Int. J. Dev. Biol. 62: 265-271 (2018)
https://doi.org/10.1387/ijdb.170276yc

THE INTERNATIONAL JOURNAL OF DEVELOPMENTAL BIOLOGY
www.intjdevbiol.com

Chicken genomics

YUANYUAN CHENG and DAVID W. BURT*

The University of Queensland, St. Lucia, Queensland, Australia

ABSTRACT As one of the most economically important species and a unique model organism for biological and medical research, the chicken represents the first non-mammalian amniotic species to have its genome sequenced; and so far, the chicken reference genome represents the best assembled and annotated avian genome. Since the release of the first draft genome sequence, the chicken genome assembly has improved greatly in coverage, contiguity and accuracy owing to the continuous efforts made by the chicken genomics community to generate extensive new data using novel sequencing technologies. Transcriptome sequencing, especially the recent effort to characterise full-length transcripts in chicken tissues, has provided key insights into the complexity of structure and function of the chicken genome. In this article, we review the progress in chicken genome assembly and annotation, and recent advances in comparative genomics in birds. Limitations of current data and plans of research are also discussed.

KEY WORDS: *chicken, genome, transcriptome, evolution, comparative genomics*

Introduction

The chicken not only provides an important source of protein around the world, but also has long served as a major model organism in the fields of medical, developmental, immunologic and nutritional research (reviewed in Burt, 2004, 2007; Stern, 2005). It also makes an excellent model system for genetics and genomics research due to the ease of breeding and maintenance, the long history of selective breeding, the resulting high phenotypic and genetic diversity, and its unique physiology and evolutionary position as a bird.

Since the release of the first draft assembly of the chicken genome (International Chicken Genome Sequencing Consortium, 2004), significant progress has been made in chicken genomics, providing informative data that has benefitted agricultural industry and biomedical research (Burt, 2005; Schmid *et al*., 2005; Schmid *et al*., 2015). The fast development of sequencing technologies and bioinformatics sciences in recent years have led to substantial improvements in the quality of the chicken reference genome and associated functional and comparative analyses. In this latest review, further progress in the completion of the genome sequence, annotation of all coding and non-coding transcripts, recent advances in comparative genomics and plans are reported.

Genome assembly

The initial draft genome assembly of a single, partially inbred, Red Jungle Fowl female was released in 2004 and represented the first non-mammalian amniotic genome to be sequenced. It provided a major advance for avian genetics, enabling a range of new "omics" analyses and technologies to be applied in poultry breeding and basic biological research (International Chicken Genome Sequencing Consortium, 2004). The first draft sequence, which had a size of 1.05 Gb (or 1.13 Gb including inferred gaps), was assembled from whole-genome shotgun sequences with 6.6x coverage, primarily consisting of paired-end plasmid (3-8 kb inserts) subclone reads generated using Sanger sequencing. The resulting genome assembly had a reasonably high contiguity, comprising 98,612 contigs with an N50 contig length of 36 kb and 32,767 supercontigs with N50 of 7.1 Mb. Further sequence scaffolding and chromosomal assignment were performed based on a BAC-clone-based physical map (Wallis *et al*., 2004) and a previously generated chicken genetic linkage map (Groenen *et al*., 2000), localising over 88% of the genome assembly to specific chromosomes or linkage groups. Additionally, 23,212 chicken mRNA sequences and 485,000 expressed sequence tags (ESTs) were used to polish the assembly and assist with sequence ordering and orientation correction. This first draft assembly of the chicken genome covered approximately 85% of the 1.23 Gb latest estimation of the genome size (Warren *et al*., 2017), providing nearly complete coverage for 26 out of 38 autosomes. However, like many draft genomes, the initial chicken assembly suffered from limitations caused by gaps and low sequence coverage in certain genomic regions. For example,

Abbreviations used in this paper: GRC, Genome Reference Consortium.

***Address correspondence to:** David W. Burt. Office 3.272, Queensland Bioscience Precinct (Building 80), The University of Queensland, St. Lucia, QLD 4072, Australia. Tel: +61 7 3446 2289. Fax: +61 7 3346 0555. E-mail: d.burt@uq.edu.au http://orcid.org/0000-0002-9991-1028

Submitted: 30 October, 2017; *Accepted:* 9 November, 2017.

ISSN: Online 1696-3547, Print 0214-6282

several chromosomes, including both sex chromosomes GGAZ and GGAW, and autosomes GGA16 (which contains the MHC region and thus many duplicated genes), GGA22, and GGA29 and smaller microchromosomes (higher G+C and CpG content than larger chromosomes), were not or only partially assembled due to insufficient sequence coverage. Genes with high GC-content were reportedly underrepresented or missing in the draft assembly (International Chicken Genome Sequencing Consortium, 2004), highlighting the need for an improved assembly quality to avoid potential bias or errors in scientific findings.

