to a particular cytoplasmic tail being switched to expression in different cell types, and the rationale for the same sequence functioning in different cell types is unclear. However, the function of BG genes is currently a mystery, with the first clue being sequence similarities of the extracellular Ig-like domain to mammalian butyrophilins, which are thought primarily to be involved in negative regulation of αβ T cells and with localisation and function of γδ T cells [Henry et al., 1999; Rhodes et al., 2016; Herrmann and Karunakaran, 2022]. However, there is evidence that the cytoplasmic tail is important, forming coiled-coils similar to tropomyosin and affecting actin-myosin interactions [Bikle et al., 1996]. Moreover, there is some evidence that suggests that the cytoplasmic tail is important for viral resistance [Goto et al., 2009]. Perhaps the apparent conundrum will be resolved when the mysteries of BG function are solved.

This report summarises evidence for the divergent properties of 2 closely linked regions of the B locus, the relatively stable BF-BL region and the highly dynamic BG region. In between is the TRIM region, whose properties have not been explored. On the centromeric side of the BG region are various genes which, along with the NOR, olfactory genes, and scavenger receptor genes, largely remain to be analysed [Delany et al., 2009; Miller et al., 2014]. However, evidence for the dynamic nature of the Rfp-Y region has recently been presented [Goto et al., 2022]. Future studies may reveal how all these regions are organised and maintained, and may give clues to how they are reorganised in the transition to the different arrangements in other species.

Acknowledgments
The author thanks Dr. Jacqueline Smith, Dr. Magda Migalska, and Ms. Lan Huynh for critical reading.

Conflict of Interest Statement
The author declares no conflicts of interest.

Funding Sources
The author thanks the Wellcome Trust (Investigator Award 110106/A/15/Z) for support.

600 Alleles and 240 Haplotypes Identified for the Chicken BF-BL Region
(Prepared by C.A. Tregaskes, R.J. Martin, L. Huynh, N. Rocos, and J. Kaufman)

The major histocompatibility complex (MHC) of mammals is a large and complex region, with hundreds of genes and much recombination, and encodes a few highly polymorphic classical class I and class II molecules that have central roles in immune responses [Kaufman, 2016]. The functional equivalent of the mammalian MHC in chickens is the BF-BL region, which is remarkably simple and compact with few genes, most of which are critical to the function of classical MHC molecules, so that this region was originally dubbed a “minimal essential MHC.” Moreover, recombination within the BF-BL region was considered to be rare, so that this region could exist as relatively stable haplotypes, with co-evolution between these closely linked genes leading to functional consequences [Kaufman et al., 1999b; Kaufman, 2018; Tregaskes and Kaufman, 2021]. However, most of these ideas arose from the analyses of a few “standard haplotypes” dating back to the original descriptions by Briles and co-workers [Miller et al., 2004; Miller and Taylor, 2016; Afrache et al., 2020].

We set out to understand more about the diversity of MHC alleles and haplotypes in different chicken populations, starting with reference strand-mediated conformational analyses (RSCA) followed by cloning and sequencing [Potts et al., 2019]. As the need for higher throughput became clear, we developed a polymerase chain reaction-next generation sequencing (PCR-NGS) system to type the classical class II B genes BLB1 and BLB2, and the classical class I genes BF1 and BF2. Taking advantage of the compact nature of chicken MHC genes, we amplified exon 2 through the intron to the end of exon 3 (roughly 750 nucleotides) from genomic DNA and used the Illumina MiSeq to paired-end sequence both exons from each gene (Fig. 36). Coupled with DNA isolation using relatively cheap reagents (which worked for most samples) and a double-barcoding system (12 pairs of barcoded primers each for BF and BL, and 96 barcoded Illumina adaptors), we were able to analyse up to 1,152 samples in a run. We developed a bespoke bioinformatics pipeline that automatically generated sequences for all the alleles present, compared them to known alleles and then assembled them into known haplotypes, leaving the unknown sequences to be analysed by inspection [Martin, 2021; C.A. Tregaskes, R.J. Martin, and J. Kaufman, unpublished].

The initial samples included DNA, blood cells or tissues primarily from experimental lines, red jungle fowl, commercial flocks, fancy birds, and African chickens provided by many collaborators. Altogether, 22 different MiSeq runs were performed covering roughly 20,000 samples. For some populations, we used the microsatellite LEI0258 [Fulton et al., 2006] and BF2-specific PCRs
to confirm and extend the assignments [Bertzbach et al., 2022; L. Huynh, C.A. Tregaskes, and J. Kaufman, unpublished]. For some samples of blood cells, flow cytometry was performed to determine the expression level of the class I (BF) molecules on erythrocytes that is known to correlate inversely with peptide repertoire [Chappell et al., 2015], which in some cases was determined by immunopeptidomics. Almost all the experimental work is now complete, except for some PacBio sequencing that has become necessary to assign BLB sequences to the appropriate loci [N. Rocos, C.A. Tregaskes, and J. Kaufman, unpublished]. We identified roughly 600 alleles and found over 240 haplotypes, but there is much analysis to complete, so only an initial overview of some preliminary results will be summarised in this report, with some details expected to change as the analyses are refined.

For understanding alleles and haplotypes, we began by assembling the data known from the scientific literature as well as from nucleotide databases (such as NCBI/GenBank) [Afrache et al., 2020]. From the literature, 16 complete BF-BL haplotypes were known from sequencing cosmids, bacterial artificial chromosomes (BACs), and long-range PCR products. Other haplotypes could be assembled from complete or partial (exon 2-exon 3 for BF, exon 2 and exon 2-exon 3 for BLB) genomic sequences, as well as complete or partial cDNA sequences. However, many gene and cDNA sequences in the databases had to be ignored, even if published, since they were deposited from single studies and differed by only 1 or 2 nucleotides, with the associated papers revealing that only one amplification had been carried out, so most of those sequences were not separate alleles but the result of nucleotide misincorporation. Altogether, 17 standard haplotypes seemed secure, and an additional 16 haplotypes were suggested [Afrache et al., 2020]. Even these data must be treated with some caution; for example, the B6 and B15 haplotypes reported in the largest haplotype-sequencing project [Hosomichi et al., 2008] have not yet been found in any sample examined by PCR-NGS.

It was very easy to assign the 339 class I sequences found by PCR-NGS (after 22 MiSeq runs) to the BF1 and BF2 loci [Martin, 2021] since they were found almost exclusively in different clades by neighbour-joining, maximum likelihood and minimum evolution tree building algorithms (Fig. 37). The exceptions include 9 alleles related to BF1*0201 and BF1*0901 which cluster together, as well as a couple of other sequences present in other clades, all in the BF2 part of the tree. The BF2 locus was more polymorphic, with 247 BF2 alleles compared to 92 BF1 alleles. Despite largely being in one clade, BF1 alleles had much sequence diversity, with deep branches in the BF1 clade tree. However, most of this variation was not obviously in the peptide-binding site, with 74% of the BF1 sequences having His9 and Asp24, which may mean a wide peptide repertoire (as recently found from the structure of BF1*1901). Moreover, 89% of the BF1 sequences had an identical or near-identical sequence in the region of the C1/C2 epitope on the α1 helix, consistent with the suggested role of BF1 molecules as natural killer (NK) ligands [Ewald and Livant, 2004; Kim T et al., 2018].

In contrast to class I loci, it became very difficult to assign all 259 BLB sequences to the BLB1 and BLB2 loci.
[Martin, 2021], largely due to finding haplotypes with 2 new class II B sequences, both of which were most closely related to known BLB1 sequences. The known BLB1 and BLB2 sequences (as well as many identified by PCR-NGS) were mixed in phylogenetic trees (Fig. 37). Ongoing experiments using “between gene” primers and PacBio sequencing to assign sequences to the BLB1 locus adjacent to Blec and to the BLB2 locus adjacent to BRD2 have resolved the ambiguities in 26 of 47 unclear BLB haplotypes [N. Roccos, C.A. Tregaskes, and J. Kaufman, unpublished]. These experiments have revealed the same sequence in one locus in one haplotype and the other locus in a second haplotype, as well as one haplotype with the same sequence in both loci. It seems likely that most haplotypes will have to be checked to ensure that new sequences most closely related to known BLB2 sequences are actually located in the BLB2 (that is BRD2-adjacent) locus. Given the facts that the BLB1 and BLB2 genes are in opposite transcriptional orientation and that most of the gene sequences are nearly identical, one possibility is that recombination between homologous sequences in the BLB1 and BLB2 loci leads to inversion; in the PacBio run done thus far, there has been no convincing evidence for such inversions.

The optimal choice of nomenclature for BF and BLB alleles continues to be unclear. The old accepted nomenclature was based on haplotypes, so the same sequence in two haplotypes would have different names [Miller et al., 2004]. Based on the nomenclature system originally described for human MHC alleles and widely used for other species [Ballingall et al., 2018; Afrache et al., 2020; Robinson et al., 2020], the gene designation would be separated from the allele designations by a star or asterisk, with distantly related alleles of a single locus differing in the first two numbers (e.g., BLB1*02 vs. BLB1*04), and with closely related alleles of a single locus having the same first two numbers and differing in the next two numbers after a colon (e.g., BLB1*02:01 and BLB2*02:02). Haplotypes would then be constructed by strings of alleles (e.g., BLB1*02:01-BLB2*02:02-BF1*02:04-BF2*02:05 or in short 2-2:02-2:04-2:05). This elegant solution ran into trouble from the criteria for close relationship, in that the number of sequence differences within clade of closely related sequences could exceed the number of sequence differences between 2 sequences from different clades. Moreover, the same BLB sequence has now been found experimentally in both the BLB1 and BLB2 loci, so how should it be named? At the moment, designations

Fig. 37. The chicken classical class I sequences mostly separate into 2 large clades (a), while classical class II B sequences are all mixed together in phylogenetic trees (b). BLB* indicates class II B sequences that could not easily be assigned to the BLB1 or BLB2 loci in the haplotypes examined. Figure from Martin [2021], derived from data available in 2020.
for many sequences are simply ad hoc, as we struggle to
develop a consistent approach.

Of the “standard haplotypes,” 7 were exhaustively an-
alysed over some years in our lab, 5 of which were includ-
ed among the 14 subsequently analysed by another lab in
a single sequencing paper [Jacob et al., 2000; Wallny et al.,
2006; Shaw et al., 2007; Hosomichi et al., 2008]. Of these
16 haplotypes, all had different BF2 alleles except for two
B15 haplotypes which differed in BF1. The original B15
haplotype described (and almost all subsequently) had no
expressed BF1 allele (like the B14 haplotype), but the B15
haplotype from a chicken line in Japan had a BF1 allele
present. Among the 242 haplotypes identified by PCR-
NGS (although there are a few from published data that
we have not found), 27 (11%) have no BF1 allele amplified
by the primers used. Originally Southern blots suggested
an insertion in the BF1 loci that was not expressed [Wall-
y et al., 2006; Shaw et al., 2007], but the latest experi-
ments with primers outside the gene have amplified this
region in those haplotypes, and identified a deletion of the
whole BF1 gene which is the result of 2 short direct (but
imperfect) repeats [N. Rocos, F. J. Coulter, and J. Kaufman,
unpublished].

Of the “standard haplotypes,” B19 was identified as a
recombinant of B12 and B15 haplotypes, and 3 haplotypes
(B5, B8, and B11) were also found to be recombinants, al-
though by “gene conversion” of long stretches of DNA
[Wallny et al., 2006; Shaw et al., 2007; Hosomichi et al.,
2008]. Among the 242 haplotypes, 128 (53%) could have
arisen by recombination between BLB2 and BF1 (some
with subsequent mutation to produce closely related al-
leles); 22 BLB haplotypes are in combination with 96 BF
haplotypes (with closely related BF alleles combined, since
they might have arisen from mutation subsequent to re-
combination). The most extreme is the BLB haplotype
5-5, which is associated with 25 different BF haplotypes
(with closely related BF alleles combined). There is also
apparent recombination between BF1 and BF2, with 15
different BLB1-BLB2-BF1 haplotypes in association with
37 BF2 alleles. As an example, 4-8-4 is found with
BF2*24:01, 53:01, and 4 closely-related 43-type alleles.

