Abstract

The importance of the chicken as a model organism and critical dietary source merits a continuation of reference improvement efforts. No mechanism exists for the user community to point out problems in the existing reference, correct errors, better annotate copy number variants and add their own local sequence information. Many labs are now sequencing chicken genomes with next generation sequencing methods and/or doing sequencing of local areas of special interest (e.g., the MHC/B-complex). The immediate ideal solution would be for chicken to be added to those species supported by the Genome Reference Consortium, which currently supports human, mouse, and zebrafish genomes. Long term solutions to the missing sequence problem and other gaps will require new technology and/or manual gap-closing efforts. Manual scaffold gap closure efforts will require focused use of this data. The current long read technology holds great promise for this purpose and it is anticipated that these problems will be solved in the coming year and, at that point, this sequencing technology should be able to break a technical hurdle encountered when entering gaps within scaffold structures, complex repeats. In summary, we present a phased approach to improve the chicken reference, format this reference for community input through established genome browsers, and significantly improve our abilities to define traits of economic importance in the chicken.

I. Introduction

Chicken meat and eggs provide a leading source of high quality protein at a time when worldwide demand for this source of nutrition is growing rapidly [1]. Beyond the importance of a safe and nutritious food supply to human life, the enormous world-wide interest in raising poultry for food provides a collateral source of scientific data that inform our understanding of biology in general. The huge commercial populations mean that large scale breeding studies can be done in the chicken with unprecedented genetic resolution. In many cases the traits that are of interest (nutrition, growth, disease resistance, reproductive success) to the poultry industry who possess the largest flocks are traits that are of similar importance to human health, so studies of chicken genetics and human medicine are often complementary. The initial genome sequence of the chicken [1] provided a quantum leap for poultry genetics, enabling a range of new "omics" analyses and technologies to be applied to poultry science and commercial breeding. However, the gaps that remain in that sequence persist as handicaps that prevent the full potential of poultry genomics from being realized.

A major example of how a complete genome sequence is hindering the full potential of the poultry industry is in genomic selection. Specifically, all the major breeding companies realize that they need to incorporate molecular genetic methods to meet the growing demands of consumers. So since the development of molecular genetic maps and especially after the release of the chicken genome sequence, they have pursued genomic selection. In brief, genomic selection uses genotypes from evenly-spaced markers spanning the entire genome on individuals to estimate their breeding value, which in theory could substantially increase the rate of genetic gain compared to traditional selection methods. The power of genomic selection increases with the density of markers and the extent of population-wide linkage disequilibrium (LD). Also, fortuitously, costs for genotyping many thousands of single nucleotide polymorphisms (SNPs) have come down dramatically and will continue to do so in the near term. Thus, for the first time, there appears to be a combined solution that addresses both theoretical and technical issues of marker-assisted selection in animal breeding.

The issue as will be discussed later in greater detail is that the most current chicken genome assembly has significant gaps, including 8 microchromosomes, which are thought to be gene-rich. As a consequence, molecular geneticists are unable to "tag" a significant portion of the chicken genome that probably includes agriculturally relevant genes. **To highlight this shortcoming, Cobb Vantress, the leader in broiler breeding accounting for 50% of worldwide sales, has committed $150,000 in additional direct funding to the funding requested herein for this project.** This financial commitment clearly demonstrates the need by the poultry industry to improve the chicken genome assembly. Furthermore, at the request of the USDA, a second Whitepaper was written in 2010 outlining the need for an improved chicken genome sequence and proposing some similar initiatives to those described herein (www.poultry.mph.msu.edu/about/Chicken_Genome%20supplemental%20sequencing.pdf). At that time, this Whitepaper received supporting emails from 84 representatives of academia and industry around the world (list at http://www.poultry.mph.msu.edu/about/about.htm).

In addition to the poultry industry, academic scientists are hindered by the incomplete assembly. It is clear that modern biology is centered on genomes. Functional and comparative genomics are highly dependent on having a complete genome. Thus, like molecular genetics, the incomplete chicken genome reference hinders the ability of scientists to further our biological knowledge of important traits such as production, disease resistance, and welfare.

