Third Report on Chicken Genes and Chromosomes 2015

M. Schmid
University of Würzburg

J. Smith
The Roslin Institute and R(D)SVS

D. W. Burt
The Roslin Institute and R(D)SVS

B. L. Aken
European Bioinformatics Institute, EMBL

P. B. Antin
University of Arizona

Follow this and additional works at: http://lib.dr.iastate.edu/ans_pubs

Part of the Agriculture Commons, Genetics and Genomics Commons, and the Poultry or Avian Science Commons

The complete bibliographic information for this item can be found at http://lib.dr.iastate.edu/ans_pubs/175. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.

This Article is brought to you for free and open access by the Animal Science at Iowa State University Digital Repository. It has been accepted for inclusion in Animal Science Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Abstract
Following on from the First Report on Chicken Genes and Chromosomes [Schmid et al., 2000] and the Second Report in 2005 [Schmid et al., 2005], we are pleased to publish this long-awaited Third Report on the latest developments in chicken genomics. The First Report highlighted the availability of genetic and physical maps, while the Second Report was published as the chicken genome sequence was released. This report comes at a time of huge technological advances (particularly in sequencing methodologies) which have allowed us to examine the chicken genome in detail not possible until now. This has also heralded an explosion in avian genomics, with the current availability of more than 48 bird genomes [Zhang G et al., 2014b; Eöry et al., 2015], with many more planned.

Disciplines
Agriculture | Animal Sciences | Genetics and Genomics | Poultry or Avian Science

Comments
This article is from Cytogenic and Genome Research 145 (2015): 78, doi:10.1159/000430927.

Rights
Works produced by employees of the U.S. Government as part of their official duties are not copyrighted within the U.S. The content of this document is not copyrighted.

Authors

This article is available at Iowa State University Digital Repository: http://lib.dr.iastate.edu/ans_pubs/175
Third Report on Chicken Genes and Chromosomes 2015

Prepared by


1Department of Human Genetics, University of Würzburg, Würzburg, and 2Bioinformatics Group, Department of Computer Science and Interdisciplinary Centre for Bioinformatics, University of Leipzig, Leipzig, Germany; 3Division of Genetics and Genomics, The Roslin Institute and R(D)SVS, and 4Division of Infection and Immunity, The Roslin Institute, University of Edinburgh, Edinburgh, 5European Bioinformatics Institute, EMBL, and 6Wellcome Trust Sanger Institute, Hinxton, 7Medical Research Council Laboratory of Molecular Biology, Cambridge, 8Department of Comparative Biomedical Sciences, Royal Veterinary College, University of London, and 9Section of Virology, Department of Medicine, Imperial College London, London, 10School of Life Sciences, University of Nottingham, Nottingham,

Supported by Biotechnology and Biological Sciences Research Council (BBSRC); USDA Agriculture and Food Research Initiative Competitive Grant; NIH Intramural Research Program of the National Institute of Allergy and Infectious Diseases; European Union FP-7 project QUANTOMICS; Deutsche Forschungsgemeinschaft (DFG); Swiss National Science Foundation; Lady Tata Memorial Trust; French ‘Agence Nationale de la Recherche’ Epilfèbre; Intramural Research Program of the NIEHS, NIH; Department for International Development (DFID).

Correspondence to:

Michael Schmid
Department of Human Genetics, University of Würzburg
Biozentrum, Am Hubland
DE-97074 Würzburg (Germany)
E-Mail m.schmid@biozentrum.uni-wuerzburg.de

Jacqueline Smith
Division of Genetics and Genomics
The Roslin Institute, University of Edinburgh
Edinburgh EH25 9RG (UK)
E-Mail jacqueline.smith@roslin.ed.ac.uk

David W. Burt
Division of Genetics and Genomics
The Roslin Institute, University of Edinburgh
Edinburgh EH25 9RG (UK)
E-Mail dave.burt@roslin.ed.ac.uk
The Chicken Leads the Way in Avian Genomics
(Prepared by J. Smith)

Following on from the First Report on Chicken Genes and Chromosomes [Schmid et al., 2000] and the Second Report in 2005 [Schmid et al., 2005], we are pleased to publish this long-awaited Third Report on the latest developments in chicken genomics. The First Report highlighted the availability of genetic and physical maps, while the Second Report was published as the chicken genome sequence was released. This report comes at a time of huge technological advances (particularly in sequencing methodologies) which have allowed us to examine the chicken genome in detail not possible until now. This has also heralded an explosion in avian genomics, with the current availability of more than 48 bird genomes [Zhang G et al., 2014b; Eöry et al., 2015], with many more planned.

The chicken has long been a model organism for genetic and developmental studies and now takes its place as a model genome, opening up the fields of phylogenetics and comparative genomics like never before. This report summarizes the current efforts to complete the gaps in the genome and describes the progress of genomic annotation, particularly with respect to noncoding RNAs and genetic variants. Reviews of comparative genomics, avian evolution and sex determination are also given. Transcriptomic case studies are described as are developments in epigenetic studies. We also report on the development of the National Avian Research Facility (http://www.narf.ac.uk/) which is home to various chicken resources including antibodies, chicken lines and transgenic protocols and which hosts information currently available on the sequenced avian genomes.
In 2004, the assembly and annotation of the draft chicken genome assembly was reported [International Chicken Genome Sequencing Consortium, 2004], which represented the first avian (and livestock) genome to be sequenced. The chicken genome assembly has proved an invaluable research tool for the avian research community, and efforts to improve the genome and its annotation have continued.

**Current Status**

The initial draft of the chicken genome was based on a single female red junglefowl bird (UDC 001) and was assembled using a whole-genome sequencing strategy based on Sanger sequencing (7× coverage), including BAC, fosmid and plasmid paired-end reads. Additional data sources have been incrementally added, including high-resolution genetic linkage maps [Groenen et al., 2009], BAC maps [Ren et al., 2003], radiation hybrid maps [Morisson et al., 2007], targeted genome sequences (gaps and a BAC-based sequence of the Z sex chromosome [Bellott et al., 2010]) and using next-generation sequencing data (14× coverage of 454 and 74× coverage of Illumina reads) [Ye et al., 2011]. In addition, improvements in genome assembly software have increased the quality of these assemblies, for example, removing artefacts due to allelic variants and multigene families [Ye et al., 2011].

The chicken karyotype comprises 39 chromosome pairs as follows: 10 pairs of large autosomes (chromosomes 1–10), 28 pairs of microchromosomes (chromosomes 11–38), and a pair of sex chromosomes (chromosomes W and Z). Chromosome Z is a large chromosome and present as a pair in males. Chromosome W is a microchromosome and present only in the heterogametic (ZW) females [Masabanda et al., 2004]. The current genome assembly, Gallus_gallus-4.0 (Galgal4, GCA_000002315.2) covers 1.03 Gb or 96% of the predicted genome size. This includes assembled sequence for all 10 large autosomes, 19 of the 28 microchromosomes, and both sex chromosomes (tables 1, 2). In addition, it includes sequence that is yet to be assigned to an assembled chromosome (linkage groups LG1–2 to LG20, LG22, and LGX) and 14,093 unplaced scaffolds. The contig N50 is 280 kb (n = 27,041), and the scaffold N50 is 12.9 Mb (n = 16,847). The contig N50 is defined as the contig length at which 50% of the total number of bases in the assembly are in contigs of that length or greater.

Overlaps with available cDNA clones suggest that less than 5% of coding genes are missing from the final assembly; however, gene duplications and GC-rich sequences are still difficult regions to assemble and pose a particular problem for chicken. Chromosome 16 (which contains the MHC region) and the W sex chromosome are a rich source of duplicated genes and repeats, and are poorly represented. In addition, sequenced contigs have not been assigned to the smallest microchromosomes (GGA29–31 and GGA33–38). Further work is needed to complete the chicken genome by sequencing gaps (so far specific gaps of 18 Mb and 27 Mb have been defined on the ordered and unordered chromosomes, respectively) and missing chromosomes.

The full utility of the chicken genome requires a comprehensive annotation; enumerating and defining all the transcribed and regulatory regions. Until recently, the main tools of gene prediction and annotation for species with few resources (such as full-length cDNA sequences in human and mouse) were based on gene homology with sequences from other species, for example the Ensembl
Table 2. Genome assembly statistics for the current chicken reference genome assembly (Galgal4, GCA_000002315.2)

<table>
<thead>
<tr>
<th>Assembly unit</th>
<th>Sequence name</th>
<th>INSDC accession</th>
<th>Total length (including gaps)</th>
<th>Present in Galgal4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome 1</td>
<td>CM000093.3</td>
<td>195,276,750</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 2</td>
<td>CM000094.3</td>
<td>148,809,762</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 3</td>
<td>CM000095.3</td>
<td>110,447,801</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 4</td>
<td>CM000096.3</td>
<td>90,216,835</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 5</td>
<td>CM000097.3</td>
<td>59,580,361</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 6</td>
<td>CM000098.3</td>
<td>34,951,654</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 7</td>
<td>CM000099.3</td>
<td>36,245,040</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 8</td>
<td>CM0000100.3</td>
<td>28,767,244</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 9</td>
<td>CM0000101.3</td>
<td>23,441,680</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 10</td>
<td>CM0000102.3</td>
<td>19,911,089</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 11</td>
<td>CM0000103.3</td>
<td>19,401,079</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 12</td>
<td>CM0000104.3</td>
<td>19,897,011</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 13</td>
<td>CM0000105.3</td>
<td>17,760,035</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 14</td>
<td>CM0000106.3</td>
<td>15,161,805</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 15</td>
<td>CM0000107.3</td>
<td>12,656,803</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 16</td>
<td>CM0000108.3</td>
<td>535,270</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 17</td>
<td>CM0000109.3</td>
<td>10,454,150</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 18</td>
<td>CM0000110.3</td>
<td>11,219,875</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 19</td>
<td>CM0000111.3</td>
<td>9,983,394</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 20</td>
<td>CM0000112.3</td>
<td>14,302,601</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 21</td>
<td>CM0000113.3</td>
<td>6,802,778</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 22</td>
<td>CM0000114.3</td>
<td>4,081,097</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 23</td>
<td>CM0000115.3</td>
<td>5,723,239</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 24</td>
<td>CM0000116.3</td>
<td>6,323,281</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 25</td>
<td>CM0000124.3</td>
<td>2,191,139</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 26</td>
<td>CM0000117.3</td>
<td>5,329,985</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 27</td>
<td>CM0000118.3</td>
<td>5,209,285</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 28</td>
<td>CM0000119.3</td>
<td>4,742,627</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 29</td>
<td>N/A</td>
<td>N/A</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Chromosome 30</td>
<td>N/A</td>
<td>N/A</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Chromosome 31</td>
<td>N/A</td>
<td>N/A</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Chromosome 32</td>
<td>CM0000120.2</td>
<td>1,028</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Chromosome 33</td>
<td>N/A</td>
<td>N/A</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Chromosome 34</td>
<td>N/A</td>
<td>N/A</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Chromosome 35</td>
<td>N/A</td>
<td>N/A</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Chromosome 36</td>
<td>N/A</td>
<td>N/A</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Chromosome 37</td>
<td>N/A</td>
<td>N/A</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Chromosome 38</td>
<td>N/A</td>
<td>N/A</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Sex chromosome W</td>
<td>CM000121.3</td>
<td>1,248,174</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Sex chromosome Z</td>
<td>CM000122.3</td>
<td>82,363,669</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Linkage group LGE22C19W28_E50C23</td>
<td>CM000123.3</td>
<td>965,146</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Linkage group LGE64</td>
<td>CM000367.2</td>
<td>799,899</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Unplaced scaffolds</td>
<td></td>
<td></td>
<td>(14,093 sequences)</td>
<td></td>
</tr>
</tbody>
</table>
gene annotation system [Cunningham et al., 2015]. This approach has been successful at defining protein-coding regions, but not so successful when translated across the large evolutionary distance from mammals to birds, in particular for defining rapidly evolving genes such as innate immune genes or noncoding genes, such as microRNAs (miRNAs) and long noncoding RNAs (lncRNAs). Developments in RNA-seq have changed this landscape, capable of both defining and quantifying small and large RNAs, either coding or noncoding [Robertson et al., 2010]. The Ensembl gene annotation pipeline has been extended to incorporate these types of data to improve gene predictions for both coding and noncoding genes [Flicek et al., 2011, 2014].

In 2012, an International Avian RNAseq Consortium [Smith et al., this report] was formed and agreed to share data from a wide range of tissues to be used for gene discovery. Collaboration between this Consortium and Ensembl completed a new annotation of the chicken protein-coding genes (table 1) and the RNA-seq data provided direct experimental evidence for 94% of the predicted transcripts (table 3). All the gene models and RNA-seq data are available to end-users from the Ensembl Genome Browser (www.ensembl.org). This combined evidence-based approach available in Ensembl release 78 predicted 15,508 protein-coding genes in the chicken, which compares with 20,364 (13,990 chicken orthologs, 15,197 human orthologs) and 22,606 (13,741 chicken orthologs, 15,072 mouse orthologs) predictions in human and mouse, respectively. These comparisons suggest that mammals may have ∼6,000 protein-coding genes not found in the chicken. This may be an artefact or a real biological finding, which remains to be tested. Interestingly, it appears that ∼274 genes are truly missing in the avian lineage [Lovell et al., 2014], yet are conserved in human, thus potentially serving as naturally evolved models to study the physiological consequence. More limited predictions are available for noncoding RNAs (ncRNAs) in the chicken, which include 1,534 small ncRNA genes in 31 distinct families: 1,049 miRNAs, 29 rRNAs, 227 small nucleolar RNAs ( snoRNAs), 79 small nuclear RNAs ( snRNAs), and a further 150 (27 families) miscellaneous RNAs. This compares with 9,673 small ncRNAs and 14,817 IncRNAs in human, and 5,860 small ncRNAs and 5,762 IncRNAs in mouse. These comparisons suggest that many more ncRNAs in the chicken are still to be discovered, in particular the IncRNA class. Only 42 pseudogenes were detected in the chicken compared to 14,415 and 8,015 predictions in human and mouse, respectively. The reason for the very low number in the chicken when compared to mammals may be the sequence specificity of reverse transcription by avian LINES [International Chicken Genome Sequencing Consortium, 2004]. Mammalian LINES are more promiscuous and able to retrotranspose most mRNAs.

**Ongoing Improvements**

To ensure the chicken’s continued utility as a model organism, we are undertaking improvements to the chicken genome assembly. Several different approaches are currently being used to further improve the assembly, including single molecule sequencing from PacBio [W. Warren, unpubl. results] to extend reads and fill in gaps, and chromosome sorting [Burt and Ferguson-Smith, unpubl. results] to target sequencing of the tiny chromosomes (GGA29–31 and GGA33–38), which are currently missing from the genome assembly. Microchromosomes are gene dense, and so even these tiny chromosomes are likely to be of significance [Burt, 2002; International Chicken Genome Sequencing Consortium, 2004].

Within the assembled chromosomes, it is important to ensure that the ordering and orientation of the contigs is correct. Optical mapping (http://en.wikipedia.org/wiki/Optical_mapping) to create high-resolution, chromosome-wide restriction maps based on rare-cutters is used to complement the creation of sequenced contigs [Burt and Dunn, unpubl. results]. Combining these methods will provide accurate data on: (1) confirmation of chromosome assemblies, (2) defining gaps and duplicated regions, (3) assembly of sequence contigs for microchromosomes, and (4) defining a route for completion of the

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Transcript models</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>13,326</td>
<td>74.2</td>
</tr>
<tr>
<td>Heart</td>
<td>10,710</td>
<td>59.7</td>
</tr>
<tr>
<td>Testes</td>
<td>14,600</td>
<td>81.3</td>
</tr>
<tr>
<td>Bone-derived macrophage</td>
<td>11,339</td>
<td>63.2</td>
</tr>
<tr>
<td>Embryo somites</td>
<td>11,350</td>
<td>63.2</td>
</tr>
<tr>
<td>DF1 cell line</td>
<td>9,266</td>
<td>51.6</td>
</tr>
<tr>
<td>Kidney</td>
<td>10,688</td>
<td>59.5</td>
</tr>
<tr>
<td>Breast muscle</td>
<td>9,161</td>
<td>51.0</td>
</tr>
<tr>
<td>Brain cerebellum</td>
<td>10,591</td>
<td>59.0</td>
</tr>
<tr>
<td>Brain</td>
<td>10,890</td>
<td>60.7</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>11,324</td>
<td>63.1</td>
</tr>
<tr>
<td><strong>Total expressed</strong></td>
<td><strong>16,915</strong></td>
<td><strong>94.2</strong></td>
</tr>
</tbody>
</table>

**Table 3. Summary of chicken RNA-seq expression profiles**
Future Directions

As the chicken genome assembly improves, we will ensure that genome annotation improves and expands. For genome assembly, long read single-molecule sequencing has already shown great future application, with de novo assembled N50 contig length reaching >1 Mb and 90% of the genome assembled in >250-kb length contigs [W. Warren, unpubl. results]. This long-range contiguity promises to achieve the goal of mostly complete copies of each chicken chromosome. In addition, the physical assignment of sequenced microchromosomes will be necessary to form a definition. For gene annotation, PacBio full-length sequences are in progress and will be used to identify missing genes, join gene fragments, define transcription start/end sites and annotate novel transcript isoforms. These annotations will include both coding and noncoding RNA genes. Significant progress has been made in establishing the Bird Gene Nomenclature Committee but more needs to be done, with a greater role for the avian research community [Burt et al., 2009]. Knowledge of gene regulation is a critical component of genomics research. We will pursue methods for identifying transcription factor binding sites, ChIP-seq histone marks, DNA methylation sites and define functional elements, such as enhancers, silencers and promoters (http://www.animalgenome.org/community/FAANG/). As more individual chicken genomes are sequenced, structural and single nucleotide variations will be identified and submitted to the public archives. With a well annotated chicken genome, these genome variants will be associated with functional elements, if any, such as genes or regulatory elements. Quantitative trait loci (QTLs) and genome-wide associations are important tools for identifying regions of the genome associated with specific phenotypes. Over 4,300 QTLs have been identified for chicken (http://www.animalgenome.org/cgi-bin/QTLdb/GG/index). It is anticipated that, with growing variation and expression data, expression QTLs will be identified for chicken in the future. The final goal will be to associate functional variants with specific phenotypes.

The Avian RNAseq Consortium: A Community Effort to Annotate the Chicken Genome

(Prepared by J. Smith, D.W. Burt, and the Avian RNAseq Consortium)

Publication of the chicken genome sequence in 2004 [International Chicken Genome Sequencing Consortium, 2004] highlighted the beginning of a revolution in avian genomics. Progression of DNA sequencing technologies and data handling capabilities has also meant that genome sequencing and assembly is now a relatively simple, fast and inexpensive procedure. The success seen with the chicken genome was soon followed by the completion of the zebra finch genome [Warren et al., 2010], an important model for neurobiology [Clayton et al., 2009], again based on Sanger sequencing. In recent years, the rapid advances in next-generation sequencing (NGS) technologies, hardware and software have meant that many more genomes can now be sequenced faster and cheaper than ever before [Metzker, 2010]. The first avian genome to be sequenced by NGS methods was the turkey [Dalloul et al., 2010], which was also integrated with genetic and physical maps, thus providing an assembly of high quality, even at the chromosome level. Recently, NGS has been used to sequence the genomes of a further 42 avian species, as part of the G10K initiative [Genome 10K Community of Scientists, 2009]. In addition, there have also been 15 other genome assemblies recently published, each with a focus on a unique aspect of avian biology, including the Japanese quail (domestication) [Kawahara-Miki et al., 2013], Puerto Rican parrot (speciation) [Oleksyak et al., 2012], scarlet macaw (speech, intelligence and longevity) [Seabury et al., 2013], medium and large ground finches (speciation) [Parker et al., 2012; Rands et al., 2013], collared and pied flycatchers (speciation) [Ellegren et al., 2012], peregrine and saker falcons (predatory lifestyle) [Zhan et al., 2013], rock pigeon (domestication) [Shapiro et al., 2013], the ground tit (adaptation to high altitude) [Cai et al., 2013], and the northern bobwhite (population history) [Halley et al., 2014]. Through November 2014, there are currently 56 avian genome sequences completed, either published or in press (table 4). A new project, B10K (web.bioinfodata.org/B10K), proposes sequencing of all avian genomes; this would include all 40 orders, 231 families, 2,268 genera and 10,476
species of birds. The chicken genome remains the best-described genome and is used as a reference upon which the annotations of other assemblies are based. Assembly and annotation of the genome continues to improve. However, gaps and unaligned regions remain (particularly for some of the smallest microchromosomes), which can cause practical problems in the analysis and annotation of important loci, especially for those representing gene families. Other approaches, such as long reads generated by Pacific Biosciences (PacBio) sequencing, chromosome sorting and optical maps are being used to resolve these assembly issues [Warren and Burt, pers. com- m Sep 2011, it now includes 50 people from 27 different institu- tions (fig. 1) who have contributed to the effort to create a detailed annotation of the chicken genome by either providing RNA-seq data or by helping to analyze the combined data.

We currently have 21 different data sets (representing more than 1.5 Tb of data) with more data being added (fig. 2; table 5). These data represent transcriptome sequences from many different chicken tissues and from many different experimental conditions, including several infection/disease cases. These data were submitted to public archives, collected at The Roslin Institute and then passed on to the Ensembl team who used the information passed on to the Ensembl team who used the information

<table>
<thead>
<tr>
<th>Table 4. Avian species with sequenced genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbreviation</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>ACACH</td>
</tr>
<tr>
<td>AMAVI</td>
</tr>
<tr>
<td>ANAPL</td>
</tr>
<tr>
<td>APAVI</td>
</tr>
<tr>
<td>APTFO</td>
</tr>
<tr>
<td>ARAMA</td>
</tr>
<tr>
<td>BALRE</td>
</tr>
<tr>
<td>BURCH</td>
</tr>
<tr>
<td>CALAN</td>
</tr>
<tr>
<td>CAPCA</td>
</tr>
<tr>
<td>CARCR</td>
</tr>
<tr>
<td>CATAU</td>
</tr>
<tr>
<td>CHAPE</td>
</tr>
<tr>
<td>CHAVO</td>
</tr>
<tr>
<td>CHLMA</td>
</tr>
<tr>
<td>COLLI</td>
</tr>
<tr>
<td>COLST</td>
</tr>
<tr>
<td>COLVI</td>
</tr>
<tr>
<td>CORBR</td>
</tr>
<tr>
<td>COTJA</td>
</tr>
<tr>
<td>CUCCA</td>
</tr>
<tr>
<td>EGRA</td>
</tr>
<tr>
<td>EURHE</td>
</tr>
<tr>
<td>FALCH</td>
</tr>
<tr>
<td>FALPE</td>
</tr>
<tr>
<td>FICAL</td>
</tr>
<tr>
<td>FICHY</td>
</tr>
<tr>
<td>FULGL</td>
</tr>
</tbody>
</table>
**Fig. 1.** Worldwide locations of current RNAseq Consortium members.

**Fig. 2.** Comparison of the different relative amounts of RNA-seq data from each tissue. Tissues from different data providers are shown separately as they have all been subject to different treatments/stimuli. Numbered data sets are as referred to in table 5.
This new annotation includes 15,495 protein-coding genes, 1,049 miRNAs, 456 ncRNAs and 42 pseudogenes. This gene build is primarily concerned with coding genes, but there are many more noncoding genes which remain un-annotated. Consortium members have analyzed the RNA-seq data for long ncRNAs (lncRNAs) [manuscript in preparation], snoRNAs [Gardner et al., 2015] and other features of interest. Around 14,000 potential lncRNA genes have thus far been identified from the RNA-seq data. Ensembl release 71 marked a significant update in the annotation of the chicken genome with gene models based on experimental data. Table 7 shows how this gene

Table 5. Details of RNA-seq data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>Description of data</th>
<th>Reads, bp</th>
<th>Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Antin whole embryo</td>
<td>35</td>
<td>Illumina SE</td>
</tr>
<tr>
<td>2</td>
<td>Blackshear LPS-stimulated macrophages vs. control CEFs</td>
<td>51</td>
<td>Illumina PE</td>
</tr>
<tr>
<td>3</td>
<td>Burgess/McCarthy miRNA from various red junglefowl tissues (adrenal gland, adipose, cerebellum, cerebrum, testis, ovary, heart, hypothalamus, kidney, liver, lung, breast muscle, sciatic nerve, proventriculus, spleen)</td>
<td>50</td>
<td>Illumina SE</td>
</tr>
<tr>
<td>4</td>
<td>Burt/Smith spleen: infectious bursal disease virus infected vs. control lung and ileum: avian influenza infected vs. control (high path H5N1 and low) lung short-read data</td>
<td>36 36 25</td>
<td>Illumina SE Illumina SE Illumina SE</td>
</tr>
<tr>
<td>5</td>
<td>de Koning/Dunn/McCormack bone from 70-week-old Leghorns</td>
<td>100</td>
<td>Illumina PE</td>
</tr>
<tr>
<td>6</td>
<td>Frésard/Pitel brain from high-mobility sperm vs. low-mobility sperm pooled whole embryos (stage HH26)</td>
<td>380–400 100</td>
<td>Roche 454 Illumina PE</td>
</tr>
<tr>
<td>7</td>
<td>Froman/Rhoads testes: roosters with high-mobility sperm vs. low-mobility sperm embryo, DF1 cell line and bone marrow-derived macrophages</td>
<td>35 100</td>
<td>Illumina SE Illumina PE</td>
</tr>
<tr>
<td>8</td>
<td>Garceau/Hume Newcastle disease virus infection vs. control (trachea and lung epithelial cells) DT40 cells</td>
<td>50 60</td>
<td>SOLiD SE Illumina PE</td>
</tr>
<tr>
<td>9</td>
<td>Kaiser bone marrow-derived dendritic cells from 6-week-old birds (control, DCs + LPS); BMDMs from 6-week-old birds (control, BMDMs + LPS); heterophils isolated from blood of day-old chicks (control, het + LPS)</td>
<td>100</td>
<td>Illumina PE</td>
</tr>
<tr>
<td>10</td>
<td>Lagarrigue/Roux abdominal adipose tissue and liver tissue from 14-week-old broilers</td>
<td>100</td>
<td>Illumina PE</td>
</tr>
<tr>
<td>11</td>
<td>Lamont livers of 8 individuals, 28-day-old broiler males: 4 controls, 4 heat-stressed</td>
<td>100</td>
<td>Illumina SE</td>
</tr>
<tr>
<td>12</td>
<td>Munsterberg/Pais somites injected with anti-mir206 vs. non-injected</td>
<td>50</td>
<td>Illumina PE</td>
</tr>
<tr>
<td>13</td>
<td>Schmidt tissues from heat-stressed and control birds (liver, brain, spleen, thymus, bursa, kidney, ileum, jejunum, duodenum, ovary, heart, breast, monocyte)</td>
<td>42–50</td>
<td>Illumina SE</td>
</tr>
<tr>
<td>14</td>
<td>Schwartz/Ulitsky whole embryo stages – HH4/5; HH11; HH14/15; HH21/22; HH25/26; HH32; HH36 – stranded</td>
<td>80/100</td>
<td>Illumina PE</td>
</tr>
<tr>
<td>15</td>
<td>Skinner CEFs</td>
<td>100</td>
<td>Illumina PE</td>
</tr>
<tr>
<td>16</td>
<td>Wang/Zhou lung from Fayoumi and Leghorn birds: control and H5N3 infected</td>
<td>75</td>
<td>Illumina SE</td>
</tr>
</tbody>
</table>

BMDMs = Bone marrow-derived macrophages; CEF = chicken embryo fibroblasts; DC = dendritic cell; het = heterophils; LPS = lipopolysaccharide; PE = paired end; SE = single end.

Table 6. Ensembl 71 annotation statistics of 17,108 genes

<table>
<thead>
<tr>
<th>Genes</th>
<th>Description</th>
<th>Biotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>15,495</td>
<td>Ensembl</td>
<td>protein-coding</td>
</tr>
<tr>
<td>42</td>
<td>Ensembl</td>
<td>pseudogene</td>
</tr>
<tr>
<td>2</td>
<td>mt_genbank_import</td>
<td>Mt_rRNA</td>
</tr>
<tr>
<td>22</td>
<td>mt_genbank_import</td>
<td>Mt_trRNA</td>
</tr>
<tr>
<td>13</td>
<td>mt_genbank_import</td>
<td>protein-coding</td>
</tr>
<tr>
<td>1,049</td>
<td>ncRNA</td>
<td>miRNA</td>
</tr>
<tr>
<td>150</td>
<td>ncRNA</td>
<td>misc_RNA</td>
</tr>
<tr>
<td>29</td>
<td>ncRNA</td>
<td>rRNA</td>
</tr>
<tr>
<td>227</td>
<td>ncRNA</td>
<td>snoRNA</td>
</tr>
<tr>
<td>79</td>
<td>ncRNA</td>
<td>snRNA</td>
</tr>
</tbody>
</table>

This new annotation includes 15,495 protein-coding genes, 1,049 miRNAs, 456 ncRNAs and 42 pseudogenes. This gene build is primarily concerned with coding genes, but there are many more noncoding genes which remain un-annotated. Consortium members have analyzed the RNA-seq data for long ncRNAs (lncRNAs) [manuscript in preparation], snoRNAs [Gardner et al., 2015] and other features of interest. Around 14,000 potential lncRNA genes have thus far been identified from the RNA-seq data. Ensembl release 71 marked a significant update in the annotation of the chicken genome with gene models based on experimental data. Table 7 shows how this gene
build was the first to use the Galgal4 assembly and, through the use of RNA-seq data, was able to help remove assembly errors and reduce the number of predicted gene transcripts by identifying incorrectly predicted genes from previous builds and improving identification of short ncRNAs. The significance of this community effort is indicated by the fact that the current Ensembl 77 gene set has not changed since Ensembl release 71, with only difference being reflected in the total number of base pairs. This is due to the correction of one particular scaffold on the Z chromosome (which was reflected in Ensembl release 74).

The availability of these data will allow for the further development of a chicken expression atlas by providing the ability to analyze transcript levels across tissues (http://geneatlas.arl.arizona.edu/). It will also enable development of exon capture technology for the chicken and has already proved to be of great use in helping annotate the other avian genomes which have now been sequenced. On-going collection of RNA-seq data will remain a valuable resource as genomic analysis of avian species continues to expand.

**Methods**

**Ensembl Gene Build**

The chicken gene build from Ensembl release 71 was done using standard Ensembl annotation procedures and pipelines, mostly focussed on protein-coding sequences. Briefly, vertebrate UniProtKB proteins were downloaded and aligned to the Galgal4 (GCA_000002315.2) assembly with Genewise (http://www.ebi.ac.uk/Tools/psa/genewise/) in order to annotate protein-coding models. UniProt assigns protein existence (PE) levels to each of their protein sequences. The PE level indicates the type of evidence that supports the existence of a protein sequence, and can range from PE 1 (‘Experimental evidence at protein level’) to PE 5 (‘Protein uncertain’). Only PE 1 and PE 2 proteins from UniProtKB were used for the Genewise step. RNA-seq models were annotated using the Ensembl RNA-seq pipeline and models from both the Genewise and the RNA-seq pipelines were used as input for the final protein-coding gene set. Chicken cDNAs and RNA-seq models were also used to add UTRs in the 5′ and 3′ regions. Some missing gene models were recovered by aligning chicken, zebra finch and turkey translations from Ensembl release 65 (December 2011) to the new chicken genome assembly.

**RNA-seq Gene Models**

Raw reads were aligned to the genome using BWA [Li and Durbin, 2009] to identify regions of the genome that are actively transcribed. The results from all tissues were used to create one set of alignment blocks roughly corresponding to exons. Read pairing information was used to group exons into approximate transcript structures called proto-transcripts. Next, partially mapped reads from both the merged (combined data from all tissue samples) and individual tissues were re-aligned to the proto-transcripts using Exonerate [Slater and Birney, 2005], to create a merged and tissue-specific sets of spliced alignments. For each gene, merged and tissue-specific transcript isoforms were computed from all observed exon-intron combinations, and only the best supported isoform was reported.

**Annotation of ncRNAs**

The following ncRNA gene types were annotated: ribosomal RNA (rRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), microRNA precursors (miRNA), miscellaneous other RNA (misc_RNA). Most ncRNA genes in Ensembl are annotated by first aligning genomic sequence against RFAM [Burge et al., 2013], using BLASTN (parameters W = 12, B = 10000, V = 10000, -hsmax 0 -gsmx 0 -kaps -cups = 1), to identify likely ncRNA loci. The BLAST [Altschul et al., 1990] hits are clustered, filtered for hits above 70% coverage, and used
to seed an Infernal [Nawrocki and Eddy, 2013] search with the corresponding RFAM covariance model, to measure the probability that these targets can fold into the structures required. Infernal’s cmsearch is used to build ncRNA models. miRNAs are predicted by BLASTN (default parameters) of genomic sequence slices against miRBase [Kozomara and Griffiths-Jones, 2014] sequences. The BLAST hits are clustered, filtered to select the alignment with the lowest p value when more than one sequence aligns at the same genomic position, and the aligned genomic sequence is checked for possible secondary structure using RNAFold [Hofacker et al., 1994]. If evidence is found that the genomic sequence could form a stable hairpin structure, the locus is used to create a miRNA gene model. Transfer RNAs (tRNAs) were annotated as part of the raw compute process using tRNAscan-SE with default parameters [Schattner et al., 2005]. All results for tRNAscan-SE are available through Ensembl; the results are not included in the Ensembl gene set because they are not annotated using the standard evidence-based approach (i.e. by aligning biological sequences to the genome) that is used to annotate other Ensembl gene models.

**Summary**

The availability of this collection of chicken RNA-seq data within the consortium has allowed:

- annotation of 17,108 chicken genes, 15,495 of which are protein-coding (Ensembl 71)
- identification of ~14,000 putative IncRNA genes (with >23,000 transcripts suggested)
- annotation of miRNAs, snoRNAs, and other ncRNAs
- future generation of an expression atlas which will allow comparisons of expression over many tissues
- an improved avian reference for comparative analyses with 48 other avian genomes [Zhang G et al., 2014b]

**Future Directions**

The next stage in progressing annotation of the avian genomes will concentrate on the analysis of data generated by PacBio sequencing, in conjunction with stranded RNA-seq data from a wide variety of tissues. PacBio technology allows for very long read lengths, producing reads with average lengths of 4,200–8,500 bp, with the longest reads over 30,000 bp. This enables sequencing of full-length transcripts. Extremely high accuracy means that de novo assembly of genomes and detection of variants with >99.999% accuracy is possible. Individual molecules can also be sequenced at 99% reliability. The high sensitivity of the method also means that minor variants can be detected even when they have a frequency of <0.1% [http://www.pacificbiosciences.com/products/smrt-technology/smrt-sequencing-advantage/]. We currently have brain transcriptomic PacBio data generated from a female Brown Leghorn J-line chicken [Blyth and Sang, 1960]. This will be analyzed alongside stranded RNA-seq data that has been generated from 21 different tissues. The advantage of using strand-specific sequence information is that it provides an insight into antisense transcripts and their potential role in regulation and strand information of ncRNAs as well as aiding in accurately quantifying overlapping transcripts. It is particularly useful for finding unannotated genes and ncRNAs. This strategy should allow us to obtain full-length transcript sequences, identify novel transcripts and low-level transcripts, map transcription start and stop sites and confirm further ncRNAs.

**Get Involved**

If you are interested in helping further the annotation of the avian genomes, and you can provide avian RNA-seq data or can help with the analysis of such data, then please contact Jacqueline Smith (jacqueline.smith@roslin.ed.ac.uk) or Dave Burt (dave.burt@roslin.ed.ac.uk).

**Avian RNAseq Consortium Members**

J. Smith, I. Dunn, V. Garceau, D. Hume, P. Kaiser, R. Kuo, H. McCormack, D. Burt (Roslin Institute); A. Cooksey, F. McCarthy, P.B. Antin, S. Burgess (University of Arizona); A. Münsterberg, H. Pais (University of East Anglia); A. Oler (NIH National Institute of Allergy and Infectious Diseases); S. Searle (Wellcome Trust Sanger Institute); P. Flicek, B.L. Aken, R. Nag (European Molecular Biology Laboratory, European Bioinformatics Institute and Wellcome Trust Sanger Institute); C. Schmidt (University of Delaware); C. Klopp (INRA Toulouse); P. Prieta Barja, I. Erb, D. Kedra, C. Notredame (CRG, Barcelona); D. Froman (Oregon State University); D.-J. de Koning (Swedish University of Agricultural Sciences, Uppsala); D. Rhoads (University of Arkansas); I. Ulitsky (Weizmann Institute of Science, Rehovot); J. Häslar, M. Neuberger (in memoriam) (MRC, Cambridge); L. Frébard, F. Pitelet (INRA, Auzville); M. Fasold, P. Stadler (University of Leipzig); M. Schwartz (Harvard Medical School); M. Skinner (Imperial College London); O. Hanotte (University of Nottingham); P. Blackshear (NIEHS, North Carolina); S. Lagarrigue, P.-F. Roux (INRA Agrocampus Ouest); T. Derrien (University of Rennes); S. Ommeh (Jomo-Kenyatta University of Agriculture and Technology, Kenya); S. Muljo (NIH NIAID, Bethesda); S.
Kemp, H. Noyes (University of Liverpool); S. Lamont, D. Fleming, D.J. Coble (Iowa State University); Y. Wang, H. Zhou (UC Davis).

Availability of RNA-seq Data

Data have been submitted to the public databases under the following accession numbers:
- Antin/Burgess/McCarthy/Schmidt data: BioProject ID: PRJNA204941 (Sequence Read Archive); Blackshear data: PRJEB1406 (European Nucleotide Archive); Burt/Smith data: E-MTAB-2908, E-MTAB-2909, E-MTAB-2910 (Array Express); de Koning/Dunn/McCormack data: E-MTAB-2737 (Array Express); Frésard/Pitel data: SRP033603 (Sequence Read Archive); Froman/Rhoads data: BioProject ID: PRJNA247673 (Sequence Read Archive); Garceau/Hume data: E-MTAB-3048 (Array Express); Hanotte/Kemp/Noyes/Ommeh data: E-MTAB-3068 (Array Express); Häsl/oner/Oler/Muljo/Neuberger data: GSE58766 (NCBI GEO); Kaiser data: E-MTAB-2996 (Array Express); Lagarrigue/Roux data: SRP042257 (Sequence Read Archive); Lambert data: GSE51035 (NCBI GEO); Münsterberg/Pais data: GSE58766 (NCBI GEO); Schwartz/Ülitsky data: SRP041863 (Sequence Read Archive); Skinner data: GSE51035 (NCBI GEO); Münsterberg/Pais data: GSE58766 (NCBI GEO); Schwartz/Ülitsky data: SRP041863 (Sequence Read Archive); Skinner data: PRJEB7620 (European Nucleotide Archive); Wang/Zhou data: GSM1385570, GSM1385571, GSM1385572, GSM1385573 (NCBI GEO).

Noncoding RNAs in the Chicken Genome

Noncoding RNAs (ncRNAs) are some of the evolutionarily most conserved biomolecules and are essential for many cellular processes [Jeffares et al., 1998]. These include the ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) involved in translation, the small nucleolar RNAs (snoRNAs) and RNase P/MRP RNAs involved in maturing ncRNAs, and the microRNAs (miRNAs) involved in gene regulation [Cech and Steitz, 2014]. The recently rediscovered and expanded class of ‘long noncoding RNAs’ (lncRNAs) are involved in diverse functions, including dosage compensation (e.g. Xist, roX1, roX2) and forming scaffolds for recruiting proteins to form functional complexes (e.g. HOTAIR) [Erdmann et al., 2000; Cech and Steitz, 2014].

The First Report on Chicken Genes and Chromosomes 2000 [Schmid et al., 2000] made high-resolution genetic and physical maps readily available and cemented the chicken as a model genome. However, coverage of ncRNAs was restricted to a brief mention of the Z-linked MHM (male hypermethylated) locus, adjacent to the candidate sex-determining gene, DMRT1 (doublesex and mab-3-related transcription factor 1), subsequently described by Teranishi et al. [2001] and reviewed in The Second Report on Chicken Genes and Chromosomes 2005 [Schmid et al., 2005]. Briefly, the MHM region is hypermethylated and transcriptionally silent in males (ZZ), but in females (ZW), MHM is hypomethylated and transcribed into an lncRNA suspected to play a role in the downregulation of DMRT1. For an up-to-date review of the role of MHM in sex determination, see the contribution by Smith et al. in this report.

Compiled prior to the publication of the draft chicken genome [International Chicken Genome Sequencing Consortium, 2004], coverage of ncRNAs in the Second Report on Chicken Genes and Chromosomes 2005 [Schmid et al., 2005] was limited to 3 ncRNA classes, but each was reviewed in detail. Namely, there was extensive coverage of MHM (described above), the rRNAs (28S, 18S, 5.8S and 5S) and telomerase RNA. A number of recent studies have since expanded the number of chicken ncRNAs. These include miRNAs [Glazov et al., 2008; Shao et al., 2012], lncRNAs [Chodroff et al., 2010; Necullea et al., 2014], and the entire complement of ncRNAs [International Chicken Genome Sequencing Consortium, 2004; Zhang et al., 2009; Gardner et al., 2015].

The ncRNAs reviewed in this contribution fall into 3 major classes: (1) housekeeping RNAs such as rRNAs, tRNAs, spliceosomal RNAs, and snoRNAs as well as a few minor classes, (2) miRNAs and related small RNAs, and (3) lncRNAs.

Housekeeping RNAs

Ribosomal RNAs. The rRNA operon comprising the 18S, 5.8S and 28S rRNAs is located in more than 100 copies on a microchromosome [Muscarella et al., 1985]. The pol-III transcribed 5S rRNA is organized as 2 tandem repeats on chromosome 9 [Daniels and Delany, 2003]. In addition, the mitogenome contains its own copy of the LSU and SSU rRNA [International Chicken Genome Sequencing Consortium, 2004].