The first analyses by RSCA were performed with high
level (elite, great-grandfather) lines of commercial breed-
ers, and we were shocked at the low diversity of these pop-
ulations; some had only a single BF-BL haplotype. In or-
ter to better understand the commercial chickens that are
actually in the field, we obtained farm-level samples from
our collaborators, examining 6 broiler lines, 15 egg-layer
lines, and 1 dual purpose line. The take-home message is
that there are typically very few haplotypes, mostly 4 or 5
haplotypes above 1% genotype frequency, usually with 1
haplotype by far the majority. In particular, a haplotype
not described before (made up of alleles previously de-
scribed, provisionally called B31) is present in 33–64%
genotypes of the 6 farm-level broiler flocks. Similarly, a
previously undescribed haplotype provisionally called
B9:02 dominates nearly all the brown egg layer flocks. If
these numbers are representative, then there are billions
of chickens in the field that are MHC homozygotes.

How do these commercial chickens survive with such
low MHC diversity? Part of the answer may be that most of
the high frequency haplotypes are those with BF2 al-
leles that have low cell surface expression and promiscu-
ous peptide binding (for those with known peptide mo-
tifs). Such so-called “promiscuous haplotypes” are known
to protect against a variety of economically important in-
fecious diseases in chickens and have been suggested to
act as generalists, in contrast to “fastidious haplotypes”
which may act as specialists [Chappel et al., 2015;
Kaufman, 2018; Tregaskes and Kaufman, 2021]. A few
haplotypes with high-expressing BF2 alleles are found in
some populations; these may function as specialists or
have some other useful attribute(s).

A wealth of information has already emerged from this
PCR-NGS typing, but there is much more to be learned by
finishing the detailed analysis of commercial, fancy,
and African chickens. Moreover, there are many chickens
worldwide that have not been examined by this kind of
analysis, particularly in South, Southeast, and East Asia.
As the typing methods benefit from longer reads that cov-
er more genes, many interesting attributes of the chicken
MHC are likely to be revealed.

Acknowledgments
We thank Dr. Jacqueline Smith and Dr. Magda Migalska for criti-
cal reading.

Conflict of Interest Statement
The authors declare no conflicts of interest.

Funding Sources
The authors thank the Wellcome Trust (Investigator Award
110106/A/15/Z to J.K.), the BBSRC (DTP PhD in Biological Sci-
ences at the University of Cambridge to R.J.M.), and the Univer-
sity of Edinburgh for support.
The Chicken Genomes and Their Diversities
(Prepared by J.-L. Han and O. Hanotte)

This review intends to provide a comprehensive summary of the status of chicken genome resources and what we have been learning from the analyses of genomic diversity of our commonest domestic poultry species.

Chicken Genome Assemblies

The chicken was the first bird and agricultural animal species to have its genome sequenced. The first reference genome of 6.6x genome coverage from a female red junglefowl, known as “RFJ #256” (inbred line, UDC001), was published in December 2004 [International Chicken Genome Sequencing Consortium, 2004]. It was obtained through Sanger sequencing and included around one billion nucleotides and 20,000–23,000 annotated genes. This genome assembly was further improved by adding 14x genome coverage next-generation sequencing (NGS) data and referred to as Galgal3 or Gallus_gallus-2.1 (GenBank RefSeq accession no. GCF_000002315.1 and GCF_000002315.2 released in November 2006) [Groenen et al., 2011]. It was followed by a de novo assembly of the “RFJ #256,” using all available Sanger and NGS reads available at the time and new Illumina short reads at 68.6x genome coverage (Galgal4, Gallus_gallus-4.0 with GCF_000002315.3 released in November 2011) [Ye et al., 2011; Schmid et al., 2015]. Then, another de novo assembly, based on third-generation sequencing data (PacBio RSII long reads at 50.6x genome coverage) of the “RFJ #256” was merged with all Galgal4 sequences and referred to as Gallus_gallus-5.0 (GCF_000002315.4 released in December 2015) [Warren et al., 2017]. This genome assembly was finally upgraded by adding PacBio RSII long reads to a sequencing depth of around 80x and by generating a high chromatin proximity map to help in the order and orientation of the assembled contigs and scaffolds. It is referred to as GRCg6a (GCF_000002315.6 released in March 2018). This assembly has a 1.07 Gb total genome size. It includes 34 chromosomes with 1,402 contigs assembled into 524 scaffolds. The N50 length for the contigs is 17.66 Mb, while the scaffold N50 is 20.79 Mb. The NCBI Gallus_gallus Annotation Release 104, including 17,477 protein-coding and 6,558 noncoding genes (released in May 2018), and the Ensembl release 94, including 16,878 coding and 7,166 noncoding genes (released in April 2018) (database version 106.6) complement the GRCg6a assembly.

Development and Application of SNP Arrays for the Assessment of Genomic Variation

Considering the details on gene expression microarrays in an early review [Gheyas and Burt, 2013], we focus on the development of genotyping arrays in this summary. The first Illumina 3K SNP array was developed based on the selection and validation of 3,072 out of the 2.8 million SNPs at one SNP each in 3,072 bins distributed evenly throughout the chicken genome. It was based on the genome sequence of WASHUC1 assembly (Galgal2 released in February 2004 at https://genome.ucsc.edu/cgi-bin/ hgGateway) with linkage information to account for the recombination rate of each chromosome [Aerts et al., 2007]. In addition, 34 SNPs in genes of interest were added [Muir et al., 2008a]. Then, a moderate density (60K) Illumina SNP BeadChip was developed using additional SNPs from the Illumina short reads of broiler and layer chicken populations aligned against Gallus_gallus-2.1.0 (released in May 2006), 454-read-based contigs of the “RFJ #256,” and the mitochondrial genome. It included 60,800 SNPs [Groenen et al., 2011]. A 600K Affymetrix Axiom high-density (HD) genotyping array was independently constructed based on 139 million SNPs identified by aligning the Illumina short reads of 243 chickens from 24 lines of commercial broilers, white-egg layers, and brown-egg layers as well as experimental inbred layers and unselected layer line against the Galgal4. They were scaled down to 1.8 million SNPs as a robust and tractable subset for HD array design, including those built in the Illumina 3K [Muir et al., 2008a] and 60K [Groenen et al., 2011] arrays, and SNPs among the 7 million SNPs identified from chicken populations of different lines [Rubin et al., 2010]. The 580,954 SNPs were validated by genotyping an additional 282 chickens, including trio samples from 3 types of commercial lines and traditional breeds [Kranis et al., 2013]. All segregating SNPs on this Affymetrix 600K array were evenly spaced across the genome, following the genetic map distances, instead of their physical distances, for both broiler and layer lines [Groenen et al., 2009], to account for the difference in recombination rates between macro- and microchromosomes [Rodionov, 1996; Groenen et al., 2000, 2009; International Chicken Genome Sequencing Consortium, 2004; Megens et al., 2009; Liu Z et al., 2021].

These SNP arrays have significantly facilitated genome-wide association studies (GWAS), paving the way for genomic selection, identification of selection signatures, fine mapping of QTLs, and detection of copy number variations (CNVs). Although the first genome-wide scan based on the WGS data at low genome coverage of 4...
birds failed to identify selective sweeps for adaptive alleles following chicken domestication [Wong et al., 2004], several selective sweeps linked with the genes associated with growth, appetite, and metabolic regulation were subsequently detected using enhanced genome-wide SNPs of broilers, while one of the most striking selective sweeps at TSHR locus has been considered as a “domestication” signature [Rubin et al., 2010; Elferink et al., 2012].

These SNP arrays were also widely used for the characterization of population genomic diversity and genetic structure of many wild, commercial, experimental, indigenous, and fancy chicken breeds/populations [Muir et al., 2008b; Elferink et al., 2012; Malomane et al., 2019; Zhang et al., 2020; Cendon et al., 2021], even though SNP ascertainment bias might be of some concern [Malomane et al., 2018; Geibel et al., 2021]. Also, population stratification can lead to false associations in GWAS. Henceforth, a careful examination of population genetic structure is warranted [Kranis et al., 2013].

The Illumina 3K array was used for genotyping 2,551 informative SNPs in 2,580 chickens from commercial male and female broilers, white- and brown-egg layer pure lines, experimental, and traditional breeds. Analysis of commercial lines showed a loss of 50% or more of the genetic variability present in ancestral breeds, a consequence of founding effects in making these lines. It raises questions about the suitability of these lines to respond to future needs (consumers, societal, and producers’ needs), as well as about their genetic repertoire for resistance to infectious disease challenges [Muir et al., 2008b]. The Illumina 60K array was used to genotype 51,076 autosomal SNPs in 67 populations, including Ceylon Junglefowl (Gallus lafayetti), red junglefowl (G. g. gallus and G. g. spadiceus), eight broiler sire lines, five broiler dam lines, 11 white-egg layer lines, 11 brown-egg layer lines, 19 traditional Dutch breeds, and 10 Chinese indigenous breeds [Elferink et al., 2012]. A phylogenetic tree rooted at the Ceylon Junglefowl separated the red junglefowls from all domestic chickens that were divided into 2 branches, one including brown-egg layers, broilers, and Chinese breeds, and the other white-egg layers and Dutch breeds. Among the broilers, the sire and dam lines were separated from each other, supporting their distinct origins.

The Affymetrix 600K array was evaluated for its utility in inferring population stratification with samples of known history through principal component analysis (PCA). It was found that birds from the same line/breed clustered together. The broiler lines were closer to the brown-egg layers than to the white-egg layers. In contrast, the two white-egg layer lineages were separated from each other, supporting different origins of the two lines [Kranis et al., 2013].

Also, it was observed that the phylogeny reconstructions using genome-wide SNP arrays mirrored those based on microsatellite data [Eding et al., 2002] and known geographic origins and breeding history of the studied populations [Elferink et al., 2012]. In particular, the broilers and brown-egg layers shared their ancestry. They were initially developed by crossing European and Asian breeds, while the white-egg layers originated from the single-combed White Leghorn of European origin [Crawford, 1990; Muir et al., 2008b; Elferink et al., 2012].

The Affymetrix 600K array had been further applied to several large-scale population genomic studies. For instance, genotyping of 1,200 chickens, ranging from 41 to 469 samples of five Chinese breeds of Beijing-You, Hongshan, Shouguang, Taihe Silkie, and Tibetan, along with White Leghorn originating from Italy, Houdan chicken from France, and Rhode Island Red from USA, identified a higher genomic diversity in most Chinese breeds compared to exotic ones, while all breeds carried some unique polymorphisms allowing successful assignment of all samples back to the breeds of origins. Local Tibetan chickens with high-altitude adaptability had a high level of genetic admixture from White Leghorn [Nie et al., 2019]. The addition of 69 samples (including five Chinese game breeds, Cornish chicken, and red junglefowls) to the 1,200 chicken dataset strengthened the distinct genomic variability present within each breed, with the highest genomic uniqueness observed in White Leghorn, Houdan, and Rhode Island Red. Cornish breed, known as “Indian Game” carrying the genetic footprint of Malay and other Oriental chicken blood, showed a close genomic relationship with five Chinese game breeds, all sharing a highly admixed genomic background of likely a single origin. As expected, the red junglefowls showed the highest genomic diversity, followed by Cornish, four of the five Chinese game breeds, other Chinese local breeds, French Houdan, and commercial breeds, based on linkage disequilibrium decay (LD), effective population size, and runs of homozygosity (ROH) estimates [Zhang et al., 2020]. A regional study on seven indigenous breeds in the Jiangxi province of China indicated a similarly higher genomic diversity but smaller ROH in most Chinese indigenous breeds relative to European and commercial breeds. Recent selection for meat and egg production have resulted in reduced genomic diversity but increased ROH in improved Chinese local breeds [Chen L et al., 2019]. The genotyping of eight Italian local chicken
breeds showed unique genomic diversity in most breeds, calling for breed-specific conservation strategies [Cendron et al., 2021]. Also, genomic variation in three Chinese indigenous breeds (Baier Yellow, Beijing-You, and Langshan) maintained by ongoing ex-situ conservation programs were evaluated at three different generations (generations 7, 10, and 15 for Baier Yellow; 7, 10, and 13 for Beijing-You, and 10, 12, and 15 for Langshan). These conserved flocks were managed by keeping 30 males and 300 females per generation and implementing a random mating within families with one son selected from one sire family and one daughter from one dam family. There was no differentiation in population genetic structure within the breeds over the three generations [Zhang et al., 2020]. The genomic diversity of a live Norwegian poultry Genebank indicated that inbreeding level was high in all lines, while the relatively more inbred white-egg layers were differentiated from the brown-egg layers that contributed more to total genetic diversity. Though distinct from other commercial populations, the newly developed Norwegian commercial lines were closely related to them. These Norwegian Genebank lines were therefore believed to be of conservation value at national and international levels [Brekke et al., 2020].