In this submission, to address the need for a more complete chicken genome assembly and to include the research community in this process, we propose the following objectives:

- Fill in known gaps using new computational approaches and existing sequences and assemblies
- Target the remaining gaps with third generation sequencing technologies
- Develop or refine linkage maps for the microchromosomes to aid the placement of sequence contigs
- Establish a proven reference maintenance infrastructure that will solicit and collate communal input

II. Genome reference history

In 2003 a single, partially inbred Red Jungle Fowl female bird (the primary wild progenitor of domestic chickens) was sequenced with NHGRI support in response to a White Paper (see http://genome.wustl.edu/genomes/view/gallus_gallus/) by The Genome Institute (GI), Washington University School of Medicine, with the first draft sequence published in 2004 [2]. That assembled draft based on 6.6X sequence (Sanger technology) coverage was then aligned to chromosomal linkage groups using comprehensive physical [3] and genetic linkage maps [4, 5]. Subsequently, a second build (WUGSC 2.1/galGal3) was generated (May, 2006). In this version an additional 198K reads focused on contig ends and regions of poor quality were added. Furthermore, the assembly order and orientation were improved using early SNP mapping data that better aligned contigs on chromosomes. Total sequence in galGal3 includes 1.1 Gb of sequence, about 95% of which is anchored to autosomes 1-28 and 32, along with the Z and W sex chromosomes. The Z and W sex chromosomes were sequenced only to ~3.3X due to their hemizygous state in the female bird used. Build galGal3 increased the size of chrZ from 33.6 to 74.6 Mb and decreased chrW from 4.9 Mb to 0.26 Mb due, in part, to several mistaken assignments of contigs to W in galGal2 that were actually on Z. A focused effort to improve the Z chromosome subsequently resulted in a nearly contiguous version of this chromosome, incorporated into galGal4 (see below) [6].

A third build of the genome (galGal4), performed by the University of Maryland in collaboration with the GI, was released in November 2011. This assembly included the use of next-generation sequencing technology (454 Titanium, 12X) in combination with previous read types. Using a combination of Sanger and 454 sequence resulted in an increase of N50 contig size by 460% to 252 Kb. The supercontig N50 increased to 17.6 Mb. In addition to the significantly improved Z chromosome sequence, the new

assembly removes about 10 Mb of artifactual duplications. The total amount of sequence mapped to the chromosomes increased by 15 Mb, after accounting for duplication errors. These corrections derive from the fact that the bird used for sequencing was incompletely inbred, leading to heterozygous alleles being mistakenly called as duplications in earlier assemblies. This galGal4 assembly is currently being annotated for gene content at Ensembl and is already available through the UCSC and NCBI browsers. In addition, the GI currently has 70X Illumina coverage of the reference genome that isn't part of the above assembly. To facilitate further evaluation of reference genome representation individual assemblies have been created for each sequencing technology, 454 and Illumina[7].

III. Need for refinement.

The current chicken genome assembly suffers from many of the same problems that have been proven to occur in other vertebrate genomes, even in highly "finished" genomes such as those of human and mouse. Being the first agricultural animal genome sequenced also means that the initial build (galGal2) was done at comparatively high cost and lower coverage with sequencing technology circa 2003. The galGal4 version addresses some of these deficits by adopting a hybrid approach that fuses old and next-generation sequence technology and by including the nearly contiguous Z chromosome. However, the galGal4 assembly still retains about 9,900 gaps on the ordered chromosomes and 21,327 gaps on chromosome-aligned, but random (sequences placed on a chromosome with low confidence) scaffolds and unaligned scaffolds (chrUn). The assembly is particularly problematic for the small microchromosomes (average size of 12 Mb) and the W sex chromosome. A high quality W chromosome sequence reference is being generated at the GI in collaboration with David Page's lab at the Whitehead Institute, and some efforts have been made to fill out the chromosome 16 sequence, at least in the MHC (B-complex) region[8]. As with other vertebrate genomes, a major problem in the current sequence relates to segmental duplications. An obvious example of this comes from Bellott et al. [6] who identified a tandem array of four testis-expressed genes that constitutes ~15% of the Z chromosome, one-fifth of all chicken segmental duplications and in total about a third of the protein-coding genes on Z. The array exists in two blocks that were unassembled in galGal2 (partly in fragments on chrUn but mostly missing). It's well known that such duplications are frequently missed ("over-collapsed") in draft quality assemblies, that they are common sites of copy number variation (CNV), and that such variation frequently has major phenotypic consequences [9].

