At the time the turkey genome sequencing project was conceived, draft genome sequences existed for swine, bovine, sheep, and chicken, but not for this fourth most economically significant food animal species. The turkey continues to be an important international agricultural commodity with the United States producing roughly one-half of the world’s commercial turkeys at an estimated value approaching $5.5 billion US (USDA National Agricultural Statistics Service, 2012). Thousands of pounds of turkey meat are exported annually, contributing significantly to the national export revenue. Within the United States, annual production (almost 300 million birds) has roughly doubled in the last 3 decades with current per capita annual consumption at more than 7 kg (National Turkey Federation, 2012). With the ever increasing popularity of turkey meat, the turkey has become the fourth major food and protein source, both in the United States and overseas.

This dramatic increase in meat production has resulted primarily from intense genetic selection for increased growth rate, meat yield, and growth efficiency. Although the industry may justifiably take pride in increased production, intense selection has also come with economically unfavorable consequences. Increased incidence of growth induced myopathies, meat quality defects, cardiac morbidity, and skeletal deformities are most prevalent (reviewed by Reed et al., 2005). In addition, turkeys selected for production traits may exhibit reduced immune response to pathogens (Nestor et al., 2012).
Consortium—A Collaborative Approach

et al., 2011). Solutions to problems associated with intense selection for heavy-muscled birds have been difficult to achieve because of our limited understanding of the complex underlying genetic factors. Providing a completed turkey genome sequence would undoubtedly be beneficial to poultry breeders and producers in terms of finding solutions to disease resistance, nutrient utilization, and reproductive success. Breeders have historically increased production efficiency and product quality through traditional selection. They now consider using genomic information for selection as the next big step for the breeder industry. Genomic selection would benefit those difficult to select traits such as health, reproduction, and feed efficiency (Reed and Dalloul, 2011).

Sequencing of eukaryotic genomes has migrated over the past several years, from a minimal tiling path approach of bacterial artificial chromosome (BAC) sequencing, to whole genome shotgun (WGS) sequencing with plasmid, fosmid, and BAC libraries, to ultimately WGS sequencing of uncloned genomic DNA with next-generation sequencing (NGS) technologies. The rapid and continuing development of NGS technologies has made it feasible to contemplate sequencing the genomes of numerous species of agronomic, evolutionary, and ecological importance, as well as biomedical interest (Genome 10K Community of Scientists, 2009). Below, we review the genome sequencing progress of the domesticated turkey (Meleagris gallopavo) accomplished using primarily NGS platforms. For this project, a combination of Roche 454 and Illumina GAII sequencing was employed. The high throughput and low cost of such NGS technologies allowed sequencing the turkey genome at a fraction of the cost of other recently reported genomes of agricultural interest (e.g., bovine and equine). This turkey genome sequence represents the second domestic avian genome to be sequenced, and thus permits a genome-level comparison of the 2 most economically important poultry species (Dodgson et al., 2011).

RESOURCES EMPLOYED

Consortium—A Collaborative Approach

Preliminary activities toward sequencing the turkey genome were initiated in the summer of 2008 by researchers at Virginia Tech University. In a community effort, an initial group of poultry scientists interested in achieving the turkey genome sequence was assembled that included researchers from Michigan State University, Utah State University, the University of Minnesota, and the Roslin Institute, in addition to the “Hokie Nation” scientists. The efforts of the consortium were officially kicked off at a mini-symposium at Virginia Tech in Blacksburg in the fall of 2008. A key decision made at this meeting was the DNA source for genome sequencing (discussed below). As work progressed with preliminary Roche 454 sequencing, the consortium was joined by a group from the USDA Agricultural Research Service (Beltsville, MD) led by Julie Long, thus bringing together 2 sequencing strategies (Roche 454 and Illumina GAII). The growing consortium was joined by University of Maryland colleagues including groups of Steven Salzberg, James York, and Aleksey Zimin, experts in genome assembly and annotation. Further analyses of the assembly brought together groups of scientists on an international scale (Dalloul et al., 2010).