Several customized SNP arrays have also been developed in recent years. For example, a 42K Illumina SNP array was developed by private funds from the EW Group (Visbeck, Germany) [Fulton, 2012]. Also, a proprietary Affymetrix 50K SNP array was built based on a subset of SNPs extracted from the Affymetrix 600K array to capture specific genetic diversity in highly selected and pedigree populations, along with additional SNPs identified in previous studies [Wolc et al., 2020]. It is expected that the genotypes of this 50K SNPs may be imputed into the 600K SNPs using the parent, grandparent, and great-grandparent 600K SNP data [Herry et al., 2018; Psifidi et al., 2021]. These two arrays were specifically designed for QTL mapping and genomic selection in commercial layer and broiler lines, and they were never made publicly available.

Another Affymetrix 55K genotyping array (IASCHICK) incorporating specific SNPs identified by aligning the WGS reads from several Chinese indigenous chicken breeds, which were poorly represented in early arrays, against the GalgaL4 [Liu et al., 2019], and an Illumina 50K BeadChip (PhenoixChip-I) containing unique SNPs from several commercial layer lines important to the Chinese market, based on the 1,846,003 SNPs screened by aligning their WGS reads against the GRCg6a [Liu et al., 2021], were recently developed. Both IASCHICK and PhenoixChip-I arrays also include many candidate SNPs associated with important economic traits in chicken [Liu et al., 2019; Liu Z et al.,2021]. The IASCHICK array proved to be effective in detecting within-population genetic diversity in nine indigenous and improved Chinese breeds and three commercial layer and broiler lines, with calling rates ranging from 97.0% to 98.7% and polymorphic SNPs ranging from 76.7% to 88.0% across the breeds/lines. But the genotyping results from the PhenoixChip-I array showed some level of ascertainment bias. Still, the genetic variation identified by both IASCHICK and PhenoixChip-I arrays had a sufficiently high resolution to support the assignment of all samples back to their expected breeds/lines of origins, validating the power of 50K genome-wide SNP arrays for studying population genetic diversity and structure.

SNP genotyping arrays have also been used to analyze copy number variations (CNVs) of genomic structural variants (SVs) in the form of segmental insertion, deletion, and duplication greater than 50 bp. They are a significant source of genomic diversity underlying phenotypic variation. For instance, a duplicated sequence close to the first intron of SOX5 has been linked to the pea-comb phenotype [Wright et al., 2009], while an inverted duplication including EDN3 is associated with dermal hyperpigmentation in chicken [Dorshorst et al., 2011]. In addition, partial duplication of PRLR has been related to late feathering [Elferink et al., 2008], whereas a CNV linked with the HOXB7 and HOXB8 genomic region has been associated with the beard phenotype [Yang KX et al., 2020] in chicken.

Using the Illumina 60K array, up to 818 CNVs were identified in 184 White Leghorns and 233 brown-egg dwarf layers, of which 315 were unique. They aggregated into 209 copy number variation regions (CNVRs) on 27 autosomal chromosomes. These CNVRs were distributed proportionately to the chromosomal length, e.g., 14.7 CNVRs on macrochromosomes but 3.65 CNVRs on microchromosomes. The CNVRs shared by the two breeds suggest their relatively ancient origins, while some CNVRs were breed specific [Jia et al., 2013]. Also, 137 CNVRs were reported in 1,310 Beijing-You chickens [Zhou W et al., 2014]. Genotypes of 475 birds after the 11th generation of divergent selection for abdominal fat content in the same grandsire line of the Arbor Acres broilers revealed 438 and 291 CNVs, of which 271 (176 loss, 68 gain, and 27 loss + gain) and 188 (143 loss, 25 gain, and 20 loss + gain) CNVRs were low- and high-fat line specific, respectively. Differences in genetic drift and selection were thought to have contributed to the variation
in the numbers and distribution patterns of both CNVs and CNVRs between the two lines [Zhang H et al., 2014]. Genotypes of 554 chickens from an F₂ full-sib population, a cross between Xinghua and White Recessive Rock chickens, identified 1,875 CNVs distributed in 209 CNVRs, of which 109 were novel [Rao et al., 2016]. A follow-up analysis of the CNVs in this F₂ population revealed a polymorphic CNV overlapping with SOX6, with the number of CNVs positively associated with the expression of SOX6, which is associated with skeletal muscle development [Lin et al., 2018].

Only large CNVRs were identified in these studies using the Illumina 60K array due to its low marker density and non-uniform marker distribution. To improve the efficiency and reliability of SNV detection, the Affymetrix 600K array has been used in several studies. For example, Chinese indigenous chicken and exotic commercial breeds carried 5.1 and 3.3 CNVs per bird, respectively; both values were much higher than that reported earlier [Jia et al., 2013]. After the calibration of the CNVs between the Galg4 and Galgal3 for comparison with those detected in early studies, 153 CNVs were found to be novel [see Table 2 in Yi et al., 2015]. Genotyping of 48 deformed-beak and 48 normal Beijing-You chickens identified LRIG2 as a candidate for beak deformity [Bai et al., 2018]. Four SNP chips (Illumina 42K, Affymetrix 600K, and two customized Affymetrix 50K chips) were successfully used to genotype and identify CNVs in 18,719 chickens from four pure lines and one commercial cross. Here, the genome landscape of CNVs was determined not only by the number of samples and genetic background of the lines, but also by the SNP density on the arrays: the higher the density of the arrays, the more the CNVs per sample [Drobik-Czwarno et al., 2018]. As many as 1,003 CNVs in 564 CNVRs were identified in 94 chickens of six local Italian breeds, contributing to their distinctiveness [Strillacci et al., 2017]. Also, 1,924 CNVs in 1,216 CNVRs were observed in 265 local Mexican chickens [Gorla et al., 2017]. These CNVRs were mapped on 28 autosomes in the Galgal4 [Gorla et al., 2017; Strillacci et al., 2017]. Using 23,214 CNVs and 5,042 CNVRs in 1,238 chickens of a Brazilian male broiler line, originating from the crossing of White Plymouth Rock and White Cornish breeds, CNV-based GWAS revealed potential candidate genes such as KCNJ11, MyoD1, and SOX6 nearby several CNVs associated with growth traits [Fernandes et al., 2021].

In most of these studies, the same parameters implemented in the PENNCNV software, based on Hidden Markov Model (HMM), were applied to call CNVs from genotyping data, including signal intensity (log R ratio, LRR), allelic intensity (B allele frequency, BAF), and marker distance and population frequency of allele B [Wang et al., 2007]. The HMM with default parameters but a cutoff for the standard deviation of LRR < 0.30 were considered [Jia et al., 2012; Zhang H et al., 2014; Zhou W et al., 2014; Rao et al., 2016]. The CNVs containing at least three consecutive SNPs were specifically chosen [Zhang H et al., 2014]. Aggregating overlapping CNVs determined CNVRs among individuals [Redon et al., 2006]. A common pattern was that most CNVs and CNVRs ranging from 68.8% to 82.9% were singletons and thus segregated among individuals within a breed/population [Yi et al., 2015; Gorla et al., 2017; Strillacci et al., 2017].

To reduce the genotyping cost, low-density SNP chips (e.g., 10K SNPs) have been used in populations previously studied using high-density SNP arrays followed by SNP imputation [Herry et al., 2018; Huang et al., 2018]. Other low-cost methods such as reduced representation sequencing based on restriction enzyme cleavage have also been attempted for SNP discovery, validation, and characterization in chicken populations. Among them, restriction-site-associated DNA sequencing (RAD-seq) [Zhai et al., 2015; Yang Z et al., 2020], specific locus amplified fragment sequencing (SLAF-seq) [Jin et al., 2015; Wang W et al., 2016, 2019; Li F et al., 2018, 2021], genotyping by genome reducing and sequencing (GGRS) [Liao et al., 2015, 2016; Zhao QB et al., 2018], and genotyping by sequencing (GBS) [Pértille et al., 2016; Zhang M et al., 2018; Zhang Y et al., 2021; Habimana et al., 2021] proved to be effective in identifying and genotyping some novel SNPs in chickens. However, these methods have not yet been widely applied.

**De Novo Assembly of the New Chicken Reference Genome**

Two de novo haplotype assemblies using a trio of samples, including an F₁ hybrid female (bGalGal1) from a broiler hen (bGalGal2) mated with a White Leghorn cock (bGalGal3), were recently released: bGalGal1.mat.broiler.GRCg7b (GenBank and RefSeq assembly accession nos. GCA_016699485.1 and GCF_016699485.2, released in January 2021) and bGalGal1.pat.whiteleghornlayer.GRCg7w (GenBank and RefSeq assembly accession nos. GCA_016700215.2 and GCF_016700215.2, released in October 2021). They were assembled using PacBio Sequel I CLR, Illumina NovaSeq, Arima Genomics HiC, and Bi-onano Genomics DLS reads at 102.01x genome coverage. bGalGal1.mat.broiler.GRCg7b has a total genome length of 1.05 Gb in 677 contigs assembled into 214 scaffolds with N50 contigs of 18.83 Mb and N50 scaffolds of 90.86
De Novo Assemblies of Indigenous and Commercial Chicken Breeds

Following the reduction in sequencing cost and the advent of new sequencing technologies, de novo genome sequences of chicken breeds of particular interest are being generated to identify the genetic control of their unique phenotypes. These are warranted for comparative genomic analyses of SNPs, insertions, and deletions (INDELS), structural variations (SVs), and coding and non-coding transcriptomes.

The first de novo draft genome of an indigenous chicken, the Korean Yeonsan Ogye, was released in December 2017 (Ogye1.0, GenBank Genome Accession no. GCA_002798355.1). This breed is characterized by having an entirely black external feature and internal organs. Following a hybrid de novo assembly pipeline, the genome was assembled into 8,241 pseudo-contigs and 1,906 scaffolds, which were further aligned and anchored to the GalGal4 chromosomes, based on high-depth Illumina short reads (GCA_024653045.1), Rhode Island Red (GCA_024652985.1), WhiteLeghorn (GCA_024652995.1), Cornish (GCA_024653035.1), Silky (GCA_024653025.1), and Tibetan chicken (GCA_025370635.1); PacBio Sequel II long reads (>87×), Illumina HiSeq short reads (>56×), and HiC data (>112×) for six genomes at chromosome level for Houdan chicken (GCA_024653045.1), Rhode Island Red (GCA_024652985.1), WhiteLeghorn (GCA_024652995.1), Cornish (GCA_024653035.1), Silky (GCA_024653025.1), and Tibetan chicken (GCA_025370635.1); PacBio Sequel II long reads (>53×) and Illumina HiSeq short reads (>45×) for three genomes at scaffold level for Asil (GCA_024686355.1), Naked Neck (GCA_024686465.1), and Thailand Gamefowl (GCA_024686285.1); Illumina HiSeq short reads (around 134×) for 10 genomes at scaffold level (GCA_024653045.1 - Silky, GCA_024679355.1 - Daweishan, GCA_024679375.1 - Liyang, GCA_024679395.1 - Tibetan, GCA_024679415.1 - White Plymouth Rock, GCA_024679765.1 - Rhode Island Red, GCA_024679905.1 - White Leghorn, GCA_024686295.1 - Chahua, GCA_024686315.1 - Langshan chicken, and GCA_024687005.1-Cornish chicken); and one at contig level based on PacBio Sequel II long reads and Illumina HiSeq short reads (GCA_024686275.1 - Fayoumi). The completeness of these 20 genomes ranged from 92.4% to 95.3% (BUSCO analysis), comparable to the GRCg6a (95.4%). The pan-genome analysis of these 20 genomes aligned to the GRCg6a identified 1,335 novel coding genes, among which many were housekeeping genes and genes involved in immune pathways. These immune-related genes had a 3-fold elevated substitution rate. There were also 3,011 new long noncoding RNAs absent in the GRCg6a [Li Y et al., 2022].