Transfer RNAs. The predicted number of tRNAs encoded in the chicken genome is relatively modest according to tRNAscan-SE [Lowe and Eddy, 1997]. Just 299 tRNA genes are predicted, which is just 4 more than in Saccharomyces cerevisiae [Chan and Lowe, 2009]. All 20 canonical tRNAs are represented with copy numbers
ranging from 6 (Trp) to 36 (Ala), there is 1 selenocysteine tRNA and 4 pseudogenized tRNAs. There are also several significant tRNA clusters. Namely, 10 Cys tRNAs are clustered in a 5-kb region on chromosome 27, 10 Arg and Tyr tRNAs are clustered in a 40-kb region on chromosome 2, and 8 Glu, Leu, Lys, and Val tRNAs can be found in a 4-kb region on chromosome 16.

Spliceosomal RNAs. The chicken genome encodes both a major and a minor spliceosome. The components of the major spliceosome (U1, U2, U4, U5, and U6) are present in multiple copies (many of which are presumably pseudogenes) as in most vertebrates, while most snRNAs of the minor spliceosome (U4atac, U6atac, U11, and U12) are single-copy genes [Marz et al., 2008].

Other Nuclear RNAs. There are single copies of the RNase P and RNase MRP RNAs on chromosomes 8 and 4, respectively; 4 copies of the RNA component of the signal recognition particle in a 1-Mb region on chromosome 5, and a cluster comprising 3 distinct Y RNAs in a 17-kb region on chromosome 2. Of the 2 vault RNA loci (chromosomes 5 and 12), the one on chromosome 5 appears to be a pseudogene. The telomerase RNA and the U7 snRNA are also represented as single copies.

Small Nucleolar RNAs. Covariance models of known snoRNAs were used to identify snoRNAs in chicken using INFERNAL. In addition, an independent homology-based approach using snoStrip was applied. Both results were merged leaving 106 box C/D and 68 box H/ACA snoRNAs, of which 90 (84.9%) and 48 (70.6%) were confirmed with RNA-seq, respectively.

miRNAs and Related Small ncRNAs

An INFERNAL search for miRNAs using those miRNA families from miRBase that have at least 1 representative in vertebrates returned 427 miRNAs. Of these, 280 genes are confirmed with RNA-seq. Only 51 miRNAs annotated in chicken seem to be species-specific, since they do not show any significant homology in other vertebrates. Only 18 of these miRNAs are organized in 8 families, while the remaining 33 miRNAs come as single sequences with no paralogs at all. Although miRNAs are lost infrequently in animal genomes, it appears that 3 miRNA families (i.e. mir-139, mir-425 and mir-1287) have been lost completely in chicken. Before the divergence of chicken, a few other miRNAs have also been lost (e.g. 8 at the branch leading to the split of Sauropsida and 4 at the ancestor of Dinosauria/Aves).

The recent availability of comprehensive strand-specific high-throughput RNA sequencing data for several developmental stages and tissues allows for further experimental confirmation of the current catalog of chicken ncRNAs. Only few traces of small ncRNA are found in total RNA libraries due to size selection for larger fragments. In small RNA-seq protocols, the size fraction of 17–25 nt is selected for library preparation. The resulting sequencing data comprises not only processed miRNA sequences, for which those protocols were originally developed, but contains also fragments from snoRNAs, tRNAs and other housekeeping RNAs. Table 8 shows the fraction of major ncRNA families that could be experimentally validated [Gardner et al., 2015].

IncRNAs

While many classes of short RNAs have been characterized in recent years, IncRNAs are less well understood. They are operationally defined as noncoding transcripts that are longer than 200 bp. Here, we only describe IncRNAs falling in intergenic regions, thus excluding intronic transcripts or those overlapping a coding transcript. There is still much debate about whether IncRNAs constitute mainly transcriptional noise or in their majority have functions such as reviewed in Rinn and Chang [2012]. Our approach revealed a high degree of sequence conservation in other bird species as well as reptiles, thus hinting at a large percentage of functional transcripts.

RNA-seq data from 20 different tissues comprising 238 read libraries was mapped to the Ensembl 71 chicken genome. Ab-initio prediction of transcripts was then performed using the software Cufflinks [Trapnell et al., 2010]. After merging predictions from individual libraries, only intergenic transcripts (at least 1 kb away from coding genes) were kept. An evaluation of coding potential as well as repeat content was used to extensively filter

<table>
<thead>
<tr>
<th>RNA family</th>
<th>Homology-based predictions, n</th>
<th>Confirmed by RNA-seq, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>tRNAs</td>
<td>300a</td>
<td>280 (93.3%)</td>
</tr>
<tr>
<td>rRNAs</td>
<td>22</td>
<td>10 (45.5%)</td>
</tr>
<tr>
<td>microRNA</td>
<td>427</td>
<td>280 (65.6%)</td>
</tr>
<tr>
<td>snoRNA box C/D</td>
<td>106</td>
<td>90 (84.9%)</td>
</tr>
<tr>
<td>snoRNA box H/ACA</td>
<td>68</td>
<td>48 (70.6%)</td>
</tr>
<tr>
<td>snoRNA – Cajal body</td>
<td>12</td>
<td>12 (100%)</td>
</tr>
<tr>
<td>spliceosomal snRNAs</td>
<td>77</td>
<td>35 (45.4%)</td>
</tr>
</tbody>
</table>

Table 8. Number of ncRNAs identified in the chicken genome to date and the proportion confirmed by RNA-seq experiments

*a Plus 4 pseudogenes.
the transcripts thus obtained. To further cut down on false positives, only multi-exonic transcripts were retained. Systematic phylogenomic profiling against 52 other bird and reptile species using the procedure described in Derrien et al. [2012] was performed to reveal the degree of conservation of the spliced sequences. Transcript expression was studied on a reduced set of 9 tissues (comprising 123 libraries) with a sufficient number of replicates (between 4 and 36) each.

We report 10,364 putative lncRNA transcripts forming 7,597 gene loci. Less than 10% of these transcripts appear to be chicken-specific, almost half of them are conserved in at least 10 species, and more than 10% are conserved in at least 40 species (table 9). Expression of ~80% of transcripts could be detected in at least one of the selected tissues. Generally we found a strong bias toward tissue-specific transcripts: more than half of the expressed transcripts are expressed in a single tissue, and only ~5% are found ubiquitously expressed in all tissues.

Synopsis
The ncRNAs in the chicken genome form a valuable collection for a number of comparative analyses as the number of ncRNA-derived pseudogenes in the avian genomes is remarkably low [International Chicken Genome Sequencing Consortium 2004; Gardner et al., 2015]. This, therefore, forms a useful way to discriminate pseudogenized ncRNAs from functional copies in related genomes where synteny is likely to have been preserved. A further use for the chicken ncRNA annotations is the identification of ‘missing’ ncRNAs. The chicken genome has the most complete collection of microchromosome sequences currently available, and has proven essential to explain the consistent ‘loss’ of essential ncRNAs in the avian lineage such that these have been traced to the poor ability of sequencing methods for capturing the microchromosomes [Gardner et al., 2015]. As more transcripts from more bird species, from more developmental stages and from more tissue types become available, new ncRNAs will undoubtedly be located. However, identifying the function of these novel ncRNAs will remain a major research challenge.

Genome Sequencing in Birds and Evolutionary Inferences from Avian Genome Sequences
(Prepared by H. Ellegren)

Developments in avian genome analysis well illustrate the overall progress in genome sequencing of non-model species offered by next-generation sequencing (NGS) technology [Ellegren, 2014]. It took 6 years between the reports of the first 2 avian genome sequences in 2004 (chicken, Gallus gallus [International Chicken Genome Sequencing Consortium, 2004]) and 2010 (zebra finch, Taeniopygia guttata [Warren et al., 2010]), respectively, both sequenced with Sanger technology. In 2010, the genome sequence of turkey (Meleagris gallopavo) was also reported, the first based on NGS [Dalloul et al., 2010]. Subsequently, there has been an exponential increase in the number of avian genomes reported, recently topping by the parallel sequencing and analysis of 45 new genomes by the Avian Phylogenomics Consortium [Jarvis et al., 2014]. By now (January 2015), a total of some 60 bird species have been subject to genome sequencing (table 10). The precise number is somewhat dependent on what criteria for coverage, sequence continuity and other assembly statistics one sets for defining a genome to have been sequenced. Moreover, as new genomes are continuously reported, the total number changes quickly. All 34 currently recognized orders of living birds in the avian tree of life are represented in the list of sequenced species. This offers an unprecedented resource to comparative and evolutionary studies, some brief examples of which will be provided here and limited to work that specifically has made use of whole-genome sequence data. For a general review of avian evolutionary genomics, see Ellegren [2013].

Molecular Evolution
From alignments of 2 or more avian genomes, it is possible to analyze a number of aspects related to the accumulation of nucleotide substitutions and the forces driving this process. For neutral sites, the rate of sequence divergence should reflect the underlying rate of mutation. Under the assumption that 4-fold degenerate sites evolve

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ab-initio predicted transcripts</td>
<td>10,364</td>
</tr>
<tr>
<td>Conserved beyond chicken</td>
<td>9,386</td>
</tr>
<tr>
<td>Conserved in at least 10 species</td>
<td>5,058</td>
</tr>
<tr>
<td>Conserved in at least 40 species</td>
<td>1,251</td>
</tr>
<tr>
<td>Expressed by at least 0.1 RPKM in 1 of 123 libraries</td>
<td>8,345</td>
</tr>
<tr>
<td>Expressed by at least 0.1 RPKM on average in at least 1 of 9 tissues</td>
<td>6,360</td>
</tr>
<tr>
<td>Of these, expressed in only 1 tissue</td>
<td>3,475</td>
</tr>
<tr>
<td>Of these, expressed in all 9 tissues</td>
<td>318</td>
</tr>
</tbody>
</table>

Table 9. Number of lncRNAs identified in the chicken genome to date along with those remaining after application of different approaches to increase the likelihood that these are functional.
neutrally, Nam et al. [2010] estimated the neutral divergence rate at 1.2–2.2 × 10^{-9} per site in 3 avian lineages (the ancestral bird lineage and the lineages leading to chicken and zebra finch, respectively), with evidence for rate variation among lineages. Zhang G et al. [2014b] found a mean rate of 1.9 × 10^{-9} at 4-fold degenerate sites across the species sequenced by the Avian Phylogenomics Consortium, again with significant variation among lineages. Intriguingly, they also found a positive correlation between rate of divergence and number of species per order, suggesting an association (though not necessarily causative) between divergence and diversification.

Selection in coding sequences has often been studied by estimating the ratio of the non-synonymous (d_N) and the synonymous (d_S) substitution rates (i.e. d_S/d_N). For example, the approach can be used to identify fast and adaptively evolving genes with codons or whole sequences showing d_S/d_N > 1. Functional annotation of genes based on gene ontology has indicated that certain categories of genes may be overrepresented among adaptively evolving genes in different bird lineages [Nam et al., 2010; Huang et al., 2013; Zhan et al., 2013; Li C et al., 2014; Zhang G et al., 2014b]. d_S/d_N can also be used to study how the intensity of selection varies among lineages in relation to life history [Weber et al., 2014b] as well as how it varies within genomes. For example, it has been suggested that the high rate of recombination in microchromosomes reduces the effect of Hill-Robertson interference, leading to lowered d_S/d_N estimates under the assumption that slightly deleterious mutations contribute to d_N [Nam et al., 2010; Zhang G et al., 2014b]. The rationale here is that selection is more efficient in the removal of slightly deleterious mutations when recombination uncouples loci and reduces the effects (interference) of linked selection. However, recombination is also related to the process of GC-biased gene conversion in avian genomes [Weber et al., 2014a], and this can easily lead to erroneous conclusions on selection in coding sequences [Ratnakumar et al., 2010]. There is clear evidence that the recombination hot-spot in the pseudoautosomal region of avian sex chromosomes has a strong effect on sequence evolution [Smeds et al., 2014].

**Genome Evolution**

It is well known that birds have the smallest genomes among amniotes. Genome assemblies are in the range of

<table>
<thead>
<tr>
<th>Latin name</th>
<th>Species name</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acantita chloris</td>
<td>rifleman</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Amazona viridissima</td>
<td>Pucero Rican amazon</td>
<td>Oleksy et al., 2012</td>
</tr>
<tr>
<td>Anas platyrhynchos</td>
<td>Pekin duck</td>
<td>Huang et al., 2013</td>
</tr>
<tr>
<td>Anseriformes</td>
<td>Chuck-will’s-widow</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Apodidae</td>
<td>bar-tailed trogon</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Apteryx forsteri</td>
<td>emperor penguin</td>
<td>Li C et al., 2014</td>
</tr>
<tr>
<td>Apus apus</td>
<td>golden eagle</td>
<td>Doyle et al., 2014</td>
</tr>
<tr>
<td>Ara macao</td>
<td>scarlet macaw</td>
<td>Seaby et al., 2013</td>
</tr>
<tr>
<td>Balaeniceps rex</td>
<td>grey crowned crane</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Bucerotidae</td>
<td>rhinoceros hornbill</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Calypte anna</td>
<td>Anna’s hummingbird</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Carpodacus edithae</td>
<td>red-legged serema</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Catharina aura</td>
<td>turkey vulture</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Chaetura pelagica</td>
<td>chimney swift</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Charadrius vociferus</td>
<td>killdeer</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Chloemys cincta</td>
<td>MacQueen’s bustard</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Colinus virginianus</td>
<td>northern bobwhite</td>
<td>Halley et al., 2014</td>
</tr>
<tr>
<td>Colius striatus</td>
<td>speckled mousebird</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Columba livia</td>
<td>pigeon</td>
<td>Shapiro et al., 2013</td>
</tr>
<tr>
<td>Corvus corax</td>
<td>American crow</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Corvus corone</td>
<td>hooded crow</td>
<td>Poletira et al., 2014</td>
</tr>
<tr>
<td>Coturnix japonica</td>
<td>Japanese quail</td>
<td>Kawaihara-Miki et al., 2013</td>
</tr>
<tr>
<td>Cuculus canorus</td>
<td>common cuckoo</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Eretroa garzetta</td>
<td>little egret</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Fregata magnificens</td>
<td>sunbittern</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Falco cherrug</td>
<td>saker falcon</td>
<td>Zhang et al., 2013</td>
</tr>
<tr>
<td>Falco peregrinus</td>
<td>peregrine falcon</td>
<td>Zhang et al., 2013</td>
</tr>
<tr>
<td>Ficedula albicollis</td>
<td>collared flycatcher</td>
<td>Ellegren et al., 2012; Smeds et al., 2015</td>
</tr>
<tr>
<td>Fulmarus glacialis</td>
<td>northern fulmar</td>
<td>Jarvis et al., 2014</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>chicken</td>
<td>ICGSC, 2004</td>
</tr>
</tbody>
</table>

ICGSC = International Chicken Genome Sequencing Consortium.
1.0–1.3 Gb, although cytogenetic analyses indicate that there might be species with somewhat larger genomes. The relative compactness of avian genomes is evident from shorter introns and less intergenic DNA than in mammals and reptiles. It seems largely to be due to a lower transposable element activity in the avian lineage but also to a high rate of deletions [International Chicken Genome Sequencing Consortium, 2004; Zhang G et al., 2014b], which might be driven by high recombination rates in avian chromosomes [Nam and Ellegren, 2012].

That the avian karyotype is unusually stable is also well established; the majority of bird species have a diploid chromosome number of 2n = 76–80. Genome sequence data now allow for more quantitative assessments of this stability and, in particular, analyses of intrachromosomal rearrangements not detected in previous cytogenetic work or in work based on low-resolution linkage maps. Kawakami et al. [2014] estimated the rate of inversion at 1.5 and 2.0 events per million years in 2 song bird lineages, respectively, with a mean inversion size of ~7 Mb. Across more distantly related groups of birds, a mean rate of 1.25 rearrangement events (mostly inversions) per million years was observed, with increased rate seen in some lineages [Romanov et al., 2014; Zhang G et al., 2014b]. This is considerably lower than that seen in mammals, particularly in rodents [Zhao and Bourque, 2009]. Karyotypic stability coupled with a low repeat activity is also manifested in highly similar amounts of assembled DNA per syntenic chromosome of divergent avian lineages (table 11).

The access to whole-genome sequences provides new possibilities to elucidate the character and extent of evolutionary conservation in avian genomes. Sequences that are conserved across distantly related species are likely to be subject to functional constraints and to evolve under purifying selection. The proportion of constrained sequence has been estimated to 7.5% [Zhang G et al., 2014b], which is at least twice as high as the proportion of exonic sequence. This demonstrates that protein-coding genes are not the only, maybe not even the dominating, functionally important category of sequences in avian genomes. Sequences involved in regulation of gene expression are likely to constitute a large part of the conserved elements, some of which are conserved across both birds and mammals.

Demography

DNA sequence data is often used in attempts to infer the demography of species and to test different demographic scenarios and estimate divergence times. A major limitation of traditional methods for such inference has

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Assembled sequence, Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>collared flycatcher</td>
</tr>
<tr>
<td>1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>119.8</td>
</tr>
<tr>
<td>1A&lt;sup&gt;a&lt;/sup&gt;</td>
<td>74.8</td>
</tr>
<tr>
<td>2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>157.4</td>
</tr>
<tr>
<td>3&lt;sup&gt;a&lt;/sup&gt;</td>
<td>115.7</td>
</tr>
<tr>
<td>4A&lt;sup&gt;b&lt;/sup&gt;</td>
<td>21.2</td>
</tr>
<tr>
<td>4B&lt;sup&gt;b&lt;/sup&gt;</td>
<td>70.3</td>
</tr>
<tr>
<td>5&lt;sup&gt;a&lt;/sup&gt;</td>
<td>64.6</td>
</tr>
<tr>
<td>6&lt;sup&gt;a&lt;/sup&gt;</td>
<td>37.2</td>
</tr>
<tr>
<td>7&lt;sup&gt;a&lt;/sup&gt;</td>
<td>39.3</td>
</tr>
<tr>
<td>8&lt;sup&gt;a&lt;/sup&gt;</td>
<td>32.0</td>
</tr>
<tr>
<td>9&lt;sup&gt;a&lt;/sup&gt;</td>
<td>26.8</td>
</tr>
<tr>
<td>10&lt;sup&gt;a&lt;/sup&gt;</td>
<td>21.3</td>
</tr>
<tr>
<td>11&lt;sup&gt;a&lt;/sup&gt;</td>
<td>21.7</td>
</tr>
<tr>
<td>12&lt;sup&gt;a&lt;/sup&gt;</td>
<td>21.9</td>
</tr>
<tr>
<td>13&lt;sup&gt;a&lt;/sup&gt;</td>
<td>18.6</td>
</tr>
<tr>
<td>14&lt;sup&gt;a&lt;/sup&gt;</td>
<td>17.4</td>
</tr>
<tr>
<td>15&lt;sup&gt;a&lt;/sup&gt;</td>
<td>14.9</td>
</tr>
<tr>
<td>16&lt;sup&gt;d&lt;/sup&gt;</td>
<td>– &lt;0.01</td>
</tr>
<tr>
<td>17&lt;sup&gt;a&lt;/sup&gt;</td>
<td>12.4</td>
</tr>
<tr>
<td>18&lt;sup&gt;a&lt;/sup&gt;</td>
<td>13.1</td>
</tr>
<tr>
<td>19&lt;sup&gt;a&lt;/sup&gt;</td>
<td>11.9</td>
</tr>
<tr>
<td>20&lt;sup&gt;a&lt;/sup&gt;</td>
<td>15.6</td>
</tr>
<tr>
<td>21&lt;sup&gt;a&lt;/sup&gt;</td>
<td>8.1</td>
</tr>
<tr>
<td>22&lt;sup&gt;a&lt;/sup&gt;</td>
<td>5.7</td>
</tr>
<tr>
<td>23&lt;sup&gt;a&lt;/sup&gt;</td>
<td>7.9</td>
</tr>
<tr>
<td>24&lt;sup&gt;a&lt;/sup&gt;</td>
<td>8.0</td>
</tr>
<tr>
<td>25&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.7</td>
</tr>
<tr>
<td>26&lt;sup&gt;a&lt;/sup&gt;</td>
<td>7.6</td>
</tr>
<tr>
<td>27&lt;sup&gt;a&lt;/sup&gt;</td>
<td>5.5</td>
</tr>
<tr>
<td>28&lt;sup&gt;a&lt;/sup&gt;</td>
<td>6.1</td>
</tr>
<tr>
<td>Z&lt;sup&gt;d&lt;/sup&gt;</td>
<td>59.7</td>
</tr>
</tbody>
</table>

<sup>a</sup> Chicken chromosome 1 corresponds to chromosomes 1 and 1A in flycatcher and zebra finch, the result of a fission in the galliform lineage.

<sup>b</sup> Chicken chromosome 4 corresponds to chromosomes 4 and 4A in zebra finch and flycatcher, the result of a fission in the passeriform (flycatcher and zebra finch) lineage.

<sup>c</sup> Chromosome 16 is for some reason very difficult to sequence in birds.

<sup>d</sup> The smaller size of the Z chromosome in flycatcher owes mainly to a relatively high proportion of scaffolds that could not be ordered with confidence on the Z chromosome by linkage analysis.
been the inability to model changes in effective population size through time. However, this has come to change by the introduction of the pairwise sequentially Markovian coalescent (PSMC) model that infers local time to the most recent common ancestor for each region of the genome and uses coalescent theory to deduce the effective population size (\(N_e\), basically the number of reproducing individuals, at equilibrium) at different points in time [Li and Durbin, 2011]. The method, in its original form, requires the access of a diploid genome sequence where sequencing has been done at sufficient depth (approx. >15×) that the vast majority of heterozygous sites have been correctly called.

PSMC-based estimates of ancestral population sizes have recently been presented for several bird species and have given us the first detailed insights into how bird populations have varied in size through time [Nadachowska-Brzyska et al., 2015]. The crested ibis (Nipponia nippon), now one of the most rare and endangered species on earth, was once common and had an \(N_e\) of \(\sim 70,000\) one million years ago [Li S et al., 2014]. It subsequently went through 2 population bottlenecks, reducing \(N_e\) to 10,000 about 10,000 years ago. During modern times, the decline has continued.

The Adélie (Pygoscelis adeliae) and emperor (Aptenodytes forsteri) penguins showed steady increases in \(N_e\) during the period between 1 million and 100,000 years ago, from <100,000 to 200,000 and 600,000, respectively [Li C et al., 2014]. Except for a decrease to 400,000 in the Adélie penguin about 60,000 years ago, the population sizes of the 2 penguins seem to have been stable during the last 100,000 years. Higher \(N_e\) estimates have been obtained for the collared flycatcher (Ficedula albicollis). It increased from 500,000 five million years ago to 1,500,000 about 200,000 years ago [Nadachowska-Brzyska et al., 2013]. Then, probably like many species in the temperate part of the world, the population decreased in size to again reach about 500,000 during the last glaciation period.

**Final Remarks**

It is not unlikely that the majority of bird species in not too many years will have been subject to genome sequencing. This will provide a most comprehensive portrait of the genetic diversity within the class Aves and allow us to pinpoint the genetic changes that underlie specific adaptations in different bird lineages. It will help us addressing topics such as convergence and what types of mutations are essential for phenotypic novelties. Genome resequencing of multiple individuals within species will also be important for characterization of avian biodiversity. It probably has its greatest potential when it comes to illuminating the processes that govern avian evolution, like local adaptation, speciation and diversification.

**The Use of Avian BAC Libraries and Clones**

(Prepared by M.N. Romanov and D.K. Griffin)

High-density gridded libraries of large-insert clones using bacterial artificial chromosome (BAC) and other vectors are essential tools for genetic and genomic research in chicken and other avian species. Earliest chicken BAC libraries [Zoorob et al., 1996; Zimmer and Ver-rinder Gibbins, 1997; Crooijmans et al., 2000] were applied to solve a range of problems at both genome-wide and chromosomal levels. Uses of BAC libraries and individual clones include physical and comparative mapping, support of whole-genome sequencing projects, positional cloning of quantitative trait loci, and isolation of genes and other genomic regions of interest [e.g. Lee et al., 2003; Ren et al., 2003; Wallis et al., 2004; Huang et al., 2006; Reed et al., 2008; Blagoveshchenskii et al., 2011]. In early studies, BACs were also used as one of the sources for the development of microsatellite markers to characterize genetic diversity in populations [e.g. Morisson et al., 1998; Romanov and Weigend, 1999; Crooijmans et al., 2000]. For the purposes of BAC library screening and cross-species BAC-based physical mapping, one of the most cost-time-effective methods is the overgo-based filter hybridization technique [Romanov et al., 2003; Romanov and Dodgson, 2006]. Pools of short sequence-specific overgo probes facilitate integration of linkage and physical maps by high-throughput assignment of genes and markers to BAC contigs.

BAC clones are the probes of choice in FISH mapping for the development of chromosome-level physical maps of genomes and (along with chromosome paints) for comparative cytogenetic mapping. Given that defining avian karyotypes (and hence whole genome assemblies) is impeded by the presence of a large number of near-indistinguishable microchromosomes, development of microchromosome-specific BAC probes has proved to be very useful in microchromosome identification [Masabanda et al., 2004; Griffin et al., 2008; Skinner et al., 2009; Völker et al., 2010], fine mapping, and characterization of inter- and intrachromosomal rearrangements [e.g. Fillon et al., 1998; Romanov et al., 2005; Griffin et al., 2008; Skinner et al., 2009; Völker et al., 2010; Lithgow et al., 2008].

**Cytogenet Genome Res 2015;145:78–179**

DOI: 10.1159/000430927

Schmid et al.
et al., 2014]. Indeed, the generation of cross-species FISH data, along with bioinformatic analysis of genome sequences, is crucial in understanding the gross genome evolution of avian species and, in particular, reconstructing ancestral karyotypes [e.g. Zhang et al., 2011; Romanov et al., 2014].

In addition to chicken, BAC libraries have been constructed and characterized for several other bird species including representatives of Anseriformes, Casuariiformes, Cathartiformes, Galliformes, Psittaciformes, Pelicaniformes, and Passeriformes. Most of these are publicly available (table 12), enabling BAC-based applications in other important research areas. These include structural analyses of the MHC chromosome (microchromosome 16 in chicken), loci and alleles in chicken [Solinhac et al., 2010; Suzuki et al., 2012], turkey [Chaves et al., 2007; Bauer and Reed, 2011] and golden pheasant [Ye et al., 2012]. These studies have contributed to a more detailed exploring of distinctive organization, function and evolution of the bird MHC, showing conserved synteny of a single microchromosome involved in the immune system in Aves and variation in MHC alleles related to disease resistance/susceptibility.

Comparative screening of BAC libraries and FISH using BAC clones have also assisted in developing genomic resources and tools for the endangered California condor and the white-throated sparrow, a model for behavioral studies in humans [Romanov et al., 2006, 2009a, b, 2011; Thomas et al., 2008; Modi et al., 2009]. Kothapalli et al. [2011] examined California condor BACs harboring the immunoglobulin lambda locus and found a region of high homology to the chicken and zebra finch orthologs.

**Table 12. Major avian BAC libraries showing approximate number of clones, insert size and coverage [after Romanov et al., 2009a, with amendments]**

<table>
<thead>
<tr>
<th>Bird (Scientific Name)</th>
<th>Library</th>
<th>Clone DB abbreviation</th>
<th>No. of clones</th>
<th>No. of clones in Clone DB</th>
<th>Average insert, kb</th>
<th>Coverage</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duck (Anas platyrhynchos)</td>
<td>China Agriculture University</td>
<td>–</td>
<td>84,480</td>
<td>–</td>
<td>117.94</td>
<td>9.84</td>
<td>Yuan et al., 2006</td>
</tr>
<tr>
<td>Golden pheasant (Chrysolophus pictus)</td>
<td>Zhejiang University reverse-4D</td>
<td>–</td>
<td>89,600</td>
<td>–</td>
<td>106.87</td>
<td>7.42</td>
<td>Ye et al., 2012</td>
</tr>
<tr>
<td>Emu (Dromaius novaehollandiae)</td>
<td>VMRC16 (LIBGSS_011154)</td>
<td>VMRC16</td>
<td>133,632</td>
<td>–</td>
<td>165</td>
<td>13.5</td>
<td>Kellner et al., 2005</td>
</tr>
<tr>
<td>California condor (Gymnogyps californianus)</td>
<td>CHORI-262</td>
<td>CH262</td>
<td>89,665</td>
<td>–</td>
<td>–</td>
<td>14</td>
<td>Romanov et al., 2006</td>
</tr>
<tr>
<td>Chicken (Gallus gallus)</td>
<td>CHORI-261</td>
<td>CH261</td>
<td>73,000</td>
<td>66,486</td>
<td>195 (182)</td>
<td>11</td>
<td>Romanov et al., 2003</td>
</tr>
<tr>
<td>Texas A&amp;M 031-JF256-BI</td>
<td>TAM31</td>
<td>38,400</td>
<td>108</td>
<td>150</td>
<td>5.2</td>
<td>Lee et al., 2003</td>
<td></td>
</tr>
<tr>
<td>Texas A&amp;M 032-JF256-BI</td>
<td>TAM32</td>
<td>38,400</td>
<td>10,792</td>
<td>152</td>
<td>5.3</td>
<td>Romanov et al., 2004</td>
<td></td>
</tr>
<tr>
<td>Texas A&amp;M 033-JF256-BI</td>
<td>TAM33</td>
<td>38,400</td>
<td>66</td>
<td>171</td>
<td>6</td>
<td>Romanov et al., 2004</td>
<td></td>
</tr>
<tr>
<td>Texas A&amp;M Wageningen 020-CHK-H3 (LIBGSS_009949)</td>
<td>WAG</td>
<td>49,920</td>
<td>8,611</td>
<td>130 (134)</td>
<td>5.4 (5.57)</td>
<td>Crooijmans et al., 2000</td>
<td></td>
</tr>
<tr>
<td>Tohoku University</td>
<td>–</td>
<td>49,152</td>
<td>–</td>
<td>149</td>
<td>3.2</td>
<td>Horii et al., 2000</td>
<td></td>
</tr>
<tr>
<td>China Agriculture University</td>
<td>LIBGSS_003202, LIBGSS_009895</td>
<td>–</td>
<td>138,240</td>
<td>–</td>
<td>118</td>
<td>13.34</td>
<td>Liu et al., 2003</td>
</tr>
<tr>
<td>Turkey (Meleagris gallopavo)</td>
<td>CHORI 260 (LIBGSS_010330)</td>
<td>CH260</td>
<td>71,000</td>
<td>10,438</td>
<td>190</td>
<td>11.1</td>
<td>Romanov and Dodgson, 2006</td>
</tr>
<tr>
<td>TKNMI</td>
<td>TKNMI</td>
<td>71,000</td>
<td>9,519</td>
<td>160</td>
<td>–</td>
<td>Romanov and Dodgson, 2006</td>
<td></td>
</tr>
<tr>
<td>Budgerigar (Melopsittacus undulatus)</td>
<td>CHORI-263</td>
<td>CH263</td>
<td>197,392</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>CHORI</td>
</tr>
<tr>
<td>Crested ibis (Nipponia nippon)</td>
<td>Zhejiang University 4D-PCR</td>
<td>–</td>
<td>129,312</td>
<td>–</td>
<td>86.5</td>
<td>7.8</td>
<td>Lan et al., 2014</td>
</tr>
<tr>
<td>Zhejiang University reverse-4D</td>
<td>–</td>
<td>1,040,000</td>
<td>–</td>
<td>100.9</td>
<td>35</td>
<td>Romanov et al., 2004</td>
<td></td>
</tr>
<tr>
<td>Zebra finch (Taeniopygia guttata)</td>
<td>TG_Ba (TG_Ba)</td>
<td>147,456</td>
<td>205</td>
<td>134</td>
<td>15.5</td>
<td>Kellner et al., 2005</td>
<td></td>
</tr>
<tr>
<td>TGMCRa</td>
<td>TGMCRa</td>
<td>–</td>
<td>132,650</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Thomas et al., 2008</td>
</tr>
<tr>
<td>White-throated sparrow (Zonotrichia albicollis)</td>
<td>CHORI 264 (LIBGSS_011743)</td>
<td>CH264</td>
<td>196,354</td>
<td>1,872</td>
<td>144</td>
<td>21.1</td>
<td>Romanov et al., 2009b</td>
</tr>
</tbody>
</table>

* As listed in the Library Resources, BACPAC Resources Center, CHORI, Oakland, Calif. (http://bacpac.chori.org/libraries.php).
In another highly endangered bird, the crested ibis, 2 genomic BAC libraries were created to target and characterize a cluster of defensin genes known to be a critical component of the innate immune system [Lan et al., 2014]. Moreover, by generating a zebra finch library TG_Ba, Luo et al. [2006] reported a genomic region that was previously difficult to sequence and involved the first bird androgen receptor gene and a conserved regulatory element.

The emu BAC library and clones have proved to be pivotal tools for investigating the structure, organization and evolution of the sex chromosomes in ratite, reptilian and other bird species [Chapus and Edwards, 2009; Janes et al., 2009, 2011]. Indeed, a gene SubA, with orthologs being expressed in mammalian gonads, has been isolated from the emu library and assigned to a pair of microchromosomes using BAC-FISH [Janes et al., 2008].

Taken together, these studies demonstrate that applications of large-insert clones and BAC libraries derived from birds are, and will continue to be, effective tools to aid high-throughput and state-of-the-art genomic efforts and the important biological insight that arises from them.

### Comparative Genomics

(Prepared by D.M. Larkin, M. Farré, and J. Damas)

Even after the chicken genome sequencing and assembly completion in 2004 [International Chicken Genome Sequencing Consortium, 2004], the avian comparative genome studies were not as advanced as studies of mammal genomes due to a very limited number of sequenced avian genomes, with only 3 of them published by 2013 [International Chicken Genome Sequencing Consortium, 2004; Dalloul et al., 2010; Warren et al., 2010]. In contrast, more than 50 mammalian genomes were published or sequenced by the end of the same year. Because of these unfortunate limitations and a lack of ordered physical maps for avian genomes, until recently, comparative genome studies of avian species were mostly focused on the comparison of karyotype evolution achieved through either a direct DNA-to-DNA hybridization (FISH) [Griffin et al., 2007] or sequence comparison of avian and mammalian genomes [Larkin et al., 2009]. Despite the limited resolution of FISH, these studies provided important insights into the unique structure and evolutionary stability of avian karyotypes, and revealed major patterns of early amniote genome evolution [Griffin et al., 2007; Larkin et al., 2009].

The situation changed drastically in 2014 when de novo sequencing of 45 and parallel comparative studies of 48 avian genomes were published in Science and other journals [e.g. Whitney et al., 2014; Zhang G et al., 2014b]. This became possible due to the Avian Genome Consortium, with members spread across 13 countries. The sequencing of 20 genomes was performed at Beijing Genome Institute with a high level of sequence coverage (>50×), which resulted in high N50 and long scaffolds for 19 of them; while the additional 25 genomes were sequenced at lower (24–32×) coverage [Zhang G et al., 2014a]. A major difference of this ‘flock of genomes’ from the vertebrate genomes sequenced and assembled in a traditional way was that the majority of the species still had no physical maps that could be used to verify the quality of the assemblies and anchor scaffolds along chromosomes [Lewin et al., 2009]. Several exceptions are the genomes of chicken, turkey and zebra finch, published previously and the assemblies of budgerigar and ostrich assisted by optical maps [Griffin et al., 2008; Warren et al., 2010; Zhang G et al., 2014a].

Due to the highly fragmented nature of the majority (25 out of 45) of the new avian genome assemblies, most comparative analyses involving this dataset were focused on gene or short sequence feature aspects of avian genome evolution. These comparisons provided new insights into the avian genome evolution, phylogeny, and unique adaptations formed in the avian lineage. Some important aspects of avian chromosome evolution were also uncovered; they are described in the cytogenetic section of this report.

### Genome Size Reduction

Among amniotes, birds have the smallest genome size ranging from 0.91 Gb in the black-chinned hummingbird to 1.3 Gb in the common ostrich [International Chicken Genome Sequencing Consortium, 2004]. The small size of avian genomes is often associated with adaptations related to flight (e.g. higher rates of oxidative metabolism) [Hughes and Hughes, 1995]; however, there is evidence that genome size reduction could have already occurred in the dinosaur ancestor of all birds [Organ et al., 2007]. One of the striking differences between the genomes of mammals and birds, which accounts for the smaller genome sizes in birds, is a lower fraction of transposable elements (TEs) and other repetitive sequences in avian genomes compared to those of mammals. In birds, TEs comprise 4–10% of the genome length in most species, while in mammals, TE fractions range from 35 to 52% [Zhang G et al., 2014b]. Among the genomes analyzed by...
the Avian Genome Consortium, only the downy wood-pecker had a large fraction of TEs (22%) resulting from either a species- or lineage-specific expansion of LINE type CR1 (long interspersed elements; chicken repeat 1) transposons [Zhang G et al., 2014b]. Compared to non-avian reptiles, birds have on average 19× less SINEs (short interspersed repeat elements), suggesting that this reduction could have happened in the common ancestor of all birds [Zhang G et al., 2014b]. Endogenous viral elements are also rare in avian genomes. Their fraction is 6–13× lower in avian than in mammalian genome sequences, suggesting that birds either are able to suppress viral DNA insertion in their genomes, or (less likely) are not subjected to viral infections to the same level as mammals [Cui et al., 2014].

In addition to a lower fraction of TEs, the comparative analysis of avian genomes demonstrated that the protein-coding genes and intergenic regions are shorter in birds compared to mammalian and non-avian reptile counterparts [Zhang G et al., 2014b]. Most likely the shorter gene length in avian genomes is a consequence of a lower fraction of TEs in introns and intergenic intervals. A similar compression of the gene and intergenic region length was observed in bats, the only flying mammal [Zhang G et al., 2013]. In both cases (birds and bats), this could be related to unique adaptations of flight [Zhang G et al., 2013, 2014b].

Interestingly, avian genomes have a large number of ancestral microdeletions absent from the genomes of other reptiles. Larger-scale structural rearrangements were also found associated with segmental deletion of 1,559 genes in birds, most of which had at least one additional paralog in the lizard or human genomes [Lovell et al., 2014]. These missed genes are not randomly distributed across amniote chromosomes, but tend to cluster into syntenic blocks with the highest fraction of missed genes located in lizard chromosome 2 (human chromosome 19) [Lovell et al., 2014]. About 90% of the genes missed in avian genomes are present in the crocodile genome suggesting that the loss occurred in the lineage leading from the archosaurian to the avian ancestor [Lovell et al., 2014]. This is the highest level of segmental loss of genes found in vertebrates so far. The fact that the majority of lost avian genes are members of gene families suggests that the remaining members of the same gene family in the avian genomes could have compensated for the loss of paralogs.

**Chromosome Structure**

While a detailed review of the karyotype evolution of birds is covered in the cytogenetic section of this report, here it is worth mentioning several aspects of this feature. A macro-synteny comparison of 6 avian genomes assembled to scaffolds with N50 > 10 Mb and the remaining genomes with N50 > 2 Mb [Zhang G et al., 2014a] confirmed previous observations based on cytogenetic studies that avian karyotypes are highly conserved [Burt et al., 1999]. An alignment of all avian genomes against chicken chromosome sequences is available from the Evolution Highway Comparative Chromosome Browser [Larkin et al., 2009] (fig. 3). Very few interchromosomal rearrangements were found in the majority of avian species confirming previous observations that the structure of avian chromosomes is stable [Romanov et al., 2014]. A micro-synteny analysis of homologous genes among avian genomes confirms the hypothesis of evolutionary stability of avian genomes, indicating that a higher fraction of genes share the same neighbours in avian genomes than in different mammals [Zhang G et al., 2014b]. However, it was observed that all vocal learning species had significantly more rearranged chromosomes than the closely related non-vocal learning species [Zhang G et al., 2014b]. This could be related to unique adaptations of the vocal learners or alternatively to a larger radiation of highly successful vocal learning clades compared to other avian groups [Zhang G et al., 2014b].

Among these genes, they identified two (YTHDC2L1 and TMRA) that are highly expressed in the vocal learning-associated nuclei of the zebra finch brain [Whitney et al., 2014].