Several key improvements were made to the chicken genome reference assembly as new datasets became available in the following years (Schmid *et al.*, 2015). A second build (Gallus_gallus 2.1; GCA_000002315.1) was released in 2006 based on the original 6.6x coverage data and an additional collection of 198,000 reads, which were generated via targeted sequencing of BACs and fosmids to cover contig ends and regions of low quality in the original assembly. This second assembly also incorporated data from a chicken radiation hybrid (RH) map constructed using 2,531 genetic markers (Morisson *et al.*, 2007) as well as updated physical and linkage maps with significantly improved resolutions (Groenen *et al.*, 2009). These additional data coupled with upgraded genome assembly methods contributed to improved assembly statistics and quality of Gallus_gallus 2.1, with a total of 78,534 contigs (N50 = 46 kb) and 17,506 scaffolds (N50 = 11.1 Mb) spanning 1.10 Gb sequence length. 95% of the sequence was anchored to autosomes GGA1-28 and 32 and sex chromosomes; particularly, GGA16, GGA22, GGAW, and GGAZ had better (though still incomplete) coverage than in the original version, containing 0.63, 4.0, 0.86, and 67.8 Mb sequences, respectively.

The third version of the chicken genome assembly (Gallus_gallus-4.0; GCA_000002315.2) released in 2011, utilised next-generation sequencing (NGS) technology and elevated genome coverage by sequencing the same animal to 12x coverage on a Roche 454 platform. Combining new 454 data with the previously generated Sanger reads and mapping data, a *de novo* assembly of 1.03 Gb (1.05 Gb with gaps) in size was built, consisting of 27,041 contigs with N50 length of 0.28 Mb and 16,847 scaffolds with N50 of 12.9 Mb. The new assembly also integrated a greatly improved GGAZ (81.8 Mb) sequence generated through a BAC-focused targeted sequencing effort (Bellott *et al.*, 2010) and addressed ~10 Mb of erroneous duplications detected in Gallus_gallus 2.1 (Rubin *et al.*, 2010), which were possibly caused by stringent assembly parameters separating more diverse allelic variants into false duplicate loci. Like the previous version, over 96% of the sequence in Gallus_gallus-4.0 was mapped to chromosomes GGA1-28, 32, Z, and W, but still leaving microchromosomes GGA29-31 and 33-38 missing.

Significant improvements in the quality of the chicken reference genome were achieved in the fourth build of the genome, Gallus_gallus-5.0, GCA_000002315.3 (Warren *et al.*, 2017). Using single molecule sequencing technology (SMRT), 50.6x coverage of new sequences were generated for the same animal on a Pacific Biosciences RSII system, providing 18.7 Gb (15.3x coverage) of error-corrected long-read data; the *de novo* assembled sequences were then merged with Gallus_gallus-4.0 sequences to produce the final assembly. Additional data including 36x of Illumina paired-end reads and 168 finished clones from the CHORI-261 chicken BAC library were used for further base error and mis-assembly correction. As long-read sequencing technologies are more powerful than short-read in resolving complex genomic structures (e.g. repeats, regions with high GC-content), this upgraded assembly increased the size of the chicken reference genome to 1.23 Gb, improving the contiguity of ungapped contigs by over 10-fold (N50 = 2.9 Mb vs. the previous 0.28 Mb). Notably, three of the nine previously missing microchromosomes were captured and to some degree represented in Gallus_gallus-5.0, including GGA30 (0.22 Mb), GGA31 (0.17 Mb), and GGA33 (3.7 Mb). Also importantly, while repeat elements such as long terminal repeat (LTR) retrotransposons and Chicken Repeat 1 (CR1; a type of long interspersed nuclear elements or LINEs) only accounted for less than 9% of the initial chicken genome assembly (International Chicken Genome Sequencing Consortium, 2004), more than 200 Mb (16.4%) of sequence in Gallus_gallus-5.0 was identified as repeats. These improvements demonstrated the benefits and importance of using long-read sequencing for achieving better genome representation and completeness. Moreover, the error correction step using complementary short-read data was also proven to have greatly contributed to a high base level accuracy of Gallus_gallus-5.0. For instance, using GGAW sequence data, which should contain no SNPs in the genome of a female chicken (in birds, females have ZW and males have ZZ chromosomes), it was estimated that the base error rate in the new assembly was 74% lower than that in the previous Gallus_gallus-4.0 genome assembly (Warren *et al.*, 2017).

Current ongoing work to complete and further polish the chicken reference genome includes generating more genome coverage of PacBio data and performing chromosome-level assembly using genome wide physical mapping technologies (e.g. Hi-C chromatin interaction mapping). The future Gallus_gallus-6.0 reference genome will have longer contigs, chromosome-level scaffolds and hopefully will cover most, if not all, the microchromosomes and like the Z chromosome, a BAC-assembled W sequence (Bellot *et al.*, 2017) will be integrated into Gallus_gallus-6.0. We all hope this latest assembly will cover all chromosomes, as 39 contigs and a W contig. Further progress beyond Gallus_gallus-6.0 will involve a targeted approach to complete specific regions, such as the challenge of completion of Chromosome 16 and the highly duplicated gene families associated with the MHC-region. These final stages will be aided by membership of the Genome Reference Consortium (GRC, https://www.ncbi.nlm.nih.gov/grc/chicken) and input from the chicken genome community.

Genome annotation

A full genome annotation involves the identification of all coding and non-coding sequences and regulatory elements in a genome, and further, the interpretation of their functions and relationships with one another. As genome sequencing becomes easier nowadays, genome annotation remains highly challenging, with most of the existing genome annotations being incomplete and prone to bias and errors. Recent developments in chicken genome annotation have illustrated the significance of having a well annotated genome in providing comprehensive insights into an animal's biology.