WGS-Based Population Genome Analyses

As mentioned above, the SNP arrays were designed for genotyping polymorphisms at hundreds or thousands of specific locations across the chicken genome. They have been used to assess the ancestry of particular breeds/populations/lines and identify signatures of selection and candidate genomic regions associated with complex and multifactorial phenotypes through GWAS, including...
production and adaptive traits. However, they have some limitations. These SNP arrays largely include SNPs common either across different breeds/populations from a relatively large geographic coverage or within specific breeds/lines of limited ancestry and/or under intensive selection. These SNP arrays will therefore perform poorly for genotyping rare genetic variants. It is well known that genetic diversity is not equally distributed across populations of different ancestries, and some of these rare variants may become common in genetically isolated populations. Together with the challenge of uneven spacing of SNPs across the genome, they contribute to SNP ascertainment bias, which leads to biased inferences (e.g., proportion of admixture) [Lachance and Tishkoff, 2013; McTavish and Hillis, 2015; Dokan et al., 2021]. There are also limitations associated with the application of the SNP arrays in detecting CNVs (e.g., chromosomal translocations and inversions).

Considering the constraints of SNP arrays and the reduction of DNA sequencing costs, NGS-based WGS approach is now the most popular tool to infer the evolutionary history of chicken genomes, to assess the genetic diversity and population genetic structure at regional, continental, and global scales, to identify the genetic controls of specific traits of interest, and to screen for the signatures of selection for productive traits and environmental adaptation [Rubin et al., 2010; Mugal et al., 2013; Wang MS et al., 2015, 2016a, 2016b, 2017, 2020, 2021; Guo X, 2016, 2022; Guo Y et al., 2016, 2021; Zhang et al., 2016, 2022; Li D et al., 2017, 2019; Boschiero et al., 2018; Derks et al., 2018; Lawal et al., 2018, 2020; Sohrabi et al., 2018; Almeida et al., 2019; Talebi et al., 2020; Tiley et al., 2020; Wang Q et al., 2020; Gheyas et al., 2021, 2022; Li YD et al., 2021; Liu Z et al., 2021; Mariadassou et al., 2021; Rostamzadeh Mahdabi et al., 2021; Tan et al., 2021; Yang Z et al., 2021; Asadollahpour Nanaei et al., 2022; Chen X et al., 2022; Gao et al., 2022; Geibel et al., 2022; Li Y et al., 2022; Liu J et al., 2022; Sun et al., 2022; Wang S et al., 2022; Xu et al., 2022; Yuan J et al., 2022; Yuan X et al., 2022; Zeng et al., 2022].

For example, a novel panel of 64 exonic SNPs screened from WGS data were applied to the genetic characterization of Italian local breeds and proved to be cost-effective for genotyping many samples to aid genetic traceability and breeding programs [Viale et al., 2017]. Also, to further explore the deficit in homozygous carriers for 77 haplotypes in four purebred white- and brown-egg lines and two crossbred lines observed using the Illumina 60K array, the WGS data of 250 white-egg layers were annotated. This analysis identified 4,219 putative deleterious variants, including 152 mutations relevant to embryonic lethality, at homozygous state. These deleterious variants present in genomic regions of low recombination rates have been subjected to purifying selection [Derks et al., 2018]. Many deleterious and stop-gain/loss SNPs were also observed in the WGS data of Brazilian meat and white-egg layer chickens [Boschiero et al., 2018].

Homozygous deleterious variants were also identified in the WGS data of red junglefowls and domestic chickens, with 2.95% more mutations in the chicken genome compared to its wild ancestor, which was interpreted as the result of the “cost of domestication.” Around 62.4% of these deleterious variants were in the heterozygote state and believed to be recessive [Wang MS et al., 2021]. Considering the possible occurrence of harmful mutations linked to favorable variants in positively selected sweeps and the positive relationship between recombination rate and purging efficiency, a genomic marker-assisted selection is highly recommended to minimize the frequency of undesirable functional mutations, while to sustainably improve both indigenous and commercial chicken genetic resources [Derks et al., 2018; Wang et al., 2021].

The WGS-based population landscape of SNPs in 10 Chinese indigenous breeds showed less diversity on the Z chromosome than autosomes that were likely under relatively strong selection pressures. Tibetan chickens were admixed with other breeds, while cockfighting Game chickens were closely related to red junglefowls. Strong signatures of selection were observed at genomic regions linked to the genes likely responsible for the rapid adaptation of Tibetan chickens to high altitudes [Li D et al., 2017, 2019]. The WGS data from 863 chickens and junglefowls from America, China, Indonesia, and Europe illustrated their general patterns of variability, ancestry, evolutionary relationship, and breeding history, and convincingly localized the main center of chicken domestication in Southeast Asia. The intensive genomic analyses of domestic chickens also detected several specific genetic admixture events with red junglefowls and other junglefowl species outside the proposed center of domestication [Wang MS et al., 2020]. Such admixture and/or introgression events have also been reported in another WGS study [Lawal et al., 2020]. Genomic footprints of admixture from indigenous large-sized Asian breeds and egg-laying Mediterranean breeds were identified in the gene pool of commercial meat and laying chickens, supporting their contributions to modern commercial meat and egg industries, respectively [Guo Y et al., 2021]. The first, and largest WGS dataset of African indigenous chickens, including 234 genomes from 24 Ethiopian chicken populations dis-
tributed in different types of climates and productive systems, was generated to call up to 15 million SNPs mapped to the GRCg6a reference. High-quality variants have been used to assess population genomic diversity and screen for genomic regions under selection for environmental adaptations. The identified candidate regulatory genes could be epigenetic machineries driving rapid adaptation of African chickens of a relatively recent origin from Asian counterparts [Gheyas et al., 2021, 2022]. The WGS analysis of eight indigenous breeds from Guangxi, China, revealed genomic diversity similar to that of red junglefowl but limited genetic differentiation and little genetic admixture from commercial chickens [Sun et al., 2022].

The WGS-based phylogeny of all four junglefowl species was reconstructed to achieve a highly confident topology of the genus, which is crucial in inferring the effects of interspecific introgression and genetic admixture on chicken domestication processes. Following an intensive evaluation of major topological inconsistencies caused by different genetic/genomic datasets and bioinformatic methods, the genetic contamination of some captive junglefowl flocks is verified, while the importance of large number of genetic variants in different genomic components of many reliable samples (e.g., collected from their original ranges in the wild) to phylogenetic inference is substantiated [Tiley et al., 2020; Wang et al., 2020; Mariadassou et al., 2021].

WGS data analysis have also provided new insights on the ROH in chicken genomes. Compared to the red junglefowls with the least number of ROH, white layers carried the largest number of ROH per bird, mostly shorter than 1 Mb, while broilers tended to have relatively more ROH >2 Mb. Both SNP-based Wright’s $F_{IS}$ and ROH-based $F_{ROH}$ metrics suggested higher inbreeding in the white layers compared to the broilers and red junglefowls [Talebi et al., 2020]. Candidate genes in the ROH islands may also be associated with QTLs responsible for productive traits [Talebi et al., 2020] and high-altitude adaptation [Yuan et al., 2022].

As expected, more CNVs and CNVRs were identified from the WGS data compared to the findings of SNP genotyping arrays. For instance, 12,955 CNVs in 5,467 CNVRs, accounting for 9.42% of the genome, were found in two Iranian indigenous and commercial chicken breeds, with 34% of these CNVRs overlapping with those identified in SNP-array-based analyses [Sohrabi et al., 2018]. Across 51 WGS datasets of Chinese indigenous breeds (Xinghua, Luxi Game, Beijing-You, and Silkie), commercial lines (Recessive White Rock and White Leghorn), and red junglefowls, 19,329 duplications and 98,736 deletions in 11,123 CNVRs, accounting for 7% of the total autosomal size, and overlapping with 2,636 protein-coding genes were identified [Chen X et al., 2022]. Like in all previous SNP-array-based studies, the vast majority of the CNVRs were singletons, with only 152 CNVRs common to all 51 birds. Around 600 CNVRs, including 90 protein-coding genes, were breed specific, suggesting their functional importance in driving chicken phenotypic and adaptive evolution [Chen X et al., 2022].

Two analyses clearly demonstrated the power of the WGS data in CNV identification and functional annotation. Last but not least, the possibility of identifying SVs, though challenging for calling accuracy, was demonstrated from the NGS-WGS of commercial chicken genomes, among which deletions were believed to be more accurately detected compared with duplications, inversions, and translocation breakpoints [Geibel et al., 2022].

Conflict of Interest Statement
The authors declare that they have no conflict of interest.

Funding Sources
This study was funded by the Bill and Melinda Gates Foundation (BMGF) and with UK aid from the UK Government’s Department for International Development (Grant Agreement OPP1127286) under the auspices of the Centre for Tropical Livestock Genetics and Health (CTLGH), established jointly by the University of Edinburgh, Scotland’s Rural College (SRUC), and the International Livestock Research Institute (ILRI). The Chinese Government’s contribution to CAAS-ILRI Joint Laboratory on Livestock and Forage Genetic Resources in Beijing (2018-GHZ-01) is appreciated. The findings and conclusions contained within this article are those of the authors and do not necessarily reflect positions or policies of the BMGF nor the UK and Chinese Governments.

Chicken Genomic Diversity Consortium: Large-Scale Genomics to Unravel the Origins and Adaptations of Chickens

Consortium Description
On October 25–26, 2019, a satellite meeting devoted to the preparation of a Chicken Genome Diversity Consor-
tium was organised after the 11th European Symposium of Poultry Genetics in Prague. Researchers involved in chicken genomics from Europe, Africa, and China, discussed the objectives of such a consortium with some presenting their data. However, the technical aspects of how to share and jointly analyse the data were not finalized, nor was the funding model for the cost of data storage and computation. In 2021, an opportunity arose with the call for projects of the SuperMUC computing cluster of the Leibniz–Rechenzentrum in Germany. A new consortium of scientists re-launched the discussion to establish a project with the aim to explore how the high-throughput genomics age can be harnessed to answer evolutionary questions surrounding the chicken. The FARMGENOMIC project (23826) was accepted for funding in autumn 2021, gathering around 20 members from 10 institutions in Europe, North America, and Africa. This newly formed Chicken Genomic Diversity Consortium brings together members from a variety of disciplines, including genomics, palaeogenetics, animal breeding, immunology, organismal biology, evolutionary biology, and archaeology. Central to the consortium are the concepts of inclusivity and openness – all data are to be made available to all members of the consortium, and later distributed to the wider community, and collaborations between groups are fostered and actively encouraged. It is hoped this state-of-the-art resource, curated in-house by bioinformaticians, will enable the community to answer previously intractable questions in chicken evolution.

Dataset Description

At the core of the consortium is a substantial genomic dataset of chickens and junglefowl. At the time of writing (September 2022), the dataset comprises 4,392 chicken and junglefowl genome sequences, of which 2,307 were derived from public databases, and the remainder provided by consortium members. In addition to domesticated chickens and red junglefowl (comprising all 5 subspecies: G. g. gallus, G. g. bankiva, G. g. jabouillei, G. g. murghi, G. g. spadiceus; total n = 291), we also included members of the congeneric Gallus species Gallus varius (n = 21), Gallus lafayettii (n = 12), and Gallus sonneratii (n = 15). Among the domesticated chickens, a wide array of geographical locations are represented (Africa, n = 1,047; East Asia, n = 856; South East Asia, n = 72; South Asia, n = 137; Middle East, n = 219; European fancy breeds, n = 462; North America, n = 835; South America, n = 15; Oceania, n = 24) as well as commercial birds (n = 329) and experimental lines (n = 42). The wide scope of the dataset aims to capture a significant proportion of the extant genetic variation in the chicken genome. Furthermore, the addition of 15 ancient chicken genomes from Europe and the Middle East will provide supplemental time-depth, including a window into past genetic variation following the arrival of chickens into Europe from Southeast Asia (Fig. 38).