DNA Source for Sequencing

The task of assembling whole genome sequences is simplified if highly inbred individuals (increased homozygosity) are available for sequencing. A female bird, “Nici” (Nicholas Inbred, donated by Nicholas Turkey Breeding Farms/Aviagen Turkeys Inc., Lewisburg, WV), was selected as the candidate source DNA. This bird was from an inbred sub line (sib-mating for 9 generations) originally derived from a commercially significant breeding line. Nici was also the source DNA for 2 BAC libraries (CHORI-260 and 078-TKN-MI) used for physical mapping. Based on pedigree analysis, Nici had an increased inbreeding coefficient of 0.624 relative to the founder breeding line. As a prelude to initial genome sequencing, heterozygosity of Nici was characterized by genotyping randomly distributed microsatellites and analysis of sequenced BAC clones (Chaves et al., 2009). These studies and subsequent SNP genotyping demonstrated that Nici still retained considerable nucleotide diversity, but had an estimated homozygosity equivalent to birds from a USDA Beltsville Small White flock closed for 30 yr (Martien Groenen, Wageningen University, Wageningen, the Netherlands, personal communication). It was determined that the genetic resources already established from Nici (BAC libraries and sequences) outweighed the concern of sequence polymorphism; hence it was adopted as the DNA source for genome sequencing. Of note, NGS depth of coverage allowed for the use of this genome that was only partially inbred.

Maps

Two genetic linkage maps (Reed et al., 2007; Kerstens et al., 2011) and a BAC contig physical map (Lee et al., 2009) for turkey were used to produce a combined map for alignment of assembled sequence to chromosomes. The combined map had ~32,000 markers that mapped both to the initial assembly (UMD2.01) and to the turkey chromosomes MGA1 through MGA30. Maps for the turkey sex chromosomes (W and Z) were not used due to fragmentary marker coverage. Instead, scaffolds that aligned only to chicken W and Z chromosomes were identified and then ordered and oriented according to the chicken coordinates. The BAC contig
Table 1. Major characteristics of the turkey and chicken genome assemblies

<table>
<thead>
<tr>
<th>Item</th>
<th>UMD 2.01</th>
<th>UMD 3.0</th>
<th>UMD 4.0</th>
<th>UMD 5.0</th>
<th>Giga 4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of scaffolds &gt;1 Kb</td>
<td>26,917</td>
<td>23,958</td>
<td>13,564</td>
<td>16,147</td>
<td>16,847</td>
</tr>
<tr>
<td>Number of contigs &gt;1 Kb</td>
<td>128,271</td>
<td>162,932</td>
<td>66,632</td>
<td>64,914</td>
<td>27,027</td>
</tr>
<tr>
<td>Scaffolded sequence, Mb</td>
<td>931</td>
<td>951</td>
<td>1,009</td>
<td>1,075</td>
<td>1,047</td>
</tr>
<tr>
<td>N50 scaffold size, kb</td>
<td>1.5</td>
<td>1.5</td>
<td>2.3</td>
<td>5.7</td>
<td>12.9</td>
</tr>
<tr>
<td>N50 contig size, kb</td>
<td>12.6</td>
<td>17.1</td>
<td>34.5</td>
<td>36.3</td>
<td>279</td>
</tr>
<tr>
<td>Contig coverage</td>
<td>&gt;30 X</td>
<td>&gt;30 X</td>
<td>&gt;30 X</td>
<td>&gt;30 X</td>
<td>&gt;30 X</td>
</tr>
</tbody>
</table>

1UMD 2.01 = Dalloul et al., 2010.
2UMD 3.0 = 2 runs (1/library on 454/Titanium) from 3 and 8 kb mate pairs plus updated assembly routine.
3Computed with the original assembly size of 935 Mb for fair comparison. UMD 4.0 = two 8-kb runs (1 library − 454/Titanium) with incorporation of the turkey bacterial artificial chromosome (BAC) physical map (Zhang et al., 2011).
4UMD 5.0 = Illumina GAII sequences from pools of BAC from both the CHORI260 and TAMU libraries with end sequences with Basic Local Alignment Search Tool hits to MGAZ (sex chromosome), ChrUn, and an additional BAC pool derived from screening of the CHORI260 library for MHC-related sequences (MGA18).