The chicken genome assembly has a particular problem that appears not to be shared with those of mammals: the missing sequence/microchromosome problem. Like most birds, chickens contain about 10 "macrochromosomes" with lengths typical of those in mammals, but the remaining 28 autosomes are "microchromosomes" that are too small to easily distinguish or order by standard cytogenetics. Moreover, as confirmed by the draft sequence [2], the microchromosomes are unusual in base composition (GC rich), recombination rate (high cM/Mb), gene density (high) and intron size (low). Of primary importance to this proposal, microchromosome sequences are underrepresented or totally missing in galGal4 and all clone libraries examined to date. The reason for the

missing sequence remains uncertain, but it is especially problematic for the smallest chromosomes, GGA16, 25 and 27-38. Representation of these chromosomes in both chicken and turkey BAC libraries is typically even less than that of the hemizygous Z chromosome, and no distinguishing BAC probes are available for GGA29 and higher. As a result, there either is no sequence alignment to these chromosomes (GGA29 and higher, except for ~1 Kb on GGA32) or the assembly is incomplete and uncertain (GGA16, 25, 27, 28). This was demonstrated by the ordered resequencing of GGA28 [10], which greatly reoriented its assembly.

While a significant fraction of the missing sequence is likely repetitive, it's clear that a substantial number of genes are also missing. For example, the great majority of chicken genes homologous to those on HSA19q cannot be found in the assembly, and, for the most part, they are also missing in chicken BAC libraries and in the Trace Archives. We recently searched for chicken sequence orthologous to the 758 genes on HSA19q. For the 723 genes with coordinates from 35 Mb to the end of HSA19q (59 Mb), no matches or only paralogous matches were found in galGal4 for virtually all. For about 20-25% of these genes, a likely orthologous chicken EST could be detected that either failed to align with galGal4 or aligned only with chrUn. Thus, at least some of the orthologous genes definitely are retained in the chicken genome but are missing or unplaced in galGal4. A similar situation exists for the 40 most distal genes on HSA8q (coordinates 144.9 Mb to 146.3 Mb). It seems likely that the orthologous segments for both of these regions of the human genome lie on one or more chicken microchromosomes.

It was initially thought that the missing microchromosome sequences arose from an inability to be cloned in *E. coli*. However, the recent turkey genome sequence assembly [11] was shown to be equally deficient in HSA19q orthologues, even though it was based on NextGen sequencing with no cloning involved. Similarly, extensive 454 and Illumina sequencing of chicken at GI has not added substantially to the microchromosome assemblies. Thus, the explanation for the missing sequence remains uncertain, but it most likely involves a combination of poor clone representation, high GC content leading to poor sequence reads, and a high density of simple tandemly repetitive sequences that interfere with assembly. Based on the results described above, at least some of the missing sequence lies within the 39 Mb that still remain on chrUn_random in galGal4. It also remains possible that some characteristic of these chromosomes makes it unusually difficult to recover DNA from them.

One potential explanation for the missing microchromosomal sequence is that they are present in raw sequencing reads that are never included in the genome assembly due to abnormally low coverage or high repeat content. The Brown lab at Michigan State recently applied a novel assembly graph coverage normalization procedure (arxiv.org/abs/1203.4802) to a 70x Illumina short-read data set and recovered 40 Mb of assembled contigs not contained within the galGal4 genome. These contigs were cross-referenced with unmappable transcripts from a spleen mRNAseq data set, and were found to contain matches to more than 70% of the transcripts. While preliminary, these results suggest that a significant portion of the missing sequence is present in the

Illumina data and could potentially be recovered with more sensitive computational techniques.