Consortium Aims

The aims of the consortium are numerous and varied, reflecting the diverse interests of the contributing groups. Specific scientific aims include:

1. **Deleterious alleles and possible inbreeding.** Breeds with high rates of inbreeding and potential health risks will be identified on the basis of genetic load and deleterious variants in sequence data. We will also investigate loss-of-function variation in relation to pseudogenes and adaptation.

2. **Structural variation.** The impact of structural variants (e.g., deletions and duplications) on trait variation will be assessed. This analysis aims to produce a catalog of structural variation and associated frequency estimates, as well as predicted functional consequences of the variants identified. We also aim to construct a graph genome from a diverse selection of breeds using a combination of long- and short-read sequencing technologies.

3. **Phenotype and trait adaptations.** We aim to identify causal gene variants that underlie adaptive traits. For instance, we are interested in covariation between genetic variants and agro-pastoral markers to shed light on the genetic basis of adaptation to different environments. We will also investigate adaptation to phenotypic and production traits, such as feather colour and egg shell quality.

4. **Distribution of extant chicken genetic diversity.** Sequence data from such diverse geographical sources permit a detailed investigation of extant chicken diversity with respect to geographical spread. Within this investigation, finer scale analyses of diversity – particularly within the continents of Africa and Asia – are to be conducted.

5. **Evolutionary history of chickens.** The introduction of chickens into Europe (when and how many times) remains unclear. By comprehensively mapping the extant variation in chicken populations, we aim to build a high-quality reference panel for variants, which can be used to phase and impute genomes, including low-coverage ancient genomes from Europe dating to ~2000 years ago (a few centuries following the introduction of chickens in Europe). Using similar ap-
proaches, we also plan to decipher the evolutionary history of chickens in Neotropical America, in which chickens underwent a much more recent (~500 years ago) introduction.

6. **Evolutionary history of Gallus spp.** Combining data from domesticated chickens and congeneric jungle-fowl is expected to help answer questions regarding the ultimate origin of domestic chickens and the contribution of junglefowl to modern chicken ancestry.

7. **Evolution and adaptation in the immune system.** A simple prediction is that chickens have had to adapt to cope with (1) exposure to novel pathogens, and (2) increased intensity of pathogen pressure due to increased flock size and density of rearing. This is likely to have left signatures of adaptation at immune loci of the chicken genome. Genes such as the Toll-like receptors and other pattern-recognition receptors at the front line of defense against pathogens will be investigated for signals of selection. We aim to conduct in vitro testing to validate bioinformatic predictions of functional change in immune receptors using methods that are well-established within the consortium.

These lines of investigation will be synthesized into several publications over the course of the consortium, led by different principal investigators depending on expertise. At the outset, the consortium has aimed to be as inclusive as possible, and as such, the studies listed above are neither exhaustive nor limited to current members of the consortium. The consortium welcomes input from any groups wishing to make the best use of this genomic resource.

**Processing Pipeline**

In order to provide complete consistency of analysis, the entire dataset was re-processed from raw reads using a state-of-the-art mapping and processing pipeline implemented on the SuperMUC computing cluster of the Leibniz-Rechenzentrum, Bavarian Academy of Science, Germany (Fig. 39). All reads underwent pre-processing (quality trimming, adaptor removing) with Fastp (v 0.21.0) then were mapped to the most recent version of the chicken genome (GRCg7b) (GCA_016699485.1) with BWA (v 0.7.17-r1188). The resulting BAM files from the same samples were then merged with samtools (v 1.9). For the variant calling, we generated gvcf using Elprep (v 5.1.1, compiled with go1.17; a reimplementation of GATK in GO langage). These gvcfs were integrated into a GATK genomicDB (gatk v4.2.3.0). To optimize performance, we built 48 databases corresponding to partitioning the genome into ~20 Mb intervals. Variant calling was performed using GATK genotypeGVCFs to obtain a VCF file by interval. We then obtained a global
VCF file using GATK GatherVCFs. GATK VariantReca-
librator was then used to recalibrate variants using known
SNPs.

Project Timeline
The project began in 2021 and is expected to conclude
(at least the first tranche of analyses) in 2023. The first
phase of the project (Q3-Q4 2021) involved data gather-
ing from both public and private sources and curation of
associated metadata. In Q1 2022, the read files were qual-
ity checked to remove low quality samples and to check
pre-processing from the variety of sequencing platforms
included in the dataset. In Q2 2022, read mapping com-
menced, soon followed by variant calling. At the time of
writing (September 2022), mapping and variant calling
have been completed and the VCF will shortly be made
available for further analyses.

Data Hosting and Availability
The SuperMUC computing cluster will provide the
processing power and storage capability to generate and
store raw read files, alignment map files, and variant call
files (VCF) for the duration of the project. The final VCF
will be made available in the first instance to members of
the consortium, and will also be provided to the commu-


Conflict of Interest Statement
The authors have no conflicts of interest to declare.

Funding Sources
The computational part of the project was funded by Super-
MUC grant ID: 23826 awarded by The Leibniz Supercomputing
Centre (LRZ), of the Bavarian Academy of Sciences and Humani-
ties. The authors were supported by European Research Council
grants (ERC-2013-StG-337574-UNDEAD and ERC-2019-StG-
853272-PALAEOFARM) and by the Wellcome Trust
(210119/Z/18/Z). The authors would like to acknowledge the Ed-
inburgh Genomics Facility (Edinburgh, UK) for generation of the
sequence data. This study was funded by the Bill and Melinda
Gates Foundation (BMGF) and with UK aid from the UK Govern-
ment’s Department for International Development (Grant Agree-
ment OPP1127286) under the auspices of the Centre for Tropical
Livestock Genetics and Health (CTLGH), established jointly by the
University of Edinburgh, SRUC (Scotland’s Rural College), and
the International Livestock Research Institute. The findings and
conclusions contained within are those of the authors and do not
necessarily reflect positions or policies of the BMGF nor the UK
Government. We thank the CGIAR livestock program (CRP) for

Fig. 39. Processing pipeline. To ensure between-sample consistency, all samples have been re-processed from raw
fastq reads. Reads underwent pre-processing and quality control before mapping to the latest version of the
chicken genome (GRCg7b), variant calling, and generating a VCF.
supporting the sampling component of the research. Sequencing data provided by INRAE teams were produced with public sources of funding, coming from INRAE, the French National Research Agency (ANR), the European Commission (SABRE project) and its HORIZON 2020 program (FEED-A-GENE, IMAGE projects). Data funded by INRAE and ANR were previously described by Tixier-Boichard et al. [2020]. https://doi.org/10.20870/productions-animales.2020.33.3.4564. Sampling of data provided by FLI was funded by the German Federal Ministry of Education and Research (BMBF) via the SYNBBREED project (FKZ 0315528E; www.synbreed.tum.de) and sequenced within the project IMAGE - Innovative Management of Animal Genetic Resources (www.image2020.eu, funded by the EU Horizon, 2020 research and innovation program No. 677533). Sequencing of the ancient genomes was funded by the Arts and Humanities Research Council (grant AH/L006979/1). O.L. is supported by the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 895107.

Biobanking and Candidate Genes for Genome Editing to Support the Tropical Poultry Improvement Programmes

(Prepared by C. Keambou Tiambo, C. Kamidi Muhonja, M. Ogugo, O. Ouko, E. Ilatsia, S. Kemp, A. Djikeng, and M. McGrew)

As the world increasingly relies on a handful of chicken breeds in intensified systems, a great diversity of tropical indigenous breeds is kept in backyard systems and on small farms, sometimes undergoing extinction threats through inbreeding and negative selection. If tropically adapted indigenous breeds disappear, the global poultry industry and the science community stand to lose access to adaptive genetic traits that have been nurtured and developed by local communities over millennia. The conservation of animal genetic resources through cryopreservation, referred to as biobanking, is an important component for the conservation and revival of rare or endangered species. While it was not previously possible to completely and efficiently biobank avian genetic material, a scientific innovation using primordial germ cells (PGCs) is changing the landscape and providing a way forward to preserve the biodiversity of tropical poultry breeds, hence also providing sustainable biological material from a more ethical 3Rs – Reduction, Refinement, Replacement – poultry biotechnology research and innovation. The tropical poultry PGC biobanking innovation includes the development and use of chickens that are devoid of their own sperm or ovum as recipients of the biobanked germ cells from a donor. With the introduction of the PGCs into the growing “sterile” embryo, the chick then develops into a surrogate fertile animal, but only produces gametes (ovum or sperm) that are 100% genetically those of the donor chicken breed. A sire-dam crossing of the surrogates then allows restoration of indigenous chicken breeds from biobanked material in a manner that is animal welfare-friendly. Most importantly, the lab-based techniques developed as part of this innovation are teachable and transferable to partner institutions, thereby enabling countries across the tropics to adopt chicken genetic resource biobanking.

Africa’s population is growing and transforming very fast, and will double by 2050, making food security the main challenge for the continent [UNDESA, 2017]. The rapid mutations in the demand for animal-source food, derived products and services, coupled with rising concerns of climate change, command a sustainable intensification of livestock production in Africa to accomplish food and nutritional security. Today, Africa must build the foundations to steer livestock on a sustainable development trajectory. This can be a short- to medium-term perspective if relying on short life cycle and locally adapted genetic resources.

Over 1,600 local chicken breeds have been identified globally [Eda, 2021], of which 126 are from Africa [Domestic Animal Genetic Resources Information System, 2022]. These breeds contain vast ranges of phenotypic and genetic diversity derived from the varied pathogenic, environmental, and selection conditions under which these ecotypes were developed.

Unfortunately, many of these local breeds are considered at risk due to the introduction and adoption of non-local breeds, the acceptance of intensive chicken production systems, changes in environment and disease conditions, and adverse development policies.

To minimize this genetic erosion, it is essential to increase knowledge of local breeds and production systems, improve planning, and raise awareness of the threat at the policy level. New innovations in genetic preservation technologies for chickens are also needed.

Conservation of poultry breeds and genetic lines pose particular challenges. One means of genetic diversity conservation in livestock species has been biobanking, whereby sperm, eggs or zygotes are preserved for future use. Cryopreservation of eggs and zygotes is not possible in avian species due to the large amount of lipid deposited in the female oocyte [Petitte, 2006; Whyte et al., 2016]. Other genetic preservation and propagation techniques such as cloning using somatic cell nuclear transfer are not possible, as embryo transfer cannot be done in avian species [Kjelland et al., 2014]. Recent developments by researchers from the International Livestock Research In-
stitute and the Centre for Tropical Livestock Genetics and Health (CTLGH), at the Roslin Institute, University of Edinburgh in the UK, have shown that the isolation and freezing of PGCs from chicken embryos can provide a new approach for biobanking poultry material. Biobanking of these PGCs will have a significant role in the conservation of African poultry genetic resources; further adding safeguards against the population and diversity losses that could threaten a breed’s survival [Hall, 2013].

Furthermore, beyond its role as an important source of protein and a valuable model for the study of developmental biology, immunology, physiology, and neurology in vertebrates [Yasugi and Nakamura, 2000; Speedy, 2003; Mozdziak and Petitte, 2004; Stern, 2004, 2005], there is an increasing interest in generating genetically modified chickens resistant to specific pathogens, benefiting from the availability of gene manipulation techniques. Hence, the full potential of new chicken breeding tools such as genome editing needs to be exploited in addition to conventional technologies. Clustered regularly interspaced short palindromic repeats (CRISPR/Cas)-based genome editing has rapidly become the most prevalent genetic engineering approach for developing improved chicken lines because of its simplicity, efficiency, specificity, and ease of use. Genome editing improves chicken breeds by conferring special attributes including specific chicken bioreactors, production of knock-in/out chickens for various production and adaptability traits, low-allergenicity eggs, or to serve as disease-resistance models [Chojnacka-Puchta and Sawicka, 2020]. In African countries with the most advanced regulations in animal biotechnologies like Kenya and Nigeria, such genome-edited animals with no foreign gene integration are not regulated as genetically modified organisms (GMOs). Researchers from the Centre for Tropical Livestock Genetics and Health (CTLGH) at the Roslin Institute and at the International Livestock Research Institute (ILRI) are using CRISPR/Cas-based genome editing for improving chicken productivity and adaptability.