Sequencing and Assembly Strategy

The genome of the turkey was the first vertebrate genome sequenced entirely using combined NGS platforms (Roche 454 and Illumina GAII). The first release of the turkey genome sequence combined data from 2 leading NGS platforms, BAC contig and genetic linkage maps (Dalloul et al., 2010). While this approach presented unique assembly challenges, the turkey sequence benefitted from the particular advantages of both platforms. In addition, this novel approach allowed us to use a BAC contig-based physical and comparative map, along with the turkey genetic map (Reed et al., 2007; Aslam et al., 2010) and the chicken genome sequence (International Chicken Genome Sequencing Consortium, 2004) to align the turkey sequence contigs and scaffolds to most of the turkey chromosomes. Such an alignment is essential for making long range evolutionary comparisons and employing the sequence to improve breeding practices using, for example, genome-based selection approaches, where chromosome locations are critical.

Sequence data from the 2 platforms (~15M 454 and ~400M GAII reads) were assembled using a modified version of the Celera Assembler 5.3 and the resulting scaffolds were ordered and oriented on turkey chromosomes. The initial sequence build (2.01) was estimated to represent ~89% genome coverage with 931 Mb of the assembled sequence. Sequence contigs were assigned to 30 of the 40 chromosomes with approximately 10% of the assembled sequence (19 Mb) corresponding to unassigned chromosomes (ChrUn). Although the initial average coverage of the turkey genome (17X) was significantly greater than that of the chicken (7X at that time), the assembly of NGS data did not produce contigs and scaffolds as large as those in assemblies based on Sanger sequencing. Costs, however, were significantly lower and the time frame much shorter.

We continue to refine the turkey genome sequence through both genome-wide and area focused sequencing, including shotgun and paired-end sequencing, and targeted sequencing of chromosomal regions with low or incomplete coverage. For example, construction of additional 3 and 8 kb libraries with sequencing on Roche/454 Titanium system improved scaffold and contig sizes and further closed many gaps. The key improvements in Assembly 3.0 were an additional 20 Mb of sequence scaffold, significantly reducing the number of independent scaffolds, which added to the total length of the chromosome sequences. This also produced larger contigs resulting in 35% improvement in the N50 contig size and higher genome coverage increasing genome redundancy to 25X and overall coverage by roughly 2%. Table 1 summarizes the major characteristics of the various turkey assemblies in comparison with the latest chicken build (4.0).

Construction and sequencing of more 3 and 8 kb Roche/454 libraries further improved the assembly. Using the original data and incorporating additional 3 and 8 kb paired-end reads, a significantly improved turkey genome assembly was produced with N50 contig size almost tripled (from 12.5 to 34.5 kb) and N50 scaffold size increasing from 1.5 to 2.3 Mb. The improvements in the assembly quality were to a significant degree due to improved assembly strategies. The assemblies beyond 2.01 were done with different builds of MSR-CA (now MaSuRCA) assembler (Zimin et al., 2013) starting with build 1.4 for 3.0 assembly. The MaSuRCA assembler error-corrects the short read Illumina data and converts the reads into longer super-reads. Many Illumina reads get replaced by the same super-read resulting in a transformed sequence data set with longer reads and lower coverage, while preserving almost all of the information contained in the original reads. The total amount of sequence in build 4.0 increased from 951 to 1,009 Mb at an overall coverage depth over 30X. These builds (3.0 and 4.0) were created to assess the value of the newly added sequences and did not undergo ex-
Tensile annotation as additional sequences were being added. As such, neither build was publically released as efforts continued to enhance coverage of the smaller microchromosomes and the sex chromosomes, detailed gene annotation, and further genomic analysis of the recently completed build (5.0). For this build (UMD 5.0), the Illumina GAII was employed for sequencing pools of BAC clones from both the CHORI260 and TAMU libraries with end sequences with BLAST hits to the Z chromosome, ChrUn, and an additional BAC pool derived from screening of the CHORI260 library for MHC-related sequences (MGA18). Build 5.0 was produced with the MSR-CA assembler version 1.8.3. Major improvement in the assembly contiguity between the 3.0 and 5.0 assemblies was due to enhanced assembly algorithm. Improvements over build 4.0 include a total sequence of 1,075 Mb and more than doubled N50 scaffold size (5.7 Mb).