Due to the lack of full-length transcript resources, which is a prerequisite for accurate transcriptional annotations (e.g. transcription start and termination sites, and alternative splicing patterns),

the initial annotation of the chicken genome (the first draft assembly) was based primarily on computational detection of sequence homology with genes mostly protein coding) and sequence models from other species, particularly human (International Chicken Genome Sequencing Consortium, 2004). Although a set of cDNA and EST transcript fragments was also used to assist with gene prediction with the Ensembl annotation pipelines, this initial version of annotation was incomplete and human centric. Only 571 non-coding RNA (ncRNA) sequences were predicted, mostly consisting of short ncRNAs, including transfer RNAs (tRNAs; n = 280), microRNAs (miRNAs; n = 121), and small nucleolar RNAs (snoRNAs; n = 83). The number of putatively protein-coding exons in the chicken was estimated to be 183,812, roughly converting into 20K-23K genes and pseudogenes based on an inferred average number of exons per gene between 9.6 and 8.0 (International Chicken Genome Sequencing Consortium, 2004). This estimated number of genes (especially non-coding genes) in the chicken is much lower than that found in the human genome, which contains 20,433 coding and 17,835 non-coding genes (NCBI Release108). However, as the authors indicated, there can be many potential errors in these initial estimations. For example, single-exon coding sequences supported solely by EST/cDNA evidence were treated as possible genomic DNA contaminations and were excluded from annotations, which potentially removed a significant number of real genes. Also, considering the long evolutionary distance between birds and mammals, the homology-based approach may be more successful at identifying relatively conserved genes, but can be less sensitive at detecting fast evolving genes (e.g. non-coding genes) or unique gene content that is specific to the avian lineage.

Developments in short-read RNAseq technology, accompanied by improvements in the contiguity and accuracy of the genome assembly, have led to great improvements in the annotation of the chicken genome. By 2015, a large collection of chicken transcriptome data from a wide range of tissue types was generated by the *International Avian RNAseq Consortium* (Schmid *et al.*, 2015; Smith *et al.*, 2015), providing more direct experimental evidence for structural and functional analysis of the genome. The RNAseq data allowed the identification of 15,495 coding genes and a large variety of ncRNAs (Ensembl release 71), representing an important advance in the chicken genome annotation. More recently, based on the fourth build of the reference genome (Gallus_gallus-5.0) and 9.2 billion transcripts (mostly generated on Illumina platforms) from 124 different tissue samples, an updated chicken gene set was released, containing 19,119 protein-coding genes and 6,839 non-coding genes (Warren *et al.*, 2017). Among the 57,960 unique transcripts identified, 80% accounted for mRNAs,