**Biobanking African Chicken Genetic Resources for the Future**

**Primordial Germ Cell Isolation, Conservation, and Recovery for Production of Chimeric Chickens**

PGCs are specialized stem cells that can be isolated from chick embryos, which – depending on the sex of the individual embryo – will eventually form a sperm or egg. Following isolation, these PGCs can be cultured and cryopreserved. Chickens are one of the few species from which PGCs can easily be propagated in vitro to increase cell numbers up to 100,000 cells from a single embryo in 4 weeks [Whyte et al., 2015] and these cells can easily be cryopreserved [Glover and McGrew, 2012; Glover et al., 2013].

Biobanked chicken material is important in cases where a specific breed might be selected in the future to be scaled up to useful production levels. The genetic material can also be used to support research to isolate certain traits that can then be introduced into existing chicken populations.

When these biobanked chicken genes are needed, the preserved PGCs can be transferred into a 2-day-old “recipient” chick embryo. Part of this PGC biobanking innovation is the development and use of sterile recipient chicks. Because the recipient 2-day-old chick embryo is sterile, it lacks its own PGCs, and this eliminates the need to manage the PGCs of the recipient bird. Implanted with biobanked PGCs, the recipient chick grows up into a fertile bird. These adult birds have only had their reproductive cells changed to the genetics of the donor bird and will now act as surrogate parents. The recipient birds will still look like and have the genetic components of their breed, but they will produce sperm or eggs that are genetically identical to the original biobanked donor breed implanted.

The development of the PGC biobanking has revolutionized the ability to preserve tropical chicken genetic resources at ILRI-Nairobi. Tables 10 and 11 present the progress in African chicken ecotypes and lines of chicken already cryopreserved.

Along with the active biobanking of African chicken ecotypes and PGC lines, poultry biobanking activities are being scaled up through active knowledge transfer to African scientists in collaboration with the African Union-Inter African Bureau for Animal Resources (AU-IBAR) Training of Trainers (ToT) workshops in Kenya, Cameroon, and the Democratic Republic of Congo.

These workshops have led to increased knowledge of the role and importance of locally adapted chicken breeds and fostered greater understanding of the need for indigenous breed conservation. The Kenyan Agricultural and Livestock Research Organization (KALRO) has embraced the technology for the conservation of indigenous chicken ecotypes and plans to use it to sustain the development of the Kinyeji (local) chicken breeding programme. Further collaboration with KALRO is creating opportunities for greater upscaling and uptake by National Research Systems across Africa.
Restoration of Poultry Biodiversity and Dissemination of Potential Elite Lines Using the Chimeras Conventional Crossbreeding

Germline chimeric chickens were produced at ILRI-Nairobi by transfer of primordial germ cells from indigenous chicken to White Leghorn. An average of 500 primordial germ cells from indigenous chicken were injected into the bloodstream through the dorsal aorta of stage 14–15 White Leghorn recipient embryos which were then incubated until hatching, and the chimera derived offspring were identified based on their feather colour (Fig. 40).

Using the Gene Edited Surrogate Host Technologies

The DDX4 Knockout (KO) Surrogate Host Technology

The creation of the DDX4 KO line was led by Mike McGrew at the Roslin Institute and originally described by Ballantyne et al. [2017a]. The DDX4 KO surrogate hosts are genetically sterile female surrogate host chickens. This sterile female surrogate provides a major advance for the creation of genetically altered chicken lines and the preservation of rare breeds. The DDX4 KO line transmits 100% of its offspring from donor female primordial germ cells (PGCs). Therefore, the DDX4 KO line can produce offspring from genetically altered PGCs or PGCs from other breeds of chicken. Woodcock et al. [2019] have shown that by injecting PGCs from a heritage breed of Vantress chicken into DDX4 KO female embryos, all the offspring produced by these DDX4 females were derived from the donor heritage breed. Subsequent, artificial insemination of the DDX4 KO female surrogate host with frozen Vantress semen produced several pure heritage breed chicks.

The iCaspase9 Surrogate Host Technology

The creation of the iCaspase9 line was led by Mike McGrew at the Roslin Institute and originally described by Ballantyne et al. [2021a]. The iCaspase9 surrogate hosts are conditionally sterile male and female surrogate host chickens. In the iCaspase9 surrogate host line the germ cell lineage of both males and females are chemically ablated. These conditionally sterile transgenic chickens provide a major advantage for creating genetically altered (GA) chicken lines. GA PGCs or PGCs from other chicken breeds can be introduced into sterile male and female iCaspase9 embryos. Once the iCaspase9 birds reach sexual maturity, mating between male and female iCaspase9 birds or Sire Dam Surrogate mating results in 100% transmission of the donors’ PGCs in the first generation of offspring. The iCaspase9 surrogate hosts cut the time re-

Table 10. African chicken ecotypes cryopreserved using the blood and blastoderm methods

<table>
<thead>
<tr>
<th>Country</th>
<th>Ecotype</th>
<th>Number of cell lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>KENYA</td>
<td>White leghorn</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Karen (Nairobi)</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Narok</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Bomet</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Siaya</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Migori</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Homabay</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Kakamega</td>
<td>13</td>
</tr>
<tr>
<td>ETHIOPIA</td>
<td>Arobe</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>Horro</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Hawassa</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Shinyanga</td>
<td>84</td>
</tr>
<tr>
<td>TANZANIA</td>
<td>Shinyanga</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>Mbeya</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Morogoro</td>
<td>69</td>
</tr>
</tbody>
</table>

Table 11. Kenyan chicken ecotypes and lines cryopreserved using the embryonic gonad method

<table>
<thead>
<tr>
<th>Kenyan ecotypes</th>
<th>Male lines</th>
<th>Female lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laikipia (LP)</td>
<td>44</td>
<td>43</td>
</tr>
<tr>
<td>Bungoma (BG)</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>Kilifi (KF)</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Kakamega (KK)</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>Bomet (BM)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Homabay (HB)</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>Siaya (SY)</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Kwale (KW)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>KALRO chicken Improved lines 1 (KC1)</td>
<td>61</td>
<td>63</td>
</tr>
<tr>
<td>KALRO chicken Improved lines 2 (KC2)</td>
<td>75</td>
<td>41</td>
</tr>
<tr>
<td>Total</td>
<td>242</td>
<td>205</td>
</tr>
</tbody>
</table>
quired to create GA chicken lines, and reduce the number of animals required to produce GA chicken lines. The iCaspase9 surrogate host line has been used to create GA chickens with altered feather traits and birds with targeted mutations in the \textit{DMRT1} gene to investigate avian sex determination.

The chicken surrogate host technology has been approved by the National Biosafety Authority of Kenya for the biobanking and revival of the African indigenous poultry biodiversity. It can be used to harness enhanced resilience and productivity, and support future response to new poultry breeding requirements.

**Primordial Germ Cells for 3Rs Tropical Poultry Research and Gene Editing**

PGC technology has brought high hopes for advancing poultry research in breeding, production, and veterinary health problems. The stem cell application in restoration and regeneration of tissue, cloning, and transgenic poultry production carries much promise for (1) booster productivity and feed efficiency, (2) disease-free or disease-resistant chickens, and (3) production of heat stress resilient chickens. McGrew has stated “Discovering a way to easily freeze avian reproductive cells and subsequently bring back a genetically diverse flock will help the preserv-
tion of endangered breeds of poultry, increase food security from disease outbreaks and reduce numbers of animals used in research.” The poultry PGC process allows restoration of indigenous chicken breeds from biobanked stem cells in a manner that supports the 3Rs—Reduction, Refinement, Replacement—and is animal welfare-friendly. Most importantly, the lab-based techniques developed as part of this innovation are teachable and transferable to partner institutions, thereby enabling countries across Africa to adopt chicken genetic resource biobanking.

PGCs for Poultry Candidate Gene Screening

Biobanked poultry PGCs as precursors of sperm and egg have the potential to transmit the complete genetic and epigenetic information to the next generation [Mat-sui et al., 1992; Han, 2009], unlike poultry cryopreserved sperm. PGCs also exhibit unique migration and settling activity which plays a pivotal role in avian genetic resource protection and stem cell research [Burt and Pour-queir, 2003; Li et al., 2004; Oishi et al., 2016]. According to Taylor et al. [2017], the ability to precisely genetically edit the chicken genome will not only allow the investigation of key developmental signaling pathways in avian species but also the examination of genes involved in egg production, disease susceptibility and resistance with a view to promoting sustainability and biosecurity in both livestock and poultry production [Tizard et al., 2016; Whyte et al., 2016].

Biobanked Chicken PGC Lines to Support Genome Editing Research and Productivity Improvement

Chicken is one of the few vertebrate species for which the long-term in vitro propagation of primordial germ cells is possible, so performing genetic manipulation of cultured PGCs is becoming a standard practice, as demonstrated in recent studies (Fig. 41a-2). Several researchers have been developing CRISPR/Cas9 and TALENS technologies to investigate multiple genetic variations into pure breeds of chicken [Dimitrov et al., 2016; Oishi et al., 2016]. Recently, TALENs were used to target the DDX4 locus in chicken PGCs [Taylor et al., 2017]. DDX4 is located on the chicken Z sex chromosome and the mRNA is only expressed in the germ cell lineage. Genome editing in chicken is an emerging field and examples of gene editing in bird species other than chicken are currently lacking [Woodcock et al., 2017]. According to previous authors, genome editing can provide additional benefit to marker-assisted selection in breeding programmes, by either producing novel markers or introducing new traits to the genome. This is the case for chickens with reduced transmission of avian influenza virus produced by lentiviral transfection of embryos to insert an RNA hairpin molecule into the genome to interfere with viral replication [Lyall et al., 2011]. Greater understanding of the pathogenicity of specific diseases could open new avenues for avian disease management, through the application of genome editing. However, most tropical countries still have a long way to go in acquiring and domesticating the genetic manipulation technologies. This is, however, in process through research from the Centre for Tropical Livestock Genetics and Health (CTL-GH), and will be facilitated through improvement of poultry PGC line culture and the advances in the adoption of the surrogate host technology at the International Livestock Research Institute (ILRI).

Collectively, rapidly developing genome-editing technology will also accelerate progress in the poultry biotechnology field, opening up new opportunities for poultry to contribute to various industries (Fig. 42).

Candidate Genes for Genome Editing in Tropical Poultry

Poultry performance in tropical regions has been restricted by environmental conditions that cause heat stress and favour the development of parasites and diseases, impairing animal health and productivity. As stated by Professor Appolinaire Dijkeng in Poultry World, “Poultry is a key livestock animal for millions of smallholder farmers in low- and middle-income countries. Any gains in efficiency, productivity and health from introducing useful traits from other poultry breeds could significantly improve the lives of these farming families through increased food production and income” [Mc-Dougal, 2021].

In a collaborative project between the Roslin Institute and ILRI under the Centre for Tropical Livestock Genetics and Health (CTLGH) and the African Chicken Genetic Gains (ACGG) program, Gheyas et al. [2021] conducted an integrated environmental genome analysis of indigenous chickens from 3 African countries (Ethiopia, Nigeria, and Tanzania) to elucidate the drivers of tropical adaptation in African indigenous chicken populations. The results from the whole-genome sequencing analysis identified some strongly supported genomic regions under selection for environmental challenges related to altitude, temperature, water scarcity, and food availability. These regions harbour several gene clusters including regulatory genes, suggesting a predominantly oligogenic control of environmental adaptation. Few candidate genes detected in relation
to heat-stress indicate likely epigenetic regulation of thermo-tolerance for a domestic species originating from a tropical Asian wild ancestor. These results provide possible explanations for the rapid past adaptation of chickens to diverse African agro-ecologies, while also representing new landmarks for sustainable breeding improvement for climate resilience.