**Turkey Genome Sequence Analysis and Usefulness**

For the initial sequence assembly, a synergistic combination of 2 NGS platforms with a detailed comparative BAC physical contig map was employed. This integrated approach has provided a model for both gene and chromosome level assemblies of other species with agricultural, ecological, and evolutionary interest, as evidenced by recent eukaryotic genome sequencing projects employing this approach. For example, to sequence the duck genome, Huang et al. (2013) used paired-end libraries of multiple size inserts that were sequenced on the Illumina NGS platform. The turkey genome sequence has afforded detailed comparative analysis of avian genomes particularly between chicken and turkey as detailed in the published genome (Dalloul et al., 2010). Comparative analyses focusing on chromosome evolution in Galliformes have identified 20 to 27 chromosomal inversions between turkey and chicken (Zhang et al., 2011). Annotation of build 2.01 identified almost 16,000 genes (15,093 representing distinct protein coding loci and 611 noncoding RNA genes).

The whole genome sequence also served as the anchor and coordinating platform for genome-wide SNP discovery and the future development of SNP-chip resources (Aslam et al., 2012). Within the single sequenced genome, 0.9 M single nucleotide variants (SNP and deletion/insertion polymorphisms, DIP) were identified and nucleotide diversity across the genome has been quantified. The initial sequence identified 920,126 SNV of which 601,490 were considered as having strong support. To expand the SNP resources, Aslam et al. (2012) recently resequenced DNA from a broad spectrum of turkeys (pure lines, closed inbred birds, heritage breeds, and nondomestic wild breeds). Their SNP discovery project resulted in the detection of 5.49 million putative SNP compared with the reference genome. The ultimate goal of these projects is to provide the community with robust tools and resources to facilitate immediate and long-term applications of genome-wide selection for specific traits including production, disease resistance, and reproduction parameters.

Other comparisons included lineage-specific changes in transposable elements, gene family copy number, immune system genes, noncoding RNA (ncRNA), and phylogeny. To this end, detailed analysis of the publicly available sequence (UMD 2.01) has identified some inconsistencies within the genome assembly. Thus, the 2.01 assembly should be treated as a work in progress until a more comprehensive and accurate sequence is published. Comparison of build 2.01 with fully sequenced BAC clones has demonstrated the overall breadth of the assembly, but also the difficulty in correctly assembling and assigning sequence blocks when multi-locus gene families are involved (Reed and Dalloul, 2011).

Although some important genes are still missing from the assembly (e.g., MHC class IIα) and several cases of incomplete or inaccurate annotation are known, the draft turkey genome sequence has become an important tool in the analysis of transcriptome sequences. Current RNA-Seq projects use alignments to both the genome sequence and the annotated turkey reference gene set (Kim et al., 2013). Projects investigating immunogenetics and toxicology are focusing on identifying important gene modulators of the immune and inflammatory response. For example, the turkey genome sequence has provided the framework for identifying undescribed regions of the avian MHC. Comparative alignments of human, reptilian, amphibian, and fish MHC loci have been used to identify sequences in the turkey genome that are homologous to genes of the extended MHC. In humans, the class III region includes the large, super-clusters of genes encoding histones, tRNA, butyrophilins, olfactory receptors, tripartite motif-containing proteins, and zinc finger proteins. In the turkey, examples of these class III genes are found in sequence contigs currently assigned to ChrUn and unassembled sequence reads. These sequences are being used as starting points to further expand knowledge of the avian MHC. Further, analyses of the turkey and other available avian genomes (chicken and zebra finch) revealed the presence of a second TCRδ locus in Galliformes that was not found in the Passerine avian lineage (Parra et al., 2012). This particular chain may potentially play a role in the functionality of γδTCR+ T cells, which are predominantly prevalent in poultry gut-associated lymphoid tissues.