IV. Genetic and Physical Mapping.

The chicken genome is organized on 38 autosomes plus the Z and W sex chromosomes; males are the homogametic sex. As a companion to the initial genome sequencing, we generated a detailed BAC clone-based chicken genome physical map [2]. This was based on extensive fingerprinting of clones from 5 different BAC libraries, along with "overgo" hybridization to identify BACs that contain specific linkage map markers and/or genes of interest and alignment of BAC end sequences (BES) with the initial sequence assembly [1]. The final chicken physical map [2] contains only 260 contigs, however, repetitive regions such as centromeres, telomeres and ribosomal RNA-encoding regions (rDNA) remain unassembled. Subsequently, a detailed BAC physical and comparative map was also generated for the turkey genome [10]. This physical-comparative map consists of 74 BAC contigs, with an average contig size of 13.6 Mb, and it defines 20 to 27 major rearrangements distinguishing turkey and chicken chromosomes, despite up to 40 million years of separate evolution between the two species. This turkey map provided the platform upon which the NextGen-based turkey genome sequence was assembled [6], and it suggests likely locations for a few of the small unassembled contigs from the chicken assembly [10]. However, as noted above, the turkey sequence is equally deficient in assigning and assembling the smallest microchromosomes. It's important to note that the turkey genome assembly depends critically on alignment with the chicken genome and thus it as well as other avians would benefit from the improvements proposed to the chicken sequence in this proposal.

In conjunction with the physical map, the accuracy and coverage of the genome sequence depends greatly on the quality of the genetic map. Dr. Cheng and colleagues curates the East Lansing (EL) genetic map, which is based on a reference panel generated by mating a single UCD001 line RJF male to a single female from the inbred UCD003 White Leghorn (WL) line [12]. Subsequently, the EL genetic map was combined with the Compton (C) genetic map[13] and the Wageningen (W; The Netherlands) map maintained by the Groenen lab to form a consensus genetic map[4]. What may not be readily apparent in this consensus map is that it combines the strength of the EL genetic map for unambiguous ordering of markers due to the use of inbred lines and a simple population structure, while allowing for the more accurate determination of map distances between closely linked markers via the W map.

In 2004, the Beijing Genome Institute led an effort that identified 2.8+ million in silico single nucleotide polymorphisms (SNPs) by sample sequencing three birds (commercial broiler, experimental WL, and Chinese Silkie) and comparing these reads to the RJF sequence; ADOL was a contributing member. Through two USDA grants (Cheng as PI), a panel of 3,072 [14] and 60K selected SNPs[15] was designed for genotyping on the Illumina platform. As a result, the current consensus map includes 10,000 genetic markers, and by targeting unmapped sequence contigs, we filled genome assembly

gaps and formed new linkage groups that may contain some of the missing microchromosomes.

V. Genome Reference Consortium (GRC)

The GRC was initially conceived as a means to improve and update the assembled human genome, however, other important model organisms have been added to the portfolio, namely mouse and zebrafish. The goal of this group is to correct the small number of regions in the reference that are currently misrepresented, to close as many of the remaining gaps as possible, and to produce alternative assemblies of structurally variant loci when necessary. The GRC consists of GI, The Wellcome Trust Sanger Institute (WTSI), The National Center for Biotechnology Information (NCBI), and The European Bioinformatics Institute (EBI).

The first assembly done by the GRC, (human) GRCh37, was released in Spring 2009. This assembly was done primarily as a proof of principal to validate the new assembly method being used. The previous build, Build36, was created from various sequencing centers submitting their finished portion of the genome to one repository, while the new method, one produced from NCBI, provided a consistent assembly across the entire genome. Since GRCh37, we have released 11 additional 'patch' releases, which are fixes of existing sequence or novel sequence that do not disrupt the existing chromosomal coordinate system in the main build. To date, 115 fix patches and 73 novel patches have been released, some of which include fixing the ABO blood group locus, as well as providing alternate haplotypes for regions like the MHC.