**Nucleotide and Gene Evolution**

Comparative analysis of avian DNA sequences demonstrated that the nucleotide substitution rates in avian genomes are generally lower than those in mammals [Zhang G et al., 2014b]. However, large interordinal differences in the rates were detected between distinct avian groups. It was found that the substitution rate correlates with the number of species in a clade. For example Passeriformes, the most specious avian order, had a neutral nucleotide substitution rate 2 times higher than any other neoavian species on average. Interestingly, significant differences in the substitution rates were also observed when comparing all landbirds as a group to waterbirds, consistent with observations that landbirds have higher diversification rates [Zhang G et al., 2014b]. Overall, these observations suggest that differences in nucleotide
<table>
<thead>
<tr>
<th>Species</th>
<th>Genus</th>
<th>Family</th>
<th>Order</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opossum</td>
<td>Opossum</td>
<td>Rodentidae</td>
<td>Rodentiformes</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Little egret</td>
<td>Egretta minimus</td>
<td>Ardeidae</td>
<td>Gruiformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Hoatzin</td>
<td>Opisthocomus</td>
<td>Anseriformes</td>
<td>Gruiformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Chimney swift</td>
<td>Chaetura pelagica</td>
<td>Apodidae</td>
<td>Apodiformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Chimney swift</td>
<td>Chaetura pelagica</td>
<td>Apodidae</td>
<td>Apodiformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Rock penguin</td>
<td>Eudyptes schlegeli</td>
<td>Spheniscidae</td>
<td>Sphenisciformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Adelie penguin</td>
<td>Pygoscelis adelia</td>
<td>Spheniscidae</td>
<td>Sphenisciformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Emperor penguin</td>
<td>Aptenodytes forsteri</td>
<td>Spheniscidae</td>
<td>Sphenisciformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Crested ibis</td>
<td>Nipponia nippon</td>
<td>Phasianidae</td>
<td>Galliformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Downy woodpecker</td>
<td>Dryobates pubescens</td>
<td>Picidae</td>
<td>Piciformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Peregrine falcon</td>
<td>Falco peregrinus</td>
<td>Falconidae</td>
<td>Falconiformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Budgerigar</td>
<td>Melopsittacus undulatus</td>
<td>Psittacidae</td>
<td>Psittaciformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Golden-collared manakin</td>
<td>Manacus auratus</td>
<td>Pipridae</td>
<td>Psittaciformes</td>
<td>Aves</td>
</tr>
<tr>
<td>American crow</td>
<td>Corvus brachyrhynchos</td>
<td>Corvidae</td>
<td>Coraciiformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Chinese alligator</td>
<td>Alligator sinensis</td>
<td>Crocodylia</td>
<td>Crocodylia</td>
<td>Reptilia</td>
</tr>
<tr>
<td>Western-painted turtle</td>
<td>Malaclemmys leucomystax</td>
<td>Testudinidae</td>
<td>Testudines</td>
<td>Reptilia</td>
</tr>
<tr>
<td>Anole lizard</td>
<td>Anolis carolinensis</td>
<td>Lacertidae</td>
<td>Squamata</td>
<td>Reptilia</td>
</tr>
<tr>
<td>Opossum</td>
<td>Opossum</td>
<td>Rodentidae</td>
<td>Rodentiformes</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Rock penguin</td>
<td>Eudyptes schlegeli</td>
<td>Spheniscidae</td>
<td>Sphenisciformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Adelie penguin</td>
<td>Pygoscelis adelia</td>
<td>Spheniscidae</td>
<td>Sphenisciformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Emperor penguin</td>
<td>Aptenodytes forsteri</td>
<td>Spheniscidae</td>
<td>Sphenisciformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Crested ibis</td>
<td>Nipponia nippon</td>
<td>Phasianidae</td>
<td>Galliformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Downy woodpecker</td>
<td>Dryobates pubescens</td>
<td>Picidae</td>
<td>Piciformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Peregrine falcon</td>
<td>Falco peregrinus</td>
<td>Falconidae</td>
<td>Falconiformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Budgerigar</td>
<td>Melopsittacus undulatus</td>
<td>Psittacidae</td>
<td>Psittaciformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Golden-collared manakin</td>
<td>Manacus auratus</td>
<td>Pipridae</td>
<td>Psittaciformes</td>
<td>Aves</td>
</tr>
<tr>
<td>American crow</td>
<td>Corvus brachyrhynchos</td>
<td>Corvidae</td>
<td>Coraciiformes</td>
<td>Aves</td>
</tr>
<tr>
<td>Chinese alligator</td>
<td>Alligator sinensis</td>
<td>Crocodylia</td>
<td>Crocodylia</td>
<td>Reptilia</td>
</tr>
<tr>
<td>Western-painted turtle</td>
<td>Malaclemmys leucomystax</td>
<td>Testudinidae</td>
<td>Testudines</td>
<td>Reptilia</td>
</tr>
<tr>
<td>Anole lizard</td>
<td>Anolis carolinensis</td>
<td>Lacertidae</td>
<td>Squamata</td>
<td>Reptilia</td>
</tr>
</tbody>
</table>

(Mbp Mbp, GGA1:100K GGA1:500K)
substitution rates between avian groups could be related to the levels of diversification and the differences in adaptive pressures the diverse groups were subjected to.

In agreement with the hypothesis of evolutionary stasis in avian genomes (compared to other vertebrates and mammals particularly) [Burt et al., 1999], orthologous genomic intervals in bird genomes were found to contain more constrained sequences than orthologous regions of mammalian genomes. It was shown that 7.5% of the avian genome on average is found in sequences that evolve slower than at the neutral substitution rate estimated for birds [Zhang G et al., 2014b]. Only ~12.5% of these sequences were associated with protein-coding genes, while the majority of other sequences were intergenic or located in introns. These sequences are candidates for regulatory elements in avian genomes and often are sites for transcription regulatory factors [Woolfe et al., 2005]. This compares to 2.2% of the human and mouse genome sequences that are constrained in both species after divergence from their common ancestor [Rands et al., 2014]. Interestingly, the avian-specific constrained noncoding sequences are enriched for transcription factors functioning in metabolism, while the ancestral amniote-specific elements (shared by birds and mammals) are associated with factors contributing to signal regulation, stimulus responses and development [Zhang G et al., 2014b]. This indicates that the regulation of genes responsible for metabolism could have changed during the evolution of the ancestral avian genome.

The availability of a large number of avian genomes distributed among the whole avian phylogenetic tree allowed for the first time the identification of signatures of convergent evolution in avian genomes. For example, signatures of selection related to vocal learning, a phenotype that appeared in the avian evolution 3 times in different lineages [Zhang G et al., 2014b], were detected comparing the sequences of protein-coding orthologous genes of vocal-learning and control non-learning species. Out of 7,909 genes present in all vocal-learning and control non-learning species’ genomes, about 200 genes contained signatures of convergent accelerated evolution in vocal learners. Strikingly, 73% of these genes were expressed in the songbird brain with the vast majority of these being expressed in song-learning nuclei [Whitney et al., 2014]. The analysis of accelerated evolution of non-coding sequences revealed 822 accelerated elements shared by all 3 vocal learning groups. Of these elements, 322 were associated with 278 genes with a high proportion (192) expressed in song nuclei. Overall, there was a 2.0–3.5-fold enrichment in accelerated evolution associated with genes expressed in vocal learning brain regions compared to other tissues [Zhang G et al., 2014b].

**Adaptive Phenotypes**

Comparative analysis of avian genomes shed light on the evolution of avian-specific adaptations. The leading one – ability to fly – requires several major skeleton and bone modifications: bones must be light and at the same time strong. In birds and other tetrapods this was achieved through the reduction of the number of bones and their pneumatization [Zhang G et al., 2014b]. Out of 89 genes related to ossification in birds, 49 showed evidences of positive selection, which is twice as high as in mammals. The highest ratio of non-synonymous to synonymous substitutions (dN/dS) was detected for the genes AHSG and P2RX7 related to bone mineral density and bone homeostasis, respectively [Zhang G et al., 2014b]. An increased metabolism related to flight adaptations requires highly-efficient gas exchange in the lungs. Birds have evolved a constant volume lung and a rigid trunk region. Five genes related to the mammalian lung development were apparently lost in birds. The development of feathers (composed of α- and β-keratin proteins) is associated with an almost 2-fold increase in the number of β-keratin genes in birds compared to reptiles, while the α-keratin gene family was significantly contracted in avian species [Zhang G et al., 2014b].

An important phenotypic feature of birds that finally was explained through the comparative genome analysis of avian genomes is the lack of teeth or ‘edentulism’. This phenotype has evolved independently in several animal lineages. Fossil records show that several extinct avian lineages had teeth [O’Connor and Chiappe, 2011]. Therefore, edentulism could have evolved independently in avian evolution [Meredith et al., 2014]. A search of avian and opossum) contain complete HSBs, while blocks in the remaining rows represent either HSBs or syntenic fragments. Evolutionary breakpoint regions are defined as white intervals in between either 2 adjacent syntenic fragments originating from the same scaffold in a target genome or 2 adjacent HSBs. A complete set of all chicken chromosomes compared to other avian and non-avian genomes is available from http://eh-demo.ncsa.uiuc.edu/birds/.
genomes for the traces of tooth-related genes recovered multiple pseudogene fossils of enamel and dentin genes with multiple frame-shift or exon-deletion mutations. The majority of these mutations differ in diverse avian genomes, suggesting that they occurred independently in evolution. However, all birds analyzed by the Avian Genome Consortium shared the same deletions in 4 enamel genes (ENAM, AMELX, AMTN, and MMP20) and 1 dentin-related gene (DSPP) suggesting that the common ancestor of all birds likely had no mineralized teeth [Meredith et al., 2014].

Another amazing feature of birds is their most advanced vertebrate visual system. They exhibit the ability to distinguish colors over a wider range of wavelengths than mammals. Unlike mammals, birds likely have retained the ancestral tetrapod set of cones [Zhang G et al., 2014b]. For the majority of vertebrate visual opsin genes, birds had a higher number of copies compared to mammals. The number of opsin gene classes (4) found in most birds suggests that birds are most likely tetrachromatic [Zhang G et al., 2014b], with the exception of penguins who had only 3 classes of opsin genes [Li C et al., 2014] suggesting a 3-chromatic vision, consistent with earlier observations in aquatic mammals who also lost 1 or 2 of cone pigments [Newman and Robinson, 2005].

In conclusion, the comparative analysis of 48 avian genomes proved to be a powerful tool to reveal multiple signatures of genome adaptations related to avian ability to fly. The evolutionary stability of avian karyotypes is likely related to the reduction of transposable and other repetitive sequences in avian genomes. Avian-specific segmental deletions of gene paralogs together with shorter genes and intergenic regions made gene regulation fast and energy-efficient. The skeleton modifications that resulted in a smaller number of light-weight bones were accompanied by accelerated evolution of genes involved in ossification. Avian genomes tend to show relatively small variation in regulatory gene sequences compared to mammals reflecting the high degree of adaptation and specialization of bird genomes, probably inherited from their dinosaur ancestor.

**Avian Cytogenetics Goes Functional**
(Prepared by D.K. Griffin, M.N. Romanov, R. O’Connor, K.E. Fowler, and D.M. Larkin)

It is now over 10 years since the first avian genome [International Chicken Genome Sequencing Consortium, 2004] and the first complete avian karyotype [Mabanda et al., 2004] were both published; however, until 2014, avian cytogenetics has focused heavily on descriptive studies [e.g. Griffin et al., 2007, 2008; Skinner et al., 2009; Völker et al., 2010] with less attention to its functional relevance. Last year, however, saw 2 landmark efforts in the chromosomal studies of birds: a special issue of Chromosome Research in April and the announcement of recently completed sequences of multiple new avian genomes in Science and the BMC journals (taking the total number sequenced to over 50) in December. Studying the chromosomes of birds is, perhaps for the first time, telling us more about avian biology, function and evolution than it ever has.

**What Do We Know So Far? Karyotypic Stability**
The near-unique nature of the avian karyotype has remained a consistently reported feature of bird biology since the first chromosome preparations were made. Although many animal groups have microchromosomes, the small size and abundant number of chromosomes in avian species set birds apart genomicsmically from other vertebrate groups. To the best of our knowledge, there are over 1,000 published avian karyotypes, most comprehensively summarized by Christidis [1990], with several hundred added since this review. All of these karyotypes are partial however, with usually only 5–10 pairs of chromosomes easily distinguished, and the rest homogeneously classified. Moreover, the vast majority of karyotypes hardly differ from each another, with rare exceptions including the stone curlew (Burhinus oedicnemus; 2n = 40), the beach thick knee (Esacus magnirostris; 2n = 40), several hornbills (2n = 42), kingfishers and hoopoes (Upupa epops; 2n > 120) at each end of the numerical spectrum [Christidis, 1990]. Indeed, even since the advent of FISH, the identification of an interchromosomal rearrangement in a bird is a relatively uncommon event [Griffin et al., 2007].

Central to our understanding of avian biology and evolution is establishing the reasons why avian karyotypes are evidently so stable. Clues to such an inquiry might lie in those rare exceptions to the rule. For instance, the Falconiformes (falcon, etc.) and Psittaciformes (parrots, etc.) have noticeably undergone numerous evolutionary changes. Moreover, it is noteworthy that when interchromosomal change occurs, it tends to recur. The best example of this is a fusion of the ancestral chromosomes 4 and 10; an event that appears to have occurred independently throughout evolution in chicken (Gallus gallus), greylag goose (Anser anser), collared dove (Streptopelia decaocto) and probably other species also [Griffin et al., 2007]. In this review, we
examine some of the latest tools and preliminary solutions that are being used to understand the underlying mechanisms that lead to chromosome rearrangements in birds (and in eukaryotes in general).

If we accept that interchromosomal change occurs only rarely in birds, then it is reasonable to assume that this happens usually only when there is an adaptive value to doing so. In most species, phenotypic diversity is usually associated with wholesale changes in karyotype structure. Aves as a phylogenetic class underwent a series of rapid speciation events beginning $\sim 65$ million years ago (Mya) and ending $\sim 50$ Mya. Chromosomal change is usually a cause or consequence of speciation (i.e. a species barrier), but until recently, the microchromosomes that constitute the majority of the avian karyotype, have not been amenable to study. The latest studies, however, have paved the way for a flurry of research activity that not only describes the avian karyotype in more detail, but might also provide functional clues as to its nature.

**New Molecular Cytogenetic Tools**

Lithgow et al. [2014] produced a set of chromosome paints and bacterial artificial chromosomes (BACs) that will start the process of characterizing the microchromosomes and their changes over evolutionary time. They reported the development of chicken microchromosomal paint pools and generation of pairs of specific microchromosome BAC clones with some success in zoo-FISH experiments. For instance, they detected a fusion of the ancestral chicken chromosome 23 orthologue to a macrochromosome in gyrfalcon (*Falco rusticolus*). A FISH image of BACs hybridized to peregrine falcon (*Falco peregrinus*) chromosomes [our unpubl. data] is shown in figure 4.

McPherson et al. [2014] examined the Japanese quail (*Coturnix japonica*). Comparing chicken and turkey BAC clones on mitotic and meiotic chromosomes, they demonstrated that high-resolution FISH is practicable. Ishi-shita et al. [2014] also assessed the distribution of centromeric repetitive sequences on both micro- and macrochromosomes. It is therefore now possible to achieve full, high-resolution characterization of all avian chromosomes in all species studied, including the elusive chromosome 16 and the D-group (smallest) chromosomes. There are several current strategies to fill the gaps; one of these is by the use of PacBio, a novel single-molecule real-time sequencing platform, targeting the sequence of smaller chromosomes using sorted chromosome preps, and assembling contigs into scaffolds and super-scaffolds from optical maps [Ganapathy et al., 2014].

**What Have Sequence Assemblies Taught Us?**

The progress of genome assembly in birds has been slow in comparison to other animal groups such as mammals. Following chicken [International Chicken Genome Sequencing Consortium, 2004], it took a further 6 years until the second and third avian genome sequences were published, namely those of the zebra finch (*Taeniopygia guttata*, a model for neurological function, especially learned vocalization) [Warren et al., 2010] and turkey (*Meleagris gallopavo*) [Dalloul et al., 2010]. More recently, the Pekin duck (*Anas platyrhynchos*) [Huang et al., 2013] was added along with 2 falcon species (*Falco peregrinus* and *F. cherrug*) [Zhan et al., 2013] and many others (table 4). The availability of these assembled genomes provided the opportunity for comparative genomics at a chromosomal level. In 2010, we made the first comparison of 2 species using genome assembly information from the macrochromosomes [Völker et al., 2010]. A similar comparison more recently was made in chicken compared to duck [Rao et al., 2012], and then a 3-way comparison (allowing studies of the direction of change) in chicken, turkey and zebra finch [Skinner and Griffin, 2012; Lithgow et al., 2014]. The principal features of chromosomal change in birds are homologous synteny blocks (HSBs), which are demarked by evolutionary breakpoint regions (EBRs). While analyzing these features, some general patterns have started to emerge. The first is that, although interchromosomal change is rare, intrachromosomal changes are commonplace. Breakpoint reuse is also commonplace, significantly more so than in mammals, and there is some evidence of an association between chromosomal breakage and non-allelic homologous recombination (NAHR) [Völker et al., 2010].
Zhang G et al. [2014b] used a whole-genome shotgun strategy to generate new whole-genome sequences from 45 bird species representing many of the major clades and at least 1 representative from over 90% of all avian orders. Around 20 species had a high (50× or greater) coverage and these were the subjects of further cytogenetic studies. These included the common ostrich (Struthio camelus) and the budgerigar (Melopsittacus undulatus), which were further assembled using data from optical mapping experiments [Ganapathy et al., 2014]. This had the effect of significantly increasing the assembly’s N50 scaffold sizes to around 15 Mb, and these were subsequently used with those already assembled by chromosome (chicken, turkey, zebra finch, and duck). Romanov et al. [2014] made use of novel whole-genome sequence information from 21 avian genome sequences available on an interactive browser (Evolution Highway). By focusing on the 6 best-assembled genomes (chicken, turkey, duck, zebra finch, ostrich, and budgerigar), a putative karyotype of the avian ancestor (probably a bipedal feathered dinosaur) was assembled for each chromosome. The evolutionary events were reconstructed that led to each of the 6 species’ genome organization. Intra- and interchromosomal changes appear best explained most parsimoniously by a series of inversions and translocations with common breakpoint reuse. Microchromosomes represent conserved blocks of synteny in most of the 21 species, and a series of interchromosomal changes in the ostrich were also described that would not have been predicted by karyotype analysis alone. These results suggest that mechanisms exist to preserve a static overall avian karyotype/genomic structure, including the microchromosomes, with rare interchromosomal change (e.g. in ostrich and budgerigar lineages); this is discussed in depth in the next section. Of the species examined, it seemed that chicken had the least number of chromosomal rearrangements compared to the dinosaur ancestor. From Evolution Highway it is also possible to assess rates of chromosomal evolution in birds. Zhang G et al. [2014b] suggest that birds have a lower chromosomal rearrangement rate than mammals but nonetheless can undergo ‘bursts’ of rearrangement, e.g. during the evolution of vocal learning. This finding corroborates those of Romanov et al. [2014] that identified the zebra finch and budgerigar as the 2 species with the most chromosomal rearrangements from the avian ancestor.

If we accept that chicken and its galliform relatives underwent the least number of chromosomal changes whilst diverging from the ancestral bird, we also must consider whether they also have undergone the fewest phenotypic changes. In other words: is the dinosaur avian ancestor more like a land fowl than any other bird? The most ancient near-certain fossil representative of modern birds (Neornithes) was almost certainly aquatic (for example, Vegavis, a genus of birds from the Late Cretaceous epoch) and has been identified as a Galloanserae. Indeed, the earliest known bird-like creatures in the fossil record (e.g. the Ornithurae Gansus) were either fully aquatic or at least amphibious, and it has been suggested that, due to the fact that they had webbed feet (as well as other traits), they were more like ducks [Romanov et al., 2014]. On the other hand, most authors agree that the dinosaur ancestors of birds were terrestrial, feathered, bipedal, relatively small and with limited flying ability – not unlike a chicken. At best we can determine therefore, the ancestral birds were most likely more phenotypically associated with the Galloanserae, and the confusion of whether they were more akin to water- or land fowl may be due to interpretations based on depositional sampling biases, limited understanding of functional anatomy, and whether the individuals that have been discovered are actually fully representative of the groups to which they belonged. Chromosomal evidence provides an independent record of the functional material of inheritance in living birds and, as such, can complement a fossil record that is always likely to be incomplete.

Of all species studied so far, it seems clear that the rearrangement of chromosomes is non-random [Pevzner and Tesler, 2003; Larkin et al., 2009]. The reasons for this non-random nature warrant deeper investigation. According to mammalian evidence, evolutionarily conserved HSBs appear to evolve in different ways from the dynamic and ever-changing EBRs; whether this is true of birds remains to be seen. In mammals, chromosomal breakpoints are correlated with sequences of segmentally duplicated or repetitive DNA [Bovine Genome Sequencing and Analysis Consortium et al., 2009; Larkin et al., 2009; Groenen et al., 2012; Ruiz-Herrera et al., 2012], and species-specific EBRs are correlated with regions enriched for transposable elements (TEs) [Bovine Genome Sequencing and Analysis Consortium et al., 2009; Groenen et al., 2012]. In mammals, EBRs and HSBs largely contain genes with notably different functional ontologies, e.g. organismal development in HSBs [Larkin et al., 2009] and lineage-specific biology and adaptive features in EBRs [Bovine Genome Sequencing and Analysis Consortium et al., 2009; Larkin et al., 2009; Groenen et al., 2012]. It has been suggested therefore that chromosome rearrangements and the respective gene ontologies con-
tained within HSBs and EBRs help to explain lineage-specific phenotypes in mammals. Mammalian and avian genomes are very different however (not least because of the interchromosomal stability of avian genomes), and thus the question remains about whether the patterns that have been observed in mammals will apply to birds also. Birds have less repetitive DNA through the elimination of repetitive sequences [International Chicken Genome Sequencing Consortium, 2004; Shedlock, 2006; Zhang G et al., 2014b] so that the avian genome is constrained by size, primarily because of gene loss as well as lineage-specific erosion of repetitive elements and large segmental deletions. In addition to their karyotypic stability, bird genomes also have a very high degree of evolutionary stasis at nucleotide sequence and gene synteny levels. Nonetheless, one of the key findings was the detection of non-neutral evolutionary changes in functional genes as well as non-coding regions. Many of these changes coincide with adaptations to different lifestyles and niches and display homoplasy [Zhang G et al., 2014b].

The non-random nature of chromosome rearrangement in birds, the reasons for the apparent interchromosomal (but not intrachromosomal) stability of avian karyotypes (see next section), the role of TEs and NAHR, the relationship to phenotype, the question of whether spatial organization of ancestral gene networks is maintained in bird and other reptile lineages, and the question of whether lineage-specific EBRs alter gene order in networks that had adaptive value, all require further investigation. Harnessing the data from over 50 avian genomes (undoubtedly with many more on the way) and employing tools such as Evolution Highway will give us unprecedented insight into avian chromosome evolution and its relationship to avian biology.

Why Is the Avian Karyotype Structure Conserved Inter- but Not Intrachromosomally?

Burt’s ‘fission-fusion’ hypothesis suggested that most avian microchromosomes became fixed in the common dinosaur ancestor with a karyotype of 2n ≈ 60 including 20 microchromosome pairs [Burt, 2002]. The remainder, including the smallest, probably was created by further fission. Romanov et al. [2014] suggested that a basic pattern of 2n = 80 (∼30 microchromosome pairs) was fixed before the Palaeognathae-Neognathae divergence 100 Mya. The subsequent paucity of intermicrochromosomal rearrangements between most Neognathae indicates an evolutionary advantage either to retaining this pattern or a lack of opportunity for change. For instance, an explanation for such evolutionary stasis might be that the underlying mutational mechanisms of chromosomal changes are fundamentally different in birds compared to other amniotes through a lack of adaptive value, rather than purifying selection, slowing down the rate of change. Much of this could be explained, in part, by a paucity of copy number variants (CNVs; including segmental duplications), recombination hotspots, TEs and/or endogenous retroviruses; however, this would not explain why interchromosomal change is rare but intrachromosomal change is common, particularly in groups that have undergone rapid speciation such as Passeriformes.

The rate of chromosome rearrangement (and subsequent speciation) depends on: (1) the mutation rate and (2) the fixation rate [Burt et al., 1999]. The first of these is related to the frequency of homologous sites [Burt, 2002]. Repeat structures in general (e.g. CNVs), and TEs in particular, provide substrates for chromosomal rearrangement. In a genome constrained by size, the opportunity for mutation is reduced and only fission (or intrachromosomal change, e.g. inversion) can occur. This provides an explanation why (1) avian genomes are more fragmented than any other vertebrate (birds have the most chromosomes) and (2) why there have been fewer interchromosomal rearrangements. There might also be advantages to retaining multiple chromosomes in a karyotype through the generation of variation, the driver of natural selection. That is, a karyotype with more chromosomes leads to a greater number of genetic variants that the gametes produce and an increase in recombination rate due to the fact that there needs to be at least 1 obligatory chiasma per chromosome. Burt [2002] proposed that a higher recombination rate has also led to the features that we most associate with microchromosomes (high GC content, low repeats, high gene density, etc.) and resulted in the formation and fixation of the archetypal avian karyotype with both macro- and microchromosomes and little interchromosomal rearrangement. Such a constraint, however, does not preclude rearrangement within the individual chromosomes. Romanov et al. [2014] and King [1995] argue that an increase in intrachromosomal rearrangement correlates with bursts of speciation in birds, perhaps mediated by an increase in localized repeat content.

Some birds nonetheless have a significantly different karyotype from the standard 2n ≈ 80. This can occur within one closely related group, e.g. Adélie penguin (Pygoscelis adeliae; 2n = 96) and the emperor penguin (Aptenodytes forsteri; 2n = 72) (but both associated with high degrees of intermicrochromosomal rearrangement), thereby suggesting that similar mechanisms can both re-
duce or increase chromosome number in relatively short time frames. Comparisons of chromosomal change in the zebra finch and the budgerigar suggest that rearrangement rates are similarly high in both groups to which they belong (Passeriformes and Psittaciformes, respectively) but that the latter is capable of fixing interchromosomal rearrangements, while the former is not. The mechanisms underpinning these differences are, as yet, unknown, but studies of the gene ontology terms of species-specific EBRs might provide clues. As more avian genomes with better assemblies are analyzed, this may indicate adaptive phenotypic features associated with specific gene ontologies typical of individual orders, families or genera.

The Sex Chromosomes

Worthy of especial consideration is the conserved sex chromosome ZW system that is present in all birds apart from the Palaeognathae. Their independent origin from the XY system does not escape the fact that similar mechanisms appear to have run in parallel, for instance genes on the Z chromosome (like the mammalian X) have undergone selection for male-advantage functions. Like the Y chromosome, the W is small (albeit medium-sized by avian standards), heterochromatic and gene poor. Graves [2014] suggests that the W chromosome is at a more advanced stage of differentiation than the Y chromosome as it has accumulated more LINEs and lost more genes during its evolution. Pokorná et al. [2014] considered multiple sex chromosomes and meiotic drive in a range of amniotes. This study noted that the single ZW system in birds contrasts with that of other reptiles and amniote groups; they raised a very exciting hypothesis that this contrast may possibly be related to the differential involvement of sex-specific sex chromosomes in female meiosis (females being the heterogametic sex). Early in the assembly of the chicken genome, the quality of the build of both the Z and W sex chromosomes was very poor, and limited studies existed on sex determination. Since this, the Z chromosome was painstakingly assembled and sequenced BAC by BAC [Bellott et al., 2010], and is now one of the best-assembled chromosomes in the chicken genome. The same is now expected for the W sex chromosome, which currently is very poorly assembled [Chen et al., 2012]. Zhou Q et al. [2014] conclude that the ancestral sex chromosome organization is closer to that of the Palaeognathae (ostrich and emu) and demonstrated that there is less degradation of the sex chromosomes and a closer synteny with non-avian reptile species.

Copy Number Variation

Redon et al. [2006] first highlighted the impact of CNV in the human genome. This seminal study heralded a new era in cytogenetics and has subsequently been applied to many other species and groups including birds. Skinner et al. [2014] provided a global overview of apparent cross-species CNVs in birds using cross-species array-CGH. Griffin and Burt [2014] pointed out issues of definition in that ‘copy number variation’, strictly speaking, refers to polymorphisms within a species. The question arises therefore whether results of cross-species array-CGH represent genuine variation in copies of orthologous genes between species. Skinner et al. [2014] stated that ‘difference in gene copy number between species is a question of gene duplication, segmental duplications etc. and may be driven by expansion and contraction of paralogs within different gene families.’ Nonetheless, this paper provided a broad appraisal of apparent cross-species CNVs in 16 avian species. Microchromosomes appear to have more apparent CNVs than macrochromosomes. Indeed, in species with microchromosomal fusions such as Falconiformes, the fused ‘former microchromosomes’ still retained their ancestral features such as a higher degree of cross-species CNVs. Skinner et al. [2014] reported that ~50% of the apparent cross-species CNVs overlap with known chicken-specific CNVs. In terms of gene ontology, there appears to be a general enrichment in immune response and antigen presentation genes as well as 5 CNV regions perfectly correlated with the unique loss of sexual dichromatism. More specifically, there were also suggestions of CNVs involved in diet in turkey (proteolytic digestion/degradation of trypsin inhibitors), and correlation of the unique migratory behaviour of common quail among fowl through the following genes: OBSCN associated with hypertrophy of myofibrils, and MAPK8IP3 implicated in respiratory gaseous exchange [Skinner et al., 2014]. There were also suggestions of an association with muscle activity in falcons through the gain of MYOZ3, preferentially expressed in fast-twitch myofibers and skeletal muscle and an association between immune function in the common quail (Coturnix coturnix) and silver pheasant (Lophura nycthemera) (LEAP2 and ITCH) as well as homeotic genes in common pheasant and California quail (SCML2 and DLX5). Finally, Skinner et al. [2014] identified cross-species CNVs associated with brain development and neuronal function in turkey (e.g. loss of CTXN1), common quail (gain of LRFN5) and duck (e.g. DLGAP2).
Conclusions

The most recent advances in avian cytogenetics have culminated in great promise not only for the study of bird karyotypes, but also for providing insight into the mechanisms of chromosome evolution in general. New avenues for investigation include gene regulation; for instance, it will become necessary to map accurately the physical location of polyadenylation and transcription start sites, important reference points that define promoters and post-transcriptional regulation. It will also become possible to sequence full-length transcripts, to allow accurate identification of alternate splicing events and their controlling elements. The ENCODE (Encyclopedia of DNA Elements) project has helped to define functional elements of the human genome, including those aforementioned as well as other chromatin signals, e.g. active chromatin, enhancers, insulators, methylation domains, etc. An effort of agENCODE is underway to include agriculturally important birds such as chicken, turkey, duck, quail, and perhaps ostrich. The study of cytogenetics will be essential here in helping to define higher-order structures in nuclear organization that show regulatory interactions within and between chromosomes. Finally, reconstruction of evolutionary events allows us to study genome organization and function not only in extant but, by extrapolation, in extinct species also. Reconstruction of avian-reptilian ancestral karyotypes will allow us to define chromosomal rearrangements in long-dead species that have captured the public imagination. Here be dragons!

Hypermethylated Chromosome Regions in Chicken and Other Birds


The advent of specific antibodies against the different nucleosides and nucleotides has promoted direct cytogenetic analyses of the various DNA classes along eukaryote chromosomes. These antibodies were first produced by the group of Bernard F. Erlanger some decades ago [Erlanger and Beiser, 1964; Garro et al., 1968; Sawicki et al., 1971; Erlanger et al., 1972]. They were produced by immunizing rabbits to bovine serum albumin (BSA) conjugated to one of the DNA bases. The antibodies are reactive with the BSA conjugate used to induce them and also with single-stranded DNA [Erlanger and Beiser, 1964]. They are highly specific for the base and show little or no cross-reaction with the other bases. Over the years, a series of such polyclonal antibodies were produced, with specificities for a number of nucleosides, nucleotides and dinucleotides [Dev et al., 1972; Erlanger et al., 1972; Miller, 1973]. In the early 1990s, the first monoclonal antibodies against 5-methylcytosine (5-MeC) and other modified nucleosides were produced [Reynaud et al., 1992] and subsequently used for chromosome staining [Barbin et al., 1994; Miniou et al., 1994; Montpellier et al., 1994; Bernardino et al., 1996].

Of special interest were antisera specific for 5-MeC which were initially applied by the group of Orlando J. Miller to the chromosomes of several mammalian species, including human, chimpanzee, gorilla, cattle, mouse, and kangaroo rat [Miller et al., 1974; Schreck et al., 1974, 1977; Schnedl et al., 1975, 1976]. Using an immunofluorescence technique and anti-5-MeC antibodies, they showed that methylated DNA can be detected in fixed metaphase chromosomes after they have been UV-irradiated to generate regions of single-stranded DNA. In these species, the methylated regions corresponded to the locations of repetitive DNA, i.e. to the heterochromatic regions of all or a subset of the chromosomes in the karyotypes. Subsequently, this technique was applied to chromosomes of further mammalian species [Vasilikaki-Baker and Nishioka, 1983; Bernardino et al., 2000] and to human chromosomes [Barbin et al., 1994; Montpellier et al., 1994; Bernardino et al., 1996; Kokalj-Vokac et al., 1998], even including cases of inherited chromosome aberrations [Breg et al., 1974] and leukemia cell lines [Ben-saada et al., 1998].

With one exception [Grützner et al., 2001], no immunofluorescence studies on the distribution of hypermethylated regions in bird chromosomes have been published. The present report is a brief summary of the results obtained for avian chromosomes in an ongoing project on the hypermethylation patterns in vertebrate chromosomes [Schmid et al., in preparation].

Mitotic chromosomes of 13 species from 7 orders, belonging to both modern (Neognathae) and primitive (Palaeognathae) birds (table 13), were prepared from embryonic or skin fibroblast cell cultures following standard techniques (colcemid treatment, exposure to hypotonic solution, fixation with methanol:acetic acid). Hypermethylated DNA was detected by indirect immunofluorescence using a monoclonal antibody against 5-MeC. In double-stranded DNA, the methyl groups are hidden in the phosphodiester backbone of the double helix and not accessible to the antibody. The anti-5-MeC antibody recognizes and binds to its target only if the DNA is in the single-stranded configuration. Therefore, the slides with
the chromosome preparations were immersed 1 cm below the level of a buffer solution (PBS) and denatured by UV-light irradiation for 2.5–3 h at a distance of 10 cm from a UV lamp (254 nm). For indirect immunofluorescence, the slides were first incubated in a coplin jar for 1 h in blocking solution (PBS, with 0.3% BSA, 0.1% Tween) and then with 50 μl of a monoclonal mouse anti-5-MeC (primary) antibody (Imprint® monoclonal anti-5-methyl-cytosine antibody 33D3, Sigma-Aldrich) diluted 1:1,000 with the blocking solution in a humified incubator at 37 °C for 1 h. A non-siliconized coverslip (22 × 60 mm) was placed on the 50-μl drop to spread the anti-5-MeC antibody over the complete slide surface. Subsequently, the slides were washed twice in PBS (with 0.3% BSA) for 3 min each, and then incubated with 70 μl of the secondary antibody (TRITC-conjugated rabbit-anti-mouse IgG, Sigma-Aldrich) diluted 1:200 with PBS. The incubation conditions were as for the primary antibody. After 2 further washes with PBS for 3 min each, the chromosome preparations were mounted in Vectashield® mounting medium (Vectashield) with DAPI. Image analysis was performed with Zeiss epifluorescence microscopes equipped with thermoelectronically cooled charge-coupled device cameras (Applied Spectral Imaging) using FISHView 6.0 software. The demonstration of constitutive heterochromatin (C-bands) in the chromosomes followed the technique of Sumner [1972]. C-banded metaphases were photographed under a Zeiss Axio ImagerA1 epifluorescence microscope employing the digital imaging system software BandView 6.0 (Applied Spectral Imaging).

The present study shows that, like in mammals, the hypermethylated regions in the karyotypes of birds are largely confined to constitutive heterochromatin (figs. 5, 6; table 13). In some preparations, faint fluorescence signals are also located outside the constitutive heterochromatin, but these are not consistent if many metaphases...
are analyzed (e.g. macroautosomes 1 in fig. 5a, b). Again, similar to mammals, there seems to be no strict rule on the distribution of hypermethylated chromosome regions in birds. In most instances, hypermethylated heterochromatic regions are located in the centromeric regions of chromosomes, but in sex chromosomes they can also be found in telomeric positions (fig. 5d, e). In most of the species examined, the centromeric heterochromatin in many, if not all, of the microchromosomes is hypermethylated (figs. 5a, b, 6a, b; table 13), but there are no-
table exceptions. Thus, in the karyotype of *Pavo cristatus*, the only detectable hypermethylated heterochromatic regions are located in the macroautosome pair 1 and in the Z chromosome. None of the microchromosomes is labeled (fig. 5d, e).

The nearly completely heterochromatic W sex chromosomes of neognathous birds, which are distinctly smaller than their corresponding Z chromosomes, show distinct hypermethylation, although in some species the W chromosomes exhibit only partial immunofluorescent labeling (fig. 5a, b). In contrast, in the extant palaeognathous birds (ostrich, emu, nandu), the primitive ZW sex chromosomes are still largely homomorphic, mostly euchromatic and contain similar genes [Fridolfsson et al., 1998; Ogawa et al., 1998; Shetty et al., 1999; Nishida-Umehara et al., 2007]. No hypermethylated heterochromatic regions are present in the W chromosomes of these species (table 13).

In the female specimens examined, visible differences in the hypermethylation patterns between the 2 Z sex chromosomes are not present. Depending on the species, both Z chromosomes contain either no hypermethylated heterochromatic regions at all, or both have clear fluorescing signals which were present in all metaphases analyzed (fig. 5d, e).

A close analysis of high-quality karyotypes of those species possessing many hypermethylated microchromosomes reveals that, even in the smallest microchromosomes, the immunofluorescence labeling is restricted to the heterochromatic centromere regions and does not extend over the whole microchromosomes (fig. 6a, b). In a previous study on neognathous birds (chicken, quail, pheasant) and palaeognathous birds (nandu, emu), Grützner et al. [2001] visualized the hypermethylated chromosome regions by the technique of indirect immunofluorescence, but used a different, well-characterized antibody against 5-MeC [Miniou et al., 1994; Mayer et al., 2000], and denatured the chromosome preparations in 70% formamide, 2× SSC for 1 min at 90°C. Contrary to the present study, they observed an apparently complete immunofluorescence labeling of the microchromosomes. They correlated this high density of methylated cytosines with the high gene density on the avian microchromosomes. Indeed, there is considerable evidence for an increased density of genes in microchromosomes [Smith J et al., 2000]. They have a

---

Figure 6. Karyotypes of female mallard duck (*Anas platyrhynchos*) after indirect immunofluorescence using a monoclonal antibody against 5-MeC (a, b) and C-banding (c). The 5-MeC-rich heterochromatic regions show red fluorescence signals, the chromosomes are stained blue with DAPI.
The figures published by Grützner et al. [2001] show dense clusters of microchromosomes, so it is not clearly evident if the immunofluorescence actually extends over the entire microchromosomes. In this context, it is important to emphasize that immunofluorescence patterns do not only reflect inherent chromosomal properties (i.e. distribution and enrichment of 5-MeC) but are also dependent on the different denaturation conditions and antibody specificities used. Moreover, the methylation status does also depend on the cell type and culture conditions. Admittedly, some sites of DNA methylation may not be recognized by the technique of UV-irradiation used here because the anti-5-MeC antibody can attach only to regions of single-stranded DNA. UV-irradiation is particularly effective in generating single-stranded regions in the DNA which is rich in AT base pairs with adjacent pyrimidines so that thymine dimer formation can occur [Schreck et al., 1974; Schnedl et al., 1975]. In DNA regions which are not denatured by the UV-irradiation technique, 5-MeC would not be recognized. However, partially denatured DNA strands in euchromatic chromosome regions do not suffice to explain the differences obtained in the present study and those published by Grützner et al. [2001].

Although karyotypes of a very limited number of avian species were examined with the present technique, it already becomes obvious that more closely related species within the same family do not share similar patterns of hypermethylated heterochromatic regions (table 13). Apparently, these patterns do not represent stable properties reflecting a common phylogenetic ancestry of different taxa. This is to be expected, because the major components of constitutive heterochromatin are repetitive DNA sequences which are characterized by repetition of relatively long monomers (of a few hundred base pairs) over many megabases of DNA [for reviews, see Brutlag, 1980; Long and Dawid, 1980; Singer, 1982; Southern, 1984; Beridze, 1986]. It is not uncommon to find up to 25% of a genome made up of different repetitive DNA families [for review, see Lohe and Roberts, 1988]. Even among closely related species, reiterated DNAs usually differ in quantity, sequence and chromosomal location [Miklos, 1985; Charlesworth et al., 1994]. Since they are not subjected to evolutionary selection pressure, sequence changes in repetitive DNAs can be accumulated and fixed in genomes very much faster than can the changes in functional (transcribed) DNA. As a consequence, constitutive heterochromatin is heterogeneous within and between species [for review, see Verma, 1988]. This heterogeneity of heterochromatin is paralleled by extremely rapid changes of its hypermethylating patterns.

The data obtained indicate that, like in mammals, the hypermethylation patterns of constitutive heterochromatin in birds are species-specific with respect to size, location and staining intensity. Because of this species-specificity, they are extremely useful as cytogenetic parameters for differentiating between closely related species that possess the same diploid number and chromosome morphology. It should, however, be emphasized that the results of this preliminary study were obtained exclusively on mitotic chromosomes prepared from fibroblast cultures. It must be confirmed whether the hypermethylation patterns of heterochromatin actually constitute a stable intraspecific feature, or if there exist tissue-specific patterns. In this context, the examination of meiotic chromosomes, especially those in the lambrush stage of female bird meiosis is important.

An Overview of Avian Evolution
(Prepared by S.B. Hedges)

Present and Past Diversity
The living birds form a large and diverse group of vertebrates, with nearly 10,000 species in ~2,000 genera, 200 families and 29 orders [Sibley and Monroe, 1990; IUCN, 2014]. New species are discovered each year, but at a low rate indicating that our knowledge of current avian diversity is quite good. The major uncertainty in the total number of species concerns disagreement over the definition of particular species, such as the recognition of subspecies as full species, rather than discovery of new species. In contrast, other groups of tetrapods (e.g. mammals, lizards, snakes, amphibians) have a higher rate of species discovery [Uetz and Hošek, 2014].

The fossil record of birds is relatively poor, probably because the avian skeleton is fragile (hollow bones), and most species are small and occur in environments (humid forests) where decomposition is rapid. However, the early history of birds has become much better known in recent decades with the discovery of exceptionally preserved fossils of birds and dinosaurs from the Cretaceous (~150 Mya, Mya), especially in China [Chiappe and Dyke, 2002]. The earliest bird is Archaeopteryx from the late Jurassic (~150 Mya) of Ger-
many, and the consensus view is that birds evolved from carnivorous (theropod) dinosaurs called coelurosaurs (e.g. *Tyrannosaurus rex*), and specifically from a group of relatively small and agile species called dromaeosaurids (e.g. *Velociraptor*). This was suggested in the 19th century when *Archaeopteryx* was first classified as a small dinosaur, before the impressions of feathers were noted, and has since been supported by detailed comparisons of anatomy and (recently) egg morphology [Varricchio et al., 1997]. To recognize this derivation from dinosaurs, paleontologists sometimes refer to birds as ‘living dinosaurs’ and refer to classical dinosaurs (e.g. *Stegosaurus*) as ‘non-avian dinosaurs’. The carnivorous (theropod) ancestry of birds is also illustrated by the fact that *Archaeopteryx* and many of the other Mesozoic (251–65 Mya) species possessed sharp teeth and raptorial claws.