Other projects focus on identifying key elements of the host immune response to several impactful poultry diseases. Transcriptomic profiling and analysis of disease-stricken birds (e.g., turkey cellulitis) and those infected with foodborne pathogens (e.g., *Salmonella*) are currently underway (R. A. Dalloul, unpublished data). Within this context, the turkey genome sequence has also afforded detailed comparative analysis of avian ge-
nomes particularly between chicken and turkey. Striking differences are seen in the number of innate immune system genes in birds versus mammals. In general, birds have fewer interleukin, cytokine, and chemokine loci (Dalloul et al., 2010). For example, humans possess 10 \(IL-1\) genes compared with 2 to 4 in the avian lineages. Surprisingly, however, 11 \(IL-1\) receptor genes are found in both the avian and mammalian genomes. Closer comparisons of the sequence-level differences in these gene families will provide important insight into the ongoing evolution of the innate immune response and the associated genes. Other immunity-related projects focus on identifying key elements of the host immune response to avian influenza viruses that specifically infect turkeys. It is well known that susceptibility and resistance to disease present the most challenging phenotypes to study. This approach will not only answer fundamental questions of host/pathogen interactions, but will also aid in further improving the transcriptional annotation of the genome.

**Potential Applications**

Increased throughput and decreased costs of NGS technologies facilitated cost- and time-effective sequencing of the turkey genome sequence, which represents the first eukaryotic genome completely sequenced and assembled de novo from data produced by a combination of 2 NGS platforms (Roche-454 and Illumina-GAII). This genome project was a first where most the production cost was invested in analysis and interpretation rather than generating sequence, with a resulting sequence that is comparable in genome coverage to the predominantly Sanger-based sequences of the chicken and zebra finch (International Chicken Genome Sequencing Consortium, 2004; Warren et al., 2010). The most updated sequence assigned to the chromosomes covers approximately 95% of the turkey genome, and with the high sequence quality, it has become a valuable resource for comparative genomics including identification of thousands of SNV amenable to whole genome analyses.

Not only will this NGS sequencing project greatly benefit the production industry, but the turkey genome sequence also serves as a tool to accelerate research in agricultural animal genomics. Due to the close homology between the chicken and turkey genomes, the turkey genome will be employed as a resource to fill in gaps currently missing from the chicken genome sequence, and as a comparative platform for additional turkey genome sequences. It will certainly enhance the gene annotation for the chicken genome and provide better coverage for poorly sequenced regions such as the W and micro-chromosomes.

Of particular interest to our research programs, the benefits to poultry health in general could be tremendous. Current research in poultry health focuses on immunology and progress in this area has been hindered by the lack of immunological reagents, particularly those specific for the turkey. These typically rely on unique epitopes that often lack cross-species reactivity among avian species. The process of acquiring such reagents is limited mainly by the annotation levels in the current chicken and turkey genomes in addition to missing sequence information. Our laboratories and many others use chickens and turkeys to understand bacterial and viral diseases in the context of host-pathogen interactions. Central to this work is elucidating differential resistance to diseases in birds of different genetic backgrounds. Unfortunately, much of the research progress is frequently impaired due to lack of reagents and many avian scientists invest significant resources in identifying and characterizing avian homologs of important genes and their protein products. Having a fully sequenced and highly annotated turkey genome is radically enhancing our ability to explore the avian immune system, study host-pathogen interactions, and better characterize major infectious diseases of economic consequences. It is providing a valuable resource for studying genetic variations underlying other economically important traits in poultry, effectively allowing comparative genomic analyses of avian and vertebrate species, and collectively accelerating our ability to positively affect the productivity and sustainability of the US and world poultry industry.

**ACKNOWLEDGMENTS**

Funding for this project was mainly provided by the USDA National Institute of Food and Agriculture Animal Genome Program (Washington, DC) Grants #2010-65205-20412 (Dalloul) and 2009-35205-05302 (Reed).

**REFERENCES**