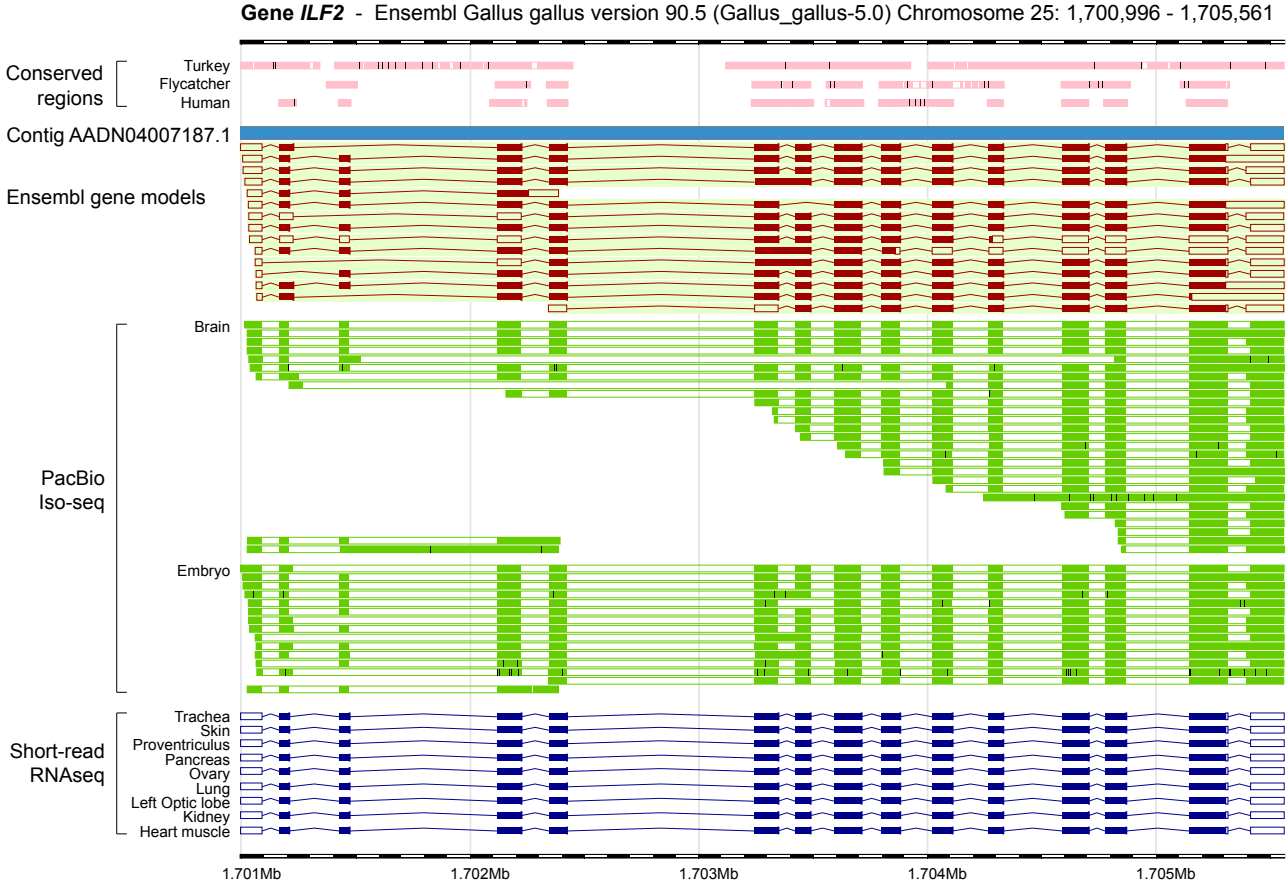


Fig. 1. Full-length transcript sequencing using the PacBio Iso-seq method revealed a high transcriptional complexity in chicken tissues. *In this example using gene* ILF2 *on chicken chromosome 25, the Iso-seq method detected multiple transcription start sites and exons, multiple transcription termination sites and exons, and diverse alternative splicing patterns (retained introns, skipped exons, and alternative exon/intron splicing sites), whereas only one transcript variant was identified with short-read RNAseq data.*

while the remaining 20% mostly comprised long non-coding RNAs (lncRNAs; 14.1%), miRNAs (1.8%), tRNAs (0.5%), and miscellaneous other RNAs (misc_RNAs; 3.6%). This annotation (NCBI Release 103) revealed a much higher gene content in the chicken genome than originally predicted in the first draft, altering the initial perception that in association with their relatively small genome sizes, birds tend to have smaller numbers of genes than other tetrapods (i.e. mammals, reptiles and amphibians) due to gene loss (Lovell *et al.*, 2014; Zhang *et al.*, 2014). The extensive RNAseq data also helped identify many protein-coding genes that were formerly thought to have been lost in the chicken (Bornelöv *et al.*, 2017; Hron *et al.*, 2015; Warren *et al.*, 2017), while providing further confirmation for a range of other reportedly missing genes as absent in the transcriptomes (Lovell and Mello, 2017; Warren *et al.*, 2017).

Although short-read based RNA sequencing has enabled significant advances in genome annotation, it comes with key limitations due to the short-read length, making it less capable of resolving complex features in the transcriptome, such as variable transcription start and termination sites, and certain alternative splicing and exon chaining patterns. Computationally assembled transcripts from short reads are often incomplete and only represent a fraction of all transcript variants within the transcriptome examined, therefore leading to an underestimation of the size and complexity of the transcriptome (Fig.1).

To overcome these issues, efforts have been made to uncover new transcript isoforms and obtain full-length transcript sequences using long-read single molecule sequencing, essentially avoiding the assembly problem. By sequencing cDNA from embryonic chicken hearts using a combination of long-read and short-read technologies, Thomas and colleagues (2014) identified 9,221 new transcripts, which mapped to 5,391 known genes and 539 new genic regions (based on Ensembl release 74). Most recently, Kuo and colleagues (2017) sequenced chicken brain and embryo cDNA libraries using the PacBio Iso-Seq method, which generates long reads spanning the entire length of transcripts. 5'-cap selection was performed for the embryo library to allow capture of intact 5' sequences, and library normalisation was carried out for both tissues to provide better coverage for less abundant transcripts and isoforms. Several key findings were made from these data, marking a major development in chicken genome annotation and transcriptome characterisation. A total of 14,421 protein-coding genes and 17,178 non-coding genes were identified to be expressed in the tissues from chicken brain and embryo, represented by 43,738 protein-coding and 20,539 ncRNA transcripts. An overall transcript per gene ratio of 2.22 was observed, which is comparable to that seen in the rat (2.24 according to NCBI Release 106) though lower than that of the human (4.04; NCBI release 108). This ratio is likely to increase with more tissue types and deeper sequencing depth being used for the analysis. Notably, 20,516 ncRNA transcripts identified were longer than 200 bp, i.e. lncRNAs, which would have been difficult to detect without the direct experimental data due to their low levels of conservation among distant lineages of species (Necsulea *et al.*, 2014). The long-read technology also allowed a comprehensive assessment of the transcriptional complexity in the chicken (Fig.1). With the start and end of transcripts for the first time reliably defined, it was observed that among genes with multiple transcripts detected, 96.4% and 91.9% had multiple transcription start sites and termination sites, respectively, the majority of which were attributed to alternative starting or termination exons. Along with highly diverse alternative splicing patterns detected, such as retained introns, skipped exons, and alternative exon/intron splicing sites, these results led to a conclusion that the chicken genome encompasses a similar level of transcriptional complexity compared to the human.

Current efforts are using the long-read Pacbio Iso-seq approach to characterise full-length cDNAs from a wide range of tissues. Current estimates based on assembly of short-read RNAseq data from more than 20 tissues (Burt *et al.*, *in preparation*) predict ~50K genes (both coding and non-coding) and more than 250K alternate but unique transcripts. So, we can be certain the transcriptome of the chicken will become even more complex, with more genes, more transcripts and more complex patterns of expression. More detailed analysis of these new data will fill in the gaps of so-called "missing" genes in birds and strengthen evidence for avian-specific genes and transcripts (Lovell and Mello, 2017; Warren *et al.*, 2017). The next challenge will be to define the function of all the coding and increasing collection of lncRNAs.