Discoveries of feathered dinosaurs in the last 2 decades have also bolstered the link between birds and dinosaurs. In the earliest-branching species of coelurosaurs, the feathers were filamentous and probably functioned more like the down of some living birds, providing insulation [Chiappe and Dyke, 2002]. However, the maniraptoran coelurosaurs (including dromaeosaurids), which were the closest relatives of birds, possessed vaned feathers that more closely resembled avian flight feathers. There is no evidence that these dinosaurs could undergo the sustained flapping flight of modern birds, but extensive development of feathers and their configuration indicate that at least some of these dinosaurs were gliders [Xu et al., 2003]. These fossils have helped to blur the distinction between dinosaurs and birds. Nonetheless, the full suite of adaptations for sustained, flapping flight, such as asymmetric feathers, alulas (wing structures for improving the airfoil function), and other aerodynamic structures, are found only in birds. Even *Archaeopteryx* probably was able to initiate flight directly from the ground [Chiappe and Dyke, 2002].

Debate continues as to how the Cretaceous birds are classified. One arrangement divides them into 2 large groups (Sauriurae and Ornithurae), whereas another places them in a ladder-like tree of lineages leading to modern birds [Chiappe and Dyke, 2002]. In either case, the closest relative of modern birds is believed to be *Ichthyornis*, a small, toothed, tern-like marine bird of the late Cretaceous. Some Cretaceous fossils have been postulated to be representatives of modern orders such as Galliformes (e.g. fowl), Anseriformes (e.g. ducks), Psittaciformes (e.g. parrots), Charadriiformes (e.g. plovers), Procellariiformes (e.g. petrels, albatrosses), Gaviiformes (e.g. loons), Gruiformes (e.g. cranes), and Pelecaniformes (e.g. pelicans). However, all of these fossils are considered to be problematic in some way [Chiappe and Dyke, 2002]. With those aside, the fossil record of birds shows a major dichotomy at the Cretaceous-Tertiary boundary (65 Mya), when a large asteroid collided with Earth causing the extinction of the dinosaurs and other life. Almost no unambiguous fossils of modern (neornithine) birds are known before that event and no fossils of non-neornithine birds are known after that event. One important exception is *Presbyornis*, an aquatic neornithine bird apparently related to ducks (Anseriformes) which is known from the late Cretaceous and the early Tertiary [Kurochkin et al., 2002]. Fossils of most orders of modern birds appear in the early part of the Cenozoic (65–0 Mya).

**Phylogenetic Relationships**

Until recently, the phylogeny of modern birds has been poorly known despite decades of attention from morphologists and molecular phylogeneticists. Although early molecular studies [Sibley and Ahlquist, 1990] identified some obvious cases of convergence in previous morphological classifications and have helped to better organize the higher-level classification of birds, it has, in general, been a slow progression of knowledge. Probably the best explanation of the slowness is data limitation (need for many genes) and closeness in time of divergence among the neoavian orders [Cracraft et al., 2004]. Recent genome analyses have started to close this knowledge gap and have finally begun to resolve the tree of bird orders [Jarvis et al., 2014].

Some of the higher-level clades have been firmly established (fig. 7). For example, there is wide agreement that modern birds (Neornithes) form 3 major clades: Paleognathae (tinamous and ratites), Galloanserae (e.g. ducks, fowl), and Neoaves (all other birds). In addition, DNA sequence analyses, primarily of nuclear genes, indicate that Galloanserae and Neoaves are closest relatives [Groth and Barrowclough, 1999; van Tuinen et al., 2000; Hackett et al., 2008; van Tuinen, 2009; Jarvis et al., 2014]. Although earlier DNA hybridization studies had correctly identified the 3 groups [Sibley and Ahlquist, 1990], those data were ambiguous regarding their relationships. Initially, even data from complete mitochondrial genomes (~16 kb) provided a conflicting signal, indicating that Passeriformes was the most basal clade of modern birds and that the paleognaths and gallinaceous birds were closer relatives [Härlid and Arnason, 1999; Mindell et al., 1999]. Later mitochondrial analyses with additional taxon found support for the paleognath-basal tree [Paton et al., 2002; Garcia-Moreno et al., 2003].
The living ratites (flightless paleognaths) include the ostrich of Africa, rheas of South America, the emu of Australia, the cassowaries of Australia and New Guinea, and the kiwis of New Zealand. Extinct species occurred on Madagascar (elephant bird), New Zealand (moas), and elsewhere. The phylogeny of these birds has garnered considerable interest, in part because of their primarily southern distribution and the expectation that the breakup of Gondwana influenced their biogeographic history. Although the moas are now extinct, complete mitochondrial genome sequences have been obtained from fossil remains, and phylogenetic trees have been constructed with large data sets of DNA sequences. Nonetheless, the relationships have proven to be difficult to resolve. Three of the Australasian species (kiwi, emu and cassowary) form a group in most analyses, but the moa usually branches basally when present in the tree [Lee et al., 1997; van Tuinen et al., 1998; Cooper et al., 2001; Haddrath and Baker, 2001; Paton et al., 2002]. In contrast, morphological trees place the kiwi as a basal lineage, separate from the other Australasian species [Lee et al., 1997]. More recently, a nuclear gene study found tinamous to be nested...
among ratites rather than as closest relatives [Hackett et al., 2008]. For these reasons, firm biogeographic conclusions must await additional data and analyses.

The Galloanserae includes 2 orders of primarily robust-bodied birds. The Galliformes consists of mostly non-aquatic, ground-dwelling birds such as game fowl, megapodes, guans, and chachalacas. The chicken (Gallus gallus; common name, domestic fowl) belongs to this order. Its wild counterpart (also G. gallus) goes by the common name red junglefowl and is native to the lowlands (<2,000 m) of southern Asia and the Malay archipelago [Sibley and Monroe, 1990]. It is placed in the family Phasianidae along with other widely domesticated species such as the common quail (Coturnix coturnix), turkey (Meleagris gallopavo) and pheasant (Phasianus colchicus). Other members of the order Galliformes include the New World quails (Odontophoridae) and guinea fowl (Numididae). The other order within Galloanserae, the Anseriformes (ducks, geese, swans, screamers), is primarily aquatic and comprises about 450 species in ~125 genera [Sibley and Monroe, 1990].

Neoaves is the third major clade of living birds and accounts for 95% of the species. Although some of the 20 or so neoavian orders are well-defined, others are not [Cracraft et al., 2004]. For example, the diving birds, wading birds and marine birds are placed by most morphologists into about 8–9 different orders, including Charadriiformes (e.g. gulls, plovers), Ciconiformes (e.g. storks), Gaviiformes (loons), Gruidae (e.g. cranes), Pelecaniformes (e.g. pelicans, boobies, tropicbirds), Phoenicopteriformes (flamingos), Podicipediformes (e.g. grebes), Procellariiformes (albatrosses, shearwaters), and Sphenisciformes (penguins). Most of these orders are recognized today in avian field guides and other reference sources. However, DNA hybridization data [Sibley and Ahlquist, 1990] and DNA sequences [Hedges and Sibley, 1994; van Tuinen et al., 2001; Cracraft et al., 2004; Jarvis et al., 2014] have revealed relationships that place into question the recognition of those orders as classically defined. For example, the pelicans are most closely related to the shoebill stork and hammerkop and not to the boobies or tropicbird, and the grebes are most closely related to the flamingos, not to the loons [Hedges and Sibley, 1994; van Tuinen et al., 2001; Cracraft et al., 2004]. Those surprising findings have been obtained with multiple genes in different laboratories, and they have strong statistical support. Other controversial findings among the waterbirds, such as a clustering of storks with New World vultures [Sibley and Ahlquist, 1990], have not been corroborated with genomic data [Jarvis et al., 2014].

Besides the water birds, progress has been made with DNA sequence analyses in understanding relationships among other neoavian groups, including the Passeriformes [Johansson et al., 2001, 2002; Barker et al., 2002; Yuri and Mindell, 2002; Cracraft et al., 2004]. One particularly problematic bird has been the hoatzin (Opisthocomus hoazin) of South America, sometimes placed in its own order (Opisthocomiformes). It is the only bird that uses microbial foregut fermentation to convert cellulose into simple sugars, as in some mammals (e.g. ruminants). It feeds on young leaves and twigs of marsh plants and has a large muscular crop for fermentation. Most earlier phylogenetic analyses have placed the hoatzin in the Galloanserae, but DNA hybridization studies and DNA sequence studies have agreed that it is a neoavian species [Sibley and Ahlquist, 1990; Hedges et al., 1995; Hughes and Baker, 1999; Sorenson et al., 2003]. However, its particular relationship within Neoaves has been hard to resolve, even in the recent complete genome analysis [Jarvis et al., 2014].

Similarly, some other nodes in the phylogeny of neoavian bird orders have been difficult to resolve. However, large numbers of nuclear genes and complete genomes [Hackett et al., 2008; Jarvis et al., 2014] have succeeded in resolving most nodes. The results show an unexpected clade of birds (Columbea), comprising flamingos, grebes, and pigeon relatives, as the closest relative of all other neoavian orders, and – among the remaining orders (Passerea) – 2 large clades comprising birds of similar ecological zone: a land clade and a water clade [Jarvis et al., 2014]. Other orders of Passerea do not fit neatly into the land and water clades, and more work will be needed to better resolve the branching pattern of the orders and understand the important traits involved in their evolutionary history.

Molecular Clocks and Biogeography

A literal reading of the fossil record of birds is that the modern orders evolved and radiated in the early Cenozoic, after the Cretaceous-Tertiary extinctions [Feduccia, 1995, 2003]. Even the problematic Cretaceous fossils of Neornithine birds noted above are mostly from the final stage of the Cretaceous (Maastrichtian; 71–65 Mya). However, molecular clocks have instead indicated that the orders of modern birds branched more deeply in the Cretaceous [Hedges et al., 1996; Cooper and Penny, 1997; Kumar and Hedges, 1998; Haddrath and Baker, 2001; van Tuinen and Hedges, 2001; Paton et al., 2002]. If these molecular clocks are correct, why is there such a large gap in the avian fossil record? Local clock methods (e.g. Bayes-
ian) have been used in addition to global clocks, and some studies [Hedges et al., 1996; Kumar and Hedges, 1998; van Tuinen and Hedges, 2001] have used only rate-constant genes calibrated from outside of the avian fossil record to avoid potential calibration biases. Despite the use of these different methods, and carefully controlling for rate, the resulting time estimates have been fairly consistent in supporting relatively deep divergences among modern birds (fig. 7).

The divergence of the chicken (Galliformes) and duck (Anseriformes) has proven to be an important anchor point for avian molecular clocks. Anseriformes is about the only order of modern birds that can be confidently placed in the Cretaceous, with the earliest fossils from the Campanian (84–71 Mya) [Chiappe and Dyke, 2002]. Calibrating outside of birds and using 12 rate-constant nuclear genes, this divergence was estimated to be 90 ± 7 Mya [van Tuinen and Hedges, 2001]. A similar date of 85 ± 17 Mya was obtained with different data (mitochondrial DNA), calibrations (internal avian calibrations), and clock methods (‘rate smoothing’ local clock) [Had-drath and Baker, 2001]. The advantage of establishing an anchor point within birds is to provide a robust calibration for obtaining other time estimates among birds [van Tuinen and Hedges, 2004]. Using that chicken-duck calibration point, several data sets were analyzed (mitochon-drial DNA), calibrations (internal avian calibrations), and clock methods (‘rate smoothing’ local clock) [Hade-rath and Baker, 2001]. The advantage of establishing an anchor point within birds is to provide a robust calibration for obtaining other time estimates among birds [van Tuinen and Hedges, 2004]. Using that chicken-duck calibration point, several data sets were analyzed (mitochondrial DNA, DNA-DNA hybridization, and transferrin immunological data), resulting in the following mean divergence times for major nodes: 119 ± 5 Mya (paleognath-neognath), 104 ± 3 (Galloanserae-Neoaves), and 89 ± 9.6 Mya (base of neovarian radiation). Similar dates for the earliest nodes were obtained in a more recent re-view [van Tuinen, 2009]. Dates among paleognaths ranged from 50 (kiwi-emu) to 83 Mya (tinamous-ratites), between orders from 76 to 80 Mya, and within orders from 39 (chicken-quail) to 80 Mya (kingfisher-hornbill). However, sequence data were limited and most neoavian orders were not represented. Recent reviews of molecular time estimates found Cretaceous dates for all nodes among ratites and tinamous [Baker and Pereira, 2009] and most nodes among Galloanserae, with a mean split time of 107 Mya between ducks and fowl [Pereira and Baker, 2009].

In another study [Paton et al., 2002], divergence time estimates were obtained using mitochondrial sequence data, local clock methods, and calibrations of 85 Mya (chicken-duck) and 35 Mya (emu-cassowary). The resulting times were similar to an earlier study in finding deep splits (Cretaceous) among neoavian orders and a deep divergence between paleognaths and neognaths (123 Mya; 156–108 Mya). However, their dates for divergences among paleognaths were slightly older: 105 Mya (tinamous vs. ratites), 89 Mya (rhea vs. ostrich), and 81 Mya (kiwi vs. emu). More recently, Pacheco et al. [2011] constructed a timetree of neoavian orders using mitochondrial genomes and found that most divergences occurred in the Cretaceous, corroborating earlier work. As more sequences are obtained, especially from nuclear genes, time estimates from molecular clocks should become more stable until the point where most or all error is from the fossil record and calibration points [Hedges and Kumar, 2003, 2004].

A timetree was presented with the recent multi-gene analysis of bird evolution [Jarvis et al., 2014]. It showed a split between Palaeognathae and Neognathae of about 100 Mya, the split of Neoaves and Galloanserae of 88 Mya, and the split of most neoavian orders about 65 Mya. These times are considerably (16%) younger than found in earlier studies, reviewed above [e.g. van Tuinen, 2009]. It should not be concluded that the more recent study [Jarvis et al., 2014], based on more data (complete genomes), is correct and the earlier studies are incorrect. The difference, instead, is likely from an error (calibration bias) in the methodology used in the recent genome study [Jarvis et al., 2014]. Those authors used a suite of fossil minimum calibrations and one maximum calibration. That maximum calibration, the split of Palaeognathae and Neognathae, was set to 99.6 Mya for an arbitrary rea-son: it is the boundary between the Early and Late Creta-ceous, with the authors qualifying that it ‘far exceeds the age of palaeontological evidence for the existence of Neornithes.’ However, because the resulting analysis returned a date for that maximum calibration node equal to the calibration for that node, it suggests that the true time is earlier and that the resulting times throughout the tree are likely underestimates. Otherwise, it would be a remark-able coincidence that the maximum calibration used, thought to be much older than the true date for the node, would exactly match the resulting time. This data set should be reanalyzed using a better calibration methodology.

With this current timescale and phylogeny of avian evolution (fig. 7), and recognizing its limitations, it is possible to draw some general inferences concerning the niche and morphology (habitus) of birds during their early history. For example, the heavy-bodied habitus and primarily ground-dwelling, non-marine habits of the Galloanserae and Palaeognathae, and their dinosaurian ancestors, suggest that these features were common among the stem neornithine birds of the Cretaceous [van
Tuinen et al., 2000]. Also, considering that shallow, marine environments are favorable for fossilization, this (the predominance of non-marine species) may in part explain the sparseness of the early fossil record of modern birds. Nonetheless, fossils indicate that Cretaceous birds also occupied other niches (e.g. perching, diving, etc.).

The early (Cretaceous) branching of ordinal lineages of modern birds found with molecular clocks ‘is compatible with the extensive speciation that occurred within orders following the sudden availability of niches in the early Tertiary period’ [Hedges et al., 1996]. Those niches were vacated by dinosaurs, pterosaurs, and other vertebrate groups that became extinct following the Cretaceous-Tertiary asteroid impact. The near simultaneous appearance of so many major lineages of neoavian birds, adapted to different environments, in the early Tertiary supports the niche-filling hypothesis.

One explanation for the early branching of ordinal and superordinal lineages of birds is that they reflect the splitting of landmasses in the Cretaceous, the Continental Breakup hypothesis [Hedges et al., 1996]. The same hypothesis was proposed for the orders of placental mammals [Hedges et al., 1996], based on molecular clock evidence, and subsequent phylogenetic analyses have provided support [Springer et al., 1997; Stanhope et al., 1998]. In birds, it has proven more difficult to associate particular orders with geographic regions (e.g. continents) because of their wider distributions and ease of dispersal. As noted above, ratites have been singled out as having an evolutionary history associated with continental breakup [Cracraft, 1973, 2001; van Tuinen et al., 1998], but biogeographic details will remain sketchy until their relationships are better resolved. A good case has been made that modern (neornithine) birds in general are Gondwana in origin, based on a review of fossils, phylogeny, and molecular divergence times [Cracraft, 2001]. However, the extent that plate tectonics and the superordinal lineages of birds is that they reflect the splitting of landmasses in the Cretaceous, the Continental Breakup hypothesis [Hedges et al., 1996]. The same hypothesis was proposed for the orders of placental mammals [Hedges et al., 1996], based on molecular clock evidence, and subsequent phylogenetic analyses have provided support [Springer et al., 1997; Stanhope et al., 1998]. In birds, it has proven more difficult to associate particular orders with geographic regions (e.g. continents) because of their wider distributions and ease of dispersal. As noted above, ratites have been singled out as having an evolutionary history associated with continental breakup [Cracraft, 1973, 2001; van Tuinen et al., 1998], but biogeographic details will remain sketchy until their relationships are better resolved. A good case has been made that modern (neornithine) birds in general are Gondwana in origin, based on a review of fossils, phylogeny, and molecular divergence times [Cracraft, 2001]. However, the extent that plate tectonics and the superordinal lineages of birds have had on avian evolution remains unanswered.

Completion of the chicken genome, and recently other bird genomes, will be a major benefit for evolutionary studies. For example, orthology determination, which is necessary for assembling phylogenetic data sets of nuclear genes, will be greatly facilitated by having avian genomes for comparison. Also, genomic comparisons will now be able to include many representative birds and therefore increase precision in analyses and questions being addressed, providing a better evolutionary context across amniote vertebrates [Hedges, 2012]. Finally, detailed studies within birds, involving molecular clocks and phylogenies, will have the benefit of avian genomes for primer design, clock calibration, and better genetic comparisons in general.

I thank Jennifer Hines for drawings of animals.

An Update on Chicken Sex Determination and Gonadal Sex Differentiation
(Prepared by C.A. Smith)

Since the publication of the Second Report on Chicken Genes and Chromosomes 2005 [Schmid et al., 2005], there have been some significant advances in our understanding of avian sex chromosomes and sex determination [Smith CA et al., 2000; Zhao et al., 2010; Chue and Smith, 2011; Ayers et al., 2013c]. Sex in chicken and other birds is governed by the inheritance of sex chromosomes, with male being ZZ and female ZW. It is still unclear whether sex is determined by the presence/absence of the W sex chromosome or whether it depends upon Z dosage or the Z to autosome ratio (2 Z’s in male; 1 Z in female). This question could be solved once and for all by determining the sexual phenotype of birds with sex chromosome aneuploidy (e.g. 2A:Z0 or ZZW), but such birds have not been definitively identified, and indeed these genotypes could be embryo lethal in avians due to potential gene dosage effects [Graves, 2003]. Although some recent studies on sex determination look beyond the gonads (discussed below), most researchers in the field still equate ‘sex determination’ with embryonic gonadal sex differentiation. While sex is set at fertilization with the differential inheritance of the sex chromosomes, one or more genes on these chromosomes direct differentiation of the gonads into testes or (unilateral) ovary during embryonic life. Such genes – or different sex-linked genes – may also direct sexual differentiation outside the gonads in a cell-autonomous fashion (discussed below). At present, most evidence favours Z dosage underlying chicken sexual differentiation, at least in the gonads [Chue and Smith, 2011; Ayers et al., 2013c].

Figure 8 shows the genes currently known or postulated to play a role in gonadal sex differentiation in the chicken embryo. The gonads are of mesodermin origin, forming on the surface of the mesonephric kidneys. Sexual differentiation begins at the morphological level on embryonic day 6.0–6.5 [Hamburger and Hamilton (HH) stage 29–30], when seminiferous cords of the inner medulla condense in males (ZZ), and the outer epithelial layer thickens in females (ZW) [Hamburger and Hamilton,
1992; Guioli et al., 2014]. The gonadal asymmetry characteristic of female gonads (only the left becomes a functional ovary) has been shown to be driven by the asymmetric expression of the PITX2 gene and, downstream, differential activity of retinoic acid signaling and cell cycle regulators [Guioli and Lovell-Badge, 2007; Ishimaru et al., 2008; Rodríguez-León et al., 2008]. There are some interesting differences between gonadal development in mouse and chicken. The key preSertoli cell lineage, for example, derives from the outer epithelial layer of the gonad in the mouse embryo, but this layer appears to generate interstitial cells in the chicken, indicating the preSertoli cell population likely comes from the mesonephric kidneys (hence a mesenchymal origin) [Sekido and Lovell-Badge, 2007].

The best candidate master regulator of chicken gonadal sex differentiation is the Z-linked gene, DMRT1 (doublesex and mab-3 related transcription factor, #1) [Chue and Smith, 2011; Matson and Zarkower, 2012]. This gene encodes a zinc-finger-like transcription factor, and it has a conserved role in testis development after birth in mammals [Matson and Zarkower, 2012]. This role includes maintenance of Sertoli and germ cells [Minkina et al., 2014]. It the mouse, DMRT1 prevents the ovarian developmental pathway from being activated, even in adults [Matson et al., 2011]. In the chicken, DMRT1 is expressed specifically in the embryonic urogenital system, where it is more highly expressed in male compared to female gonads [Raymond et al., 1999; Smith et al., 1999; Shan et al., 2000]. Knockdown of DMRT1 expression using RNA interference results in feminization of the gonads, characterized by loss of seminiferous cord structure and male markers, and activation of the female marker, aromatase [Smith et al., 2009]. Most recently, using in ovo electroporation of DNA constructs into developing gonads, we over-expressed DMRT1 in the embryonic chicken urogenital system and found that it can induce male pathway genes (SOX9 and AMH) in genetically female (ZW) gonads.

Fig. 8. Gonadal sex differentiation in the chicken embryo. The undifferentiated gonad comprises a medulla (blue) and outer epithelial layer (pink). Germ cells are shown as black circles. The medulla proliferates in males during testis formation, while the epithelial layer proliferates in females to generate the cortex of the ovary. Male expressed genes are shown in blue, female expressed genes in pink. The size of the circles represents relative expression levels compared to the opposite sex. DMRT1 is expressed in both sexes, but more highly in males. GATA4 is also expressed in both sexes and may activate DMRT1. In males (ZZ), DMRT1 can activate HEMOGEN (HEMGN), AMH and SOX9 expression, either directly or indirectly. HEMOGEN can activate SOX9 and also DMRT1. AMH mRNA is expressed prior to SOX9, and may be activated by SF1 together with DMRT1. Prostaglandin D2 (PGD2) can stimulate SOX9. DMRT1 can repress FOXL2 and aromatase (CYP19A1). In females (ZW), MHM may contribute to lower DMRT1 expression. FOXL2 likely activates aromatase (CYP19A1). SF1 may also contribute to CYP19A1 expression. R-SPO1 and WNT4 stabilize β-catenin. Both aromatase (estrogen) and β-catenin regulate ovarian development.
nads [Lambeth et al., 2014]. This further strengthens the proposal that Z-linked DMRT1 initiates testis development in the chicken (fig. 8). Downstream of DMRT1, the conserved AMH and SOX9 genes are male-upregulated and likely play a role in testis development, as in mammals (fig. 8). In mammals, SOX9 activates AMH during testicular development, but in the chicken, AMH mRNA expression precedes that of SOX9, indicating that the AMH gene is activated by other factors in birds. However, at the protein level, we detect SOX9 before or at the same stage as AMH protein, suggesting delayed translation of AMH mRNA until the onset of SOX9 expression [Lambeth et al., submitted]. In mammals, prostaglandin D 2  is synthesized male-specifically, and it augments SOX9 expression in preSertoli cells, and this process has been shown to be conserved in the chicken embryo [Moniot et al., 2008].

Recently, another Z-linked gene, HEMOGEN, has been implicated in chicken testicular development [Nakata et al., 2013]. As for DMRT1, the gene encodes a transcription factor. It is involved in haematopoiesis in mammals, but is not expressed in the embryonic mouse gonad. In contrast, it is expressed in male but not female embryonic chicken gonads. Its overexpression induces both DMRT1 and SOX9 expression in female (ZW) embryos, and the gonads appear masculinized (sex cords, male-type arrangement of germ cells). Given its location on the Z sex chromosome, HEMOGEN could lie upstream or act codominantly with DMRT1. If it lies upstream, then a W-linked factor would be implicated in its suppression in females (directly or indirectly). However, there is still no evidence for a W-linked ovary determinant in chicken, despite our recent transcript profiling analysis of embryonic chicken gonads using RNA-seq [Ayers et al., 2013a]. More likely is the possibility that Z-linked DMRT1 activates the HEMOGEN gene during testis development. This is supported by the fact that DMRT1 is expressed before HEMOGEN in developing male (ZZ) gonads, and overexpression of DMRT1 in female gonads can induce HEMOGEN expression [Lambeth et al., 2014]. As shown in figure 8, DMRT1 may activate HEMOGEN, which in turn activates SOX9 during testicular development. Knockdown of HEMOGEN expression would clarify this point, but this experiment has not yet been reported.

Unilateral ovary development occurs in female chicken embryos (ZW). It has been known for over 30 years that estrogen synthesis is essential for ovarian development in birds [Smith et al., 1997; Vaillant et al., 2001; Lambeth et al., 2013]. The terminal enzyme responsible for estrogen synthesis, aromatase (encoded by the CYP19A1 gene), is expressed female-specifically from the earliest stage of ovarian differentiation (HH29), while inhibition of aromatase causes female-to-male sex reversal (fig. 8). The regulation of gonadal CYP19A1 is therefore a pivotal event in chicken gonadal sex differentiation. The conserved FOXL2 transcription factor colocalizes with aromatase, and evidence from chicken and other vertebrates indicates that it activates aromatase during ovarian differentiation [Govoroun et al., 2004; Pannetier et al., 2006; Wang DS et al., 2007; Chue and Smith, 2011]. Other genes may also participate in the regulation of CYP19A1 in chicken, such as SF1, GATA4 and DAX1 (fig. 8). Recently, epigenetic modifications of key gonad genes such as CYP19A1 have come under scrutiny, in chicken and in other species. Chicken CYP19A1 shows sexually dimorphic methylation in a region of its gonadal promoter (45% methylated in ZZ males vs. 20% in ZW females), a pattern that is partially reversed when gonads are feminized with exogenous estrogens [Ellis et al., 2012]. This suggests that part of the mechanism of chicken gonadal development may involve differential gene methylation. A similar idea has been put forward for reptiles with temperature-dependent sex determination, in which temperature influences gene methylation status [Matsumoto et al., 2013]. The estrogen synthesized by aromatase enzyme must act locally in the embryonic gonads, activating estrogen receptor α, leading to transactivation of target genes. Those targets in embryonic chicken gonads have not been defined, despite the key role for estradiol.

In the past 5 years, several lines of evidence have shown that the cell adhesion and signaling factor, β-catenin, is required for ovarian development in mammals [Maatouk et al., 2008; Nicol and Yao, 2014]. In the developing mouse ovary, but not the testis, the signaling molecules R-spondin1 and Wnt4 stabilize β-catenin, leading to its nuclear translocation and activation of target genes. All 3 molecules are required for proper ovarian differentiation in mammals [reviewed in Chassot et al., 2008, 2014]. This pathway appears to be conserved in the chicken embryo, as R-spondin1, Wnt4 and β-catenin are all enriched in the cortical region of the female gonad (fig. 8) [Smith et al., 2008; Ayers et al., 2013b]. In mammals, loss of both FOXL2 and components of the R-SPO1/WNT4/β-catenin pathway lead to potent female-to-male sex reversal [Ot-tolenghi et al., 2007; Auguste et al., 2011]. Knockdown and overexpression of these genes in chicken will confirm a role in gonadal sex differentiation. How the R-SPO1/WNT4/β-catenin and FOXL2/CYP19A1 pathways intersect to control ovarian differentiation is currently not known. Indeed, in the chicken embryo, there are current-
ly few proven functional links between the known or putative gonadal sex differentiation genes.

Teranishi et al. [2001] described a reiterated sequence on the chicken Z sex chromosome that they called MHM (male hypermethylated). As the name indicates, this sequence is hypermethylated and transcriptionally silent in ZZ male cells (in the gonad and throughout the body). It is hypomethylated in female cells (ZW), where it encodes a long noncoding RNA that coats the Z close to the site of its transcription. This is reminiscent of the mammalian Xist RNA, which coats one X chromosome as part of the X-inactivation phenomenon (dosage compensation). MHM may be involved in localized dosage compensation in chicken (although widespread dosage compensation appears to be absent, discussed below). In chicken, this would involve upregulation of neighbouring genes in the single Z of females, as opposed to repression by Xist on the mammalian X. However, MHM RNA coats the female chicken Z sex chromosome very close to the DMRT1 locus, and it has been suggested that it may contribute to the lower level of DMRT1 expression in females as part of the avian sex (gonadal) determination mechanism (fig. 8) [Teranishi et al., 2001]. Indeed, more recently, we found that overexpression of MHM in male chicken embryos could disturb embryonic development, including reducing the expression of gonadal DMRT1 [Roeszler et al., 2012]. The potential role of MHM in localized dosage compensation and perhaps gonadal sex differentiation requires further study. The fact that the sequence is expressed from the single Z of females and in ZZW triploid birds, but not in males nor in ZZZ triploid males, points to the existence of a W-linked factor in triggering its demethylation and expression. No such factor has been identified on the W chromosome as yet [Ayers et al., 2013a].

One of the major issues that has arisen in the field is the definition of sex determination and the relative contribution of the sex chromosomes and sex steroid hormones [Arnold et al., 2012]. The traditional view of sex determination in birds and other vertebrates is that the sex chromosome complement controls embryonic differentiation of the gonads into ovaries or testes, and these organs then secrete hormones that feminize or masculinize the rest of the body (estrogen and progesterone in females, and testosterone and anti-Müllerian hormone in males). However, a growing body of evidence challenges this view, supporting the idea that the sex chromosomes can have direct, cell-autonomous effects upon sexual phenotype. This has been described in mice, in which the sex chromosome constitution, independent of gonadal sex, can cause sex differences in metabolism. It has been argued that the ‘gonad-centric’ view of sex determination is inappropriate and that multiple parallel pathways may regulate sexual development in different tissues, integrating to yield the overall sex of an individual [Arnold, 2012]. Such pathways might be controlled by a sex chromosome that is present in one sex and not the other. An example is the Y chromosome of therian mammals, expressing the Sry gene that induces testis formation during embryonic life. Other genetic differences between the sexes can come from other sources, such as X-linked genes that escape X-inactivation in mammals, or epigenetic differences such as genomic imprinting. Sex determination may therefore occur in different ways in different tissues, with gonadal sex differentiation being one aspect – one of the most visible aspects – of sexual differentiation [Arnold et al., 2013].

This idea is especially relevant to avian sex, because there is no chromosome-wide dosage compensation mechanism, such that males (ZZ) have, on average, higher levels of Z-linked gene expression compared to females [Ellegren et al., 2007; Itoh et al., 2007; Melamed et al., 2009]. However, this varies on a gene-by-gene basis, so that there is variable dosage compensation or no compensation of Z-linked genes that differs between tissues and life cycle stage [Mank and Ellegren, 2009; McQueen and Clinton, 2009]. The chicken Z sex chromosome is enriched for genes associated with sex and reproduction [Storchová and Divina, 2006], while genes associated with testis expression are overrepresented in the chicken Z chromosome [Ellegren, 2011]. These observations could explain avian sex determination being dependent upon sex-biased expression of different Z-linked genes in different tissues, consistent with the modern view of sex determination outlined above and expounded by Arnold et al. [2012]. Recent findings support the idea that sex determination in the chicken may be at least partly cell autonomous, involving direct effects of the sex chromosomes in different tissues, and supporting the idea of multiple parallel pathways underlying sexual development. Zhao et al. [2010] described 3 gynandromorphic chickens showing striking bilateral asymmetry, with male features on one side of the body (large wattle, spur and large muscle mass) and female features on the other (small spur and wattle, less breast muscle mass) (fig. 9). The ‘male side’ contained primarily ZZ cells, while the female side contained 40–60% ZW cells. The gonads of these birds reflected the relative contributions of ZZ and ZW cells, so that gonads with predominantly ZZ cells were testis-like, while those with predominantly ZW cells were ovarian-like [Zhao et al., 2010].
These observations suggest direct cell-autonomous effects of the sex chromosomes within the various tissues. The authors also generated chimeric embryos by explanting GFP-labeled presumptive gonadal tissue into opposite sex recipients and found that donor cells retained their sexual identity (expressing male or female marker genes) and were not incorporated into the gonadal tissues of the hosts. This again supports the notion that avian somatic cells have an inherent cell-autonomous cell identity (labeled CASI) that is largely independent of hormonal signaling. The factor/s driving CASI are not known, but it cannot involve DMRT1 in all tissues, as the expression of this gene is limited to the urogenital system. It could be driven by one (or more) of the hundreds of Z-linked genes in chicken that are not subject to dosage compensation and therefore show sexually dimorphic expression. Alternatively, a W-linked gene/s could be involved. Early chicken embryos show sexually dimorphic gene expression (from both sex chromosomes and autosomes) well prior to gonadal sex differentiation [Lee et al., 2009; Zhao et al., 2010; Ayers et al., 2013a]. This supports CASI, and the idea that sexual differentiation pathways are engaged in avian embryos very early in development. As mentioned above, there is still no clear evidence that the W sex chromosome encodes a ‘female’ or ovary determinant, although genes associated with female fertility are reported to be enriched on the chicken W chromosome [Moghadam et al., 2012].

It seems clear that hormones can override CASI, because exogenous estrogens or estrogen antagonists can quite extensively sex reverse avian embryos [reviewed in Vaillant et al., 2001]. This suggests a potential interaction between cell-autonomous factors and hormonal signaling. In the case of the gynandromorphic birds, it is possible that factors responding to sex hormones are sex-linked, hence accounting for the different development of sex characters on the left and right sides of the body, which had mainly ZZ or ZW genotypes (fig. 9). Androgen and estrogen receptors are not sex-linked in chicken, although other factors influencing the response to hormonal signaling might be located on the Z and/or W sex chromosomes.

In summary, the past 10 years have seen several advances in our understanding of sex determination and gonadal sex differentiation in the chicken embryo. Although the basic mechanism (Z dosage or dominant W) remains elusive, an increasing body of evidence points to the importance of the DMRT1 gene for testicular differentiation. Some new genes have emerged such as HEMOGEN, and some interactions in the pathway have been developed (DMRT1 inhibition of CYP19A1, for example). Despite several efforts in several labs, no clear W-linked female/ovary determinant has been identified, beyond those previously found to be expressed but for which a function has not been proven (HINTW, FET1, FAF) [Ayers et al., 2013a, b]. We carried out a thorough RNA-seq study of very early chicken blastoderms and day-4.5 (HH25) gonads, but did not detect a clear candidate ovary determinant [Ayers et al., 2013a]. Most of the 40 or so W-linked genes have Z homologues, to which they have very high sequence homology. The so-called dominant W hypothesis therefore remains open-ended. The discovery that birds lack a chromosome-wide system of dosage compensation together with the discovery of gynandromorphic chickens points to the possibility that different sex-linked genes, in different combinations, may underlie sexual differentiation in different tissues during development. The lack of dosage compensation provides a good basis for CASI, if one or a combination of Z-linked factors tips sexual development towards the male versus female pathways.

Fig. 9. Gynandromorphic chicken, showing male features on the left (large spur, wattle and breast muscle) and female features on the right (smaller spur, smaller wattle and smaller breast muscle). Reproduced with permission from Zhao et al. [2010].
Avian Epigenetics
(Prepared by H. Zhou)

The draft reference sequence of the chicken (Gallus gallus) genome was completed in 2004 [Wallis et al., 2004]. Spatial, temporal and state-specific gene expression differences are a major contributor to phenotypic variation in organisms [Pastinen, 2010]; however, a decade later, like in humans and mice, a mechanistic understanding of gene expression regulation (when, where, and how much) in chickens remains to be elucidated. While genetic variations across individuals in a population may partially contribute to the gene expression differences among them, a major question remains: how can an almost identical copy of the genome within an individual be able to have cell-specific expression profiles in more than 200 different cell types in the chicken?

It became clear that the epigenome plays a significant role in directing the unique gene expression programming in each cell type at different development stages. The epigenome, a second dimension to the genome, provides pivotal insights to gene regulation besides the genomic sequence [Rivera and Ren, 2013]. The epigenome refers to sequence-independent processes that can modulate the functional output of the genome [Sarda and Hannenhalli, 2014]. It consists of chemical changes to DNA or covalent modifications of histone proteins, chromatin accessibility, noncoding RNA localization, and higher-order chromatin architecture (nucleosome positioning and occupancy, 3D chromatin structure) [Bernstein et al., 2007; Berger et al., 2009; Bonasio et al., 2010].

Fueled by the rapid development of next-generation sequencing (NGS) and the significant drop of cost rate in the past decade (even faster than Moore’s Law used in computer technologies) [Koboldt et al., 2013], the rate of growth in datasets and publications of epigenomics have drastically increased, especially in humans, mouse and model organisms. Consequently, a few international consortia in the field of epigenomics have been established including the Encyclopedia of DNA Elements (ENCODer project Consortium, 2012), the Roadmap Epigenomics Project (http://www.epigenomewebbrowser.org/) [Bernstein et al., 2010], the International Human Epigenome Consortium (IHEC; http://ihec-epigenomes.org), and modENCODE [modENCODE Consortium et al., 2010]. Although progress in the avian epigenome is very limited, recent international efforts on the functional annotation of animal genomes (FAANG) including chickens have been initiated [The FAANG Consortium, 2015]. It is expected that epigenome data generated in the next few years will greatly contribute to our understanding of the genome-phenome relationship and eventually enhance genome information, enabling approaches to improve the efficiency, sustainability and biosecurity of poultry production systems.

In this review, recent advancements of technologies, advantages and disadvantages of each technology and progress in different areas of avian epigenomics as well as challenges and future directions in this field will be discussed.

DNA Methylation

One of the well-studied DNA modifications is methylation at its 5th carbon (5mC), which plays an important role in imprinting, retrotransposon silencing and X chromosome inactivation in mammals [Bird, 2002]. There are 3 main techniques in measuring DNA methylation: (1) methyl-sensitive restriction enzyme digestion (MRE-seq), (2) affinity enrichment of methylated DNA fragments (MeDIP-seq with an antibody and MBD-seq with a methyl-binding domain), and (3) bisulfite sequencing (MethylC-seq) and reduced representation bisulfite sequencing (RRBS) with enzyme digestion [Bock, 2012]. Both MRE-seq and MeDIP-seq and MBD-seq have a limited resolution and relatively low cost compared to MethylC-seq. The resolution of MRE-seq is dependent on the frequency of enzyme cut sites and can be improved with multiple enzymes, while the resolution of MeDIP-seq is highly related to the DNA fragment size, CpG density and quality of antibody. MethylC-seq with single-base resolution is generally accepted as the gold standard for DNA methylome profiling, although potential PCR biases may skew 5mC quantitation [Laird, 2010].

Chicken methylomes at different tissues have been characterized using MeDIP- and MRE-seq [Li Q et al., 2011; Hu et al., 2013; Tian F et al., 2013]. Liver and muscle methylomes in both red junglefowl and broiler were generated using MeDIP. The hypermethylation were observed in gene body regions and repetitive sequences, while promoter regions and 5′ end regions were hypomethylated [Li Q et al., 2011]. Whole-genome methylation profiling between non-infected and Marek’s disease virus (MDV)-infected thymus in 2 genetically distinct genetic lines that confer different resistances to MDV infection were studied using MRE-seq [Tian F et al., 2013]. A higher DNA methylation level was observed in the MD-susceptible line, while a lower DNA methylation level was found in the MD-resistant line with MDV infection.
infection [Tian F et al., 2013]. Recently, MethylC-seq has been used to generate single-base resolution methylome in the lungs of 2 distinct genetic lines (Fayoumi and Leghorn). Like in other species, DNA methylation was found throughout the chicken genome but rarely occurred in CpG island promoters [H. Zhou, pers. commun.]. More than 90% CpG sites were detected using MethylC-seq compared to MRE-seq (32%) [Tian F et al., 2013].

In addition, the role of DNA methylation on dosage compensation has been investigated. Genome-wide methylation and gene expression analysis revealed few significant differences in the promoter DNA methylation, although hundreds of genes were differentially expressed between male and female brains. This result suggests that, unlike mammals, DNA methylation was not associated with dosage compensation in chickens [Natt et al., 2014].

DNA methylation is also crucial in primordial germ cell (PGC) reprogramming. Jang et al. [2013] revealed more than 200 differentially methylated regions within mammalian homologues of imprinting and X-linked genes between male and female PGCs. By comparing PGCs with chicken embryonic fibroblasts (CEFs), the promoter regions of a dozen of differentially expressed genes between PGCs and CEFs were either hyper- or hypomethylated, indicating the potential role of methylation in lineage differentiation [Jang et al., 2013].

DNA methyltransferase genes (DNMTs) play important roles in the epigenetic regulation of DNA methylation at CpG sites [Okano et al., 1999]. The methylation levels of 3 chicken DNMTs (DNMT1, DNMT3a and DNMT3b) and their genetic variations across different genetic lines, generations, tissues, and age were studied, and the results suggested that a CpG to TpG transition in DNMT3b may be associated with genetic susceptibility to MDV infection in chickens [Yu et al., 2008].

Recent studies have shown that there are several intermediates during demethylation of 5mC to cytosine [Tahiliani et al., 2009]. The 5mC can be oxidized to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) before being converted to cytosine by the ten-eleven translocation (TET) family of proteins (TET1, 2, and 3) [Pastor et al., 2013]. Subsequently, several sequencing methods have been developed to distinguish these intermediates. Oxidative bisulfite sequencing (oxBS-seq) and TET-assisted bisulfite sequencing (TAB-seq) were used to identify and quantify 5hmC [Booth et al., 2012; Yu et al., 2012]; fC-seq, fCAB-seq and 5fC-DP-seq for measuring 5fC; and 5caC-seq to detect 5caC [Raiber et al., 2012; Shen et al., 2012; Song et al., 2013]. It is expected that these new techniques will contribute new insights into the avian epigenome in coming years.