On the other hand, Marchesi et al. [2021] also highlighted relevant candidate genes such as *ATRNL1*, *PIK3C2A*, *PTPRN2*, *SORCS3*, and *gga-mir-1759* that could help to elucidate the genetics of feed efficiency traits, hence providing new insights on the mechanisms underlying the consumption and utilization of food in chickens. Harnessing these genes revealed by numerous researchers, and the potential of genome editing could be a powerful tool to use tropical poultry as a key driver for global food security and poverty alleviation.

![Fig. 41. Protocols for the restoration of biobanked tropical poultry genetic resources and the potential dissemination of potential elite lines using chimeras (a-1) and gene edited surrogate host technology (a-2 and b).](http://karger.com/cgr/article-pdf/162/8-9/405/3975130/000529376.pdf)
Scaling Up of Poultry Biobanking for Better Livelihoods

The scaling potential of poultry conservation is illustrated by the ongoing close collaboration between CTLP-GH, the Tropical Poultry Genetic Solutions (TPGS) programme, African Union – InterAfrican Bureau for Animal Resources (AU-IBAR) and the African National Agricultural Research Systems (NARS) like the Kenyan Agricultural Research Organization (KALRO). The collaboration is intensifying at various levels on ca-

Fig. 42. Schematic illustration for future application of genome-edited poultry to industries. Genome editing in poultry can improve disease resistance and meat productivity. By targeting egg white protein genes, genome edited poultry can economically produce protein drugs with improved biological efficacy. When the reported genes are targeted to the Z chromosome, the male embryo can be screened out before hatching by detecting fluorescence during incubation.
Capacity development and effective biobanking across Africa and Southeast Asia. In Kenya, this is mostly supported by the regulatory authorities (the National Biosafety Authority – NBA) that facilitates the adoption of the surrogate host technology for rapid recovery of the biobanked indigenous chicken ecotypes and the intensive dissemination of the elite local and locally adapted and improved breed for the betterment of the livelihood in local communities.

**Conclusion**

An extensive PGC biobank for indigenous avian breeds will not only support research and development to prevent problems with inbreeding and preserve at-risk poultry breeds, but will also reduce the large number of live animals needed to be kept for research across the world. It could also have an important role within poultry breeding companies to maintain important parental lines of mainstream poultry breeds used in commercial poultry production, without the need to keep large populations of live birds. Having cracked the difficult problem of chicken biobanking, this innovation of PGC preservation, coupled with sterile surrogate use, will revolutionize the preservation and future use of diverse chicken genetics.

Stem cell technologies coupled with genome editing have a prominent role to play in improving poultry production in Africa. There is a need for creating an enabling environment in Africa in general with science-based regulatory guidelines for the release and adoption of chickens developed using CRISPR/Cas9-mediated genome editing. Some progress has been made in this regard.

Considering the high demand coming from various African national agricultural research systems (NARS), the Centre for Tropical Livestock Genetics and Health (C TLGH) and the Tropical Poultry Genetics Solutions (TPGS) project have projected to scale out activities in Africa in collaboration with the Animal seed centres of Excellence of African Union-Inter African Bureau for Animal Resources (AU-IBAR), and to southeast Asia. Several countries in Central and West Africa have already identified the poultry value chain as their priority for food security and poverty alleviation in rural communities. It is further recommended that the technology be used by the Animal Seed Working group of the African Seed and Biotechnology Partnership Platform; a continental program led by the African Union Commission that frames the development of the seed sector in Africa through improved decision making and policy formulation, supporting evidence-based advocacy, and enhanced knowledge sharing. The critical factors of success of this initiative will be the effective operationalization of the African regional animal seed centres of excellence, continuous investment from national, bilateral, and multilateral partners for development of spearheading research, and a harmonized legal framework for access to poultry genetic resources and knowledge sharing between African countries and their scientific collaborators.

Prime beneficiaries of tropical chicken biobanking will be the local National Agricultural Systems (NARS), which will have the capability and technology to cryopreserve all avian genetic resources. Similarly, this avian genetic biobanking could benefit other stakeholders in poultry value chains worldwide. Scientific and research communities, as well as public and private breeding organizations will have access to genetic material from biorepositories across Africa. However, further work is needed to support countries to enforce regulations outlined in the Nagoya Protocol related to the fair and equitable sharing of benefits arising from the utilization of genetic resources.

**Conflict of Interest Statement**

The authors of this document declare that they have no affiliations with or involvement in any organization or entity with any financial interest in the subject matter or materials discussed in this manuscript.

**Funding Sources**

The research leading to this document has received funding from the Center of Tropical Livestock Genetics and Health (CTLGH) program and the Kenya Agricultural and Livestock Research Organization (KALRO).

**Application of Genomic Information in Layer Breeding Programs**

(Prepared by A. Wolc and J.E. Fulton)

As we are approaching 2 decades since the first draft sequence of the chicken genome was produced [International Chicken Genome Sequencing Consortium, 2004], we can reflect on the profound impact of this project on poultry breeding. That reference sequence, plus the simultaneous release of 2.8 million SNP (single nucleotide polymorphisms) identified in multiple chicken breeds [Wong et al., 2004] provided the essential resources needed to initiate genomic selection in poultry breeding programs. This information has been utilized in multiple ways with the most impactful being the development of high-, medium-, and low-density SNP chips that can provide SNP information for hundreds of thousands of SNP that cover the entire genome and individual assays concentrated in specific locations of interest. Combination of...
low- and high-density SNP panels with imputation [Wang C et al., 2013] or the use of medium density panels has enabled implementation of genomic selection in poultry breeding [Wolc et al., 2016]. Genomic selection has been used in layers to increase genetic gain through all parts of the “breeder’s equation”:

- Increasing accuracy of estimated breeding values for better choice of parents to reproduce the next generation
- Increasing selection intensity through the ability to hatch more and preselect male candidates using genotype information
- Preserving genetic variation by utilizing mendelian sampling information captured by an individual’s genotypes as opposed to relying purely on family information
- Shortening generation interval by reducing the need to wait for own phenotypes or those of the closest relatives at the point of selection.

The increase in accuracy and shortened generation time has been particularly valuable for selection in egg layer programs. The important traits of egg production and shell quality cannot be measured in males, and extended production cycles require 100-plus weeks of trait measurement for the females. The ability to apply genomic selection for these traits to the males and to younger females allows the selection program to make large gains. Furthermore, genomic selection can be applied to improvement in many difficult or expensive-to-measure traits including health and resilience under challenging conditions, crossbred performance, and behaviour. While the analysis methods are still evolving, the genomic and phenotyping information is accumulating to provide expanded training sets. Studies report increases of accuracy from addition of genomic information in a range of 20% to over 100% depending on trait, population, training size, and methodology [Wolc et al., 2011; Alemu et al., 2016; Picard Druet et al., 2020]. In addition to genomic prediction, genotyping enables parentage verification and assignment (for example in identification of hens laying floor eggs [Wolc et al., 2021]) or in the use of cross-classified mating with mixed semen [Hsu et al., 2015; Wolc et al., 2015] and improved product traceability (verification of correctness of line crosses throughout the breeding pyramid). With decreasing costs of genotyping, multiple genome-wide association studies (GWAS) have been performed, identifying 16,656 QTL across 370 traits as curated in the www.animalgenome.org database (as of 05/11/22). These studies, combined with additional -omics data are starting to give insights into the complex biology of chickens, their health, response to environment, feed utilization, and production traits.

All these developments were made possible by the availability of SNP chips whose contents were based on genomic sequence obtained from a small number of individuals or from DNA pools. These sequence and SNP chips have also been utilized to identify copy number variants (CNV) [Kranis et al., 2013; Boschiiero et al., 2015; Drobik-Czwarno et al., 2018]. This information also has “non-genomic” uses including line characterization, identification of variation within specific genes of interest, development of SNP sets to identify MHC haplotypes [Fulton et al., 2016], identification of retroviral inserts in the chicken genome [Mason et al., 2020b] and their effects on phenotypes [Fulton et al., 2021]. Until recently, deep sequencing had limited use in applied breeding due to its prohibitive cost. Sequence information from high-quality references combined with bioinformatic analysis of conserved genomic regions and homology can also be used for fine mapping of important genes such a blood types [Fulton et al., 2022]. Moreover, the development of low-pass sequencing methods with significant cost reduction relative to deep sequencing enables wider use of sequencing for basic and applied projects. Thousands of birds have been sequenced with data used for GWAS with increased resolution [Li J et al., 2022] which has allowed the exploration of low-pass sequencing to potentially improve the accuracy of genomic prediction [Wolc et al., 2022].

The initial chicken reference genome was from a Red Jungle Fowl bird, which is one of the progenitors of the modern chicken. Recently 2 additional genomes have been produced which better represent the modern chicken; one genome (GRCg7b) was from a commercial meat production bird (broiler) and the other (GRCg7w) was from a commercial egg production bird (White Leghorn) (see article by Warren et al., this report). Additional genome sequences obtained from a wide variety of other chicken breeds (pan-genomes) will provide further insights into variation that exists within the chicken and how this variation can impact health, performance, and behaviour of the birds.

The use of increasingly accurate and affordable genomic data is expected to be critical for breeding modern layers to improve bird welfare, production, and optimal use of resources for the sustainable future of egg production.

Conflict of Interest Statement
The authors have no conflicts of interest to declare.
The National Avian Research Facility
(Prepared by L.J. Henderson, A.B. Diack, L. Vervelde, M.P. Stevens, A. Balic, and M.J. McGrew)

The National Avian Research Facility (NARF) provides a range of resources and expertise for the avian research community in the UK and internationally. Areas of research supported by the NARF include avian immunology, host-pathogen interactions, physiology and behaviour, developmental biology, and genetics. The NARF was founded in 2013, and is based at The Roslin Institute, on the University of Edinburgh’s Easter Bush Campus, UK. The NARF was established via capital support from The Roslin Foundation and the University of Edinburgh, in addition to funding from the Biotechnology and Biological Sciences Research Council (BBSRC) and the Wellcome Trust. The facility continues to be supported by the University of Edinburgh and the BBSRC.

The facility consists of 2 units; the Greenwood building, a conventional biosecure facility and the Bumstead building that has specified pathogen-free (SPF) status. Both facilities include accommodation for the maintenance and breeding of poultry flocks for research purposes. The NARF is at the forefront of genome engineering technologies in poultry and is one of the few resources globally able to produce genetically altered (GA) chicken lines under both conventional and SPF conditions. Currently the NARF provides resources and expertise in 3 main areas: (1) Curation of unique poultry lines. The NARF maintains a wide range of transgenic chicken lines, wild-type layer lines, a broiler line, Japanese quail, and chicken lines with defined genetic characteristics. (2) Genome modification in chicken; the production and maintenance of genome-edited and transgenic chicken lines, and (3) the cryopreservation of research chicken lines, and rare or endangered chicken breeds.

Curator of Research Poultry Flocks

The NARF is home to one of the largest collections of transgenic chicken lines, including multiple ubiquitous and gene-specific fluorescent reporter lines [Davey et al., 2018] and sterile surrogate host chicken lines [Taylor et al., 2017; Ballantyne et al., 2021a]. In the Greenwood facility, the NARF also maintains a range of wild-type poultry lines, such as commercially relevant layer lines, a broiler line, and Japanese quail. In 2014, the inbred chicken lines that were maintained at The Pirbright Institute, and formerly held at the Institute for Animal Health (IAH) Compton, were relocated to the Bumstead SPF facility.

Nine inbred White Leghorn lines (Lines 61, 72, 15L, C.B4, C.B12, N, 0, P2a, and W) and 2 closed outbred chicken lines (Rhode Island Red and Light Sussex) are housed within the Bumstead facility [Kaspers and Schat, 2022]. These lines have been studied for their susceptibility and resistance to various avian pathogens based on their known MHC I haplotypes [Bacon et al., 2000; Alber et al., 2019; Chintoon-Uta et al., 2020; Bremner et al., 2021; Russell et al., 2021; Mountford et al., 2022]. Some of the inbred lines represent a considerable investment by BBSRC and other governmental departments over many years, with some tracing their origins to the late 1920s. Information regarding available resources and poultry lines can be found on the NARF website (www.narf.ac.uk).