The annotation of regulatory elements within a genome can be an even more challenging task than gene annotation, but is as important because it provides critical clues for understanding the mechanisms controlling tissue or cell-type specific and time-dependent expression of genes that underlie various biological processes (e.g. development, diseases). While the identification of promotors (e.g. TATA box and CpG islands) and proximal elements is somewhat less complex due to their close spatial association (usually within 200 bp) with genes (though can be confounded in the case of multiple transcription start sites), confident annotation of distal elements, such as silencers, enhancers, and insulators, can be difficult (Maston *et al.*, 2006). Significant efforts have been made to annotate regulatory components in the human and mouse genomes (Zerbino *et al.*, 2016); by comparison, the regulatome of the chicken is still poorly characterised. A prediction strategy integrating multiple computational tools has been developed for the systematic discovery of *cis*-acting regulatory elements and transcription factor binding sites in the chicken genome (Khan *et al.*, 2013). However, many of the computational tools rely heavily on the detection of conserved synteny or sequence homology in comparison with the human and mouse genomes, and therefore can be problematic for their use in distantly related species, such as the chicken. This has been illustrated by a comparative analysis of genome-wide CTCF (CCCTC-binding factor; a transcription factor that binds and affects activity of insulators) bound sites in human, mouse and chicken using chromatin immunoprecipitation sequencing (ChIP-seq), which revealed that only 7% of the detected CTCF sites in the chicken were located in syntenic positions as in the human and mouse genomes, most of which were not conserved at the sequence level (Martin *et al.*, 2011).

An international effort to fill the current gaps has been initiated. The *Functional Annotation of Animal Genomes* (FAANG) consortium aims to generate comprehensive maps of functional elements in the genomes of domesticated species, including the chicken (Andersson *et al.*, 2015). Using experimental protocols adapted from those developed by the *Encyclopedia of DNA Elements* (ENCODE) project and the *International Human Epigenome Consortium*, the project plans to coordinate the generation and analysis of a variety of data, such as RNAseq, ChIP-seq, and ATAC-seq (assay for transposase-accessible chromatin sequenc-

ing) data, which will provide extensive information on tissue/cell/state specific gene expression, genome-wide chromatin accessibility, histone modification, methylation, and transcription factor binding profiles in domesticated animals.

Comparative genomics

Comparative analysis of genomes of diverse species enables the detection of functionally significant genes and sequence elements that have been conserved through their extensive evolutionary history, and facilitates the identification of adaptive changes in gene content and features in specific lineage of species that are associated with unique phenotypes. Availability of whole-genome sequences for more and more species provides powerful data allowing confident resolution of close or ambiguous phylogenetic relationships among species, while also providing valuable resources for targeted analysis of gene function and evolution.

As the first avian genome to be sequenced, the chicken genome assembly enabled the first comparison to be made between the genomes of the mammalian and non-mammalian amniote lineages (International Chicken Genome Sequencing Consortium, 2004). A group of 7,606 genes were found to have been conserved as 1:1 orthologues across vertebrates. Adaptive gene expansion gave rise to a lineage-specific family of keratin genes (β-keratins), responsible for forming scales, claws, and feathers, whereas genes encoding vomeronasal receptors, casein milk proteins, salivary-associated proteins, and enamel proteins are absent in the chicken. Alignments between chicken and human genomic sequences revealed a range of conserved elements with no evidence of expression, which tend to distantly associate with genes involved in transcriptional regulation, DNA binding, and metabolic and developmental functions. The functional significance of most of these highly conserved non-coding elements remains to be determined.

Following the chicken genome, draft genome sequences of three other avian species were generated, including the zebra finch *Taeniopygia guttata* (Warren *et al.*, 2010), which represents the highly diverse lineage of songbirds (Passeriformes) and a unique model organism in neurophysiology, the turkey *Meleagris gallopavo* (Dalloul *et al.*, 2010), another agriculturally important species belonging to the same order (Galliformes) as the chicken, and the duck *Anas platyrhynchos* (Huang *et al.*, 2013), an economically and medically important waterfowl (Anseriformes) which is the major natural reservoir of influenza A viruses. While the zebra finch genome was sequenced using the shotgun-based Sanger sequencing approach, the turkey and duck sequences were generated on short-read sequencing platforms. Initial comparative analyses of these avian genomes with the chicken and other vertebrate genomes revealed a range of lineage-specific features. For instance, the high resemblance of turkey and chicken genomes indicated a relatively high stability of Galliform genomes as measured by rearrangements, showing a higher proportion of sequences under evolutionary constraint than placental mammals (Dalloul *et al.*, 2010). In the zebra finch, many neurobiologically relevant gene families were found to have expanded and several genes regulated by song behaviour showed evidence of positive selection in comparison to their counterparts in the chicken, which may be associated with the species' ability of sophisticated vocal learning and communication (Warren *et al.*, 2010). Also compared to the chicken, lineage-specific duplications were detected in certain immune gene families in the duck, such as defensins and butyrophilin-like genes, which may have contributed to its natural resistance to many influenza strains (Huang *et al.*, 2013).

Between 2013 and 2014, the *Avian Phylogenomics Project* released the draft genome sequences for another 44-bird species, which marked another major milestone in avian genomics, with most extant avian orders represented (Jarvis *et al.*, 2014; Zhang *et al.*, 2014). The whole effort has facilitated the evolutionary analysis of avian genomes, including the identification of evolutionary constrained regions and functional elements, the dynamic evolution of genome organisation and speciation, and spawned a whole new area of phylogenomics and the study of adaptation and speciation in birds.