Chromatin Modification

Chromosomal DNA is wrapped around histone octamers comprised of H2A, H2B, H3, and H4 subunits. These histone tails at specific residues are subject to covalent modifications. More than 130 post-translational modifications and 700 distinct histone isoforms have been identified in humans [Tan et al., 2011; Tian et al., 2012]. There are a few well-studied modifications including methylation, acetylation, phosphorylation, and ubiquitination. These modifications can play both activation and repression roles in transcription regulation by controlling DNA accessibility or recruitment/elimination of other protein complexes [Kouzarides, 2007]. The Mouse ENCODE generated a map of nearly 300,000 murine cis-regulatory sequences from 19 tissues and cell types [Shen et al., 2012]. Analyses that have integrated all this information together indicate that histone H3 lysine 4 trimethylation (H3K4me3), H3 lysine 27 trimethylation (H3K27me3), H3 lysine 27 acetylation (H3K27ac), and H3 lysine 4 monomethylation (H3K4me1) are the most informative histone modifications [ENCODE Project Consortium, 2012]. A high abundance of H3K4me3 correlates with promoters of active genes and transcription starts [Shen et al., 2012], while H3K27me3 is a repressive mark, and increased levels of H3K27me3 are associated with promoters of inactive genes [ENCODE Project Consortium, 2012]. H3K27ac is a mark of active regulatory elements and may distinguish active enhancers and promoters from their inactive counterparts. H3K4me1 is a mark of regulatory elements associated with enhancers and other distal elements, but is also enriched downstream of transcription starts [ENCODE Project Consortium, 2012]. High levels of H3K27ac and H3K4me1 are associated with enhancer regions, and also correlate with areas of DNase I hypersensitive sites (DHSs) [Greer and Shi, 2012; Shen et al., 2012].

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has been employed to map genome-wide DNA-associated proteins with modified histones by specific antibodies. Only a couple of studies have been conducted in chickens. Tissue-specific genome-wide landscapes of H3K4me3 and H3K27me3 in chicken liver, spleen and breast muscle were generated, and the correlation between enrichment of these 2 marks and mRNA expression in 3 tissues was analyzed. The results
suggested a conserved relationship like in mammals: H3K4me3 positively correlated with gene expression, while H3K27me3 negatively correlated with gene expression [Saelao and Zhou, 2015]. Luo et al. [2012] provided H3K4me3 and H3K27me3 profilings in MD-resistant and -susceptible chicken lines. Line-specific changes of these 2 histone modification marks with MDV infection were identified, which support the role of these 2 marks in the MD resistance in chickens.

**Nucleosome Positioning and Occupancy**

Chromatin is tightly packaged into basic repeat units of nucleosomes (147 bp of DNA and linker DNA). Nucleosome positioning can prevent or display specific DNA sequences for the regulation of transcription such as transcription factor recruitment and progression of DNA sequences for the regulation of transcription such as transcription factor proteins to specific biochemically select activities among linearly distant elements can be accomplished through short- and long-range DNA interactions. Like histone modifications and chromatin accessibility, 3D conformation/looping can regulate gene expression [Rivera and Ren, 2013]. Several assays have been developed in mapping high-order chromatin architecture depending on the number of loci examined and selected. Dekker et al. [2002] first introduced the chromosome conformation capture (3C) assay, then the circular chromosome conformation capture (4C), the chromosome conformation capture carbon copy (5C), the tethered conformation capture (TCC), a derivative of 3C, called Hi-C, and a variation of Hi-C, namely chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [van Steensel and Dekker, 2010; de Wit and de Laat, 2012; Sajan and Hawkins, 2012]. 3C and 5C are based on priori loci of interest. 3C detects the interaction frequency of one locus with another (one vs. one), while 4C measures the genome-wide interaction frequency of a single locus (one vs. all) [Zhao et al., 2006]. 5C measures thousands of anchors and baits (many vs. many) [Dostie et al., 2006]. Hi-C measures genome-wide interaction frequency without loci-specific enrichment (all vs. all) [Lieberman-Aiden et al., 2009]. Hi-C is considered as one of the most powerful and unbiased assays in measuring 3D structure.

**Chromosome Accessibility**

Chromatin openness is a prerequisite for the binding of transcription factor proteins to specific biochemically active regulatory elements in the genome such as promoters, enhancers, silencers, and insulators. A few high-throughput assays have been developed to map open chromatin including DNase hypersensitivity sequencing (DNase-seq) [Boyle et al., 2008] or formaldehyde-assisted identification of regulatory elements sequencing (FAIRE-seq) [Giresi et al., 2007]. Both assays require 100,000–10,000,000 cells. DNase-seq is time-consuming, labor-intensive, and requires a complicated sample preparation process, and FAIRE-seq has the disadvantage of a low signal-to-noise ratio [Tsompana and Buck, 2014]. Recently, a high-quality DHS map was generated in the MSB1 line, an MDV-transformed CD4+ T-cell line in chickens [He et al., 2014]. There were fewer DHS enrichments in CpG island-abundant regions. Furthermore, DHSs were strongly correlated with highly expressed genes in MSB1 cells [He et al., 2014]. To circumvent the weaknesses of MNase-seq and DNase-seq, Buenrostro et al. [2013] recently developed a novel method, ATAC-seq, that has several advantages over MNase-seq and DNase-seq. ATAC-seq is primarily based on hyperactive Tn5 transposase in probing open chromatin. First, it only requires 500–50,000 unixed nuclei; furthermore, it has single-base resolution. Finally, it can map nucleosome positioning, chromatin accessibility, and TF footprints. The FAANG consortium is planning to use ATAC-seq as a standard for measuring chromatin accessibility, although the protocol in tissues using this assay requires more optimization.
Genomic Imprinting

As one of the most important germ-line-specific epigenetic modifications that usually results in imbalanced allelic expression between 2 parental alleles, genomic imprinting has been intensively studied in mammals. However, this phenomenon in chickens is still in debate. Several known imprinting genes such as insulin-like growth factor type 2 (IGF2) and IGF2 receptor (IGF2R) in eutherians and marsupials have been tested in chickens and revealed that both genes were biallelically expressed [Yokomine et al., 2001]. Subsequently, a 0.5-Mb ortholog of the mammalian imprinting region in the chicken was investigated and further showed that several imprinting regulatory elements in mammals such as H19 imprinting center and local regulatory elements were not identified in chickens, which suggests that imprinting may have evolved after the divergence of mammals and birds [Yokomine et al., 2005]. A recent study of transcripts of 20 embryos in 2 genetically distinct chicken populations by RNA-seq identified 79 potential imprinting SNPs. Unfortunately, none of them could be confirmed [Frésard et al., 2014].

Perspectives and Challenges

In summary, avian epigenomics is still in its infancy. With more advanced high-throughput technologies, robust assays developed from human, mouse and model organism ENCODE projects, and significant cost reductions in NGS, in addition to the launching of the international FAANG consortium and great interest in functionally annotating, the chicken genome will prompt the next wave of exponential growth of epigenome data in the chicken. However, significant challenges remain. Firstly, data analysis is still one of the most difficult hurdles that the community must face and work on. In order to decipher functional information underlying the genetic blueprint in the genome, new bioinformatics tools need to be developed to integrate a variety of data types such as ChIP-seq, RNA-seq, DNase-seq, MethylC-seq, Hi-C, etc. Computational algorithms and novel statistical models also need to be implemented to appropriately analyze these ‘omic’ data. More importantly, few scientists with the required expertise available have signified an urgent need to train next-generation computational biologists, especially in the agricultural field. Secondly, new techniques need to be invented to address some limitations of each of the assays discussed above. An assay can be developed to characterize more than one mark. For example, nucleosome occupancy and methylome sequencing (NOMe-seq) provides dual epigenomic information in both nucleosome position and DNA methylation abundance from a single molecule [Kelly et al., 2012]. Single-molecule and single-cell assay is another direction of technology development. A specific cell type has a distinct epigenomic state. In order to dissect cell-specific regulatory mechanisms, single-cell populations with limited quantity need to be purified from tissues usually consisting of heterogeneous cell populations. A couple of new techniques are already underway to tackle this issue. Readers can have more details in a recent excellent review by Hyun et al. [2015]: ChIP-nano (nano-scale ChIP, only requires less than 50,000 cells) [Adli and Bernstein, 2011], SCAN (single chromatin molecule analysis in nanochannels) [Murphy et al., 2013], SMRT (single-molecule real-time without amplification step) sequencing [Flusberg et al., 2010]. Thirdly, many studies have supported the histone code hypothesis ‘that multiple histone modifications, acting in a combinatorial or sequential fashion on one or multiple histone tails, specify unique downstream functions’ [Strahl and Allis, 2000]. Re-ChIP or SeqChIP (sequential ChIP) has been developed to map the genomic locations of combinations in chromatin features [Furlan-Magaril et al., 2009; de Medeiros, 2011]. Finally, great coordination within the international scientific community is crucial to maximize ‘outputs’ with minimum ‘inputs’ as each group has limited resources and funding support. Indeed, a couple of projects under the umbrella of the FAANG consortium are already underway to profile more epigenomes from different tissues at different development stages in different breeds (UC Davis group funded by the USDA NIFA, and FR-AgENCODE funded by the INRA). We expect hundreds of epigenome maps with features of key epigenetic marks from a variety of chicken tissues will be generated in the coming years, which will contribute vital insights into the second and third dimensions of the chicken genomic sequence, and fulfill our promise in fully understanding phenotypes and mechanisms of inheritance not yet revealed by primary genomic sequences.

Structural Variation and Copy Number Variation in Poultry

(Prepared by R.P.M.A. Crooijmans and M.A.M. Groenen)

To understand phenotypic variation in farm animals and in poultry in particular, it is essential to define all potential genomic variation within a genome. Most of the emphasis has been on SNP variation [Wong et al., 122

Cytogenet Genome Res 2015;145:78–179
DOI: 10.1159/000430927

Schmid et al.
especially in chicken, the first farm animal whose genome was sequenced [International Chicken Genome Sequencing Consortium, 2004]. This resulted in the development of medium-high density SNP chips such as the Illumina 60K beadchip [Groenen et al., 2011] and, more recently, the 600K Affymetrix Axiom Genome-Wide Chicken Genotyping Array [Kranis et al., 2013]. In addition to single base alterations, genomes contain numerous other variations collectively referred to as structural variants (SVs). SVs include small insertions and deletions (InDels) of a few base pairs in size, large InDels of thousands to sometimes hundreds of thousands base pairs in size, inversions and translocations. Rather arbitrary, copy number variation (CNV) has been defined as duplications and losses of DNA segments of 50 bp and longer. CNVs have been implicated to play a role in a wide variety of phenotypic traits in many species including chicken as discussed below.

### Detection of SVs

There are basically 3 different approaches to analyze CNVs, which are based on (1) medium-high density SNP arrays, (2) high-density arrays of genomic probes (aCGH) and (3) next-generation sequencing. For the first method, so far, only the Illumina chicken 60K SNP beadarray has been used to detect CNVs by identifying groups of adjacent SNPs by means of the software PENNCNV [Wang K et al., 2007]. The number of CNVs detected in these studies (table 14) varied from 137 to 271. The limited number of CNVs is mainly due to the low SNP density (60K) of this SNP array. A more direct way to analyze CNVs is by using specifically designed comparative hybridization arrays (aCGH), where the difference in hybridization signal between a test sample and a reference sample is scored. Two commercial aCGH platforms are available, the NimbleGen and Agilent platforms, respectively. The NimbleGen arrays available for chicken have a probe density of 180K or 385K, and the Agilent arrays for chicken have a probe density of 244K or 400K, respectively (table 14). The number of CNVs detected not only depends on the number of probes on the array but also on the number of samples and the diversity of populations under study. The number of detected CNVs with aCGH varied from 45 to 1,156 and is higher than that obtained with the 60K SNP chip analysis (table 14). The third method uses whole-genome re-sequencing data, and is a very powerful tool to detect CNVs. However, this method is very dependent on the quality of the reference genome and the available depth of the sequence data. The number of CNVs detected by the 2 studies that used this approach exceeds 8,000, a number that is mainly the result of an increase in the smaller size-range fraction of CNVs, due to the higher resolution of this method (table 14). One of the main advantages of the re-sequencing method is that all the different types of variation (both SNPs and SVs) can be identified simultaneously.

### Impact of SVs on Phenotype

In many species, the impact of SVs and CNVs on specific phenotypes is still poorly studied. Nevertheless, a number of studies have reported CNVs as the underlying variation for specific qualitative traits. In pigs, the...
color is related to a CNV in the KIT gene [Wiseman, 1986; Giuffra et al., 2002], and in chicken several CNVs have been identified that affect specific traits. The partial duplication of a 180-kb fragment containing the genes PRLR and SPEF2 on the Z chromosome is responsible for the late feathering phenotype [Elferink et al., 2008], while the Pea-comb phenotype is associated with a CNV in the SOX5 gene [Wright et al., 2009]. Besides these 2 examples, 3 more CNV-trait relations have been identified (table 15) and many more are expected to be identified in the future.

### SNPs and InDels – The Most Abundant Sources of Genetic Variations

(Prepared by A.A. Gheyas, C. Boschiero, and D.W. Burt)

Single nucleotide polymorphisms (SNPs) constitute the most abundant source of genetic variation in the vertebrate genome followed by short-length (≤100 nucleotides) insertions and deletions (InDels). These genetic variants are important drivers of genomic evolution. Since the publication of the Second Report on Chicken Genes and Chromosomes [Schmid et al., 2005], the number of SNPs and InDels reported for chicken has increased many folds. This rapid increase in the discovery rate of variants has been possible thanks to the development of a reference genome assembly for chicken and the advancements in high-throughput sequencing technologies with simultaneous development of bioinformatics approaches to mine large volumes of sequence data. Nevertheless, not only the discovery but also the characterization of these variants is essential to understand their physiological impacts, if any. In the following sections, we discuss the progress in the discovery and characterization of SNP and InDel variants and their current and future applications in research and poultry breeding.

#### Surge of Discovery of SNPs and InDel Variants from the Chicken Genome

During the write-up of this article, the dbSNP (build 140) reported about 11 million SNPs and InDels from the chicken genome. Major proportions of these variants were contributed by a few studies only. The first major genome-wide discovery of variants in chicken was reported by the International Chicken Polymorphism Map Consortium (ICPMC), and the seminal paper [Wong et al., 2004] was published in the same issue of the journal Nature that released the draft assembly of the chicken genome based on the sequencing of a red junglefowl [International Chicken Genome Sequencing Consortium, 2004]. The ICPMC reported a genetic variation map of 2.8 million SNPs and short InDels discovered by comparing the sequences of 3 domestic chicken breeds – a male broiler, a female layer and a female Chinese Silkie – with the reference genome from red junglefowl. To facilitate the application of these variants to avian genetics, genomics and evolutionary studies, Beijing Genomics Institute created a Chicken Variation Database (ChickVD) to display these in relation to various genomic contexts such as genes, cDNAs, chicken orthologs of human disease genes, quantitative trait loci (QTLs), and raw sequence traces [Wang et al., 2005].

While the ICPMC study performed the sequencing with shotgun method with coverage of only 0.25×, later studies reporting the discovery of a large catalogue of variants applied next-generation sequencing (NGS) approaches to re-sequence many more individuals with much higher depth of coverage. One major study reported over 7 million SNPs and 1,300 InDels discovered by re-sequencing individuals from 8 different breeds – 2 commercial broiler lines, 1 high-growth line, 1 low-growth line, 1 line of Rhode Island Red, 1 obese strain, and 2 White Leghorn lines of which one was used for commercial purpose [Rubin et al., 2010]. To allow sequencing of multiple individuals from each of these lines and thus to obtain the frequency spectrum of the variants, the study adopted a

### Table 15. Structural variation affecting traits in chicken

<table>
<thead>
<tr>
<th>Trait</th>
<th>Gene</th>
<th>CNV, kb</th>
<th>Kind</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Late feathering</td>
<td>PRLR and SPEF2</td>
<td>176</td>
<td>duplication and rearrangements</td>
<td>Elferink et al., 2008</td>
</tr>
<tr>
<td>Dermal hyperpigmentation</td>
<td>EDN3</td>
<td>129 and 172</td>
<td>duplication and rearrangements</td>
<td>Dorshorst et al., 2011</td>
</tr>
<tr>
<td>Pea-comb</td>
<td>SOX5</td>
<td>3.2</td>
<td>partial duplication</td>
<td>Wright et al., 2009</td>
</tr>
<tr>
<td>Dark brown plumage color</td>
<td>SOX10</td>
<td>8.3</td>
<td>deletion</td>
<td>Gunnarsson et al., 2011</td>
</tr>
<tr>
<td>Rose-comb</td>
<td>MNR2</td>
<td>7.4</td>
<td>inversion</td>
<td>Imsland et al., 2012</td>
</tr>
</tbody>
</table>
pool-seq approach where DNAs from several individuals (n = 8–11) within lines were pooled prior to sequencing. The sequencing was performed with coverage of 44.5×. The variants detected in this study were used to investigate signatures of domestication selection in chickens.

In order to develop a moderate-density genotyping array (60K), a USDA-funded project created reduced representation libraries from 4 distinct chicken populations – 2 broiler, 1 white-egg layer (WEL) and 1 brown-egg layer (BEL) – by pooling DNAs from 25 individuals within each line [Groenen et al., 2011]. Variant discovery was performed based on Illumina sequencing of these reduced representation libraries at an intended coverage of 25–40× and resulted in the detection of over 352,000 SNPs. Another recent study reported the discovery of about 7.6 million SNPs by whole-genome re-sequencing of only 2 domestic chickens (a Silkie and a Taiwanese native chicken) using Illumina sequencing technology with sequence coverage of 23–25× [Fan et al., 2013]. The identified variants were characterized in relation to their position and potential effects on genes and known QTLs and were used to identify regions of selective sweep.

So far, the largest re-sequencing and variant discovery project on chicken has been undertaken by the Roslin Institute of the University of Edinburgh that sequenced 243 chickens from 25 commercial and experimental lines from diverse breeds including 4 commercial broiler lines of composite breed origin; 6 WEL lines of White Leghorn origin; 6 BEL lines originating from several breeds such as White Plymouth Rock, Rhode Island Red and White Rock; 8 experimental inbred lines from the Institute of Animal Health (currently known as the Pirbright Institute); and 1 unselected experimental line of Brown Leghorn origin from the Roslin Institute [Kranis et al., 2013]. This study also applied the pool-seq approach to combine DNAs from several individuals (n = 10–15) within lines and performed the sequencing on Illumina platform with coverage of 8–17×. From the millions of SNPs discovered in this work, a high-density (600K) genotyping array was developed and made commercially available to facilitate genome-wide analyses in chicken. The project has also characterized over 15 million SNPs adopting a number of approaches with the goal to delineate potentially functional variants [Gheyas et al., 2015].

Although the advent of NGS approaches has accelerated the pace of variant discovery, the discovery of InDels is still lagging behind that of SNPs due to a number of intrinsic difficulties. Most of these difficulties are associated with mapping reads against the reference genome. For instance, the short reads covering InDel sites from NGS are often difficult to map to the correct genomic locations and require complex algorithms for gapped-alignment [Li et al., 2008; Neuman et al., 2013]. Moreover, since the mapping algorithms align each read independently, variant callers tend to consider InDels as stretches of SNPs unless a local realignment of reads around the suspected InDel sites is performed [Albers et al., 2011; GATK, https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_indels_IndelRealigner.php]. Furthermore, the majority of the InDels occur in short tandem repeats which are difficult to map [Messer and Arndt, 2007; Madsen et al., 2008]. Due to these issues, along with the relatively low frequency of InDels, a higher sequencing coverage is needed to detect these variants [Mullaney et al., 2010].

The current dbSNP release (build 140) reports only about 439,000 InDels of which ∼273,000 were generated from the study by the ICPMC [Wong et al., 2004]. Only recently, a study has reported about 1.3 million InDels by NGS analysis of single individuals from 12 diverse breeds including the red junglefowl, 7 Chinese indigenous breeds, and 4 commercial breeds (Cornish, Rhode Island Red, White Leghorn and White Plymouth Rock) [Yan et al., 2014]. In order to minimize the rate of false positives in InDel detection, this study retained only the consensus InDels called by 2 different variant callers, viz. SAMtools and GATK. Another recent study [Boschiero et al., in press] has characterized over 883,000 InDels based on the sequenced data from 16 commercial and experimental layer lines described in Kranis et al. [2013]. The same approach of taking consensus from 2 variant callers, SAMtools and Dindel, was adopted by this study as well. A large proportion of the
InDels (80%) detected in this study overlapped with the variants from Yan et al. [2014] serving as a validation for the majority of the InDels. Unlike the previous study, however, this work screened multiple individuals from each line by sequencing pooled DNA and is the first study to report the frequency spectrum of detected InDels. The study observed that with low sequencing coverage, NGS tends to detect mostly high-frequency variants. This result, along with a large proportion of common InDels across these 2 studies, indicates that most of the known InDels are high-frequency old variants, and it would take sequencing of many more individuals with higher coverage to detect most of the low-frequency variants. Analysis of the size distribution of the InDels suggests that most of the known InDels (89% insertion and 88% deletion) are small in size ($\leq$5 nucleotides; fig. 10).

**Genomic Distribution of SNPs and InDels**

The chicken genome consists of 38 pairs of autosomes and 2 sex chromosomes. The current build of the chicken reference genome (Gallus_gallus_4.0), however, represents only 29 autosomes, 2 linkage groups, the sex chromosomes and 14,093 unplaced scaffolds (http://www.ensembl.org/Gallus_gallus/Info/Annotation/assembly, accessed 10/11/2014). Since variant discovery in most cases was performed by aligning sequence reads against the reference genome, the detected variants mostly represent regions covered by the reference genome.

Combining the chicken variants from different sources – i.e. the ones already in the dbSNP and those yet to be released in public domain – we found about 17 million SNPs ($n = 16,890,433$) and 2.5 million ($n = 2,540,719$) InDels. These numbers are comparable to those reported for other species in the public domain such as human ($\sim$63 million), mouse ($\sim$71 million) and cattle ($\sim$74 million), taking into account the fact that the chicken genome is about 2.5–3 times smaller than the genomes of these species. Counting the number of variants in each 1-kb non-overlapping window over the genome reveals the average densities of variants to be $16.22 \pm 10.01$ SNPs per kb and $2.43 \pm 2.14$ InDels per kb. About 4% of the 1-kb windows or $\sim$51 Mb of the reference genome, however, did not have any variants at all.

Discovery of millions of variants has improved our understanding of the patterns of their genomic distribution across chromosomes. In chicken, the sizes of the autosomes are highly variable with 5 macro-, 5 intermediate and 28 microchromosomes [Burt, 2005]. The study by the ICPMC [Wong et al., 2004] suggested that SNP and InDel rates are independent of chromosome size, except for chromosome 16. However, when we recalculated the densities based on the known list of SNPs and InDels, a different picture emerged. This showed that the microchromosomes had the highest mean SNP density but the lowest InDel density among the 3 groups of chromosomes (fig. 11). The higher SNP density in microchromo-
somes is expected due to their higher recombination rates [Burt, 2005] as studies on other species revealed that the polymorphism rate was positively correlated with the recombination rate [Begun and Aquadro, 1992; Nachman, 2001]. The lower rate of InDels in microchromosomes, however, is contrary to this paradigm but has been observed by several studies on chicken. This has been attributed to the high gene content of these chromosomes [Brandström and Ellegren, 2007; Yan et al., 2014; Boschiero et al., in press]. Since presence of InDels in the coding regions can have a more detrimental effect than that of SNPs, purifying selection against InDels possibly counteracts the effects of high recombination rates in microchromosomes. Chromosome 16 is another exception to the `recombination-polymorphism rates` paradigm. This chromosome has one of the highest SNP densities (26 SNPs/kb) (fig. 11) and the lowest InDel density among all the chromosomes. The presence of highly variable major histocompatibility complex (MHC) genes is the probable cause of the high SNP rate [Wong et al., 2014]. On the other hand, the W chromosome is the sex-limiting chromosome in chicken, and its reduced genetic diversity is the result of selection and the complete lack of recombination outside the pseudoautosomal region [Berlin and Ellegren, 2004].

**Functional Characterization of SNPs and InDels**

The first tier of functional characterization of variants often entails investigating their location in relation to coding and noncoding functional elements in the genome. This, however, demands a comprehensive annotation of the reference genome for functional elements. Although the chicken genome is well-annotated for protein-coding genes, the annotation of functional noncoding elements (fNCEs) is still poor. Even though protein-coding genes cover ~45% of the chicken reference genome, the actual coding part constitutes only ~1.5%. For human, a massive research consortium (ENCODE) has recently predicted that >80% of the human genome is involved in some biochemical functions, and most of these elements belong to noncoding regions, thereby emphasizing their importance [ENCODE Project Consortium, 2012]. Until a similar database of fNCEs for chicken is developed, only a small proportion of variants can be annotated.

Most studies on chicken have, therefore, attempted to annotate variants in the functional context of protein-coding genes and predict the effects of only coding vari-

---

**Table 16. Summary of functional annotation of SNP and InDel variants from major studies**

<table>
<thead>
<tr>
<th></th>
<th>Wong et al., 2004</th>
<th>Fan et al., 2013</th>
<th>Gheyas et al., 2015</th>
<th>Yan et al., 2014</th>
<th>Boschiero et al., in press</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of total variants, in millions</td>
<td>2.8</td>
<td>7.6</td>
<td>15.3</td>
<td>1.3</td>
<td>0.9</td>
</tr>
<tr>
<td>Number of variants from individual lines, in millions</td>
<td>0.9–1.2</td>
<td>5.1–5.4</td>
<td>3.6–5</td>
<td>0.37–0.52</td>
<td>0.17–0.39</td>
</tr>
<tr>
<td>Proportion, %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>in genic regions</td>
<td>37</td>
<td>43</td>
<td>47</td>
<td>45</td>
<td>48</td>
</tr>
<tr>
<td>in coding regions</td>
<td>1.2</td>
<td>1.1–1.2</td>
<td>2.3</td>
<td>0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>of nonsynonymous and stop-gain/loss SNPs</td>
<td>0.4a</td>
<td>0.3–0.4a</td>
<td>1.2b</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>of frameshift and stop-gain/loss InDels</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>of non-frameshift InDels</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>of nonsynonymous SNPs and non-frameshift InDels</td>
<td>26c</td>
<td>11c</td>
<td>49c, d</td>
<td>29d</td>
<td></td>
</tr>
</tbody>
</table>

a The estimates are based on variants detected within single lines.

b The estimate is based on all the variants detected, many of which are private to single lines.

c The predictions were performed with SIFT package.

d The predictions were performed with PROVEAN package.
The study by ICPMC [Wong et al., 2014] showed that ~37% of the variants fell within genes – mostly protein coding (based on Ensembl database) – while only 1.2% fell within coding regions. Later studies, however, predicted a larger proportion (43–48%) of detected variants to be present within genomic regions, possibly reflecting the continual improvement of annotation of the chicken genome (table 16). Only about 1.1–2.3% of the SNPs and 0.12–0.13% of the InDels were reported to be within coding regions by different studies (table 16).

The SNPs within coding regions can be categorized as synonymous (does not alter amino acid), nonsynonymous (alters amino-acid sequence) and stop-gain/loss (causing the gain or loss of stop codons) types. The estimated proportion of amino-acid altering variants ranged between 0.3 and 1.2% in different studies (table 16). In comparison, the proportion of InDels within coding regions has been found to be very small irrespective of their categories such as: non-frameshift InDels, which are multiples of 3 nucleotides and do not change the protein-coding frame, constituted only 0.05–0.06% of variants; frameshift InDels, which are not multiples of 3 nucleotides and hence disrupt the reading frame, account for about 0.07–0.09% along with stop-gain/loss types. This indicates that the InDels in coding regions can be much more disruptive compared to SNPs and are therefore selected against.

Amino-acid altering variants are not always deleterious or radical in impact. Numerous approaches have been described to classify these variants, particularly the nonsynonymous SNPs, on both quantitative and discrete scale, by taking into account the evolutionary conservation and other properties of the proteins [Cooper and Shendure, 2011]. Major variant discovery studies on chicken have predicted the effects of nonsynonymous SNPs using various tools such as SIFT [Ng and Henikoff, 2003] and PROVEAN [Choi et al., 2012], and these studies predicted about 11–49% of the nonsynonymous variants may have radical impacts on protein functions (table 16).

One recent study by Boschiero et al. [in press] has also predicted the effect of non-frameshift InDels using the PROVEAN package and has shown that ~29% of these InDels were expected to be radical or deleterious in their effects. Such categorization can be very useful in shortlisting candidate functional mutations in studies involving specific disease or traits.

To go beyond the functional characterization of coding variants only, different studies have also utilized other resources and methods to delineate potentially functional mutations. Several studies, for instance, have attempted to analyse signatures of selection in the genome using the detected variants, not only to identify regions under selective sweep but also as a means to delineate variants that may be functionally important [Rubin et al., 2010; Qanbari et al., 2012; Fan et al., 2013; Gheyas et al., 2015]. Yan et al. [2014] have annotated InDels in the context of known QTLs. Few studies have overlaid the variants with most-conserved elements in the chicken genome as mutations within these elements are likely to have functional effects [Gheyas et al., 2015; Boschiero et al., in press]. Furthermore, the study by Gheyas et al. [2015] also attempted to predict the effects of variants from different categories, i.e. UTR, intronic, ncRNA and coding, on the secondary structures of RNAs – both mRNA and ncRNAs – as these may have functional implications. Several studies have looked for enrichment of the potentially functional variants within gene clusters/networks to associate these variants with specific physiological functions [Fan et al., 2013; Yan et al., 2014; Gheyas et al., 2015; Boschiero et al., in press]. With rapid advancements in the transcriptome profiling (e.g. RNA-seq) and continuous improvements in our understanding of INCEs in the genome, it will soon be possible to perform a comprehensive annotation of noncoding variants.

Current and Future Applications of SNP and InDel Variants

The millions of SNP and InDel variants discovered over the last few years from chicken genome have found many applications in both research and breeding. One major outcome of the availability of a large catalogue of variants has been the development of SNP genotyping arrays to facilitate high-throughput screening of many individuals. Although genotyping arrays of various densities have been developed by different groups [Gheyas and Burt, 2013], detailed descriptions are available for only 3 arrays: a low-density 3K Illumina array [Muir et al., 2008a, b], a medium-density 60K Illumina array [Groenen et al., 2011], and a high-density 600K Affymetrix® Axiom® array [Kranis et al., 2013]. Only the 600K array is commercially available, while all others are proprietary and are not available for wider use. Below, we describe some of the major areas where high-throughput variant data, either from genotyping arrays or directly from re-sequencing, has been or can be used.

Phenotype-Genotype Association

A major use of molecular markers has been in the detection of QTLs associated with various traits or diseases. Traditionally, QTL analysis in chicken has been performed using only a handful of markers, generally micro-
satellites. Since most of these studies used low-density marker maps, the QTLs were often detected with very large confidence intervals, making it difficult to identify important genes or causal mutations [Zhang H et al., 2012]. Availability of a dense set of markers has opened up the opportunity to map QTLs or locate putative causal mutations with greater accuracy by fine mapping of QTLs or applying genome-wide association studies (GWAS). As a result, we have seen a surge of research papers on GWAS published just in the last few years on a wide range of traits in chicken, e.g. growth [Xie et al., 2012], fatness [Abasht and Lamont, 2007], meat quality [Sun et al., 2013], egg production and quality [Wolc et al., 2014], egg shell strength [Ghebrewold et al., 2014], resistance towards diseases and pathogens [Luo CL et al., 2013, 2014; Wolc et al., 2013], pigmentation [Park et al., 2013], etc.

Although genotyping arrays have so far been used for GWAS, they often fall short in identifying the actual causal mutations. This is because the genotyping arrays only search for association from among pre-defined panels of markers, which generally includes only common SNPs leaving out the low-frequency variants or mutations of other types that might be important [Gheyas and Burt, 2013]. As a consequence, it has been observed that the regions detected through GWAS often explain only a small part of the genetic variance of a trait [Maher, 2008]. With increasing affordability of NGS, it is now possible to perform association analysis on all the variants detected from whole-genome or targeted sequencing of candidate regions. A number of recent studies have used NGS to identify candidate functional mutations from known QTL regions such as those affecting adiposity and fatness [Roux et al., 2014; Moreira et al., 2015], meat quality [de Koning et al., 2012] and muscle deposition [Godoy et al., 2015] in chicken.

A number of recent studies have used the variant data either from genotyping arrays or from NGS platforms to identify genomic regions that are under selection pressure [Rubin et al., 2010; Elferink et al., 2012; Qanbari et al., 2012; Fan et al., 2013; Gholami et al., 2014]. Although such analyses do not require the use of phenotypic information, they can generate valuable understanding about genotype-phenotype relationship.

Genomic Selection

Genomic selection is the latest advancement in marker-assisted selection technology where genotypes from dense marker panels are used to estimate total genomic breeding values (GBEVs) of animals for various traits, and these GEBVs are used for selection of breeding candidates [Meuwissen et al., 2001]. Genomic selection promises to increase the rate of genetic gains substantially by capturing all or most of the genetic variance associated with a trait, allowing accurate estimation of breeding values without phenotypic records or progeny information, and facilitating selection at an early stage leading to shortening of the generation interval [Meuwissen et al., 2001]. Several studies have reported the application of genomic selection in commercial poultry breeding using medium- to high-density genotyping arrays [Avendaño et al., 2010; Chen et al., 2011; Fulton, 2012; Preisinger, 2012].

Population Genetics

High-throughput genotype data on a large panel of markers can be very useful in answering a wide range of questions related to population genetics. A number of studies have utilized dense marker sets to investigate the linkage disequilibrium (LD) profiles in chicken [Aerts et al., 2007; Andreescu et al., 2007; Megens et al., 2009; Elferink et al., 2010; Qanbari et al., 2010]. The LD and haplotype maps are very useful resources as they can provide unique insights into the genetic characteristics of a population, shed lights on the recombination patterns in the genome, help identify regions under positive selection, and guide the design and analysis of GWAS and GS [International HapMap Consortium, 2005]. For example, a comprehensive study on several layer populations revealed that the magnitude of LD in WEL populations was greater and extended over a longer distance compared to that in BELs [Qanbari et al., 2010]. This was attributed to the intense long-term selection pressures on WEL lines and their narrow genetic base in contrast to the BELs. Another study on commercial broiler lines observed a significant overlap of LDs of marker pairs between related populations suggesting that haplotypes are shared among these lines [Andreescu et al., 2007]. Such information has important implications as it can help predict the transferability of genetic parameters (e.g. GEBVs, or QTL effects) from one population to another. Further studies have shown that LD in macrochromosomes extended over longer genomic distances compared to that in microchromosomes, implying that a denser panel of markers would be needed to identify QTLs on micro- than on macrochromosomes by association mapping [Megens et al., 2009; Qanbari et al., 2010].

SNP markers have also been reliably used to investigate genetic diversity within populations and assess biodiversity. Using about 3K SNPs distributed over the genome, Muir et al. [2008a, b] showed that pure chicken
lines from commercial stocks lacked 50% or more of the genetic diversity observed in ancestral breeds and suggested that SNP markers can be used as sentinels of biodiversity. A number of other studies on chicken have confirmed the usefulness of SNPs in biodiversity analyses by showing their ability to reliably cluster related individuals into groups or in expected phylogenetic orders [Hillel et al., 2007; Twito et al., 2007; Kranis et al., 2013; Mekchay et al., 2014].

Concluding Remarks
With the availability of the chicken reference genome and advancements in sequencing technologies, we have seen a surge of discovery of SNP and InDel variants in just a few years. These variants are being used for a wide range of purposes, both in research and breeding. The full benefits of these variants, however, can only be reaped with their better characterization such as having knowledge about their frequency profiles in different populations and possible functional effects. With the rapid improvements of the annotation of the chicken genome, interpretation of functional impacts of variants, especially of the noncoding ones, would be possible in the very near future.

Genetic Diversity of Village Chickens
(Prepared by T.T. Desta, R.A. Lawal, and O. Hanotte)

Domestic chickens are represented by 3 main gene pools: village, commercial and fancy (traditional) chicken. Population size of village chicken comes second to commercial ones with a total of at least 3 billion birds [Roberts, 1997]. Across the world, they represent a major economical and sociocultural asset making the livelihood of millions of households. Village chickens are the most ancient types of domestic chicken. They are found and adapted to nearly all agro-ecological systems. Village chickens represent the main repository of domestic chicken genetic diversity. Untangling this diversity is increasingly contributing to our understanding of the history of our agricultural societies, while the breeding structure and ancient history of village chicken make them ideal resource populations to unravel at genome level the impact of natural and human selections and to identify the genetic control of Mendelian and quantitative traits.

Village chickens have been named under a range of synonyms. They are commonly referred as native, local, scavenging, free range, non-descript, smallholders’ or simply indigenous fowl. These synonyms invariably indicate the traditional and free-range management systems under which they have been maintained since domestication. In this respect, village chickens are mainly reared by subsistence or smallholder farmers across Africa, South America and Asia where they fulfil a range of sociocultural and economic roles [e.g. Desta and Wakeyo, 2012]. In Europe, village chickens largely do not exist anymore having been replaced by traditional chicken populations with specific plumage and/or morphological standards [e.g. Dana et al., 2011].

Following domestication, village chickens have been shaped by natural and human selection, with survival (e.g. following infectious diseases and predation challenges) a major adaptive trait. Scavenging is the main feed strategy, and mating is polygamous and often uncontrolled. Population structure of village chicken is therefore largely panmictic in contrast to commercial and fancy breeds. Within village chickens gene flow is expected to be high but influenced by several demographic factors including sex ratio, density, flock size, and foraging behaviour. Genetic divergence between village chicken populations in the same geographic area might be low due to the easiness of transporting such a small livestock species. Also, the use of common marketsheds (villages sharing a common trading network) facilitates genetic exchange between populations [Dest, 2015]. It has been shown that genetic divergence and diversity of village chicken populations within a geographic region is rather shaped by historical, ecological and sociocultural factors than geographic distances [e.g. Leroy et al., 2012].
growth rate) have been reported indicating their potential for breeding improvement [e.g. Dana et al., 2010]. This is well illustrated by the commercial breeds, which were initially developed from village chicken populations (e.g. the commercial white egg chicken lines developed from the Mediterranean Leghorn) or their crossbreeds (e.g. commercial broiler) [Crawford, 1990].

The main ancestor of domestic chicken, the red junglefowl Gallus gallus, shows uniformity in its plumage and morphology with however a strong sexual dimorphism [Delacour, 1977]. Indeed, only subtle morphological and plumage differentiation are distinguishing the different subspecies of the red junglefowl Gallus gallus ssp [Delacour, 1977; McGowan and Madge, 2010]. So the extensive plumage and morphological phenotypic diversity found in village chicken is not observed in the red junglefowl. It may be expected that the process of domestication and selection for behavioural traits may have ‘relaxed’ the natural selection constraints on the morphology and plumage of the birds with human selection for phenotypic novelties allowing survival of birds displaying non-wild phenotypes. Increasingly, the genetic control of this diversity is being unravelled. It includes nonsynonymous mutation in coding regions (e.g. lavender plumage) [Vaez et al., 2008], mutation with regulatory effect (silky-feather) [Feng et al., 2014], chromosomal inversion (e.g. rosecomb) [Imsland et al., 2012], retroviral insertion (oocyan egg phenotype) [Wang et al., 2013; Wragg et al., 2013], etc. Whether or not shared phenotypes observed across village chicken populations, including from different continents, do have a common genetic control (‘founder phenotype’) [Megens and Groenen, 2012] remains unknown, but it may be expected that the same loci are involved. However, in one case, unique so far, it has been shown that a domestic chicken phenotype (oocyan; blue/green egg shelf colour) has evolved independently on the Asian and South American continents [Wang et al., 2013; Wragg et al., 2013]. Also for the yellow skin phenotype, commonly found in village chicken but also in commercial and fancy chicken, we know now that it originates from the genetic introgression of the grey junglefowl G. sonneratii [Eriksson et al., 2008]. This result is in agreement with earlier observations that reported that F1 birds might be obtained from crosses among all junglefowl species and domestic chickens [Crawford, 1990]. Also, F1 females between grey junglefowl males and red junglefowl females produce fertile offspring when backcrossed to either male grey or red junglefowl [Morejohn, 1968], and accordingly Nishibori et al. [2005] have shown the introgression of red junglefowl mtDNA into the grey junglefowl. Whether non-red junglefowl species introgression makes the genetic basis of domestic chicken phenotypes beside the yellow skin remains unknown. So far, there are no documented genetic evidences of introgression from the green junglefowl G. varius or the Sri Lanka junglefowl G. lafayettii into village chicken. However, worth mentioning is the common practice of breeding of F1 hybrids (called bekisar) between male green junglefowl and female domestic chicken on the Indonesian island of Java [Delacour, 1977], as well as some recent evidences of genetic introgression of domestic chicken into green junglefowl [Sawai et al., 2010].

Origin of the Genetic Diversity of Village Chicken

The publication of the chicken genome has opened the door to genome-wide analysis of DNA polymorphism in the species [International Chicken Genome Sequencing Consortium, 2004]. In a companion paper, ~2.8 million single nucleotide polymorphisms (SNPs) were identified comparing only 3 domestic chickens (a broiler, a layer and a Chinese Silkie) to a red junglefowl hen [International Chicken Polymorphism Map Consortium, 2004]. The mean nucleotide diversity reported was about 5 SNPs per kb between domestic chickens and red junglefowl as well as within chickens. It is 6–7 times larger to what has been observed, for example, in human or in dog and in contrast to the expectation that domestic animals are inbred compared to their wild ancestors [International Chicken Polymorphism Map Consortium, 2004]. Moreover, the segre-
gation of these polymorphisms across breeds indicated that most of these were present before the divergence of modern breeds and might have originated before domestication [International Chicken Polymorphism Map Consortium, 2004]. These findings underline the importance of village chicken as a reservoir of diversity.