Since the inception of the project in 2008, The GI has closed 155 of the 467 total human reference issues in our territory. These issues range in severity from a single base change needed to fix an error in the sequence, to completely retiling through a complex region with a single haplotype of clones. 83 of 285 total mouse reference issues assigned to GI have been resolved since beginning this project. It's important to note that these issues existed even in highly "finished" genome sequences. Draft sequences, especially those from WGS-based approaches, even high quality drafts like that of the chicken, will likely require many more fixes. Some examples of the fixes (patches) that have been added to the human genome are noted in Figure 1 (figure courtesy of Deanna Church, NCBI).

No mechanism exists for the chicken user community to point out problems in the existing assembly, correct errors, better annotate CNVs and add their own local sequence information. Many labs are now sequencing chicken genomes with next generation methods and/or doing sequencing of local areas of special interest (e.g., the MHC/B-complex). The ideal solution would be for chicken to be added to those species supported by the GRC (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/), which currently supports human, mouse, and zebrafish genomes. The GI has years of experience working within this infrastructure to significantly improve the human and mouse genome references.

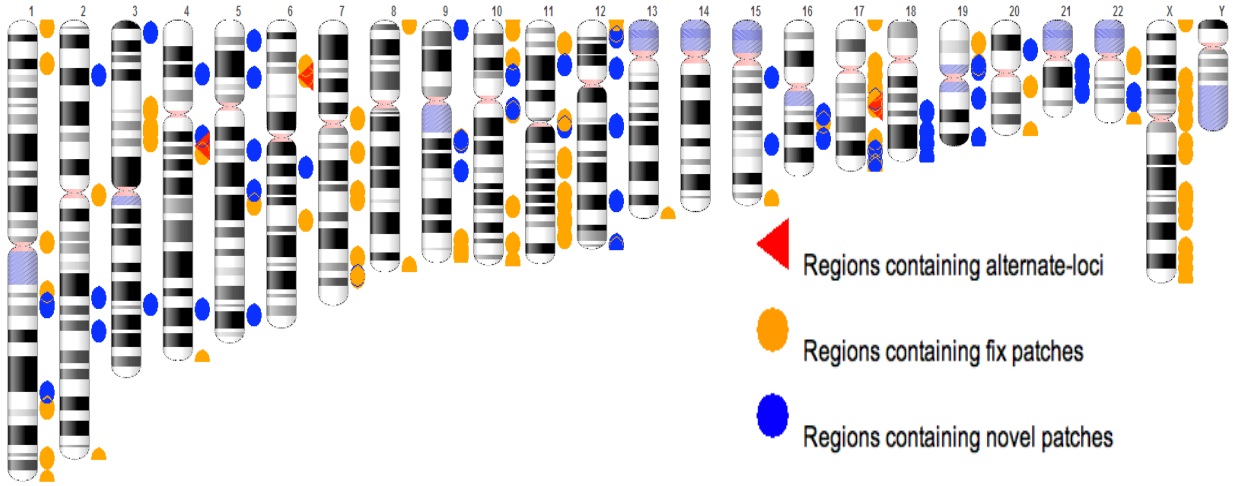


Figure 1. GRC overview of the human genome improvements.

Much of the above has been possible because our Institute has been a longtime leader in the development, implementation, and application of new methods, technology and computational tools for large-scale improvement of reference genomes and regions of biological interest. While the engineering of new systems (e.g., Illumina, 454, etc.) has been crucial to this sequencing revolution, it is important to also credit developments in laboratory methodology and applications (e.g., sample preparation, sequencing library construction, hybridization-based fragment capture), as well as software tool development to process, manage and analyze the tremendous quantity of data that is produced by next-gen sequencing systems.

VI. Community outreach

Genome sequencing is only a starting point in our efforts to disseminate critically needed genomic information to our main audience - scientists - who use these templates to test hypotheses that are at the core of our understanding of biology. We sequence and assemble species over a broad phylogenetic continuum and analyze many of these species in great depth, reporting our findings in high impact journals. These manuscripts serve as outreach to a very broad audience, exciting the public about the return on an investment in genomics. A recent success story is the zebra finch genome project, which exemplifies our efforts to disseminate information to a new community of researchers, including many focused on neurogenetics and vocal learning. Over 20 labs participated, some with previous genome consortium experience, and others with little to no experience. These diverse collaborators were identified from the associated international research communities and brought together for analysis of this avian genome, resulting in publication of the primary manuscript in Nature. Our leadership in this project was key to bringing this exciting story to the world.