Genome Modification in Chickens

The NARF provides world leading resources and expertise in genome modification in chickens. The production of GA birds has been more challenging compared with other model organisms. This was in part due to the complex structure of the avian zygote and the organisation of the avian embryo [Love et al., 1994; Sang, 2004]. However, methods like lentiviral vectors enabled the creation of a number of transgenic chicken lines that are valuable tools for developmental biology [McGrew et al., 2004; Davey et al., 2018], avian immunology [Balic et al., 2014], and biotechnology [Herron et al., 2018]. These methods have some limitations, for example to produce a stable GA chicken line, in which all offspring carry the introduced gene, requires multiple crosses and generations of chickens. This process is time-consuming and requires many animals. In addition to this, the size of the transgene introduced is limited by the vector capacity, and precise modification of genes is not possible with these methods.

Refinement of the culture conditions for chicken primordial germ cells (PGCs) [Van De Lavoir et al., 2006; Whyte et al., 2015], and the development of chicken sterile surrogate hosts [Taylor et al., 2017; Ballantyne et al., 2021a], combined with CRISPR/Cas9 methods that enable precise and targeted edits [Ballantyne et al., 2021a; Ioannidis et al., 2021], have improved the efficiency of genome editing in the chicken and consequently the creation of GA chicken lines. PGCs are stem cells that give rise to sperm or eggs and can be harvested from the blood supply of early-stage chicken embryos. PGCs can be grown in culture [Whyte et al., 2015] and modified via transgenesis [Macdonald et al., 2012] or gene-editing [Taylor et al., 2017]. These modified PGCs can be re-introduced into the blood supply of a wild-type chicken em-
bryo, where they migrate to the testes or ovaries, and give rise to sperm or eggs once the animal is sexually mature. However, because the introduced PGCs must compete with endogenous PGCs, this results in variable and low rates of chicks produced from the modified PGCs. To avoid this, sterile surrogate host chickens have been developed, where the germ cell lineage (PGCs) of both males and females can be ablated after the introduction of a chemical compound using the iCaspase 9 system [Ballantyne et al., 2021a]. When modified PGCs and the chemical compound are microinjected into the blood supply of a developing sterile surrogate host embryo [Ballantyne et al., 2021a], the endogenous PGCs are ablated and only the modified PGCs migrate to the testes or ovaries. Mating of male and female sterile surrogate hosts that have received modified PGCs results in 100% transmission of introduced PGCs in the first generation of offspring [Ballantyne et al., 2021a]. Compared to previous methods, this vastly reduces the number of animals used to create a GA chicken line [Panda and McGrew, 2022] and reduces the time required to create a GA chicken line from over 2 years to less than 1 year.

The NARF has sterile surrogate host chicken lines available for use by researchers, including the transgenic iCaspase9 chicken line that is conditionally sterile [Ballantyne et al., 2021a], and the DDX4 knockout chicken line, in which all females are genetically sterile [Taylor et al., 2017]. Sterile surrogate host chicken lines are also maintained in the SPF Bumstead facility, providing the capability to generate transgenic or gene-edited chickens under SPF conditions. Recently, the NARF’s genome editing technologies have been used to investigate avian sex determination using targeted mutations in the DMRT1 gene [Ioannidis et al., 2021], and the modification of feather traits [Ballantyne et al., 2021a]. Precision editing and transgenesis has been greatly facilitated by advances in the quality of avian reference genomes and functional annotations, including transcriptome atlases spanning cells, tissues, and developmental stages and the mapping of regulatory elements. For example, having identified genes and transcripts specific to chicken conventional dendritic cells (cDCs) [Wu et al., 2022], the NARF recently produced a novel fluorescent reporter chicken line that reports on cDCs via XCR1 gene expression. In addition, using the iCaspase 9 system [Straathof et al., 2005], this line also enables inducible ablation of cDCs. This transgenic chicken line will provide insights into the role of cDCs in natural and vaccine-mediated immunity, which is poorly understood in avian species.

Cryopreservation of Avian Genetic Resources

The NARF is in the process of creating a biobank to maintain avian genetic resources. This resource will be a repository for cryopreserved genetic material from valuable research chicken lines and rare or endangered poultry breeds. A biobank allows for the storage and protection of this genetic diversity, by mitigating against loss of resources in the event of disease outbreak or genetic bottlenecks in closed populations. Furthermore, cryopreserving chicken lines that are not currently being used for research will reduce the number of live animals that need to be maintained annually for research purposes, thereby addressing the principles of Replacement, Refinement, and Reduction (The 3Rs) of animal use in research.

Advances in PGC culture and the development of sterile surrogate host chickens supports the NARF’s ability to cryopreserve avian germplasm and re-derive cryopreserved lines. In concert with these technological advancements, the NARF is in the process of designing robust and time-efficient procedures to cryopreserve avian germplasm via PGCs [Nandi et al., 2016; Ballantyne et al., 2021b] or embryonic gonads [Hu et al., 2022]. As the biobank develops it will be important to create information sharing systems, to itemise which chicken lines have been cryopreserved, so that information is shared with the scientific community, thus preventing effort and resources being used to replicate GA lines that are already available. This is in line with similar resources that exist for other model organisms, such as mice, for which there are centralised repository systems (www.findmice.org) [Jackson et al., 2021].

Change in Focus

Previously the NARF aimed to collate and share biological tools such as antibodies, primers, and reagents for avian species, and build annotation of, and informatics-based resources for avian genomes. As the NARF’s research focus has evolved, these areas are no longer directly supported by the NARF. However, The Roslin Institute continues to support the development of avian antibodies, primers, and reagents through “The Immunological Toolbox” (www.immunologicaltoolbox.co.uk), in addition to a wide array of poultry research. Avian genomics and bio-informatics resources are now supported by projects undertaken by the “Functional Annotation of Animal Genomes” (FAANG) initiative (www.faang.org). Moving forward, NARF intends to build on its own genome engineering technologies, the production of novel GA lines, and the
development of an avian biobank. We are keen to hear from researchers in both academic and industry sectors to explore how NARF can further meet the needs of the avian research community.

Conflict of Interest Statement
The authors have no conflicts of interest to declare.

References


Garrison E, Marth G. Haplotype-based variant de-
tection from short-read sequencing. arXiv:
2017;1:207.39907v2.
Gavora JS, Kuhnlein U, Crittenden LB, Spencer
JL, Sabour MP. Endogenous viral genes: asso-
ciation with reduced egg production rate and egg
size in White Leghorns. Poult Sci.
2021;16(3):e204578.
Geibel J, Praefke NP, Weigend S, Simianer H, Re-
imer C. Assessment of linkage disequilibrium
patterns between structural variants and sin-
gle nucleotide polymorphisms in three com-
mercial chicken populations. BMC Genom-
Gheyas AA, Burt DW. Microarray resources for
genetic and genomic studies in chicken: a re-
Gheyas AA, Vallejo-Trujillo A, Kebede A, Loza-
analyses based on 1 million SNPs in commer-
Ghurye J, Rhie A, Walenz BP, Veiga-Fernandes
VF, Vera-Murillo VE, Cerolini S, et al. Ge-
nomical variability in Mexican chicken popula-
tion using SNP number variants. BMC Genet.
Gorlov IP, Gorlov OA. Cost-benefit analysis of
recombination and its application for under-
standing of chiasma interference. J Theor
Goto RM, Wang Y, Taylor RL Jr, Wakenell PS,
Hoischmidt K, Shiina T, et al. BGI has a major
role in MHC-linked resistance to malignant
lymphoma in the chicken. Proc Natl Acad Sci
Goto RM, Warden CD, Shiina T, Hoischmidt
reference genome reveals an MHCY haplo-
type organized in gene blocks that contain 107
loci including 45 specialized, polymorphic
MHC class I loci, 41 C-type lectin-like loci,
and other loci amid hundreds of transposable
Grandori C, Gomez-Roman N, Felton-Edkins
ZA, Ngouenet C, Galloway DA, Eisenman
RN, et al. c-Myc binds to human ribosomal
DNA and stimulates transcription of rRNA
Gray S, Cohen PE. Control of meiotic crossovers:
from double-strand break formation to design-
Gromen MAAM, Cheng HH, Bumstead N, Benkel
BF, Briles WE, Burke T, et al. A consensus
linkage map of the chicken genome. Genome
Gromen MAAM, Wahlberg P, Foglio M, Cheng
HH, Megens HJ, Cooimdt RM. A high-density
SNP-based linkage map of the chicken
genome reveals sequence features correlated
with recombination rate. Genome
Gromen MAAM, Megens HJ, Zarey L, Warren
WC, Hillier LW, Cooimdt RM. A high-
density marker panel for the chicken
e1002173.
Gibbons HR, Shaginurova G, Kim LC, Chapman
N, Spurlock CF, Aune TM. Divergent Inc-
RNA GATA3-AS1 regulates GATA3 tran-
2021;11(3):834.
Glover JD, Taylor L, Sherman A, Zeiger-Poli C,
Livianos C, Campbell BC. Analysis of the
Gordon S, Battagin A, Stevens E, Sokolov
O, Smith J. Whole genome sequenc-
Hanotte O, Smith J. Cost-benefit analysis of
endogenous viral genes: as-
1-10.
Hanotte O, Smith J. Whole genome sequenc-
Hastie YH, Chu HP. Functions and properties of nuclear IncRNAs—from sys-
tematically mapping the interactomes of
44.
Guiblet WM, Cremona MA, Harris RS, Chen D,
Eckert KA, Chiaromonte F, et al. Non-B
DNA: a major contributor to small- and
large-scale variation in nucleotide substitu-
tion frequencies across the genome. Nucleic
Grunder AA, Benkel BF, Chambers JR, Sabour
MP, Gavora JS. Characterization of eight en-
dogenous viral (ev) genes of meat chickens in
different commercial lines. Poult Sci.
Guo C, Geng M, Li Z. Knockdown of IncRNA
MCM3AP-AS1 attenuates chemoresistance of
Burkitt lymphoma to doxorubicin treat-
ment via targeting the miR-15a/EIF4E axis.
Ultraconserved element uc.372 drives he-
patic lipid accumulation by suppressing miR-
Guo S, Zhang C, Le A. The limitless applications
of single-cell metabolomics. Curr Opin Bio-
Guo X, Fang Q, Ma C, Zhou B, Wan Y, Jiang R.
Whole-genome resequencing of Xishuang-
banna fighting chicken to identify signatures
C, Wang JX, et al. Genetic signature of hybrid-
ization between Chinese spot-billed ducks and
domesticated ducks. Anim Genet. 2020;
51(6):866–75.


Kint T, Hunt HD, Parcells MS, van Santen V, Ewald SJ. Two class I genes of the chicken MHC have different functions: BF1 is recognized by NK cells while BF2 is recognized by CTLS. Immunogenetics. 2018;70(9):599–611.


Maruoka T, Tanabe H, Chiba M, Kasahara M. Chicken CD1 genes are located in the MHC: CD1 and endothelial protein C receptor genes constitute a distinct subfamily of class-I-like genes that predate the emergence of mammals. Immunogenetics. 2005;57(8):590–600.


McTavish EJ, Hills DM. How do SNP ascertainment schemes and population demographics affect inferences about population history? BMC Genomics. 2015;16(1):266.


Salomonsen J, Sørensen MR, Marston DA, Rodgers SL, Collen T, van Hateren A, et al. Two CD1 genes map to the chicken MHC, indicating that CD1 genes are ancient and likely to have been present in the prymordial MHC. *Proc Natl Acad Sci USA*. 2005; 102(24):8668–73.


Fourth Report on Chicken Genes and Chromosomes 2022


Wicker T, Robertson JS, Schulze SR, Feltus FA, Windau K, Viertlboeck BC, Göbel TW. The tur-

Whyte J, Blesbois E, McGrew MJ. Increased sus-


Whyte J, Blesbois E, McGrew MJ. Increased sus-


Fourth Report on Chicken Genes and Chromosomes 2022