Today, it is possible to reconcile the high genetic diversity of village chicken with our current but still fragmentary knowledge on the history of the species. Mitochondrial DNA studies indicate that the main maternal ancestor of domestic chicken is the red junglefowl G. gallus. The today geographic range of the red junglefowl extends from the north of the Indian subcontinent to western China, reaching in the south the Indonesian islands of Sumatra and Java [Delacour, 1977; McGowan and Madge, 2010]. It comprises 5 subspecies (G. g. gallus, G. g. jaboil- lei, G. g. murghi, G. g. bankiva, G. g. spadiceus). All or several of these might have contributed to the genetic background and the diversity of the domestic chicken at the time of domestication assuming several domestication centers [Liu et al., 2006; Kanginakudru et al., 2008] or subsequently through introgression into the domestic gene pools. Even today, wild red junglefowls are regularly visiting villages and crossbreeding between domestic chicken and wild red junglefowl have been documented [Berthouly et al., 2009]. Any introgression from other Gallus species will also increase the diversity of the village chicken genetic pool.

Outside the geographic ranges of the wild ancestor(s), the genetic diversity of village chicken will depend on the demographic history of the population, including founding events, multiple introductions, crossbreeding, bottleneck, genetic drift, selection, and mutation. The largely uncontrolled breeding system of village chickens is making them less vulnerable to genetic erosion compared to fancy and commercial chicken breeds. Commercial and improved chicken breeds, which are being introduced worldwide in attempts to improve village chicken productivity, may now impact at short term the genetic diversity of village chicken.

**Characterization of the Genetic Diversity of Village Chicken**

Besides the phenotypic diversity, several studies have now reported the genetic diversity of village chicken. In particular, many studies have now addressed the genetic diversity of village chicken at country or population level [e.g. Silva et al., 2009], sometimes including commercial birds as references [e.g. Leroy et al., 2012]. Large-scale studies involving different geographic regions and/or breeds using autosomal markers are few [e.g. Muir et al., 2008a; Lyimo et al., 2014]. However, several studies using modern and/or ancient mtDNA sequence information have now examined diversity of populations across several countries [Liu et al., 2006; Storey et al., 2012; Miao et al., 2012; Thomson et al., 2014]. No studies have reported so far the genome-wide functional diversity of village chicken, but a few have assessed polymorphism at candidate genes [e.g. Downing et al., 2009a, b].

From a breeding and a conservation point of view, we are interested in 3 main objectives: (1) The assessment of local village chicken genetic diversity, as indigenous chicken populations represent a reservoir of local adaptation to environmental challenges (e.g. climate, diseases) and therefore an indispensable genetic resource for local chicken production; (2) the assessment of village chicken diversity in comparison to the commercial lines, with indigenous chicken populations representing a reservoir of genetic variation for further improvement of productivity; and (3) the mapping of village chicken genetic diversity to develop short- and long-term strategies for the conservation of chicken genetic resources and to assist in population choices for crossbreeding programs.

The most extensive study that assessed the genetic diversity of commercial lines in comparison to other chicken breeds is the one of Muir et al. [2008a]. They did a genome-wide assessment of chicken SNP genetic diversity using 2,551 autosomal markers. More than half of the birds were of commercial origin including broiler, white and brown egg-layers, while the remaining ones included experimental and standard breeds of chicken. Reconstructing allele frequencies from a hypothetical ancestral population, they showed that >50% of the genetic diversity in ancestral breeds is absent in commercial lines. The missing diversity is primarily associated with founding effect attributed to small numbers of breeds and populations used to develop the commercial lines, rather than selection performed by the breeding companies. This loss of diversity observed in commercial chickens is primarily due to loss of rare alleles. The adaptive implication of this loss of diversity was not directly investigated but it has been argued that such loss might be of particular importance for local adaption (e.g. disease resistances). More recently, the study of Lyimo et al. [2014] included 9 populations of commercial lines, 3 populations of red junglefowl as well as a large number of populations from the African, Asian and European continents. They used 29 autosomal microsatellites. The highest diversity (heterozygosity and high mean numbers of alleles) was observed
in the red junglefowl populations and village chicken populations from Africa and Asia. From the latter, Southeast Asia is the region with the highest domestic chicken genetic diversity. At the opposite, chicken populations from Northwest Europe show a genetic diversity comparable to commercial lines, with the notable exception of the white egg-layer chicken lines which have the lowest diversity of all populations.

The most exhaustive studies so far in terms of geographic coverage of village chicken diversity have examined mtDNA diversity, and in particular D-loop sequence polymorphisms, with some recent studies including also full mtDNA sequences [e.g. Miao et al., 2012]. These studies primarily aim to unravel the origin of domestic chicken and their dispersion across the world as a proxy to our understanding of human past migrations and trading contacts. Village chicken diversity is an important source of information assuming that their modern mtDNA still carries genetic signatures of past demographic events. To which extent this is true may be examined through ancient DNA studies. Although restricted to a maternally inherited and a small portion of the genome, with no recombination, these studies provide an important framework to assess the origin and the today geographic distribution of village chicken diversity.

Liu et al. [2006] were the first to perform a large-scale study on mtDNA diversity of village chicken. Through the analysis of more than 800 birds across Eurasia, they revealed 9 different clades (haplogroups) with different geographic distribution patterns and expansion signatures suggesting that they originated in different geographic regions. Their findings led them to propose multiple origins of domestic chicken involving South and Southeast Asia as opposed to the previous findings of Fuji-mihito et al. [1994, 1996] which suggested that domestic chicken originated from a single domestication event in Thailand and/or adjacent regions. Subsequently, several studies confirmed the phylogeographic substructure of mtDNA in village chickens both within and outside of the geographic range of the wild ancestor [e.g. Kanginakudu et al., 2008; Muchadeyi et al., 2008; Silva et al., 2009; Mwacharo et al., 2011; Thomson et al., 2014]. These studies indicate that the diversity of village chicken mtDNA is distinct from one geographic region to another and that it has been shaped by human migration and trading networks. It is exemplified by the most likely Indonesian origin, through the Austronesian expansion of a specific mtDNA subhaplogroup in Madagascar, Southern and Eastern Africa [Razafindraibe et al., 2008; Muchadeyi et al., 2008; Mwacharo et al., 2011], as well as by the haplogroup diversity of village chicken from Sri Lanka, an ancient trading crossroad island on the Indian Ocean trading network [Silva et al., 2009]. These maritime dispersion events resulted in mtDNA diversity bottlenecks, which might have been compensated by subsequent admixture and gene flow between chicken populations originating from different dispersion events [e.g. Mwacharo et al., 2011, 2013]. Importantly, all these mitochondrial studies illustrate that village chicken diversity is distinct from one region to another with genetic drift being a major factor in the shaping of this diversity.

**Village Chicken as a Reservoir of Genetic and Phenotypic Diversity for Genomics Studies**

The high and ancient diversity as well as the breeding history of village chicken make them a treasure-trove for genome association studies of Mendelian and quantitative traits [Megens and Groenen, 2012; Wragg et al., 2012]. Linkage disequilibrium extends to short physical distances [Desta, 2015], and comparison among village chicken populations may offer the opportunity to fine-map the regions of interest through partial overlaps of haplotype block structures. Currently, the chicken genomics community has access to 2 genome-wide screening tools, a genotyping assay with around 55,000 SNPs [Groenen et al., 2011] and a high-density one including 600K SNPs [Kranis et al., 2013], well above the recommended number of 100,000 markers for genome-wide association studies [Wragg et al., 2012]. Traits like earlobes, skin colour, egg pigmentation, and comb shapes have now been mapped using the 55,000 SNP chips [Wragg et al., 2012; Siwek et al., 2013]. A study investigating the genetic control of plumage colours and patterns using the high-density 600K SNP chips shows promising results [Desta, 2015]. In the future, it is expected that full-genome sequence data will become increasingly available to unravel in-depth the genome diversity of chickens. Also, beside mapping and identification of the genetic control of morphological and production traits, village chicken populations will be increasingly studied as a model to understand how adaptation to environmental challenges may have shaped the diversity of an avian genome.

**Mendelian Traits**

(Prepared by D. Wragg)

Mendel’s ‘Versuche über Pflanzenhybriden’ [Mendel, 1866] detailed his experiments on inheritance in peas, the results of which provided evidence for his Principles of...
Segregation and of Independent Assortment. In eukaryotes, both of these principles occur in meiosis during which homologous chromosomes exchange genetic information through recombination. However, the Principle of Independent Assortment does not always hold true for genes located proximal to one another, as these are often inherited together during meiosis, and so there exists a measure of genetic linkage known as recombination frequency. This in turn gives rise to the phenomenon of linkage disequilibrium (LD) – the non-random association of alleles at multiple loci along a chromosome. The extent of LD can be influenced by several factors, including for example genetic linkage, selection, and population structure, which combine to form combinations of alleles that occur more frequently than would be expected under a random distribution – in which case they would be considered in linkage equilibrium. Thus, closely related individuals, having witnessed few recombination events, exhibit a greater extent of LD than might be observed in distantly related individuals. When a characteristic has been selected, for instance a particular plumage variant, the alleles proximal to the loci responsible for the variant will exhibit high LD relative to regions of the genome subject to little or no selection.

These concepts, together with the availability of high-density single nucleotide polymorphism (SNP) arrays [Groenen et al., 2011; Krans et al., 2013] and the declining costs of next-generation sequencing, have enabled many phenotypic characteristics of the chicken to be mapped, and subsequently to have their causal genetic mutation identified. As of September 2014, the Online Mendelian Inheritance in Animals (OMIA) database (http://omia.angis.org.au) lists a total of 127 Mendelian mutation identified. As of September 2014, the Online Mendelian Inheritance in Animals (OMIA) database (http://omia.angis.org.au) lists a total of 127 Mendelian traits/disorders in the chicken; for 40 of which the key

Another plumage variant, the autosomal dominant naked neck (Na), affects the distribution of feathers. Na was mapped to a 200-kb region identical-by-descent (IBD) on chromosome 3 [Mou et al., 2011]. Tiling the region with overlapping PCRs revealed that one specific region could not be amplified in naked-neck individuals, suggesting a possible genomic rearrangement. This was subsequently confirmed by sequencing to be a large insertion of sequence from chromosome 1, resulting in a long-range cis-regulatory effect on BMP12. The inserted sequence, originating from between WNT11 and UVRAG, contains conserved noncoding elements, but no sequence that is predicted to be transcribed. The precise mechanism of the mutation remains to be determined, although the authors suggest that BMP12 upregulation might be due to possible WNT11 enhancers in the inserted sequence. This is supported by the nature of the phenotype
and the observation that WNT11 expression was found to be significantly stronger on the neck than the body.

In addition to the absence of neck feathers, another mutation exists resulting in the near complete absence of scales and feathers, known as scaleless (sc) [Abbott and Asmundson, 1957]. Using the 60K SNP array, a significant association for sc was mapped to a 1.25-Mb region on chromosome 4 containing 11 genes [Wells et al., 2012]. Sequencing of these genes identified a SNP in exon 3 of FGF20 leading to a premature stop codon and the predicted subsequent loss of receptor and HSPG binding sites. In situ hybridization demonstrated that FGF20 was expressed in the developing feather placodes, confirming its role in early feather development, with qRT-PCR indicating expression to be specific to the epidermis.

Combs

Following the rediscovery of Mendel’s results, the pea comb was one of the first traits to have its mode of inheritance documented [Bateson, 1902]. Prior to the sequencing of the chicken genome, linkage analysis had mapped the autosomal dominant pea comb (P) to chromosome 1 [Bitgood et al., 1980; Bartlett, 1996]. After further refining the localization of the locus through SNP genotyping, Wright et al. [2009] detected a copy number variation in intron 1 of the SOX5 gene. Bateson [1902] also described the dominant inheritance of the rose comb (R), and together with Punnet [Bateson and Punnet, 1908] presented the first case of an epistatic interaction between genes, whereby individuals possessing both P and R alleles exhibit a walnut comb. The R locus was mapped to chromosome 7 [Dorshorst et al., 2010; Wragg et al., 2012] and subsequently identified to be the result of a 7.4-Mb inversion [Imsland et al., 2012]. The inversion results in the re-localization of MN12, leading to its ectopic expression during comb development. In addition, Imsland et al. [2012] postulate that the reduced fertility observed in homozygotes for R might be due to one of the breakpoints of the inversion disrupting CCDC108, resulting in a truncated transcript in the testis. Another common comb type is the autosomal incompletely dominant duplex (D), which has been mapped to chromosome 2 [Dorshorst et al., 2010; Wragg et al., 2012], although the causal mutation remains to be reported.

Skin/Shank Colour

Despite being recessive, the yellow skin phenotype is abundant throughout commercial chicken. The localization of the yellow skin locus (W) to chromosome 24 was published in the First Report on Chicken Genes and Chromosomes [Schmid et al., 2000]. Eriksson et al. [2008] used a back-cross pedigree to refine its mapping, revealing close linkage at the distal end of the chromosome. IBD mapping across a range of breeds revealed a single SNP in BCO2 in complete LD with the trait, and a minimum haplotype size of 23.8 kb. Subsequent sequencing and RT-PCR analysis suggest that tissue-specific regulatory mutation(s) alter BCO2 expression in the skin. However, the precise causal mutation remains to be determined. Also of note, arising from the same study was a phylogenetic analysis of the different junglefowl species using sequences derived from BCO2. The results of this analysis suggest that yellow skin most likely originates from the grey junglefowl, indicating a hybrid origin of the domestic chicken, although the ubiquity of the phenotype did not emerge until much later during the rapid worldwide spread of commercial chickens [Flink et al., 2014].

Black skin, also known as fibromelanosis and dermal hyperpigmentation, is a defining characteristic of the Silkie breed, amongst several others. Dorshorst et al. [2010] mapped the autosomal dominant locus (Fm) to a 2.8-Mb region of chromosome 20 using a back-cross mapping population, and subsequently refined the region to 483 kb in the same population using additional markers [Dorshorst et al., 2011]. IBD analysis identified a 75-kb haplotype within this region which contained 5 SNPs observed to be heterozygous in all black skin chickens. This observation prompted an analysis of copy number variation from the Log R ratio data from the 60K SNP array. A group-wise analysis comparing Fm and wild-type individuals revealed 2 putative duplicated regions, each larger than 100 kb and separated by 417 kb on the reference genome. Further analyses by PCR confirmed the duplications, and revealed each duplicated region to be joined to the other in an inverted orientation. The result is an increased expression of EDN3, located within the first duplicated region, leading to dermal hyperpigmentation. The duplication and overexpression of EDN3 was independently discovered and reported by Shinomiya et al. [2012].

Green shanks are the result of an interaction between the sex-linked incompletely dominant inhibitor of dermal melanin (Id) and the W loci [Smyth, 1990], and are characteristic of the green-legged partridge fowl (GP) of Poland. The GP was subject to a study by Siwek et al. [2013] to assess its genetic history using mtDNA and the 60K SNP array. In particular, significant associations were found for both Id and W, indicating that the red sea green leg phenotype is a result of recessive alleles at these 2 loci. The results of the study also highlighted the pos-
sibility of multiple alleles for \( \text{Id} \), as noted by Smyth [1990]. As the causal mutation underlying \( \text{Id} \) has not yet been identified, the mechanisms behind the interaction between this and the \( W \) locus, leading to green shanks, remain to be established.

**Blue-Shelled Eggs**

As with \( Fm \), the molecular basis underlying the autosomal dominant blue eggshell (\( O \)) phenotype has been independently discovered and reported by 2 groups, namely Wang et al. [2013] and Wragg et al. [2012, 2013]. In summary, Wang et al. investigated the trait through classical linkage analysis in an F2 resource population of Dongxiang chickens, mapping the locus to a \( \sim 120 \)-kb region on chromosome 1. Whereas Wragg et al. focussed on a diversity of chicken breeds and the Mapuche fowl of South America, from which the modern Araucana breed descends, applying an across-breed mapping strategy using the 60K SNP array to identify a 313-kb haplotype that was IBD in blue-shelled egg breeds. Both groups confirmed the tissue-specific overexpression of \( \text{SLCO1B3} \) in blue-shelled hens as a result of an endogenous retrovirus (EAV-HP) insertion upstream of the promoter; together with independent integration events for the South American and Asian continents, supported by different integration sites and target-site duplications. \( \text{SLCO1B3} \) is involved in bile acid transport, and its overexpression in the uterus of blue-shelled hens likely enhances the transportation of the bile salt biliverdin to the egg shell; biliverdin being the principal pigment in blue-shelled eggs.

Endogenous retroviruses have also been found to affect other phenotypic traits. For instance, an insertion of avian leukosis virus (ALV) in intron 4 of \( \text{TYR} \) results in the recessive white plumage of some breeds [Chang et al., 2006]. In addition, late-feathering, which is widely used to sex the chicks of some breeds, was originally linked to the promoter activity of a retroviral long-terminal repeat [Matsumine et al., 1991]; however, more recently it has been associated with a 176-kb tandem duplication resulting in the partial duplication of \( \text{PRLR} \) and \( \text{SPEF2} \) [Elferink et al., 2008]. In particular, experiments in mice have shown \( \text{PRLR} \) to influence the timing mechanism of the hair follicle [Craven et al., 2001], and it is therefore a strong candidate for late-feathering.

**Foot Development**

Several mutations affecting limb development have been documented, from which two that affect foot development have been characterized: oligozeugodactyly (\( ozd \)) and preaxial polydactyly (\( Po \)). Birds possessing an allele for \( Po \) have an additional digit on one or both feet and/or wings; however, the expression is variable and temperature-sensitive [Robb and Delany, 2012a]. A highly conserved long-range \( cis \)-regulatory \( \text{SHH} \) enhancer, known as the zone of polarizing activity regulatory sequence (ZRS), is present in intron 5 of \( \text{LMBRI} \). In the case of \( Po \) in the Silkie breed, an association with a SNP within the ZRS has been reported in several studies [Dorshorst et al., 2010; Dunn et al., 2011; Maas et al., 2011] and found to result in an allelic imbalance of \( \text{SHH} \) [Dunn et al., 2011]. In contrast to \( Po, ozd \) results in the absence of the ulna and fibula and all digits with the exception of one in the foot. However, in common with \( Po, ozd \) is the result of a mutation in the ZRS, in this case a 1,654-bp deletion causing the loss of all but the first 135 bp of the ZRS [Maas et al., 2011]. These results demonstrate the utility of the chicken as a valuable model system for the study of both limb development and gene regulation.

**Model for Disease**

A review on the chicken as an animal model for eye disease [Hocking and Guggenheim, 2013] outlines a number of recessive mutations associated with blindness. Those for which the causal variants have been identified include chicken albino (\( ca \)), arising from a 6-bp deletion in \( \text{TYR} \) [Tobita-Teramoto et al., 2000]; retinal degeneration (\( rd \)), caused by an InDel in \( \text{GCI} \) [Semple-Rowland et al., 1998]; retinal dysplasia and degeneration (\( rdd \)), resulting from a premature stop codon in \( \text{MPDZ} \) [Ali et al., 2011]; and retina globe enlarged (\( rge \)), which is due to a 3-bp deletion in \( \text{GNB3} \) [Tummala et al., 2006]. The extensive genetic resources available and the large eye size make the chicken an important ophthalmologic model.

Photosensitive reflex epilepsy has been characterized as an autosomal recessive trait in the Fepi chicken strain [Douaud et al., 2011]. In this instance, a mutation in the second intron of \( \text{SV2A} \) results in abnormal splicing, significantly reducing its expression. Upon identifying the mutation, the chickens were treated with an anti-epileptic drug (levetiracetam) which was found to significantly reduce the number and duration of seizures, indicating its anticonvulsant effect in the chicken. The mechanism of action of \( \text{SV2A} \) and its interaction with levetiracetam remain to be fully elucidated, and so the Fepi chicken provides an important model for future research.

The degenerative muscle disorder muscular dystrophy has been identified as an autosomal co-dominant trait in the chicken, for which a missense mutation in \( \text{WWP1} \) has been implicated [Matsumoto et al., 2008]. Comparative genomics indicates the missense mutation to be present...
only in dystrophic chickens, and that the region is highly conserved across the tetrapods analyzed (7 mammals, 4 reptiles, 1 amphibian, 2 birds). The mutation itself causes an amino acid replacement from arginine with a basic side chain, to glutamine with an uncharged polar side chain, and may therefore affect the function of the WWP1 protein; although no significant difference in the expression levels of WWP1 was observed between mutant and wild-type chickens. Further investigation is required to clarify the mechanism by which WWP1 triggers muscular dystrophy in chicken, the results of which could provide new insights for understanding human muscular dystrophies.

Reproductive Biology

In addition to its utility as a model for disease, and for characterizing the genetics of phenotypic diversity, the chicken is also a valuable resource for understanding reproductive biology. A review of research on the sex-linked recessive restricted ovulator (ro) phenotype [Elkin et al., 2012] indicates it to result from a naturally occurring point mutation in VLDLR located on the Z chromosome. The ovaries of R/O chickens lack follicular hierarchy and contain many small preovulatory follicles of various colours, shapes and sizes; hens also exhibit hyperestrogenemia, hypoprogesteronemia, elevated circulating gonadotropins, and upregulated pituitary progesterone receptor mRNA and isoforms. Another model of compromised female fertility, the Watanabe heritable hyperlipidemic (WHHL) rabbit [Shiomi and Ito, 2009], is widely used for studies of human autosomal dominant familial hypercholesterolemia. WHHL rabbits and R/O hens each possess defective LDL receptor supergene family members; however, the etiology of reproductive dysfunction in the 2 models is likely to be different. Nonetheless, further studies of the R/O model to improve an understanding of the molecular physiology and biochemistry concerned, may lead to new analogies between avian and mammalian systems and help to address important questions in reproductive biology.

Future Opportunities

As indicated from the number of Mendelian traits in OMIA for which the causal mutation has not been identified, many opportunities remain for further investigation. For instance ear-tufts, muffs and beard, and vulture hocks to name but a few. The Araucana and Mapuche fowl, which have already provided an important model for studying blue-shelled eggs, could also present a valuable resource for investigating ear-tufts, muffs and beard. Indeed, the Araucana has recently been the subject of study to identify candidate genes for rumpless and ear-tufts using the 60K SNP array, mapping them to chromosomes 2 and 15, respectively [Noorai et al., 2012]. The chicken is also an important model for studies concerning reproductive biology and disease, as illustrated in the small number of examples presented here, and is a potential model for numerous human diseases including for example hypothyroidism [Cole, 1966] and systemic sclerosis [Beyer et al., 2010] amongst many others. The recent availability of a 600K SNP array and the declining costs of whole-genome re-sequencing present opportunities to economically map traits at a finer resolution than previously possible. Furthermore, the widespread use of these technologies, combined with accurate phenotypic records, facilitates the incorporation of data from different studies with little difficulty.

Treasure the Exceptions: Utilizing Chicken Mutant Lines and Advanced Genetic Technologies to Uncover Genes Involved in Developmental Processes (Prepared by E.A. O’Hare and M.E. Delany)

Much of the knowledge underpinning cellular processes obtained by biologists to date has been contributed by the study of mutations, either natural variants or induced. Such discoveries shed light on many developmental mechanisms and pathways and in turn provide immeasurable benefits to the world through impacts in health and medicine.

The chicken was one of the first animals used for the analysis of human development (aka animal model) when it was studied by Aristotle (ca. 400 B.C.) to understand chicken embryo formation. Considerable use has been made of chicken mutants/genetic variants in developmental biology. Similarly, the chicken also served as an early genetics model during the seminal studies of Bateson (ca. 1900), wherein by studying comb inheritance, he proved that Mendel’s laws operated in animals.

The beauty of the classic chicken model is in its in ovo embryogenesis allowing for easy access and analysis of all stages of development in addition to a relatively short generation interval and large number of progeny. Given this, it was possible to identify a series of developmental abnormalities involving morphogenesis; this in turn led to the discovery and establishment of hundreds of unique genetic resources across the world by both university and government institutions during the 20th century [Somes, 1990a, b; Pisenti et al., 1999; Delany, 2004; Delany and Gessaro, 2008]. To date, many chicken genetic lines have...
been well-characterized for phenotype and mode of inheritance and were developed into inbred and/or congenic lines, thereby presenting mapping opportunities, candidate gene analysis and the determination of causative gene(s) and involved pathways [Pisenti et al., 1999; Delany, 2004; Robb et al., 2011].

The advent of the chicken genome sequence [International Chicken Genome Sequencing Consortium, 2004], high-throughput technologies, and genetic manipulation techniques provide an opportunity to quickly study priority candidate elements in order to better understand chicken, avian and vertebrate biology, the developmental mechanisms and pathways involved, and overall contribution to phenotype and/or disease onset [Muir et al., 2008a; Robb et al., 2011; Robb and Delany, 2012a, b; Stern, 2004]. Moreover, since many of the malformations observed in the chicken show similarity with human conditions, identification of the causative variant(s) is relevant to human health and development. Thus, the chicken is a valuable model for the scientific community and as such has been recognized as a model organism for biomedical research by the National Institutes of Health in 2009 (http://www.nih.gov/science/models/gallus/). In this section, we discuss the use of advanced technologies for the identification of those genetic variants perturbing normal development through the use of unique chicken genetic resources.

Mechanisms of ‘Gene Hunting’

The resounding success of the chicken as the premier non-mammalian vertebrate model can be attributed to the tools, strategies and genetic lines created by poultry researchers dating back over a century. The release of the chicken genome sequence and subsequent genome browser interfaces marked a historical point in this research timeline by allowing scientists to quickly and easily find a gene of interest and identify its corresponding nucleotide or protein sequence [International Chicken Genome Sequencing Consortium, 2004]. The pairing of genomic tools and sequencing strategies with unique, naturally-occurring developmental mutations in the chicken allows one to hone in on the causative gene shedding light onto the mechanisms regulating development. To date a variety of methods have been utilized to identify the causative variant, gene function and biological pathways and processes of a number of developmental mutations in the chicken. Below we describe how different techniques and technologies were utilized to uncover genes involved in development.

Mapping Techniques

The chicken chromosome map has evolved greatly over the past century to become a more fine-tuned genetic map with 97–98% of the total genomic sequence identified and assigned to specific chromosomes. However, even with the current status of the chicken genome, finding the causative variant for a particular phenotype would be nearly impossible without the knowledge of basic genetics and inheritance. As early as the beginning of the 20th century, chicken researchers showed that barring sex-linked [Spillman, 1908]. Since that time numerous developmental traits have been pinpointed to a particular chromosome type (i.e. autosomal vs. sex) and in many cases, linkage analyses and complementation studies have helped to further map phenotypes to unique chromosomes [Bitgood and Somes, 1990] aiding in the fine-mapping of these unique developmental mutations. In fact, the well-known *talpid* polydactyly mutation was localized to chromosome 5 using linkage analyses, and additional sequencing revealed a frameshift mutation in *KIAA0586* (aka TALPID3) as the causative mutation [Davey et al., 2006].

Analysis of chromosome structure is also utilized to pinpoint the region harboring causative developmental mutations. Langhorst and Fechheimer [1985] identified an autosomal recessive skeletal mutation, designated *shankless* (*shl*), which is characterized by malformed metacarpals, the lack of tarsometatarsal shanks, increased tibia length, extra bones in the digits, and fusion of the phalanges. These skeletal malformations were the result of an X-ray-induced pericentric inversion in chromosome 2 visualized by karyotype analyses [Langhorst and Fechheimer, 1985]. More recently, the causative element responsible for the classic monogenic trait *rose-comb* was determined to be the result of a chromosomal inversion. Imsland et al. [2012] elegantly utilized FISH and labeled BACs to show the presence and location of the chromosomal rearrangement, while Robb and Delany [2012a] employed the same technique to show that an inversion event was not responsible for the polydactyly phenotype. Thus, although basic in technique, classical methodologies (cytogenetics) combined with genomics (e.g. BACs) provide valuable opportunity in understanding the basis of genetic variants.

Genotyping Arrays

A single nucleotide polymorphism (SNP) array is an extremely valuable tool for studying variation between whole genomes. Exploiting this technology can shed light on linkage associations and fine-mapping, hetero-
zygosity, signatures of selection, and can even pinpoint causative variants for complex and simple genetic traits. Over the past decade, the poultry research community has successfully employed this technology to hone in on the gene(s) causing developmental mutations in the chicken. The availability of the 3K chicken SNP array in 2008 [Muir et al., 2008a] paired with specialized inbred lines allowed researchers to, for the first time, associate a specific chromosomal region with several developmental mutations. Dorshorst et al. [2010] used the 3K array to map 8 different traits in the Silkie chicken including the identification of a casual SNP present within a highly conserved cis-regulatory region of Sonic hedgehog for the polydactyly phenotype. Furthermore, despite the small number of SNPs, the 3K chicken SNP array successfully localized 7 of the 9 University of California, Davis-held developmental mutations to a unique chromosome (see suppl. table 1 in Robb et al. [2011]). Several years later, the utilization of the 60K SNP chip [Groenen et al., 2011] and an updated genome analysis allowed for further fine-mapping of those mutations with subsequent assignment of priority candidate genes [Robb et al., 2011] and the gene of interest for one, Polydactyly [Robb and Delany, 2012a]. The 60K SNP array was also useful in fine-mapping the region of the scaleless mutation to chromosome 4. After exome sequencing and subsequent analyses of the 11 genes in the region, a missense mutation in the FGF20 gene was identified as causative [Wells et al., 2012].

Copy number variants (CNVs) are small deletions and duplications of genomic DNA that impact gene expression usually identified using array-CGH. Jia et al. [2013] successfully employed the 60K SNP chip to identify new CNVs. As CNVs are commonly strongly associated with developmental malformations, utilization of SNP arrays in such a manner opens the door to investigations of associations of CNVs causing developmental defects. In 2012, a new commercially available high-density 600K SNP array was released [Kranis et al., 2013] and is being used by poultry researchers worldwide. This new chip will prove to be useful for identifying trait-associated loci, fine-mapping genomic regions, and detecting CNVs amongst all the other useful applications. With currently 3.5 million SNPs identified in the chicken to date, the chicken arrays will no doubt expand to rival those arrays used in human genome-wide association studies, extending their usefulness and allowing for a finer mapping of causative gene loci.

Sequencing
Massively parallel sequencing systems serve as a means to rapidly generate whole or targeted-region genomic sequence, gene expression profiles (e.g. RNA-seq), and protein-DNA/RNA interactions (e.g. ChIP-seq). Each application has been uniquely developed to fulfill different biological purposes. The use of these advanced technologies has revolutionized our understanding of genetics, genomics, and biology as a whole and has helped to accelerate the fields of health and disease in animals, plants and human.

Exome sequencing seeks to identify coding variants in specific, targeted genes of interest or of all known genes in the genome [Mamanova et al., 2010]. Although cheaper than whole-genome sequencing (WGS), targeted sequencing of only coding regions or lists of genes could miss genetic variants in intergenic and intronic regions or within an unknown gene. Targeted-exome sequencing has, however, been successful in the chicken in identifying the causative element for a number of developmental mutations. As previously noted, Wells et al. [2012] identified the causative element of the scaleless mutation after sequencing 11 genes on chromosome 4. Although variant identification through WGS is now becoming more cost-effective, zooming in on the region of interest, when possible, remains an attractive and cost-saving option. Robb and Delany utilized a targeted sequence capture enrichment array to uncover the causative element for 3 unique inherited mutations (coloboma, diplopodia-1, wingless-2) disrupting limb, craniofacial and/or organ development [Robb, 2012; Robb and Delany, 2012b; Robb et al., 2013].

WGS is the most all-encompassing method for identifying elements causing developmental defects in any organism. This method can detect genetic variation in a genome including SNPs, large and small insertions and deletions, CNVs, and chromosome rearrangements [International Chicken Genome Sequencing Consortium, 2004; Yan et al., 2014]. Once the data is sorted and possible causative elements are prioritized, various techniques (e.g. qRT-PCR, in situ hybridization, microarrays, RNA-seq, siRNA, protein analysis, ChIP-seq, relevant functional assays) can be utilized to confirm or eliminate variants. Chang et al. [2014] successfully paired both the 60K chicken SNP array with WGS to identify C2CD3 as the causative gene for the long-studied talpid2 mutation and provided additional evidence as to its mechanistic role using RNA-seq. Similarly, WGS and analyses have been employed for a number of the developmental mutations maintained at the University of California, Davis including diplopodia-3, diplopodia-4, eudiplopodia, limbless, and stumpy [Delany, unpubl. data].
**Future Directions for Discovering the Underlying Factors and Functions of Developmental Mutations**

The many naturally-occurring developmental mutations in the chicken offer opportunities for the study of unique aspects of developmental biology. Although many of these mutations were spontaneously derived (naturally-occurring variants), mutations can also be induced by mutagens (e.g. chemical, radiation) or targeted (e.g. TALEN, CRISPR). Gene function can also be assessed by repression of gene transcript (e.g. morpholino technology). Technology-driven mutagenesis provides a unique opportunity to conduct reverse genetics by assessment of specific gene function and will help revolutionize the field of functional genomics by allowing for the functional analysis of specific genes in vivo.

**Transient Gene Perturbation**

Morpholinos (MOs), antisense oligonucleotides used to downregulate a target gene, are a quick and easy way to assess the function of a gene, in a specific tissue, during embryonic development (GeneTools, LLC; http://www.gene-tools.com). MOs can be uniquely designed to inhibit translation (by blocking the translation initiation complex) or by modifying a normal splicing event resulting in an abnormal protein or splice-modification-triggered nonsense-mediated decay. This technology has been used in the chicken to show gene function in ovo for over 10 years [Kos et al., 2001]. Briefly, chick embryos are incubated to the developmental stage of interest, the egg is windowed on its side, and India ink is utilized to visualize the embryo. A micropipette containing the fluorescein-labeled MO (at a pre-determined concentration) is microinjected into the tissue of interest followed immediately by electroporation [Kos et al., 2003]. Uptake of the MO can be confirmed visually by fluorescent microscopy and effect of gene function can be assessed by morphological analysis [Kos et al., 2001, 2003]. This method of gene assessment can be particularly useful in eliminating possible causative genes and determining ‘the gene’ in those regions fine-mapped by other advanced technologies (e.g. SNP arrays, capture arrays, WGS). Shepard et al. [2008] elegantly showed the function of the gene versican (VCAN) in limb formation by using gene-specific MOs. This research helped to settle the gene’s long-controversial role in embryogenesis. Assessment of the developing embryo indicated that VCAN is involved in skeletogenesis as microinjection of the VCAN MO into the precartilage limb core and distal limb mesenchyme resulted in abnormal development [Shepard et al., 2008]. To date, MOs have been used widely in targeted loss-of-function experiments, and in many cases, the use of MOs has allowed for researchers to circumvent early lethality issues and identify function of genes in vivo (in ovo). Studies in the chick utilizing MOs will allow researchers to assess the effects of temporal loss-of-function thereby contributing to our knowledge of gene function in development. Assessment of the resulting morphology will elucidate the function and resulting phenotype of a gene of interest allowing researchers to discover new gene-interaction pathways, developmental processes, etc. Moreover, coinjection of a MO and the corresponding mutant transcript [containing variant(s) of interest] will shed light onto the functional role of unique variants in development allowing researchers to gain a better understanding of single variants and gene function.

**Genetically Modified Animals**

Targeted genome editing of any vertebrate system is an extensive process which involves the introduction or deletion of nucleotides in genomic DNA. Gene targeting/transgenic technology is a valuable approach to aid in the investigation of specific gene function in vivo. Transgenic and knockout animals can be generated using a number of different technologies including zinc finger nucleases (ZF), transactivator-like effector nuclease (TALEN), or RNA-guided mutagenesis using the CRISPR/Cas (clustered regularly interspaced short palindromic repeats-Cas) system; however, each strategy has its complications and limitations [Sanjana et al., 2012; Hsu et al., 2014]. Unfortunately, the difficulty in obtaining 1-celled embryos for injection and the lack of germ-line-competent embryonic stem cells has made conventional gene-targeting techniques established in other model systems difficult to employ in the chicken. However, these technical difficulties can be overcome through the use of primordial germ cells (PGCs), the precursors of eggs and sperm. Genetically modified chickens have been successfully developed by poultry researchers for over a decade. This has been accomplished through the transferring of specific genes to cultured PGCs with subsequent transplantation or microinjection to the developing embryo. Germline chimeras can then be used to produce knockout mutant offspring for future studies [van de Lavoir et al., 2006; Park and Han, 2012; Park et al., 2014a].

Genome editing, through the introduction of double-stranded breaks, is an effective way of introducing mutations into the genome. There are several new genome-editing tools available for use in the chicken. The first method is TALEN utilization, which employs an artifi-
cial restriction enzyme to generate DNA double-strand breaks in vivo ultimately leading to the introduction of an error or new DNA sequence into the gene of interest. This enzyme was generated by fusing a DNA sequence recognition domain called transcription activator-like effector (TALE) with the catalytic DNA cleavage domain of the FokI nuclease (fused complex: TALEN) [Christian et al., 2010]. Briefly, the TALE domain sequence, engineered to identify a specific DNA sequence, is inserted into a plasmid and then introduced into a target cell (e.g. PGC), wherein it is translated to produce a functional TALEN enzyme. The enzyme then enters the nucleus and, upon recognition, binds to and cleaves the target sequence. TALEN-mediated gene targeting has been successfully utilized in the chicken to mutate a number of genes including those on microchromosomes, macrochromosomes and sex chromosomes [Park et al., 2014b].

The CRISPR/Cas is a fast and cost-effective system for gene modification. This new genome-editing system exploits the immune defense mechanism used by bacteria for destroying foreign genetic material [Barrangou et al., 2007]. The highly specific, genetically engineered RNA-guided CRISPR/Cas nuclease system utilizes Watson-Crick base pairing to target genomic DNA. Upon identification of the complementary sequence, the Cas nuclease mediates a double-stranded break at a targeted genomic locus initiating subsequent error-prone cellular DNA damage repair pathways which commonly result in a gene knockout caused by frameshift mutations, premature stop codons, or even large deletions [Ran et al., 2013]. Although CRISPR/Cas system usage has not been reported in chicken to date, it is speculated to be more successful in generating genetically modified birds through injection into PGCs with subsequent transplantation.

Genome modification through use of either TALEN or CRISPR technologies will help shed light on the mechanisms behind developmental anomalies. Moreover, the development of germline mutant embryos will mimic the complete development and growth (i.e. from fertilization to hatch) of those naturally-occurring developmental mutations in the chicken as compared to assessment of gene function using MOs. Any of these tools, however, will serve as a powerful method for gene function identification in the chicken.

Conclusions

The chicken has long been a treasured resource for the scientific community contributing to advances in many fields, in particular, developmental biology. This valuable model is used to gain insight into the processes and mechanisms regulating normal development as well as shed light onto the etiology of both human and animal defects and syndromes and provide insight into possible therapeutic targets. Moreover, the advancement of gene-mapping, sequencing and mutagenesis technologies will further our knowledge of gene function in vivo. The scientific benefits of the classic chicken embryo vertebrate model system also include an expansive literature on developmental processes created by a large community of developmental, genetic and biomedical scientists. These benefits, coupled with the occurrence of inherited mutations affecting the body plan and organ systems [Pisenti et al., 1999; Delany, 2004; Robb et al., 2011, 2013; Robb and Delany, 2012a], now provide the opportunity to enhance our knowledge of the genes and sequences involved in vertebrate development. William Bateson [1908] coined the phrase ‘Treasure your exceptions’, emphasizing that naturally occurring variations, i.e. ‘exceptions’ to normal phenotypes, provide insight into the fundamental processes regulating growth and development. We now have the advantage of advanced tools to employ with either naturally-occurring or technology-induced variations to better understand processes regulating vertebrate development.

Genomic Landscape of the Chicken DT40 Cell Line
(Prepared by A. Motegi and M. Takata)

DT40 is a chicken B cell lymphoma cell line that is widely used as an excellent genetic model system with high gene-targeting efficiencies [Buerstedde and Takeda, 2006]. Along with its relatively stable karyotype and cytogenetic characters, DT40 cells provide a unique opportunity to study various cellular processes in vertebrate cells. In particular, DT40 cells contributed to our understanding of the molecular mechanisms of genome maintenance and DNA damage responses by allowing extensive genetic analyses [Motegi and Takata, 2014]. In this section, we summarize the latest state of knowledge of DT40 karyotype, genome, transcriptome, and possible molecular steps of DT40 cell development. We then briefly review the major insights into molecular mechanisms of preserving genomic integrity obtained through the reverse genetic analyses of DT40 cells.

DT40 Karyotype and Genome

DT40 cells were derived from an avian leukovirus (ALV)-transformed lymphoma experimentally devel-
oped in a domestic bred female chicken *Gallus gallus domesticus* [Baba et al., 1985]. By the time of the Second Report on Chicken Genes and Chromosomes [Schmid et al., 2005], the release of reference genome sequences of red junglefowl *G. gallus* had much improved our understanding of the chicken genome, while strain- and cell lineage-specific sequence diversities in DT40 cells remained to be determined. During this period, several genome-wide analyses of DT40 cells, including recent whole-genome shotgun sequencing and single nucleotide variation (SNV) array hybridization [Molnár et al., 2014], had been reported. By combining observations from these genome-wide analyses and from conventional cytogenetic studies, we have extracted several important cytogenetic and genomic features of DT40 cells as listed below.