The GI has made numerous and substantial contributions to the NHGRI sequencing program that have resulted, directly and indirectly, in significant advances in biology and biomedical knowledge. We have played an important role in the NHGRI Research Network since its official inception in 2003, and have been directly responsible for

initiating, formulating, contributing data to, and/or helping to manage several signature community resource projects including the 1000 Genomes Project, The Cancer Genome Atlas (TCGA), ENCODE, modENCODE, and the Knockout Mouse project. In addition, we have sequenced the genomes of many metazoan organisms including mouse (in collaboration with the Broad Institute (BI), Baylor College of Medicine (BCM) and Sanger centers), chimpanzee (with BI), orangutan (with BCM), rhesus macaque (with BCM), marmoset (with BCM), platypus, chicken, zebra finch, and many others. In an ongoing collaboration with our colleagues in the chicken community we fully intend to continue this tradition of providing high quality genome references.

Along with traditional whole genome sequencing and assembly analysis projects, we have also documented fascinating stories of sex chromosome evolution in the bird Z chromosomes [5]. This success was the result of a multi-lab collaboration, with weekly conference calls to discuss sex chromosome biology and the status of each sex chromosome, a collaboration that continues today. At present we are ahead of our expected dates of completion for the W sex chromosome. We feel such successful high profile projects demonstrate our outreach and dissemination strength at the whole genome, individual chromosome and regional genome levels. To further strengthen our community interactions, GI faculty and staff members frequently attend research community meetings, workshops, and/or visit investigators' labs in order to better understand the community needs, to speak about project goals and progress, to disseminate data, and to solicit input.

Equally critical to our dissemination of information is the use of website portals. Importantly, we have extensively reorganized our website to meet the growing demand for real time updates (http://genome.wustl.edu/). We have detailed descriptions of our projects that include alternative portals for additional information. We also post the software and tools we use to complete our projects. We have announcements and news sections that keep the community abreast of current projects. We are making using of social networking sites such as Twitter, Facebook and Wikipedia to encourage open dialogue among scientists and the public regarding the latest genomic research news. Finally we provide contact information that allows for easy access to our scientists to address all questions and comments promptly.

VII. Rationale and Significance

As noted in the original White Paper for chicken sequencing (http://genome.wustl.edu/genomes/view/gallus_gallus/), the chicken is the premier non-mammalian vertebrate model organism, as well as a growing and leading source of protein, world-wide. The impact of the chicken genome sequence has been enormous, and it continues to be routinely used by poultry geneticists and other scientists on a regular basis [16]. The chicken sequence clearly delivered on its promise as an outgroup for comparative analysis of the human genome. Background (non-selected) homology between the chicken and human genomes is essentially nil, so the chicken-human comparison helped identify the ~5% of the human genome under selective pressure. Remarkably, more than half of this sequence is not within protein coding genes and, on average, is located farther away from genes than it would be in a random distribution [2]. The sequence formed the framework for SNP discovery [5] and the

development of SNP chips. This provided the opportunity to assess the worldwide genetic diversity of the chicken[14], a first for any agricultural animal and done to a depth greater than perhaps any animal species other than human at that time. The development of larger SNP chips (now up to 600K) has enabled whole genome association studies for quantitative traits and is currently allowing the major poultry breeders to implement whole genome-based selection strategies. Due to the relatively short generation time of chickens, the effectiveness of such approaches can be experimentally assessed much more easily than in cattle or swine. More recently, the reference sequence provided the basis for extensive resequencing of a variety of chicken genomes [17] [18] to examine the diversity in even greater detail and to reveal the many marks of selection that have occurred in both the domestication of the chicken and in experimental breeding populations such as high and low body weight lines. Many other resequencing projects have been completed or are currently underway, all of which both expand interest in the best possible “reference” genome and provide additional data that could improve the assembly, should appropriate avenues exist. A higher quality reference will be critical to the successful continuation of these efforts to selectively improve domestic chicken germplasm hardiness.