Through genome-scale sequence alignments using all 48 available avian genomes, Jarvis and colleagues (2014) extracted an orthologous sequence dataset comprising 8,251 syntenic protein-coding genes and 3,769 ultraconserved elements across species, spanning 41.8 Mb of sequence. Collectively, these data gave rise to a high-resolution total evidence nucleotide tree with high statistical support for most branches, confidently resolving many previously controversial branch placements. Also, importantly, this study demonstrated that bootstrap supports on the more challenging deeper branches of the species tree, positively correlate with the amount of data used for phylogeny reconstruction, and furthermore, protein-coding gene data alone cannot provide accurate estimation of phylogenetic relationships among species due to the impact of convergent evolution driven by shared ecological or life history traits. These results highlighted the advantages of having whole-genome data for understanding the evolution of species and their genomes.

Comparative study of the 48 avian genomes also allowed the identification of several genomic characteristics unique to birds (Zhang *et al.*, 2014). A range of genes showed evidence of adaptive evolution in association with avian-specific physiological and developmental features, such as diversification of genes involved in ossification driven by positive selection (flight-related bone pneumatisation), pseudogenization of enamel and dentin formation genes (absence of teeth), and expansion of opsin genes (advanced visual system). Contractions in repeat elements, deletions of 118 large syntenic blocks compared to reptiles and mammals and associated gene loss (see also Lovell *et al.*, 2014) were suggested to have led to a characteristic compact genome size in all birds examined. A high stability of avian genomes relative to other vertebrate genomes was evidenced by higher levels of conservation of chromosomal arrangements and gene synteny, lower rates of gene turnover in multigene families, a lower overall nucleotide substitution rate, and a higher abundancy in highly conserved elements across species. Some of these findings may be worth re-examination because of the potential bias that may have been introduced due to the limited coverage and contiguity of the majority of the avian genome assemblies used. For example, reviewing various cytogenetic evidence and emerging long-read based data (e.g. chicken Gallus_gallus-5.0), Kapusta and Suh (2017) pointed out that most of the draft avian genome assemblies likely have incomplete representation of repetitive elements (e.g. transposable elements, endogenous viral elements, centromeres, telomeres), resulting in underestimations of the size of the genomes and size variation across species. Similarly, due to the lack of chromosome-level assemblies, the full scale of chromosomal rearrangements

(Farré *et al.*, 2016) and gene gains and losses (Hron *et al.*, 2015; Korlach *et al.*, 2017) during avian evolution is yet to be revealed.

Despite the current limitations, the extensive sequencing effort of the *Avian Phylogenomics Project* has provided numerous novel insights into avian evolution and biology (a detailed list of the publications available is at http://avian.genomics.cn/en). Aspirations have been raised to sequence all 10,500 extant (even some extinct) avian species as part of the international Bird 10K (B10K) genome project (Zhang, 2015). So far, about 300 species have been sequenced, marking the completion of the second phase (all avian families) of the project, and the third phase (all genera) is now ongoing. Future efforts will include the generation and incorporation of long-read sequence data and genome wide physical maps for representative species, which will substantially improve the assembly quality and enable more comprehensive unbiased comparative genome analyses in birds.

Conclusions

The chicken reference genome has improved significantly in recent years as sequencing and assembly technologies evolve and innovative approaches emerge. Accompanying each major update of the genome assembly and annotation, new insights were provided on the structure, function, and evolution of the chicken genome. It has become apparent that to ultimately have a complete chromosome-scale assembly and fully catalogued tissue and state specific transcription and regulatory modification profiles of the genome will be important for unravelling the complexity of mechanisms underlying various biological processes in the chicken and all other species.

Acknowledgements

I am grateful to my colleagues and many collaborators who continue to contribute to the development of chicken genome and acknowledge the financial support of the Biotechnology and Biological Sciences Research Council (BBSRC), European Commission (EC), Wellcome Trust, National Human Genome Research Institute (NHGRI) and many other funders of this research.

References

ANDERSSON, L., ARCHIBALD, A.L., BOTTEMA, C.D., BRAUNING, R., BURGESS, S.C., BURT, D.W., CASAS, E., CHENG, H.H., CLARKE, L., COULDREY, C. *et al.,* (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol* 16: 57.

BELLOTT, D.W., SKALETSKY, H., PYNTIKOVA, T., MARDIS, E.R., GRAVES, T., KREMITZKI, C., BROWN, L.G., ROZEN, S., WARREN, W.C., WILSON, R.K. *et al.,* (2010). Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* 466: 612-616.

BORNELÖV, S., SEROUSSI, E., YOSEFI, S., PENDAVIS, K., BURGESS, S.C., GRABHERR, M., FRIEDMAN-EINAT, M. and ANDERSSON, L. (2017). Correspondence on Lovell *et al.*: identification of chicken genes previously assumed to be evolutionarily lost. *Genome Biol* 18: 112.

BURT, D.W. (2004). The chicken genome and the developmental biologist. *Mech Dev* 121: 1129-1135.

BURT, D.W. (2005). Chicken genome: Current status and future opportunities. *Genome Res* 15: 1692-1698.

BURT, D.W. (2007). Emergence of the Chicken as a Model Organism: Implications for Agriculture and Biology1. *Poult Sci* 86: 1460-1471.