First, by evaluating the read depth of whole-genome sequencing, the original DT40 cells (from the Pirbright Institute, UK) were shown to have an almost normal chicken karyotype (38 pairs of autosomes and the ZW sex chromosomes [Masabanda et al., 2004]), only except for the trisomic macrochromosome 2 and possibly tetrasomic microchromosome 24 [Molnár et al., 2014]. We consistently observe the same karyotype of macrochromosomes and encounter aneuploidies at most less than several percent in our routine inspections into the mitotic spreads [Sonoda et al., 1998], albeit somewhat higher cytogenetic mosaicism of macrochromosomes was also reported in different stocks/conditions [Chang and Delany, 2004]. Tetrasomy of microchromosome 24 agrees with the existence of 4 alleles of the *Cbl* locus [Yasuda et al., 2000], which is now allocated on chromosome 24. We also note that wild type ‘clone 18’ from Sale Laboratory (MRC, UK) was found to have additional trisomies of microchromosomes 14 and 20 [Molnár et al., 2014], while the ‘clone 18’ stock in Kyoto was not trisomic in chromosome 14 [Yamamoto et al., 2011], suggesting that some level of microchromosomal copy number variation does exist among so-called ‘clone 18’ stocks.

Second, frequencies and spectrums (transitions and transversions) of all as well as unique SNVs in DT40 cells were very similar to those in normal blood cells from 2 domestic chicken breeds, L2 and Silkie [Fan et al., 2013] (different from Hyline SC breed, from which DT40 cells were derived), in comparison with the *G. gallus* reference genome [Molnár et al., 2014]. These observations substantiate the long-held notion that the genome of DT40 cells is fairly stable at a nucleotide level and that no nucleotide instability mechanisms have been operating in this cell line. In contrast, the numbers of insertions/deletions (InDels) unique to DT40 cells were modestly higher than those in L2- and Silkie-derived cells (156.5% compared with the average from L2 and Silkie), while those in common among the 3 strains were indifferent [Molnár et al., 2014]. Closer inspection revealed that deletions of unit(s) in the 6–16-bp range at repeat sequences, but not insertions in all ranges, were evidently higher, although no obvious mutations or defects were identified in the mismatch repair pathway [Campo et al., 2013; Molnár et al., 2014], mostly responsible for deletion of this range. Interestingly, the ratio of homozygous versus heterozygous InDels unique to DT40 cells was somewhat higher, conceivably reflecting the enhanced level of loss of heterozygosity (LOH) in the 0.1–0.2-Mb range by homologous recombination (HR).

Third, integration sites of ALV were identified at 4 different loci [Molnár et al., 2014], including the previously known *c-myc* locus [Baba et al., 1985]. No apparent contribution of ALV integration at 3 new loci on B-cell transformation, if any, can be dictated from the currently available information. There were no gross chromosomal rearrangements involving the *c-myc* locus, which is frequently translocated with immunoglobulin loci in human B-cell lymphomas [Janz, 2006]. Since the *c-myc* locus is located within the region of extensive LOH, and all reads covering the integration site of *c-myc* were identical, ALV seems to have been integrated at one of the *c-myc* alleles first, and then LOH and whole copy number gain events occurred. Consistent with the previous cDNA microarray and array-comparative genomic hybridization (array-CGH) analyses [Neiman et al., 2001, 2006], the *c-myc* gene itself does not involve gene amplification, but integration of ALV at the first intron of *c-myc* juxtaposes the strong promoter in the long terminal repeat of ALV immediate upstream of exon 2 (the first coding exon) and drives overexpression of the *c-myc* mRNA (estimated 5~18-fold).

Lastly, the telomeric status is one of the critical features of cell lines with stable genomic transmission and infinite proliferation [Murnane, 2012]. DT40 cells exhibit telomerase activity comparable to that seen in chicken embryos, adult germline tissues and other transformed cell lines [Venkatesan and Price, 1998; Swanberg and Delany, 2003]. The average length of telomeres in DT40 cells was estimated to be 3–20 kb [Wei et al., 2002; Swanberg and Delany, 2003; Cooley et al., 2009; O’Hare and Delany, 2009], although some extent of range variations exists in reports at least partly due to the clone-to-clone and end-to-end differences and methodological biases [Aubert et al., 2012]. This range of telomeric repeat
length is similar to that observed in normal chicken tissues and chicken embryonic fibroblasts [Venkatesan and Price, 1998], suggesting that telomerase expression in DT40 cells might be only reactive to maintain telomeres above a critical threshold. Cooley et al. [2009] estimated the lengths of telomeres in macrochromosomes and microchromosomes to be 17–43 and 70–1,000 kb in size, respectively, by using quantitative fluorescent in situ hybridization. Further studies would be required addressing whether the estimated ~3-fold increase in telomeric signals of microchromosomes reflects the actual difference in physical lengths or some biased factor(s) such as interstitial telomeric repeats, which were reported to be enriched in many avian microchromosomes [Nanda et al., 2002].

**DT40 Transcriptome**

The gene expression profile of DT40 cells has been studied by using custom-made microarrays [Neiman et al., 2001, 2006; Koskela et al., 2003]. One study focused on a limited number of bursa-specific ESTs and demonstrated that DT40 cells are more similar to normal bursal cells than B cells at other developmental stages [Koskela et al., 2003]. This observation matches the idea that DT40 cells are likely blocked in B cell differentiation at a stage of bursal cells, both of which already express rearranged cell surface IgM and retain stem cell function. Another study compared DT40 cells with experimentally induced chicken B-cell lymphomas and showed that the transcriptional signature of DT40 cells is more similar to those in bursal preneoplastic transformed follicles than to those in late stage metastatic tumors [Neiman et al., 2001]. In this study, the numbers of upregulated genes in myc-induced lymphomas, including DT40 cells, were positively correlated with the expression levels of myc, and those changes were diminished as myc expression was reduced in metastatic tumors. These observations are consistent with the notion that extensive myc-driven transcriptional changes in transformed follicles likely promote additional mutagenic events that fix transformed phenotypes [Neiman et al., 2001]. DT40 cells persistently express high levels of myc comparable to those in ALV-transformed lymphomas and another lymphoma-derived cell line, RP9 (5–18-fold compared with normal bursa cells), suggesting that the DT40 genome could be effectively protected from the myc-driven genomic stress [Hayward et al., 1981; Baba et al., 1985; Neiman et al., 2001, 2006]. How aberrant expression of myc is tolerated in DT40 cells, however, remains to be solved. Since the above microarrays contained limited numbers of genes with least annotations, many genes might have escaped from attention. Further studies by using more standardized microarrays such as 33K Affymetrix GenChip and 44K Agilent microarray [Gheyas and Burt, 2013] could allow finer delineations of transcriptional signatures of B-cell lymphomas and DT40 cells.

MicroRNAs (miRNAs) constitute another layer of transcriptional regulations during B-cell development and differentiation [Fernando et al., 2012; de Yébenes et al., 2013; Baumjohann and Ansel, 2014]. miRNA expression profiles of naïve embryonic B cells, CD40L-stimulated B cells and DT40 cells have been studied by deep sequencing [Yao et al., 2013]. Around 20 miRNAs were identified to be upregulated (miR-18a, miR-18b, miR-20b, miR-21, miR-101, miR-103, miR-106, miR-148a, miR-221, and miR-222) or downregulated (miR-15b, miR-16, miR-26a, miR-29a, miR-30a-5p, miR-30b, miR-30c, miR-30d, miR-30e, and miR-147) in both CD40L-stimulated and DT40 cells. Although most of miRNAs in these groups have not been well characterized in B-cell lineage, those are likely involved in cell proliferation and plasma cell differentiation, respectively. The overall expression profile of DT40 cells, however, was quite distinct from that of naïve and CD40L-stimulated cells. In particular, the most striking difference was observed in miR-155 (bic), which has been shown to facilitate ALV-induced lymphomagenesis [Tam et al., 2002]. Although the miR-155 locus was originally identified as a frequent target of ALV integration [Clurman and Hayward, 1989], no integration was found in DT40 cells [Molnár et al., 2014]. While miR-155 was the most upregulated miRNA in CD40L-activated B cells, it was rather the most downregulated one in DT40 cells [Neiman et al., 2006; Yao et al., 2013]. Intriguingly, miR-155 has been shown to repress expression of activation-induced deaminase [Dorsett et al., 2008; Teng et al., 2008], a master enzyme controlling immunoglobulin gene diversifications by somatic hypermutation and gene conversion [Arakawa and Buerstedde, 2009]. Therefore, it could be interesting to address the genomic status of miR-155 and to what extent the absence of miR-155 could influence the expression level and function of activation-induced deaminase in DT40 cells.

**Insights into the Potential Molecular Steps of DT40 Cell Development**

In addition to the myc-induced modulation of the transcriptome, DT40 cells are supposed to have developed by additional steps, including inactivation of the pro-apoptotic pathway(s) and activation of the pro-sur-
vival pathway(s). Repression of p53 is one of the earliest observations [Takao et al., 1999] that fits to the former idea and was regarded to be a good explanation for the weak G1 DNA-damage checkpoint and the relatively normal G2/M checkpoint in DT40 cells [Rainey et al., 2006]. We also did not detect p53 in wild-type DT40 cells and were not able to obtain stable clones expressing p53 transgene [M.Takata, unpubl. data]. However, another line of study showed that the topoisomerase II inhibitor adriamycin did induce p53 expression and that p53 appeared to be functional at least in suppressing the expression of heat shock proteins [Tanikawa et al., 2000]. Therefore, the functionality of p53 in DT40 cells may need further evaluations.

The best clue that fits to the latter idea would be a mutation found in PIK3R1, which encodes the regulatory subunit α of the phosphatidylinositol 3-kinase (PI3-kinase) [Molnár et al., 2014]. A deletion of 6 bp in exon 11 of the only allele of PIK3R1 on the Z chromosome resulted in loss of 2 amino acids, E451 and Y452, located at the well-conserved interSH2 domain. Mutations in this domain in Burkitt’s lymphoma, a human germinal center-derived B-cell lymphoma, were shown to be competent in binding to the catalytic subunit of PI3-kinase p110, yet specifically deficient in restricting the catalytic activity of p110 [Mellor et al., 2012]. Interestingly, transgenic overexpression of MYC and the constitutive active mutant of p110 was recently shown to suffice to develop Burkitt’s lymphoma-like tumors in mice [Sander et al., 2012]. This raised the possibility that activation of the PI3-kinase signaling pathway by mutation in PIK3R1 had been one of the earliest events in DT40-cell development. However, this activation could be only transient, because DT40 cells were also shown to be competent in inducible activation of PI3-kinase by oxidative stresses [Qin and Chock, 2003]. One possible explanation for this could be the counterbalancing upregulation of the PTEN phosphatase [Neiman et al., 2001], which directly dephosphorylates lipid products of PI3-kinase.

In the same report, homozygous frameshift mutation of the ATRX gene resulting in truncation at K1392 was reported to be another possible contributor in the transformational process of DT40 cells [Molnár et al., 2014]. Since ATRX helicase was shown to facilitate various transactions at tandem repeat-rich genomic loci [Law et al., 2010; Clynes and Gibbons, 2013], it is tempting to associate ATRX defects with enhanced deletions at repeats in DT40 cells. However, we noticed that the K1392-encoding ‘exon’ is only called in the EMBL, but not in NCBI, databases and that BLAST search with this ‘exon’ sequence in the NCBI EST database retrieved no chicken ESTs. Therefore, further evaluation would be required to determine whether the observed mutations really affect the function of ATRX in DT40 cells.

**Mechanisms Maintaining Genomic Integrity of DT40 Cells**

In order to understand the molecular mechanisms how vertebrate cells maintain genomic integrity, numerous DT40 mutant cells have been generated and analyzed. DNA repair pathways so far extensively studied include 2 major DNA double-strand repair pathways, i.e. HR and non-homologous end joining (NHEJ), the translesion synthesis (TLS) repair pathway and the Fanconi anemia (FA) pathway [Sale et al., 2006; Sonoda et al., 2006; Takata et al., 2009]. Essentiality of the Rad51 recombinase for vertebrate cellular viability [Sonoda et al., 1998] and differential preference of HR versus NHEJ in phases of the cell cycle [Takata et al., 1998] are the well-recognized observations that came from early DT40 studies. TLS facilitates the restoration of stalled replication forks at damaged templates by extending DNA synthesis past damage by specialized DNA polymerases [Sale et al., 2006]. The FA pathway is mainly composed of 2 major protein complexes and believed to coordinate the selection and/or order of recruitment of DNA repair factors to resolve interstrand DNA crosslinks [Takata et al., 2009]. We demonstrated that DT40 cells singly mutated in FA genes are partly defective in HR and TLS and that double mutants defective in both FA and HR/TLS are epistatic in interstrand crosslink repair [Takata et al., 2009]. Although other DNA repair pathways such as mismatch repair, nucleotide excision repair and base excision repair are much less explored, competency of DT40 cells in these pathways was verified [Wakasugi et al., 2007; Asagoshi et al., 2010; Campo et al., 2013]. Underscoring the accumulating evidence from these reverse genetic studies, no apparent loss-of-function mutations were observed in the whole-genome sequencing analysis [Molnár et al., 2014]. Even if thousands of nonsynonymous SNVs and hundreds of coding InDels were identified to be unique in DT40 cells compared with cells from different breeds, many of them likely had pre-existed in the original animal. Therefore, comparison of unique mutations found in DT40 cells with those in Hyline SC progenitors might allow better defining DT40-specific mutagenic events. Further evaluation of such mutations could shed lights on yet unclear mechanism(s) explaining the high targeting efficiency in DT40 cells.
RNA-seq: Primary Cells, Cell Lines and Heat Stress

Transcriptome analysis by RNA-seq has emerged as a high-throughput, cost-effective means to evaluate the expression pattern of genes in organisms. Unlike other methods, such as microarrays or quantitative PCR, RNA-seq is a target-free method that permits analysis of essentially any RNA that can be amplified from a cell or tissue. At its most basic, RNA-seq can determine individual gene expression levels by counting the number of times a particular transcript was found in the sequence data. Transcript levels can be compared across multiple samples to identify differentially expressed genes and infer differences in biological states between the samples. We have used this approach to examine gene expression patterns in chicken and human cells, with particular interest in determining response to heat stress.

Three separate in vitro cell culture experiments were conducted using the White Leghorn chicken hepatocarcinoma LMH cell line, the primary chicken liver (e15) cells and the human colorectal cancer epithelial Caco-2 cell line. Cells were freshly plated at subconfluent density and grown for 18 h at 37°C. Six biological replicates were prepared for the LMH cells, while 3 biological replicates were prepared for the primary chicken hepatocytes and Caco-2 cells. After 18 h, half of the plates were transferred to 43°C for 2.5 h after which RNA was isolated, Illumina RNA-seq libraries prepared [Coble et al., 2014; Sun et al., 2015] and sequenced to a depth of 10 million or more reads by the University of Delaware DNA sequencing facility. In addition, RNA-seq libraries were prepared from 6 freshly isolated 7-day-old chicken livers. Sequence data were analyzed using the Tuxedo software suite (Gal4 reference build and annotation) and differentially enriched genes identified as indicated below. Signaling and metabolic pathways affected by enriched genes were identified using GallusReactome+(http://gallus.reactome.org/) and eGIFT [Tudor et al., 2010].

RNA-seq data were combined using chicken:human orthologs to merge the data [Burt et al., 2009] yielding a total of 5,333 pairs. Initially, hierarchical clustering was used to compare overall gene expression patterns across the cells and tissues (fig. 13). Although prepared at 2 different developmental time points (e15 vs. d7), the gene expression patterns of the primary liver culture and freshly prepared liver samples cluster together. The chicken LMH and human Caco-2 cells segregate into a second cluster. Overall, the clustering indicates that the primary cultured liver cells are more similar to fresh liver samples than the established cell lines. The clustering of the cell lines likely reflects the transformed nature and long-term culturing of the LMH and Caco-2 lines. Comparing iTerms [Tudor et al., 2010] for genes enriched in either the cell line samples or the liver samples (fig. 14) indicates enrichment for genes involved in cell proliferation in the cell lines. In contrast, the primary liver and cultured liver samples are enriched for genes involved in functions associated with primary hepatic metabolic processes such as coagulation, cholesterol metabolism and bile production. The latter observation suggests also that the short-term in vitro cultured liver cells may model responses of the intact liver.

Heat Stress
One of the objectives of our work is identifying pathways responsive to heat stress, with a particular interest in identifying evolutionarily conserved responses. To this end, genes were identified that were responsive to heat stress in the cultured liver, LMH and Caco-2 cells. The genes were then mapped to GallusReactome+ [Sun, manuscript in preparation] to identify signaling and metabolic pathways that were modulated (p value <0.05) by heat stress in all 3 cell types. A total of 280 genes mapped to GallusReactome+ and affected 22 pathways (table 17). These pathways impact many cellular activities, including...
the immune system, basic metabolism, response to growth factors and other stimuli, membrane transport of small molecules along with the integrity of the extracellular matrix. Presumably, these represent a core set of pathways that are fundamental to the overall heat stress response.

To identify a conserved set of genes that were responsive to heat stress in all 3 cultured cell types, pairwise comparisons were conducted of the 3 experiments to identify genes that were regulated in the same direction (up or down) whose difference was statistically significant (p < 0.05) (table 18). The transcription levels of 17 genes were differentially regulated by heat stress across all 3 cell types; 16 were upregulated by heat stress while 1 was downregulated. Eleven of the upregulated genes function directly as molecular chaperones or as co-chaperones (HSPB8, SERPINH1, BAG3, HSPH1, HSP90AA1, HSPA5, DNAJA1, CHORDC1, HSPA4, HSPA4L, and possibly TPPP). Several of the gene products, such as HSP90AA1, HSPA4 and HSPB8, serve as chaperones for multiple client proteins with broad effects on cellular pathways. Some of these gene products function in specific responses. For example, HSPA5 functions during endoplasmic reticulum stress, while SERPINH1 affects collagen folding and assembly. CHORDC1 binds the nucleotide-binding domain of HSP90 when the ADP:ATP ratio is high suggesting CHORDC1 may modulate HSP90 activity as a function of energy balance [Gano and Simon, 2010].

The hypothesis that TPPP may function as a chaperone or co-chaperone arises from several observations.

Fig. 14. Stacked histogram of iTerms describing biological processes affected by genes enriched in the primary cultured liver samples plus fresh liver (red) compared with genes enriched in the LKH plus Caco-2 cell lines (blue).
Other products encoded by these heat-responsive genes may not function as chaperones. OTUD1 is a deubiquitinase that regulates the level of type 1 interferon in human cells and functions as a suppressor of the innate immune response [Kagayaki et al., 2007]. The deubiquitinase activity may allow OTUD1 to serve as a broad regulator of the ubiquitin protein degradation pathway during the unfolded protein response of heat stress [TRA2 is a nuclear protein, originally identified as controlling sexing the unfolded protein response of heat stress. TRA2 is a regulator of the ubiquitin protein degradation pathway during the unfolded protein response of heat stress]. The great depth of reads provided by RNA-seq can challenge assertions that a gene is only expressed in a specific cell or tissue. For example the OTOA gene was originally identified as expressed only in inner ear, while the IMPG2 gene has been identified as specific to retinal and pineal tissue [Acharya et al., 2000]. However, we detect low levels of both OTOA and IMPG2 transcripts in all 3 cell types. Inspection of the RNA-seq coverage for both genes revealed multiple reads mapping to the corresponding transcripts with the preponderance coming from exons (fig. 15). Also, some of the reads span an intron. Taken together, these observations suggest that the detected reads arise from mature transcripts rather than precursor RNAs or genomic contamination. The conclusion is that the reads from these genes in our experiments arose from true OTOA or IMPG2 transcripts. However, there is no current evidence that the OTOA or IMPG2 transcripts give rise to protein. This demonstrates an important current challenge: linking data from high-throughput sequencing and proteomics analysis to verify translation of such mRNAs.

**Future Directions**

A major application of RNA-seq transcriptome is determining gene expression profiles. Currently, many such studies focus on identifying genes based upon the annotated genomic sequence for the target organism. However, RNA-seq data can provide a rich source of data for transcripts that are currently not recognized in a genome’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report]. Combining data from short-read (such as data generated by Illumina’s annotation file [Smith et al., this report].
mina) and long-read (PacBio) RNA-seq technologies can provide an even better understanding of transcript structure, significantly improving genome annotation. Alternative splicing, allele-specific expression, microRNAs, and lncRNAs can all be identified with RNA-seq data, providing a rich catalog of the diversity of transcripts found in samples. Given the (relative) ease with which this data can be collected, an important current need is improving and extending bioinformatics tools. While many such tools exist, the community needs to work together to develop intuitive web-based bioinformatics pipelines and platforms that are appropriate to users with varying levels of computer sophistication. Ultimately, combining an incredibly rich set of sequence data with user-friendly bioinformatics tools will provide complete gene expression profiles of organisms in many different biological states.

### Table 18. Differentially expressed genes responsive to heat stress

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>p value*</th>
<th>Mean RPMK values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>TC_Liver</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>control</td>
</tr>
<tr>
<td>HSPA5</td>
<td>heat shock 70-kDa protein 5</td>
<td>0.0002</td>
<td>0.0088 &lt;0.0001</td>
</tr>
<tr>
<td>HSPB8</td>
<td>heat shock 22-kDa protein 8</td>
<td>0.0004</td>
<td>0.0145 &lt;0.0001</td>
</tr>
<tr>
<td>HSPA4A</td>
<td>heat shock protein 90-kDa alpha 1</td>
<td>&lt;0.0001</td>
<td>0.0423 &lt;0.0001</td>
</tr>
<tr>
<td>BAG3</td>
<td>BCL2-associated athanogene 3</td>
<td>0.0002</td>
<td>0.0307 &lt;0.0001</td>
</tr>
<tr>
<td>HSPH1</td>
<td>heat shock 105-kDa/110-kDa protein 1</td>
<td>&lt;0.0001</td>
<td>0.0180 &lt;0.0001</td>
</tr>
<tr>
<td>TRA2A</td>
<td>transformer 2 alpha</td>
<td>0.0002</td>
<td>0.0135 &lt;0.0001</td>
</tr>
<tr>
<td>HSPAA4</td>
<td>heat shock 70-kDa protein 4</td>
<td>&lt;0.0001</td>
<td>0.0417 &lt;0.0001</td>
</tr>
<tr>
<td>DNAJAI</td>
<td>DnaJ (Hsp40)</td>
<td>&lt;0.0001</td>
<td>0.0099 &lt;0.0001</td>
</tr>
<tr>
<td>CHORDC1</td>
<td>cyssteine and histidine-rich domain containing 1</td>
<td>0.0009</td>
<td>0.0366 &lt;0.0001</td>
</tr>
<tr>
<td>OTUD1</td>
<td>OTU domain containing 1</td>
<td>0.0055</td>
<td>0.0146 0.0021</td>
</tr>
<tr>
<td>HSPA4</td>
<td>heat shock 70-kDa protein 4</td>
<td>&lt;0.0001</td>
<td>0.0342 &lt;0.0001</td>
</tr>
<tr>
<td>TPPP</td>
<td>tubulin polymerization promoting protein</td>
<td>0.0064</td>
<td>0.0451 0.0076</td>
</tr>
<tr>
<td>SNTB2</td>
<td>syntrophin, beta 2</td>
<td>0.0018</td>
<td>0.0072 0.0014</td>
</tr>
<tr>
<td>VRK1</td>
<td>vaccinia-related kinase 1</td>
<td>&lt;0.0001</td>
<td>0.0106 0.0010</td>
</tr>
<tr>
<td>SERPINH1</td>
<td>serpin peptidase inhibitor, clade H-1</td>
<td>0.0005</td>
<td>0.0346 &lt;0.0001</td>
</tr>
<tr>
<td>OTOA</td>
<td>otoancorin</td>
<td>0.0018</td>
<td>0.0289 &lt;0.0001</td>
</tr>
<tr>
<td>IMPG2</td>
<td>interphotoreceptor matrix proteoglycan 2</td>
<td>0.0002</td>
<td>0.0315 0.0024</td>
</tr>
</tbody>
</table>

* p value for difference between heat-stressed and control samples derived from primary cultured liver (TC_Liver), LMH cells and Caco-2 cells.

---

Host-Viral Genome Interactions in Marek’s Disease

(Prepared by M.C. McPherson, C.M. Robinson, and M.E. Delany)

Marek’s Disease and Marek’s Disease Virus

Marek’s disease (MD) encompasses a variety of symptoms in its host, including paralysis, blindness, immunosuppression, and lymphoma formation, depending on viral serotype, bird genotype and immune status of the affected chicken. In the 1960s, the symptoms of MD were attributed to a herpesvirus, the Marek’s disease virus (MDV). The virus was named after József Marek who first identified a widespread polyneuritis common amongst birds in 1907 [Gimeno et al., 1999]. In 1982, electron-microscopy data identified MDV as an alphaherpesvirus, and sequencing efforts in 2000 further confirmed its alphaherpesvirus designation [Lee et al., 2000; Tulman et al., 2000]. We currently understand that virulent MDV undergoes 4 overlapping infection stages that contribute to viral persistence in a host bird; early cytolytic, latent, late cytolytic, and cell transformation [Addinger and Calnek, 1973; Osterrieder et al., 2006; Gimeno et al., 2011]. During the early cytolytic infection, the virus is actively spreading among host immune cells, and the MDV genome is replicated and packaged into infectious virion particles, to be shed into the environment with the host feather dander. The viral load in the immune tissues peaks during this initial stage. Following the lytic phase...

---

Fig. 15. Track diagrams from UCSC browser plotting chicken LMH RNA-seq reads from OTOA (a, b) and IMPG2 (c, d). a and c correspond to the entire gene region, while b and d have been zoomed in to show coverage of individual exons along with reads that span the intron indicating splicing events.

(For figure see next page.)
of infection, latent virus (non-replicating) is established in activated T lymphocytes around 7 days post-infection (dpi) [Baigent and Davison, 2004; Trapp et al., 2006; Arumugaswami et al., 2009]. These T lymphocytes are capable of being transformed, typically around 21 dpi, resulting in lymphoma formation [Witter, 1997; Trapp et al., 2006].

**MVD Vaccines and the Poultry Industry**

The uncovering of the oncogenic herpesvirus etiology was followed by development of effective vaccines in the early 1970s. The MD vaccines were the first vaccines capable of preventing a cancer in both veterinary and human medicine, making this aspect of MDV research particularly exciting. The Rispens vaccine, developed via serotype 1 Rispens virus attenuation, has proven to be the most effective against very virulent (vv) and very virulent plus (vv+) MDV. Inoculation of birds with herpesvirus of turkey (HVT), a serotype 3 virus, and SB-1, a serotype 2 virus, also prevent MD, but are generally less effective against vv and vv+ strains. Vaccination is currently administered either in ovo or at hatch. Vaccines are surmised to prime the chicken immune system, although they do not prevent MDV infection [Islam et al., 2002]. This process is not well understood and supplementary mechanisms in preventing and/or reducing MD-lymphoma development are doubtless involved [Witter, 1984; Osterrieder et al., 2006]. It is a well-studied feature that bird genotype for the major histocompatibility complex confers resistance (e.g. B21) or susceptibility (e.g. B19) to MDV. The poultry industry applies variable MD immunization strategies, administering either monovalent (e.g. Rispens) or polyvalent (e.g. SB-1/HVT) vaccines. Nonetheless, concerns are growing as increasingly virulent MDV strains emerge [Nair, 2005], heightening the need for an enhanced mechanistic understanding of all serotypes of MDV and further vaccine development.

**Cytogenetic Investigations of Host-Virus Interactions**

Here, we will describe MDV and host genome interactions and review current methods that incorporate genetic tools to more precisely investigate genome-level host-virus interactions. These studies and others have helped elucidate the role of chromosomal integration of MDV in disease pathogenesis.

**Initial Studies Detect MDV Integration in Host Telomeres**

It was initially understood that all MD-transformed cells maintained a latent infection with the virus; however, the status of viral DNA in the transformed cells was unknown. Delecluse and Hammerschmidt [1993] were the first to determine that MDV integrated into host chromosomal DNA in transformed cell lines developed from infected birds. Integration into host DNA was supported by detection of doublet MDV FISH signals from sister chromatids, indicating viral DNA was replicated along with the host genome. Integration loci mapping through FISH was limited to the largest macrochromosomes due to a lack of identifying features for the microchromosomes. Integration sites were typically distally-located at the telomeres on the macrochromosomes. Integration profiles were recurrent from cell-to-cell within a line, yet differed from the profiles found in other cell lines. Free linear virion, representing processed MDV genomic DNA during the lytic stage, was also detected at low levels in some of the transformed cell lines.

**MDV FISH – Current Methods**

The availability of chicken BAC clones, many of which have been assigned to a particular macro- or microchromosome [Lee et al., 2003], as well as MDV cosmid and BAC clones, which contain the viral genome or a segment of it [Lee et al., 2000; Tulman et al., 2000; Petherbridge et al., 2003, 2004; Baigent et al., 2006; Niikura et al., 2006; Silva et al., 2010], unlocked new possibilities in cytogenetic research. The former allowed the identification of specific chromosomes [Romanov et al., 2005; Delany et al., 2007], which could not be distinguished by their morphological features alone (i.e. GGA6–28), while the latter was utilized to investigate interactions between the host and viral DNA. Importantly, a number of MDV clones are currently available, including those of non-oncogenic MDV strains (i.e. HVT, SB-1, Meq-deleted). Multi-color FISH techniques, wherein a labeled MDV probe (of one color) was hybridized to interphase nuclei and metaphase chromosomes along with differentially labeled chromosome- and/or telomere-specific probe(s) (of a 2nd and 3rd color), were used for the chromosome mapping of viral DNA integrations and analysis of viral DNA status over time and in different tissues [Robinson et al., 2010].

**Cytogenomic Analysis**

Robinson et al. [2014] reported 4 viral genome status phenotypes in dividing lymphocytes interrogated by FISH with an MDV-BAC probe. The 4 phenotypes are (1) MDV-null, characterized by the complete absence of fluorescence signals in the nucleus or around the chromosomes, (2) MDV chromosome-associated, represented by distinct flu-
orescence signals around and over the host chromosomes, (3) MDV chromosome-associated/integrated, showing the associated signals as well as punctate fluorescence signals integrated into the sister chromatids at one or more telomeres, and (4) MDV chromosome-integrated only, distinguished by the presence of the distinct, punctate signals at the telomeres with no other signals. Figure 16 shows representative examples of the 4 phenotypes. The interpretations of the phenotypes are that MDV-null cells are free of MDV DNA, indicating the absence of the virus in these cells. The MDV chromosome-associated phenotype suggests cells in the cytophitic stage (free virus evident), while the chromosome-associated/integrated phenotype indicates both the presence of free episomal virus and viral DNA that has integrated into the host genome. The integrated-only phenotype is interpreted to reflect a latently-infected cell, cleared of replicating virus. The interpretations are based both on the appearances and known timing of the disease stages. The appearance of the associated-only MDV phenotype in host immune tissues is clearly linked with the timing and expected viral DNA ‘behavior’ of the lytic replication stage (1–7 dpi). However, the biological implications of the associated/integrated MDV phenotype are less clear, as it is not known whether viral replication is still occurring after copies of the MDV genome become integrated into the host DNA and precisely how a latent infection is established. The appearance of the MDV chromosome-integrated only phenotype shows significant correlation with the timing of viral latency (7 dpi) and is the predominate phenotype ultimately found in MD-induced lymphomas [Robinson et al., 2010, 2014].

**Host-Virus Genome Interactions: Investigations of MDV and the Chicken Genome**

MDV-infected chickens provide a unique research model for investigation of herpesvirus infection, virus-induced lymphoma formation and viral latency and cellular persistence, as well as interface between these events. Cytogenetic investigations provide a precise assay of the physical behavior of the virus at the level of the genome.

**Fig. 16.** Representative images of the MDV and chicken host genome interaction phenotypes in mitotically dividing cells of the spleen of virulent GA-challenged birds. **a** MDV null (no virus detected); **b** MDV chromosome-associated (interpreted as lytic stage of virus replication); **c** MDV chromosome-associated and telomere-integrated (transitional state with virus replicating and integrated into the chicken genome); **d** MDV telomere-integrated only (interpreted as latent stage). Chicken chromosomes are counterstained with DAPI (blue), and the FITC (green) signals represent MDV DNA.
Host-MDV genome interactions appear to have a significant role in the virus’ achievement of latency, cellular transformation and reemergence to lytic infection. Understanding the role of these interactions may also help identify new vaccine candidates and characterize emergent, highly virulent MDV strains.

Chromosomal Association during Timing of Cytolytic Replication

Viral genomic DNA reaches considerable levels in host tissues as early as 4 dpi based on quantitative data [Baigent and Davison, 1999]. However, Robinson et al. [2014] detected MDV chromosome-associated FISH signals as early as 1 dpi within the immune organs (bursa, thymus and spleen) of GA-infected birds, thus providing earlier evidence for the lytic replication in host organs even before 4 dpi. Immune tissues, which represented a population of primarily T and B lymphocytes (levels varying by tissue type), showed a peak level of MDV chromosome-associated phenotype cells around 4 dpi, which significantly and continuously declined at 14 and 21 dpi. This is consistent with data indicating a peak of lytic replication between 4 and 7 dpi [Baigent and Davison, 2004].

Telomere Integration by MDV: Mechanism and Mapping

Before the experimental detection of MDV integration into chicken chromosomes, segments of the MDV genome with sequence homology to chicken telomeric repeats were reported in the literature [Kishi et al., 1991]. Shortly thereafter, Delecluse and Hammerschmidt [1993] provided clear evidence of MDV integration into host telomeres. Later studies offered further evidence of a physical interaction between viral and host DNA by detecting viral genes with homology to chicken, including viral interleukin-8 and telomerase RNA [Parcells et al., 2001; Fragnet et al., 2003], and providing direct proof of MDV’s capacity to acquire host DNA into its own genome [Niikura et al., 2006]. In 2010, Robinson et al. [2010] determined that oncogenic strains were integrated specifically at the telomeres within chicken MD-induced lymphomas. In corroboration with Delecluse and Hammerschmidt’s [1993] findings, doublet FISH signals were observed, indicative of integrated MDV in sister chromatids, and the vast majority of mitoses displayed the MDV chromosome-integrated only phenotype. Notable fluorescence signal strength variation was detected within and across cells and lymphomas, suggesting that a variable number of tandemly repeated MDV genomes integrated into host DNA differing considerably from one integration event to the next. With regard to random, preferred or targeted integration into specific chromosomes, temporal analysis showed strong evidence for random integration initially (although with specific targeting to the terminal telomeric DNA) [Robinson et al., 2014], but there appear to be some ‘preferred’ sites for viral integration as per the MD tumor data [Robinson et al., 2010], e.g. GGA4, 6, 9, 12, and 20. This fine-tuning may be related to the cell-selection process in transformation and/or tumorigenesis. The mapping analysis in the same cytogenetic study examined the hypothesis as to preferential integration into chromosomes with mega-telomeres, which was not supported. It was later established that herpesvirus telomeric repeats facilitate viral genome integration into host telomeres [Kaufer et al., 2011]. These authors found DNA-junction fragments containing both host-telomeric and MDV DNA from infected lymphoblastoid cells. These unique fragments were absent when the birds were challenged with the mutant- or deleted-telomeric repeat MDV strains, suggesting such mutant viral genomes exhibit a failure to integrate at the host telomeres. Combined, these data suggest that the mechanism of integration by MDV may involve the host telomere extension pathway or homologous recombination events.

MDV Integration and Latency

As a herpesvirus, MDV enters the stage of latency becoming quiescent and thus evading further host immune attack. The latent stage of infection is characterized by reduced gene expression, decline in viral load (as compared to earlier timepoint in infection) and persistence in the host lymphocytes [Stevens, 1989]. Delecluse and Hammerschmidt [1993] were the first to establish that MDV integration is associated with latently-infected chicken cells, although the nature of the relationship between these 2 events remained unclear. Latently-infected cell lines also maintain very little to no extrachromosomal circular or linear viral DNA. Robinson et al. [2014] confirmed that the telomere-integrated only form of MDV first emerges around 7 dpi (at the timing of latency) and increases in frequency at 14 and 21 dpi in bursa, thymus and spleen. However, it is unclear as to whether the act of integration into host telomeric DNA is necessary to achieve the latent stage or, conversely, if viral integration requires latency-related events to occur. Interestingly, lymphoblastoid cells infected with mutant MDV strains, which cannot integrate at host telomeres, do not show significant changes in latency-associated gene expression (e.g. Meq and vTR) as compared to cells infected with oncogenic wild-type MDV [Kaufer et al., 2011].
MDV Integration and Oncogenesis

Multiple studies have indicated a biological relationship between MDV integration into host telomeres and cellular transformation events leading to lymphoma formation in infected, MD-susceptible birds. MDV shares this trait with a number of other oncogenic viruses, including human hepatitis B and papilloma virus-18, which integrate into the genomic DNA of the host cells they later transform [Popescu et al., 1990]. Nearly all cells within primary MD-lymphomas or transformed cell populations have MDV integrated into one or more telomeric sites [Delecluse and Hammerschmidt, 1993; Robinson et al., 2010, 2014]. Furthermore, very low levels of lytically replicating or free viral DNA are detected from these cell populations [Delecluse and Hammerschmidt, 1993]. Examination of the MDV integration profiles or the distinct telomeric sites of integrated MDV in a nucleus, revealed evidence of both monoclonal and polyclonal tumors [Robinson et al., 2010]. Profiling of MD-induced tumors indicated more heterogeneity in viral-integration profiles within and between early-stage (21 dpi) tumors, whereas late-stage (61 and 73 dpi) tumors demonstrated more homogeneity [Robinson et al., 2014]. The proposed model is that early-stage tumors represent a collection of recently transformed T lymphocytes, many of which do not persist to generate an enhanced cell lineage. The late-stage tumors may mostly represent one transformation event that was selected further by the processes of tumorigenesis to become the predominant cell lineage in the initial tumor, and leading to metastases, resulting in monoclonal tumors in other tissues. In vivo experiments demonstrated that birds infected with non-integrating MDV strains have significantly decreased lymphoma development, providing further evidence of the link between MDV integration into host telomeres and achievement of cellular transformation [Kaufer et al., 2011]. The remaining question is; how does MDV integration contribute to oncogenesis? One hypothetical answer is that integration helps to establish latency, which in turn is concomitant with viral oncogene expression.

Telomeric Integration by Non-Oncogenic MDV Strains

Recently, we characterized in vivo interactions of MDV vaccine-viruses, specifically Rispens, SB-1 and HVT, with the host genome between 1 and 21 dpi [McPherson et al., 2014, manuscript in preparation]. The MDV chromosome-associated/integrated cell phenotype was detected across all vaccinated birds, indicating that MD vaccine-viruses are capable of integrating into chicken telomeres. In contrast with the oncogenic MDV-challenged birds, the vaccinated birds indicated a sustained and higher cell population with free, replicating virus (chromosome-associated virus phenotype) from 4 to 21 dpi. The MDV telomere-integrated only phenotype was not established within spleen cell populations of vaccinated birds during early infection. Interestingly, the non-oncogenic Meq-deleted MDV strain (ΔMeq) [Silva et al., 2010] also maintains a consistently high level of lytic chromosome-associated signals and does not establish the MDV chromosome-integrated only phenotype in host cells from 1 to 21 dpi [Robinson et al., 2014]. The MDV chromosome-associated/integrated phenotype was detected by FISH in the spleen, bursa and thymus of ΔMeq-infected birds, indicating that the strain is capable of telomeric integration, yet does not transform host lymphocytes in vivo and may not have a capacity for latent infection. These results provided further evidence that Meq expression and telomeric integration by MDV play crucial roles in the establishment of a latent infection and transformation of host cells. Collectively, the data on non-oncogenic MDV strains clarifies that telomeric integration is not sufficient for cellular transformation, although it does appear to be necessary.

Reactivation of Lytic Replication in Latently-Infected Host Cells

MD-induced lymphomas principally consist of latently-infected and transformed host cells [Delecluse et al., 1993; Robinson et al., 2010, 2014]. However, an increased number of cells with lytic DNA replication (free linear virion) are detected from MD primary lymphomas in culture conditions and, particularly, after exposure to bromodeoxyuridine [Delecluse et al., 1993; Robinson, 2013], UA04, a unique latently-infected T lymphocyte line [Dienglewicz and Parcells, 1999], was used to study the behavior of reactivated MDV to investigate the hypothesis that integrated MDV was capable of emerging from its chromosome-embedded locus. Analysis of the cell line showed 2 MDV integration loci. Upon incubation with bromodeoxyuridine, which induces the latent virus to become lytic [Dunn and Nazerian, 1977], multiple, speckled, smaller FISH signals surrounding and associated with the integration loci (‘starbursts of virus signal’) appeared. The FISH signals in close proximity to MDV integration sites are interpreted as having ‘escaped’ from their integration loci and are free replicating virus (fig. 17). It has also been found that the presence of telomeric repeats in the MDV genome is crucial to viral reactivation [Kaufer et al., 2011]. These combined data suggest that...
MDV may not only use integration as a means for avoiding immune detection (latency), but also as a mode of persistence in the host with the means to ‘escape’ integration within the host genome and return to the replication phase of its life cycle. The ability to return to a state of productive replication after integration implies that the MDV genome is integrating intact (without rearrangement) into the host DNA. The mechanism by which the integrated MDV genome is excised or generates free virus involves the viral telomeric repeats; however, the exact host and/or viral machinery involved offers opportunity for future study.

**Transcriptome Variation in Response to Marek’s Disease Virus Acute Infection**

Marek’s disease (MD) is an economically significant chicken disease that affects the poultry industry worldwide with an estimated annual cost of USD 2 billion [Morrow and Fehler, 2004]. The disease is caused by the highly oncogenic Marek’s disease virus (MDV), an alphaherpesvirus that induces T-cell lymphomas in susceptible birds. Vaccination is the primary control measure, which is effective in reducing incidence of tumor formation. However, since MD vaccines are not sterilizing, they do not prevent infection or horizontal spread of the virus. As a consequence, MDV field strains that overcome vaccinal protection have arisen repeatedly over time [Atkins et al., 2013]. Therefore, sustainable alternative control measures, such as improving genetic resistance, are needed.

Many studies have reported strong associations between MHC alleles and resistance or susceptibility to MD. For example, chickens with the MHC allele B21 are resistant in contrast to chickens with the B19 allele, which are susceptible. ADOL lines 6 and 7, both share the same MHC B2 allele, yet exhibit different phenotypic responses; e.g. challenge with the JM/102W strain typically results in 0 and 100% MD incidence for lines 6 and 7, respectively. Thus, the major unanswered questions are what genetic factors, especially those that are non-MHC, contribute to susceptibility and resistance to the disease, and what are the main contributing mechanisms?