The galGal2 sequence and associated maps provided the framework for the sequence of the chicken Z chromosome which generated critical insights into the evolution of sex determination [6]. The chicken sequence also fulfilled its promise as the sequence of the model bird (and even the model dinosaur!) and was critical to the subsequent sequencing of the zebra finch [19] and the turkey [11], as it will be to all the avian genome models generated as part of the Genome 10K initiative (already 50 avian genomes have been sequenced and assembled by the Beijing Genome Institute). The gene annotation of these many avian genomes will each rely on defined gene content in our chicken reference thus errors or missing genes will limit the successful annotation of all. The sequence was also critical in the recent study of chicken centromeres that suggests that at least some chicken centromeres are among the smallest known in vertebrates, of a size easily amenable to *in vitro* mutagenesis. As with the Z chromosome [6], both studies demonstrate the importance of finished quality sequence.

VII. Research Methods

a. Overall Strategy

The needed improvement in the chicken genome reference can be compartmentalized into three areas 1) closing gaps within scaffolds, 2) closing gaps between scaffolds that are accurately placed along chromosome boundaries and 3) attaining novel microchromosome sequence assemblies currently not present in the reference. A variety of approaches can be used to address the issue of missing microchromosomes in the reference. Bioinformatics approaches such as digital normalization can be used to obtain more sensitive contig assemblies, but do not yet extend to scaffolding, and assembly merging can be used to integrate multiple assemblies into a single reference. Experimental approaches also could be used to preferentially recover microchromosomal DNA prior to sequencing. For example, one could enrich for microchromosomal DNA by flow sorting or microdissection prior to amplification and

NextGen sequencing. An alternate approach might involve the use of pulsed field gel electrophoresis to enrich for the smallest chromosomes. We have considered both of these approaches, but, at this time, we feel that neither of them is able to collect sufficient material to generate DNA libraries for third-generation sequencing. The only feasible approaches to utilize these methods would also require amplification steps that would introduce a representation bias. Therefore, at the current time, we feel the other alternatives described below provide the most cost effective options to address this third problem area.

Starting with the most recent version of the chicken genome assembly (galGal4) we plan to follow a series of iterative build processes to provide the community the best possible reference, attempting to rival the mouse in quality. galGal4 still retains about 9,900 gaps (estimated total size: 17.7 Mb or ~1.8%) on the ordered chromosomes and 21,327 gaps (estimated 27.2 Mb, ~2.7%) on chromosome-aligned, but random scaffolds and unaligned scaffolds (chrUn). Although the reasons for the gaps aren't fully understood, they likely include: high GC content, lack of suitable FISH and/or genetic markers, unclonable BACs and CNVs or high repetitive content. In this proposal we will 1) fill in known gaps with new computational approaches relying on existing independent assemblies, 2) target remaining gaps with third generation sequencing technology using clone and whole genome-based approaches, 3) further refine microchromosome linkage maps for localization of unplaced sequence, and 4) establish a reference maintenance infrastructure, successfully used to improve human, mouse and zebrafish genomes and involve the community as a whole in an iterative genome improvement process.

To improve the galGal4 version we plan to pursue these steps in chronological order as follows: 1) align previous assemblies of Illumina and 454 read technology to close gaps and capture unique data not found in galGal4, 2) produce Pacific Biosciences (PacBio®) long reads (>5 Kb) for alignment and reference integration, 3) mapping of BAC sequence ends to a revised (steps 1 and 2) galGal4 to identify gap spanning BACs that can be sequenced and assembled, 4) initiate linkage mapping studies that target microchromosomes, and 5) implement a web-based interface to field community feedback for further improvement of regions of high biological interest. The iterative steps that outline our reference improvements are graphically presented in Figure 2.