DALLOUL, R.A., LONG, J.A., ZIMIN, A.V., ASLAM, L., BEAL, K., ANN BLOMBERG, L., BOUFFARD, P., BURT, D.W., CRASTA, O., CROOIJMANS, R.P.M.A. *et al.,* (2010). Multi-Platform Next-Generation Sequencing of the Domestic Turkey (*Meleagris gallopavo*): Genome Assembly and Analysis. *PLoS Biol* 8: e1000475.

FARRÉ, M., NARAYAN, J., SLAVOV, G.T., DAMAS, J., AUVIL, L., LI, C., JARVIS, E.D., BURT, D.W., GRIFFIN, D.K. and LARKIN, D.M. (2016). Novel Insights into Chromosome Evolution in Birds, Archosaurs, and Reptiles. *Genome Biol Evol* 8: 2442-2451.

GROENEN, M.A.M., CHENG, H.H., BUMSTEAD, N., BENKEL, B.F., BRILES, W.E., BURKE, T., BURT, D.W., CRITTENDEN, L.B., DODGSON, J., HILLEL, J. *et al.,* (2000). A Consensus Linkage Map of the Chicken Genome. *Genome Res* 10: 137-147.

GROENEN, M.A.M., WAHLBERG, P., FOGLIO, M., CHENG, H.H., MEGENS, H.-J., CROOIJMANS, R.P.M.A., BESNIER, F., LATHROP, M., MUIR, W.M., WONG, G.K.-S. *et al.,* (2009). A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Res* 19: 510-519.

HRON, T., PAJER, P., PAČES, J., BARTŮNĚK, P. and ELLEDER, D. (2015). Hidden genes in birds. *Genome Biol* 16: 164.

HUANG, Y., LI, Y., BURT, D.W., CHEN, H., ZHANG, Y., QIAN, W., KIM, H., GAN, S., ZHAO, Y., LI, J. *et al.,* (2013). The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nat Genet* 45: 776-783.

INTERNATIONAL CHICKEN GENOME SEQUENCING CONSORTIUM. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432: 695-716.

JARVIS, E.D.MIRARAB, S.ABERER, A.J.LI, B.HOUDE, P.LI, C.HO, S.Y.W.FAIRCLOTH, B.C.NABHOLZ, B.HOWARD, J.T. *et al.,* (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346: 1320-1331.

KHAN, M.A.F., SOTO-JIMENEZ, L.M., HOWE, T., STREIT, A., SOSINSKY, A. and STERN, C.D. (2013). Computational tools and resources for prediction and analysis of gene regulatory regions in the chick genome. *Genesis* 51: 311-324.

KORLACH, J., GEDMAN, G., KINGAN, S., CHIN, J., HOWARD, J., CANTIN, L. and JARVIS, E.D. (2017). De Novo PacBio long-read and phased avian genome assemblies correct and add to genes important in neuroscience research. *bioRxiv*.

KUO, R.I., TSENG, E., EORY, L., PATON, I.R., ARCHIBALD, A.L. and BURT, D.W. (2017). Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics* 18: 323.

LOVELL, P.V. and MELLO, C.V. (2017). Correspondence on Lovell *et al.*: response to Bornelöv *et al., Genome Biol* 18: 113.

LOVELL, P.V., WIRTHLIN, M., WILHELM, L., MINX, P., LAZAR, N.H., CARBONE, L., WARREN, W.C. and MELLO, C.V. (2014). Conserved syntenic clusters of protein coding genes are missing in birds. *Genome Biol* 15: 565.

MARTIN, D., PANTOJA, C., MIÑÁN, A.F., VALDES-QUEZADA, C., MOLTÓ, E., MATESANZ, F., BOGDANOVIĆ, O., DE LA CALLE-MUSTIENES, E., DOMÍNGUEZ, O., TAHER, L. *et al.,* (2011). Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nat Struct Mol Biol* 18: 708-714.

MASTON, G.A., EVANS, S.K. and GREEN, M.R. (2006). Transcriptional Regulatory Elements in the Human Genome. *Annu Rev Genomics Hum Genet* 7: 29-59.

MORISSON, M., DENIS, M., MILAN, D., KLOPP, C., LEROUX, S., BARDES, S., PITEL, F., VIGNOLES, F., GÉRUS, M., FILLON, V. *et al.,* (2007). The chicken RH map: current state of progress and microchromosome mapping. *Cytogenet Genome Res* 117: 14-21.

NECSULEA, A., SOUMILLON, M., WARNEFORS, M., LIECHTI, A., DAISH, T., ZELLER, U., BAKER, J.C., GRUTZNER, F. and KAESSMANN, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505: 635-640.

RUBIN, C.-J., ZODY, M.C., ERIKSSON, J., MEADOWS, J.R.S., SHERWOOD, E., WEBSTER, M.T., JIANG, L., INGMAN, M., SHARPE, T., KA, S. *et al.,* (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464: 587-591.

SCHMID, M., NANDA, I. and BURT, D.W. (2005). Second report on chicken genes and chromosomes 2005. *Cytogenet Genome Res* 109: 415-479.

SCHMID, M.SMITH, J.BURT, D.W.AKEN, B.L.ANTIN, P.B.ARCHIBALD, A.L.ASHWELL, C.BLACKSHEAR, P.J.BOSCHIERO, C.BROWN, C.T. *et al.,* (2015). Third Report on Chicken Genes and Chromosomes 2015. *Cytogenet Genome Res* 145: 78-179.