Significant efforts have been made to study variations in global gene expression between MD-resistant and susceptible birds using microarray and RNA-seq methods in order to identify non-MHC genes that contribute to resistance to MD [Bumstead, 1998; Vallejo et al., 1998; Yonash et al., 1999; Morgan et al., 2001; Sarson et al., 2008; Smith et al., 2011]. However, none of the studies have investigated differential expression of alternative isoforms, which are known to play a significant role in many biological events including immune responses.
In addition, studies have shown that isoform expression levels can provide better signatures for some diseases [Zhang et al., 2013]. Changes of isoform expression levels are governed partly by 2 types of cis-regulatory elements: exonic splicing enhancer (ESE) and exonic splicing silencer (ESS), both located within an exon sequence. A number of sequence motifs of ESE and ESS have been identified in human and other organisms and can be predicted in silico. Mutations that disrupt or create those motifs can alter splicing patterns leading to aberrant alternative splicing. A number of disease-associated single nucleotide polymorphisms (SNPs) in coding regions that affect ESEs and ESSs have been well characterized [Blencowe, 2000; Wang and Cooper, 2007]. Therefore, variations in isoform expression could lead to identification of SNPs that underlie genetic resistance to MD. Here, we report differentially expressed genes and isoforms that may contribute to resistance to MD as well as SNPs that can potentially affect isoform expression levels.

![Fig. 18. Differentially expressed genes in response to MDV infection. The number of unique or common genes showing up- or downregulation in line 6 (MD resistant) or line 7 (MD susceptible) chicks at 4 days post-infection using splenic RNA is given in each Venn diagram.](image)

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Change in expression*</th>
</tr>
</thead>
<tbody>
<tr>
<td>LL</td>
<td>–3.36 +8.71</td>
</tr>
<tr>
<td>GIF</td>
<td>–2.15 +3.11</td>
</tr>
<tr>
<td>SFTP A1</td>
<td>–4.84 +3.73</td>
</tr>
<tr>
<td>SCAF8</td>
<td>–8.70 +8.24</td>
</tr>
<tr>
<td>C14ORF1</td>
<td>–3.11 +2.71</td>
</tr>
<tr>
<td>PPARG</td>
<td>–6.99 +2.06</td>
</tr>
<tr>
<td>NDUF A4</td>
<td>+1.66 –1.03</td>
</tr>
<tr>
<td>RAD17</td>
<td>+1.45 –1.80</td>
</tr>
<tr>
<td>RPL 39</td>
<td>+3.34 –1.63</td>
</tr>
<tr>
<td>ATP8 A2</td>
<td>+7.66 –7.80</td>
</tr>
<tr>
<td>NDUF B3</td>
<td>+1.03 –1.34</td>
</tr>
<tr>
<td>PSMG3</td>
<td>+1.52 –1.26</td>
</tr>
<tr>
<td>MED 9</td>
<td>+8.23 –3.15</td>
</tr>
<tr>
<td>PNN SR</td>
<td>+5.58 –1.81</td>
</tr>
<tr>
<td>SIPR 1</td>
<td>+1.31 –6.96</td>
</tr>
<tr>
<td>CD 7</td>
<td>+2.88 –1.38</td>
</tr>
<tr>
<td>LOC1008 5878</td>
<td>+1.26 –1.75</td>
</tr>
<tr>
<td>THOC 7</td>
<td>+7.01 –6.90</td>
</tr>
<tr>
<td>DNAJ 2</td>
<td>+2.34 –6.02</td>
</tr>
</tbody>
</table>

*The – and + symbols represent the log2 fold change down- and up-regulated genes in response to MDV infection, respectively.

Table 19. Splenic genes regulated in opposite directions between lines 6 (MD resistant) and 7 (MD susceptible) in response to MDV infection at 4 days post-infection

---

[Lynch, 2004]. In addition, studies have shown that isoform expression levels can provide better signatures for some diseases [Zhang et al., 2013]. Changes of isoform expression levels are governed partly by 2 types of cis-regulatory elements: exonic splicing enhancer (ESE) and exonic splicing silencer (ESS), both located within an exon sequence. A number of sequence motifs of ESE and ESS have been identified in human and other organisms and can be predicted in silico. Mutations that disrupt or create those motifs can alter splicing patterns leading to aberrant alternative splicing. A number of disease-associated single nucleotide polymorphisms (SNPs) in coding regions that affect ESEs and ESSs have been well characterized [Blencowe, 2000; Wang and Cooper, 2007]. Therefore, variations in isoform expression could lead to identification of SNPs that underlie genetic resistance to MD. Here, we report differentially expressed genes and isoforms that may contribute to resistance to MD as well as SNPs that can potentially affect isoform expression levels.
Differential Gene Expression Indicates Active Immune Responses to Ongoing Lytic Infection in the Susceptible Line

To study gene and isoform expression, we incorporated Ensembl gene models (release 73) with de novo and a reference-guided transcriptome assembly to build custom chicken gene models. The models, therefore, include both Ensembl-annotated transcripts and putative genes and isoforms. The advantage of using custom gene models is that it allows an investigation of unannotated genes and isoforms, which is necessary for in-depth study of gene expression. Based on these models, many genes were differentially expressed between control and infected chickens in both lines at 4 days post-infection. While the number of unique downregulated genes in both lines was approximately equal, the number of unique upregulated genes in the susceptible line was much greater compared to the resistant line (fig. 18). Interestingly, some genes that were differentially expressed in both lines were regulated in the opposite direction (table 19). Among genes downregulated in the resistant line but upregulated in the susceptible line were \textit{LL} (lung lectin) and \textit{SFTPA1}, which encode a calcium-dependent C-type lectin and a lung surfactant protein, respectively, both important in innate immunity [Kingma and Whitsett, 2006; Hogenkamp et al., 2008], and \textit{PPARG}, a suppressor of the NFκB-mediated proinflammatory response. On the other hand, nearly all genes upregulated in the resistant line but downregulated in the susceptible line are involved in cell survival such as mRNA splicing, cell growth and protein synthesis, except \textit{CD7} whose function is involved in T cell-B cell interaction. This difference suggests that even at this stage of infection, in the resistant line, the lytic phase could be repressed. Therefore, only genes involved in cell division are upregulated possibly to repair the initial damage due to infection in the resistant line. In comparison, the lytic phase in the susceptible line may still continue and as a result, genes involved in immune responses are still upregulated. In addition, type I interferons (IFN-γ and IFN-β) as well as IFN-α3 were found to be highly upregulated in infected chickens in both lines (table 20). However, genes encoding their corresponding receptors were not differentially expressed in the resistant line, but upregulated in the susceptible line. This could also reflect the ongoing immune response in the susceptible line.

Functional Analysis of Differentially Expressed Genes Indicates Inactive Adaptive Immune Responses in the Resistant Line

To determine pathways that were perturbed during the infection, data were analyzed by GOSeq, which accounts for gene length bias unique to RNA-seq data [Young et al., 2010]. Significantly perturbed pathways (FDR < 0.1) from both lines that are involved in immune response include the TLR signaling pathway, cytokine-cytokine receptor interaction, intestinal immune network for IgA produc-

### Table 20. Cytokine-related gene expression in response to MDV infection in lines 6 (MD resistant) and 7 (MD susceptible) chicks

<table>
<thead>
<tr>
<th>Gene</th>
<th>Change in expression*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
<td>name</td>
</tr>
<tr>
<td></td>
<td>line 6</td>
</tr>
<tr>
<td>IL2RG</td>
<td>interleukin 2 receptor, γ</td>
</tr>
<tr>
<td>IL6</td>
<td>interleukin 6 (interferon, β2)</td>
</tr>
<tr>
<td>IL6ST</td>
<td>interleukin signal transducer (gp130, oncostatin M receptor)</td>
</tr>
<tr>
<td>IL18</td>
<td>interleukin 18-like 1</td>
</tr>
<tr>
<td>IL19</td>
<td>interleukin 18 (interferon-γ-inducing factor)</td>
</tr>
<tr>
<td>IL15</td>
<td>interleukin 15</td>
</tr>
<tr>
<td>IL18R1</td>
<td>interleukin 18 receptor 1</td>
</tr>
<tr>
<td>IFNG</td>
<td>interferon-γ</td>
</tr>
<tr>
<td>IFNB</td>
<td>interferon-β</td>
</tr>
<tr>
<td>IFNA3</td>
<td>interferon-α3</td>
</tr>
<tr>
<td>IFNGR1</td>
<td>interferon-γ receptor 1</td>
</tr>
<tr>
<td>IFNGR2</td>
<td>interferon-γ receptor 2</td>
</tr>
<tr>
<td>IFNAR1</td>
<td>interferon-α,β receptor 1</td>
</tr>
<tr>
<td>IFNAR2</td>
<td>interferon-α,β receptor 2</td>
</tr>
</tbody>
</table>

* The log2 expression fold-change in response to MDV infection.
tion, and cell adhesion molecules (fig. 19). Some other pathways important in response to viral infection and only significantly enriched in the susceptible line include phagosome, apoptosis, RIG-I-like receptor signaling pathway, NOD-like receptor signaling pathway, and lysosome. Although MHC class I (BF1) was differentially expressed in both lines, other genes involved in expressing newly synthesized MHC class I were only upregulated in line 7 (MD susceptible) suggesting that new MHC I molecules were actively produced. Furthermore, gene ontology analysis of biological processes (GO:BP) shows that categories involved in both adaptive and innate immune responses were enriched in the susceptible line (data not shown). On the other hand, only categories involved in innate immune responses were enriched in line 6 (MD resistant). In addition, enrichment of the apoptosis pathway in the susceptible line suggests that the programmed cell death could be induced by the cytotoxic T lymphocyte response to eliminate ongoing viral infection.

At this stage of infection, our results suggest that lytic infection of MDV stimulates both innate and adaptive immune responses, which leads to activation of T cells in the susceptible line. Only activated T cells are believed to be infected by MDV; therefore, the lytic phase could facilitate the spread of the virus by enhancing expansion of activated T cells. Due to the cell-associated nature of MDV, the viruses transfer to T cells via cell-to-cell contact between B cells and T cells during antigen presentation, or B cell activation by T helper cells. Therefore, it is beneficial for the host to restrain such contact. However, it is not clear how chickens in the resistant line control the lytic infection of MDV. Two mechanisms have been speculated to contribute to MD resistance. First, innate immune responses could be highly effective and could activate strong adaptive immune responses that rapidly control viral replication and force the viruses to enter into the latent phase. Second, the innate immune response itself could be highly effective in limiting viral replication [Smith et al., 2011].

Fig. 19. KEGG pathways enriched in response to MDV infection. Splenic RNAs from both line 6 (MD resistant, top group) and line 7 (MD susceptible, bottom group) chicks were analyzed for differential gene expression in response to MDV infection with the resulting pathways (y-axis) and corresponding p values (x-axis) shown.
Genes with Differential Exon Usage in Response to MDV Infection Can Be Divided into 4 Groups Based on Their Patterns of Expression

The immune system is isoform-rich and many genes express different isoforms with distinctive functions in response to stimuli such as stress, chemicals and infection [Stamm, 2002; Hagiwara, 2005]. Changes in expression of splice forms of immune-related genes have been reported to be associated with increased susceptibility to and poor prognosis of diseases [Lynch, 2004]. Studying differential isoform expression could therefore shed light into inherent differences between lines that confer resistance or susceptibility to MD.

In the past, microarray technology has been used to study gene and isoform expression in many studies, but its sensitivity for detecting structurally similar isoforms is low, and known or predicted annotations are required to design probes [Kane et al., 2000]. Although the RNA-seq method can provide a reliable estimate of exon expression compared to microarrays [Pan et al., 2008] and is not constrained to the same limitations, studying isoform expressions using RNA-seq is still not straightforward because the read lengths are not long enough to span across all exons in an isoform. In most cases, only exons in close proximity are covered by the same read, which makes it difficult to accurately predict a full structure of the isoform. In addition, some genes are fused due to overlapping untranslated regions, which can also result in erroneous predicted isoform structures.

Due to those issues, it is not feasible to accurately estimate expression of isoforms, especially when gene annotation is constructed from de novo assembly [Trapnell et al., 2013]. To avoid these issues, we chose to study exon expression instead of isoform expression. Using MISO [Katz et al., 2010] with the exon-centric method, only reads spanning across exons and used in splicing event are examined. The expression of exon inclusion is calculated as ‘percent spliced in’ (Psi or ψ), which can be used to infer the portion of transcripts that include the exon in each sample [Katz et al., 2010]. In this study, we investigated the 3 most common alternative splicing events in vertebrates, which are skipped exons, an alternative 3′ and 5′ splice site.

Lists of differential exon usage genes from the resistant line that show differences in ψ >0.20 when compared to the susceptible line in infected chickens are shown in tables 21–24. Genes can be categorized roughly into 4 groups based on the pattern of ψ across control and infected birds in both lines.

Group I (table 21) includes genes with ψ values that were up- or downregulated in infected chickens in line 6 (MD resistant) only. This group includes *BCL11B* (B-cell CLL/lymphoma 11B zinc finger proteins), a B-cell lymphoma associated C2H2-type zinc finger protein encoding a tumor-suppressor for T-cell lymphoma in humans. The skipped exon was expressed 30% in MDV-infected line 6 chickens; whereas it was rarely expressed (4–7%) in the control line 6 birds and both groups in line 7 (MD susceptible). The skipped exon was not found to encode any known protein domain; however, it is in the middle of 2 adjacent C2H2-type finger protein domains.

<table>
<thead>
<tr>
<th>Type</th>
<th>Ensembl ID</th>
<th>Symbol</th>
<th>Line 6</th>
<th>Line 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE</td>
<td>ENSGALG00000011127</td>
<td><em>BCL11B</em></td>
<td>0.07</td>
<td>0.30</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG00000013137</td>
<td>INO80C</td>
<td>0.15</td>
<td>0.35</td>
</tr>
<tr>
<td>A5SS</td>
<td>ENSGALG000000009824</td>
<td><em>C7H20RF77</em></td>
<td>0.49</td>
<td>0.26</td>
</tr>
<tr>
<td>A5SS</td>
<td>ENSGALG00000013821</td>
<td><em>GEMIN6</em></td>
<td>0.84</td>
<td>0.61</td>
</tr>
<tr>
<td>A5SS</td>
<td>ENSGALG00000002144</td>
<td><em>THRAP3</em></td>
<td>0.31</td>
<td>0.51</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG00000027665</td>
<td><em>SYNGR1</em></td>
<td>0.46</td>
<td>0.23</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG000000005685</td>
<td><em>KSR1</em></td>
<td>0.77</td>
<td>0.44</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSG00000163875</td>
<td><em>MEAF6</em></td>
<td>0.28</td>
<td>0.57</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG00000020987</td>
<td><em>ZDHHC7</em></td>
<td>0.42</td>
<td>0.23</td>
</tr>
</tbody>
</table>

SE = Skipped exon; A5SS = alternative 5′ splice site; A3SS = alternative 3′ splice site.

*Human homolog.*
**Table 22.** Group II: Genes showing similar differential exon usage (ψ) values within line between uninfected and MDV-infected birds but different ψ values between lines 6 (MD resistant) and 7 (MD susceptible)

<table>
<thead>
<tr>
<th>Type</th>
<th>Ensembl ID</th>
<th>Symbol</th>
<th>Line 6</th>
<th>Line 7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>control</td>
<td>infected</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG0000000005522</td>
<td>DYNLL2</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000004971</td>
<td>URM1</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000015709</td>
<td>TACC3</td>
<td>0.87</td>
<td>0.93</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000014642</td>
<td>LOC374195</td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000011682</td>
<td>CNOT4</td>
<td>0.57</td>
<td>0.62</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000007511</td>
<td>ITGB2</td>
<td>0.17</td>
<td>0.22</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000006522</td>
<td>HCK</td>
<td>0.47</td>
<td>0.59</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG00000000904</td>
<td>C11H16ORF57</td>
<td>0.92</td>
<td>0.98</td>
</tr>
<tr>
<td>A5SS</td>
<td>ENSGALG000000010836</td>
<td>AHR</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>A5SS</td>
<td>ENSGALG000000011488</td>
<td>CMTM7</td>
<td>0.55</td>
<td>0.67</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG000000008939</td>
<td>FUBP1</td>
<td>0.42</td>
<td>0.26</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG000000008507</td>
<td>THOC2</td>
<td>0.52</td>
<td>0.53</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG000000002859</td>
<td>RAC3</td>
<td>0.69</td>
<td>0.84</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG000000012050</td>
<td>TNRC6B</td>
<td>0.57</td>
<td>0.39</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG000000010410</td>
<td>PEN2</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG000000027908</td>
<td>LOCA22528</td>
<td>0.28</td>
<td>0.39</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG000000011476</td>
<td>SEPT11</td>
<td>0.78</td>
<td>0.86</td>
</tr>
</tbody>
</table>

SE = Skipped exon; A5SS = alternative 5′ splice site; A3SS = alternative 3′ splice site.

**Table 23.** Group III: Genes showing differential exon usage (ψ) values only in line 7 (MD susceptible) between uninfected and MDV-infected birds

<table>
<thead>
<tr>
<th>Type</th>
<th>Ensembl ID</th>
<th>Symbol</th>
<th>Line 6</th>
<th>Line 7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>control</td>
<td>infected</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000003861</td>
<td>HERC4</td>
<td>0.31</td>
<td>0.37</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000009029</td>
<td>TSPAN12</td>
<td>0.08</td>
<td>0.15</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000009520</td>
<td>MARCH1</td>
<td>0.42</td>
<td>0.54</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000008320</td>
<td>EDEM1</td>
<td>0.95</td>
<td>0.99</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG00000000533</td>
<td>SRSF3</td>
<td>0.36</td>
<td>0.38</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000023199</td>
<td>HNRNPD1</td>
<td>0.39</td>
<td>0.40</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000006157</td>
<td>DDX26B</td>
<td>0.67</td>
<td>0.60</td>
</tr>
<tr>
<td>SE</td>
<td>ENSGALG000000001745</td>
<td>PSTPIP2</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>SE</td>
<td>ENSG00000175029a</td>
<td>CTBP2</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td>A5SS</td>
<td>ENSGALG000000008038</td>
<td>SF3B1</td>
<td>0.41</td>
<td>0.57</td>
</tr>
<tr>
<td>A5SS</td>
<td>ENSGALG000000002487</td>
<td>SFSWAP</td>
<td>0.58</td>
<td>0.73</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG000000005162</td>
<td>RNPC3</td>
<td>0.55</td>
<td>0.36</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG000000000720</td>
<td>LOC419563</td>
<td>0.88</td>
<td>0.94</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG000000009421</td>
<td>SRSF5</td>
<td>0.55</td>
<td>0.72</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG00000014915</td>
<td>THOC1</td>
<td>0.33</td>
<td>0.48</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG00000000189</td>
<td>YTHDC2</td>
<td>0.44</td>
<td>0.59</td>
</tr>
</tbody>
</table>

SE = Skipped exon; A5SS = alternative 5′ splice site; A3SS = alternative 3′ splice site.

a Human homolog.
In group II (table 22), ψ values were relatively stable in control and infected chickens within line, but not between lines. Genes that could play an important role in immune responses are RAC3, ITGB2 and HCK. RAC3 (Ras-related C3 botulinum toxin substrate 3) encodes small GTPases, belonging to the Ras family, that regulate a wide variety of cellular events including cell growth, cytoskeletal reorganization, and the activation of protein kinases. HCK transmits signals from cell surface receptors such as FCGR1A, FCGR2A, IL2, IL6, IL18, and integrins (ITGB1, ITGB2). ITGB2 (CD18) encodes subunit β2 integrin of LFA-1 and CR3 receptors. LFA-1 plays an important role in adhesion of lymphocytes with other cells. CR3 binds to a vast array of ligands and molecules including complement C3bi, microbial proteins, ICAM-1 and -2, ECM proteins, and coagulation proteins. It plays a significant role in neutrophil and monocyte activation including phagocytosis, adhesion and migration [Ehlers, 2000]. DYNLL2, SEPT11 and PFN2 are also involved in cell rearrangement and cytokinesis. In particular, DYNLL2 are dynein proteins that have been demonstrated to regulate T cell activation by driving T-cell receptor microclusters toward the center of an immune synapse [Hashimoto-Tane et al., 2011].

Group III (table 23) includes genes that exhibit differential isoform expression only in response to infection in the susceptible line. A number of genes in this group encode proteins that are parts of the spliceosome: SRSF3, HNRNPD1, SF3SWAP, THOC1, RNP3, and SRSF5.

The last group (group IV, table 24) only has 3 genes: GOSR2, SRSF6 and ENSGALG00000026498. The ψ value differences of these genes were greater than 0.20 the cutoff between control and infected chickens in the resistant and susceptible line and were significantly different between infected chickens in the resistant and susceptible lines. GOSR2 encodes a trafficking membrane protein important for transporting proteins from cis- to trans-golgi network and SRSF6 (serine/arginine-rich splicing factor 6) encodes a protein involved in mRNA splicing.

Roles of LFA-1 and Actin Cytoskeleton in T Cell Activation

By grouping genes based on patterns of ψ values, we found that many genes in group II (ITGB2, PFN2, DYNLL2, SEPT11, and RAC3) are involved in cytokinesis or cell synapse, which are important for T cell activation. As described above, ITGB2 encodes the β-subunit of integrins including LFA-1, which is exclusively expressed in lymphocytes and plays a major role in lymphoproliferation, antigen presentation, T cell activation, and cytotoxicity. Integrins are a special kind of receptors that transmit signals bidirectionally across the cell membrane. They are heterodimeric composed of α (large) and β (small) subunits [Wang H et al., 2010]. The β (CD18) subunit encoded by ITGB2 is expressed on lymphocytes and antigen-presenting cells (APCs) as a component of LFA-1 and CR3 receptors. LFA-1 binds to its ligand ICAM-1 to help form a synapse that brings APCs and T cells together to initiate antigen presentation leading to T cell activation [Dustin and Cooper, 2000].

Absence of LFA-1 leads to impaired functions of lymphocytes in proliferation and tumor rejection [Schmits et al., 1996; Scharffetter-Kochanek et al., 1998]. Mutations in the ITGB2 gene have been associated with type 1 leukocyte adhesion deficiency (LAD-1), a human autosomal-recessive inherited disease. The disease is characterized by impairment of lymphocytes in adherent-dependent functions, lack of accumulation to the site of infection and recurrent bacterial and fungal infection [Springer et al., 1987]. In addition, the response of lymphocytes to mitogens is decreased in patients with LAD. The decrease in

<table>
<thead>
<tr>
<th>Type</th>
<th>Ensembl ID</th>
<th>Symbol</th>
<th>Line 6 control</th>
<th>Line 6 infected</th>
<th>Line 7 control</th>
<th>Line 7 infected</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE</td>
<td>ENSGALG00000001107</td>
<td>GOSR2</td>
<td>0.73</td>
<td>0.92</td>
<td>0.38</td>
<td>0.59</td>
</tr>
<tr>
<td>SE</td>
<td>ENSG00000124193a</td>
<td>SRSF6</td>
<td>0.43</td>
<td>0.71</td>
<td>0.54</td>
<td>0.34</td>
</tr>
<tr>
<td>A3SS</td>
<td>ENSGALG00000026498</td>
<td>unknown</td>
<td>0.12</td>
<td>0.70</td>
<td>0.10</td>
<td>0.34</td>
</tr>
</tbody>
</table>

SE = Skipped exon; A3SS = alternative 3′ splice site.

a Human homolog.
responsiveness to mitogens has been shown to correlate with resistance to MD by Lee and Bacon [1983], who illustrated that resistant birds (MD-resistant lines 6 and N) were less responsive to phytohemagglutinin than MD-susceptible birds (line 7 and P).

The actin cytoskeleton is very important in T cell activation because it enhances the activity of LFA-1 by increasing its avidity and recruiting signaling molecules necessary for downstream signaling [Dustin and Cooper, 2000; van Kooyk and Figdor, 2000]. Cytoskeleton proteins binding to the cytoplasmic domain of LFA-1 are thought to play an important role in driving LFA-1 to aggregate on the cell surface, resulting in increased avidity. Aggregation of LFA-1 has been demonstrated to be essential for lymphocytes to bind to the ligand [van Kooyk et al., 1994]. Interestingly, RAC3 and PFN2, which are involved in the actin cytoskeleton pathway, also expressed different ratios of alternative splice forms between lines. Some of these gene products are also found in other pathways that are involved in immune responses. It could be speculated that pre-mRNA splicing of these genes is co-regulated by splicing regulators or some genetic factors.

Prediction of Functional Domains of Splice Forms of Genes in the Actin Cytoskeleton Pathway

To predict the function of the alternative splice forms of genes in the actin cytoskeleton pathway, transcript sequences were translated to protein sequences by ESTscan [Iseli et al., 1999]. Protein sequences were then searched for annotated protein domains using InterPro Scan [Quevillon et al., 2005]. Besides ITGB2, other genes have alternative exons located in coding regions that could potentially affect functional protein domains in some ways. The exon with an alternative 3’ splice site of RAC3 encodes part of a protein domain identified as a small GTPase of the Ras subfamily (ProSiteProfiles:PS5142 and SMART:SM00173). Rac3 is highly homologous to Rac1 and has been reported to possess the ability to promote membrane ruffling, transformation, activation of c-Jun transcriptional activity, and co-activation of NFκB [Werbajh et al., 2000]. Activated Rac also regulates production of superoxide in neutrophils and macrophages.

The alternative exon of PFN2 seems to disrupt the coding sequence that encodes the profilin domain (Pfam: PF00235). The profilin domain is essential for almost all organisms, and its functions include regulating actin polymerization, controlling complex networks of molecular interaction and transmitting signals from small-GTPase pathways. It also binds to Rac effector molecules and a number of other ligands [Witke, 2004]. Even though the exact mechanism is not known, lack of responsiveness of T cells to stimuli appears to benefit resistant birds because in these birds, MDV cannot induce T cells to proliferate and cause them to undergo neoplastic transformation. It has also been suggested that the mechanism that controls both lymphocyte proliferation induced by MDV and lymphocyte proliferation induced by the immune response is the same [Pazderka et al., 1975]. Therefore, it may be useful to consider a link between the deficiency of lymphocytes in the resistant line and the alternative splice form of ITGB2 that is only expressed in the resistant line. Although the exon included in the alternative splice form is noncoding, it could serve important functions in translation or posttranscriptional regulation.

Prediction of cis-Regulatory Elements in Alternative Splicing Exons of Genes in Group II

Among all groups, alternative splicing of genes in the group II is most likely to be regulated by genetic factors because the ratios of isoform expression in this group were relatively stable within line, but were significantly different between lines. Investigation of nucleotide differences within exons of both lines could reveal a possible role of SNPs in regulating alternative splicing in this group. We obtained a sequence of alternative exons from the resistant line and used Human Splicing Finder to determine whether SNPs from the susceptible line could alter predicted ESEs or ESSs. Results from some genes involved in cytokinesis are discussed below.

For ITGB2, a SNP at position 26 of the cDNA from the resistant line located at position 7,183,696 on chromosome 7 is in a predicted binding site for SC35, which is an exon enhancer. Although the exon is not expressed in the susceptible line, we found that there are no polymorphisms between lines based on genome resequencing. Therefore, this SNP may not account for exclusion of the exon in the susceptible line (fig. 20). Exon sequences of PFN2 from the lines 6 and 7 differ at position 23,221,934 on chromosome 9. A small insertion of 2 AA nucleotides is predicted to create a new binding site for Tra2-β splicing regulator (table 25), which serves as a stabilizer of an enhancer complex [Lopez, 1998]. From exon expression data, ψ of an exon with alternative splicing increases from 0.20–0.30 to about 0.50 in the susceptible line. Tra2 could possibly increase inclusion of the exon with alternative 3’ splice site via ESE-dependent 3’ splice site activation.

Replacement of an A with a G nucleotide in the skipped exon of DYNLL2 from line 6 is predicted to slightly alter the binding site of several ESEs as well as to create a new
binding site for 9G8 (table 25). This exon is upregulated in the susceptible line compared to the resistant line, therefore, the presence of the new binding site for 9G8 exon enhancer helps support the expression results. In addition, the G nucleotide in this position matches the reference nucleotide; therefore, we could expect this exon to be expressed in other datasets. According to EST tags on the UCSC genome browser, the exon has been found and sequenced from chicken eyes (EST sequence: DR424100).

Conclusions
Custom gene models built by combining gene models from de novo assembly, reference-based assembly, and Ensembl have allowed us to identify genes and isoforms that might play an important role in resistance to MD. Results from gene expression analysis indicated that adaptive immune responses were more highly activated during lytic infection in the susceptible line than in the resistant line. Because only activated T cells are thought to be infected by MDV, we speculate that enhancement of adaptive immune responses could help spread the viruses by recruiting and activating more T cells. In contrast, the delay or reduction of adaptive immune responses could benefit the host by limiting infection of activated T cells.

To elucidate the molecular mechanism of MD resistance, we investigated differential isoform expression between lines and identified a number of genes that could be responsible for differences in immune responses. Notably, this includes several genes involved in actin cyto-

Fig. 20. An SNP in an ITGB2 exon can explain alternate splicing. A single-nucleotide substitution (T>C) at position 7,183,696 of an ITGB2 exon is predicted to disrupt a putative ESE motif. The numbers not on the axes in the figure on the left side reflect the actual counts for each exon splice junction read for each sample, and the ψ values given on the right side are the percent of variants containing the potentially skipped exon.
skeleton structure and cytokinesis, which are important for the functions of lymphocytes and immune cells but have not been of great interest in the field of MD research. Even though we mainly discuss the possible role of ITGB2 in MD resistance, other genes cannot be precluded and should be candidates for further investigation and experimental studies. Moreover, the full mechanism of MD resistance is highly complex and more data from different stages of infection as well as greater sequencing depth will be required to identify all genes and isoforms involved.

To enhance the study of unannotated gene and isoform expression, our approach of constructing gene models from RNA-seq should be iteratively used to extend the Ensembl data to construct more complete gene models.

### Materials and Methods

Detailed computational analysis and methods are available at https://github.com/likit/mdv_rnaseq_paper.

### The National Avian Research Facility

(Prepared by A. Hart, R. Kuo, L. Eöry, P. Kaiser, and D.W. Burt)

The National Avian Research Facility (NARF) was created as a partnership between The Roslin Institute and The Pirbright Institute to provide a wide range of resources for the avian research community including access to avian lines, biological tools and genomic databases. We have received capital support from The Roslin Foundation, BBSRC, University of Edinburgh and The Roslin Institute as well as funding from the Wellcome Trust.

NARF is currently comprised of avian rearing facilities and a large group of collaborators contributing to the services that we aim to provide for the community. There are 2 avian units, The Greenwood Building, a conventional avian facility, and The Bumstead Building, a National Avian Specific Pathogen Free (SPF) Research Facility. The purpose of the Bumstead Building is to maintain unique inbred genetic lines of poultry currently housed at The Pirbright Institute Compton Laboratories. Both facilities will support research programmes currently running at both institutes, and NARF will also offer UK and international investigators access to these unique bioresources.

The Greenwood Building is named after Dr Alan Greenwood. Alan Greenwood moved to Edinburgh in the mid 1920’s from Australia to study with Dr Francis Albert Crew at the Institute of Animal Genetics, The University of Edinburgh. In the late 1920’s, Alan Greenwood started an inbreeding programme with Brown Leghorn chickens to study a variety of traits, egg laying, plumage, vigour, etc. Through this programme he generated 9 lines denoted A through to I. During 1947 Dr Greenwood founded and was Director of the Edinburgh Poultry Research Centre (PRC); he continued in this role until his retirement in the 1960’s. His inbred chicken

### Table 25. Selected group II genes (see table 22) showing differential exon usage ($\psi$) between lines 6 and 7 with potential sequence variants in motifs that may account for the alternative splicing

<table>
<thead>
<tr>
<th>Gene</th>
<th>cDNA position</th>
<th>Linked SR protein</th>
<th>Type</th>
<th>Reference motif</th>
<th>Mutant motif</th>
<th>Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DYNLL2</td>
<td>2</td>
<td>SF2/ASF (IgM-BRCA1)</td>
<td>ESE^a</td>
<td>CTCCGGG (86.38)</td>
<td>CTCCGAG (72.69)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>SF2/ASF, SF2/ASF (IgM-BRCA1)</td>
<td>ESE^a</td>
<td>CTCCGGG (76.91)</td>
<td>CTCCGAG (72.69)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>SF2/ASF (IgM-BRCA1)</td>
<td>ESE^a</td>
<td>CGGGGGT (73.00)</td>
<td>CGGAGGT (86.23)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>SF2/ASF, SF2/ASF (IgM-BRCA1)</td>
<td>ESE^a</td>
<td>CGGGGGT (73.00)</td>
<td>CGGAGGT (82.94)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>9G8</td>
<td>ESE^b</td>
<td>GAGGFG (60.67)</td>
<td>GAGGTF (74.05)</td>
<td>new site</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>hnRNP A1</td>
<td>ESE^d</td>
<td>new site</td>
<td>new site</td>
<td></td>
</tr>
<tr>
<td>PFN2</td>
<td>2068</td>
<td>Tra2-β</td>
<td>ESE^a</td>
<td>AAAAT (81.02)</td>
<td>AAAAa</td>
<td>+16.19%</td>
</tr>
<tr>
<td></td>
<td>2069</td>
<td>Tra2-β</td>
<td>ESE^a</td>
<td>AAAAaa (94.14)</td>
<td>AAAAaa (94.14)</td>
<td>new site</td>
</tr>
<tr>
<td></td>
<td>2070</td>
<td>Tra2-β</td>
<td>ESE^a</td>
<td>AAAaT (81.02)</td>
<td>AAAaT (81.02)</td>
<td>new site</td>
</tr>
<tr>
<td></td>
<td>2066</td>
<td>ESS</td>
<td>ESE^c</td>
<td>ACAAAAa (38.13)</td>
<td>CAAAAAAa (38.13)</td>
<td>new site</td>
</tr>
<tr>
<td></td>
<td>2067</td>
<td>ESS</td>
<td>ESS^d</td>
<td>CAAAAAaT (28.85)</td>
<td>CAAAAAaT (28.85)</td>
<td>new site</td>
</tr>
</tbody>
</table>

^b ESE motifs from Human Splicing Finder.
^c Predicted PESS octamers from Zhang and Chasin [2004].
^d hnRNP motif.
lines were intercrossed at this time to generate what is known as the J-line where this line continues to be studied at NARF.

The Bumstead Building is named after Dr Nat Bumstead, former Head of Avian Genetics Group, BBSRC Institute for Animal Health. Nat Bumstead was a leading avian geneticist who published extensively on the genetics of immunity and disease resistance and susceptibility using chicken as the model organism. He was instrumental in establishing worldwide research on the genome of the chicken. The inbred chicken lines that Dr Bumstead studied throughout his research career continue to be studied today and will move to NARF from late 2014. These birds will be kept in SPF conditions, and NARF will supply these to researchers to gain greater insights into gene function in relation to disease susceptibility and resistance.

Considerable investment by BBSRC and other governmental departments over several years has supported several of these chicken lines, with some tracing their origins to the 1930’s. All of the lines, including those that are currently housed at The Roslin Institute and The Pirbright Institute offer a unique resource backed with a wealth of data. From late 2014, NARF will become fully operational taking over the role as curator for these unique chicken lines.

NARF has also been working on creating an avian biobank. The advantages of developing an avian biobank are many:

- Cryopreserve rare or endangered breeds of chicken to maintain genetic diversity for future generations;
- In the event of disease outbreak, flocks could be rederived from frozen material;
- The cost of maintaining flocks increases year-on-year, and therefore, if we were able to cryopreserve lines using primordial germ cells and/or semen, then this could free up space in avian facilities for those lines that are being actively researched.

Robust procedures that would allow us to freeze, maintain and rederive lines in the future would provide confidence that unique genetic resources will not be lost if the live flock is no longer ‘on the shelf’.

In addition to providing various animal resources, NARF collaborators are developing biological tools such as antibodies, primers and reagents. Information regarding available chicken lines and biological tools are on the NARF website (www.narf.ac.uk). The NARF website also provides links to relevant resources as well as our own contributions to the annotation of avian genomes.

At present we are developing several informatics-based resources including a variation database, alternative splice variant annotation, and avian comparative genomics annotations.

The variation database will contain single nucleotide polymorphisms (SNPs) and short insertions/deletions (InDels) detected in the inbred lines at NARF. Where appropriate, specific SNPs and InDels will be linked to relevant annotations and biological tools to provide a more information-rich experience.

The alternative splice variant annotation will contain full-length transcript sequences obtained using long-read sequencing technology. This information will greatly expand our current knowledge of isoforms occurring in chicken as well as expanding the list of putative noncoding RNA. Using this annotation, researchers will be able to run RNA-seq analyses on the isoform level.

To facilitate genetics and genomics research, NARF has been making considerable efforts to improve and share the genome annotations currently available for poultry genomes. Having good quality, functional annotations are crucial for the understanding of the link between genetic variations and phenotypic differences between individuals. Although annotations from large-scale functional assays are at present only available for mammals (see e.g. the human ENCODE project [ENCODE Project Consortium, 2012]), annotations can still be improved by incorporating information on selective forces from comparative genome analyses of recently sequenced bird genomes [Zhang G et al., 2014b]. For example, regions in the chicken genome that are known to be under purifying selection can serve as an extra layer of annotation as these regions can be predictive for function [Lindblad-Toh et al., 2011].

While the primary focus at NARF is on poultry genomics, integrating and sharing genome sequences and annotations for bird species in general can also benefit the diverse bird research communities. Therefore, we provided access to our integrated genome resource through a web portal called Avianbase (http://avianbase.narf.ac.uk) [Eöry et al., 2015], an Ensembl-based [Flicek et al., 2014] genome browser instance, which provides visual displays of the genome sequences, gene models, transcripts, and translations along with the BLAST search facility for 44 newly sequenced bird genomes [Zhang G et al., 2014b].

The goal of NARF is to become a hub for the avian research community. We are working to provide new resources to help researchers find the information and tools they need to progress.
References


Clynes D, Gibbons RJ: ATRX and the replication
Chue J, Smith CA: Sex determination and sexual
Christidis L: Aves, in John B, Kayano H, Levan A
Coble DJ, Fleming D, Persia ME, Ashwell CM,
Clurman BE, Hayward WS: Multiple proto-onco-
Clayton DF, Balakrishnan CN, London SE: Inte-
Cole RK: Hereditary hypothyroidism in the do-
Craven AJ, Ormandy CJ, Robertson FG, Wilkins
Cracraft J, Barker FK, Braun M, Harshman J,
and the evolution and biogeography of birds.
2001).
2533–2539 (2013).
3 B (Gebrüder Borntraeger, Berlin 1990).
(eds): Animal Cytogenetics, vol. 4: Chordata
15: 539 (2014).
RNA-seq analysis of broiler liver transcriptome reveals novel re-
2008).
acr gene encodes a
man metaphase chromosomes. Exp Cell Res
1973).
and the Cretaceous-Tertiary mass extinc-
Avian evolution, Gondwana biogeo-
and the Cretaceous-Tertiary mass extinc-
Continental drift, paleoclimatology, and the evolution and biogeography of birds.
Prolactin gene-expression and next-generation se-
sequences to unravel a QTL affecting meat quality in chicken. Adv Anim Biosci 3:184–
Dengiegwicz RL, Parcells MS: Establishment of a lymphoblastoid cell line using a mutant MDV containing a green fluorescent protein expres-
Dorsey T, Wakeyo O: Uses and flock manage-
350 (2010).
350 (2010).
1084 (2014).
1089 (1990).
350 (2010).

Cytogenet Genome Res 2015;145:78–179
DOI: 10.1159/000430927

Third Report on Chicken Genes and Chromosomes 2015


Meggens HJ, Groenen D: Mammated species form a treasure-trove for molecular characteristics of Mendelian traits by exploiting the specific genetic structure of these species in across-breed genome wide association studies. Heredity 109:1–3 (2012).


Robb EA, Antin PB, Delany ME: Defining the sequence elements and candidate genes for the coloboma mutation. Plos One 8:e60267 (2013).


Scrivener D: Rare Poultry Breeds (The Crowood Press, Ramsbury 2009).


van Tuinen M: Birds (Aves), in Hedges SB, Kuo-


V"olker M, Backström N, Skinner BM, Langley EJ, Schmid et al.


Wang DS, Kobayashi T, Zhou LY, Paul-Prasanth B, Ijiri S, et al: FoxL2 up-regulates aromatase gene transcription in a female-specific manner by binding to the promoter as well as interacting with Ad4 binding protein/steroi-

Wang GS, Cooper TA: Splicing in disease: disruption of the splicing code and the decoding ma-


Wang X, Nahashon S, Feaster T, Bohannon-Stew-

ers, Ellegren H: Genomic footprints reveal traveling chromosome architecture. Nat Bio-

Wang GS, Cooper TA: Splicing in disease: disruption of the splicing code and the decoding ma-


Wisevan J: A History of the British Pig (Duck-


Witter RL: A new strategy for Marek’s disease immu-


Wolffe A, Goodson M, Goode DK, Snell P, Mc-
Ewen GK, et al: Highland conserved non-coding sequences are associated with vertebrate de-

Wragg D, Mwcharo JM, Alcalle JA, Hocking PM, Hanotte O: Analysis of genome-wide structure, diversity and fine mapping of Mendelian traits in traditional and village chick-


Wright D, Boije H, Meadows IRS, Bedhom B, Gour-

rich D, et al: Copy number variation in in-


Yamamoto KN, Kobayashi S, Tsuda M, Kurumizaka H, Takata M, et al: Involvement of SLX4 in interstrand cross-link repair is regu-

Yang Y, Yi G, Sun C, Qu L, Yang N: Genome-wide characterization of insertion and deletion variation in chicken using next generation se-

Yao Y, Charlesworth J, Nair V, Watson M: Mi-
croRNA expression profiles in avian haemo-