SMITH, J., BURT, D.W. and THE AVIAN, R.C. (2015). The Avian RNAseq Consortium: a community effort to annotate the chicken genome. *Cytogenet Genome Res* 145: 78-179.

STERN, C.D. (2005). The Chick: A Great Model System Becomes Even Greater. *Dev Cell* 8: 9-17.

THOMAS, S., UNDERWOOD, J.G., TSENG, E., HOLLOWAY, A.K. and ON BEHALF OF THE BENCH TO BASINET CV, D.C.I.S. (2014). Long-Read Sequencing of Chicken Transcripts and Identification of New Transcript Isoforms. *PLoS One* 9: e94650.

WALLIS, J.W., AERTS, J., GROENEN, M.A.M., CROOIJMANS, R.P.M.A., LAYMAN, D., GRAVES, T.A., SCHEER, D.E., KREMITZKI, C., FEDELE, M.J., MUDD, N.K. *et al.,* (2004). A physical map of the chicken genome. *Nature* 432: 761-764.

WARREN, W.C., CLAYTON, D.F., ELLEGREN, H., ARNOLD, A.P., HILLIER, L.W., KUNSTNER, A., SEARLE, S., WHITE, S., VILELLA, A.J., FAIRLEY, S. *et al.,* (2010). The genome of a songbird. *Nature* 464: 757-762.

WARREN, W.C., HILLIER, L.W., TOMLINSON, C., MINX, P., KREMITZKI, M., GRAVES, T., MARKOVIC, C., BOUK, N., PRUITT, K.D., THIBAUD-NISSEN, F. *et al.,* (2017). A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. *G3: Genes|Genomes|Genetics* 7: 109-117.

ZERBINO, D.R., JOHNSON, N., JUETTEMAN, T., SHEPPARD, D., WILDER, S.P., LAVIDAS, I., NUHN, M., PERRY, E., RAFFAILLAC-DESFOSSES, Q., SOBRAL, D. *et al.,* (2016). Ensembl regulation resources. *Database* 2016: bav119-bav119.

ZHANG, G. (2015). Genomics: Bird sequencing project takes off. *Nature* 522: 34-34.

ZHANG, G., LI, C., LI, Q., LI, B., LARKIN, D.M., LEE, C., STORZ, J.F., ANTUNES, A., GREENWOLD, J.M., MEREDITH, R.W. *et al.,* (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 80: 346.

Further Related Reading, published previously in the *Int. J. Dev. Biol.*

The surface ectoderm of the chick embryo exhibits dynamic variation in its response to neurogenic signals
Vineeta-Bhasker Tripathi, Yasuo Ishii, Muhammad M. Abu-Elmagd and Paul J. Scotting
Int. J. Dev. Biol. (2009) 53: 1023-1033
https://doi.org/10.1387/ijdb.082780vt

Cellular dynamics and molecular control of the development of organizer-derived cells in quail-chick chimeras
Jean-Baptiste Charrier, Martin Catala, Françoise Lapointe, Nicole Le Douarin and Marie-Aimée Teillet
Int. J. Dev. Biol. (2005) 49: 181-191
http://www.intjdevbiol.com/web/paper/041962jc

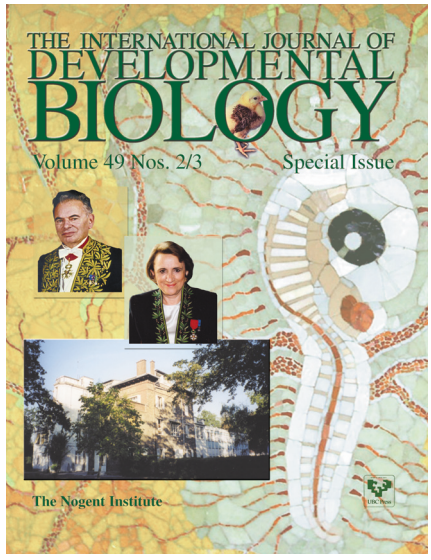
Retinal stem cells and regeneration
Ala Moshiri, Jennie Close and Thomas A. Reh
Int. J. Dev. Biol. (2004) 48: 1003-1014
http://www.intjdevbiol.com/web/paper/041870am

Notch activity is required to maintain floorplate identity and to control neurogenesis in the chick hindbrain and spinal cord
Isabelle le Roux, Julian Lewis and David Ish-Horowicz
Int. J. Dev. Biol. (2003) 47: 263-272
http://www.intjdevbiol.com/web/paper/12755331

Early neurogenesis in Amniote vertebrates
N M Le Douarin
Int. J. Dev. Biol. (2001) 45: 373-378
http://www.intjdevbiol.com/web/paper/11291868

Complementary roles of the insulin family of factors and receptors in early development and neurogenesis
F De Pablo, C Alarcón, B Díaz, M García-De Lacoba, A López-Carranza, A V Morales, B Pimentel, J Serna and E J De la Rosa
Int. J. Dev. Biol. (1996) 40: S109-S110
http://www.intjdevbiol.com/web/paper/9087719

Transplantations of the chick eye anlage reveal an early determination of nasotemporal polarity
D Dütting and S U Meyer
Int. J. Dev. Biol. (1995) 39: 921-931
http://www.intjdevbiol.com/web/paper/8901194



5 yr ISI Impact Factor (2016) = 2.421