The development of new sequencing technologies is rapid, and at least some of these may help in addressing the missing sequence problem. A prime example is the long reads described for PacBio® instruments [20]. Long reads across repetitive regions may allow for assembly of sequence contigs that currently either cannot be assembled at all or, at best, are consigned to chrUn. New assembly and integration software will help in addressing this special problem and is needed to map long reads to the existing framework sequence and deal with the different structure and error profiles of these third generation reads. If these could be captured, even in very small sequence contigs, they could provide useful starting points for more directed approaches. Other technologies may also be utilized should they become commercially viable and cost effective, such as Oxford Nanopore or Moleculo.

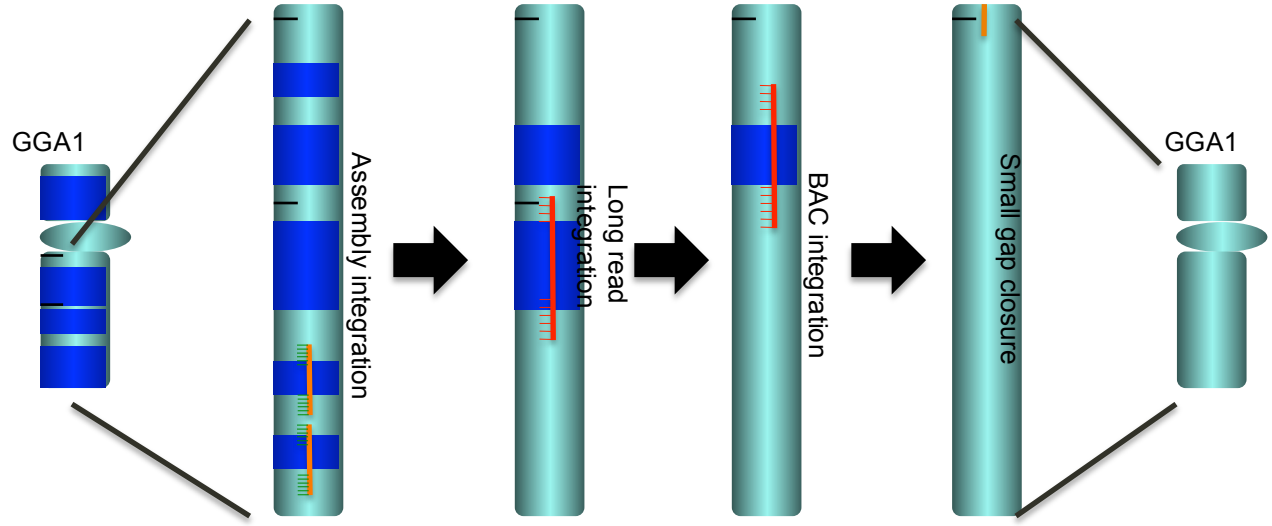


Figure 2. Iterative sequence improvement options, blue bars represent gaps.

1. Objective 1 - Improving the chicken genome with existing data.
Draft assemblies created from Illumina and 454 sequencing technology [7] were aligned to the galGal4 reference to assess gap closure and unique sequence not represented in the current galGal4 assembly. The Newbler assembly of 454 data contained 2 Mb not present in galGal4, and a SOAPdenovo assembly of Illumina data contained 7 Mb not present in galGal4. The Velvet assembly of the digitally normalized Illumina data contained only 400 Kb not present in the Newbler and SOAPdenovo assemblies, again suggesting that multiple different assembly strategies can provide significant gains in sensitivity, a critcal point if we are to discover not found microchromosome sequences.

A major obstacle to improving the chick genome with this additional data has been the lack of software capable of comparing and merging multiple vertebrate genome assemblies, as well as performing large-scale post-merge quality evaluation with the raw data. An example of the complexity for the assembly to assembly integration process is seen when aligning target and query contigs with tips (non-aligned segments; Figure 3) from independent assemblies. CAP3[21] and Minimus[22] are effectively incapable of building merged assemblies on this scale, while GAA[23] can merge any two assemblies but consumes considerable compute resources in doing so; GAA has not yet been applied to iterative merging of more than two assemblies.

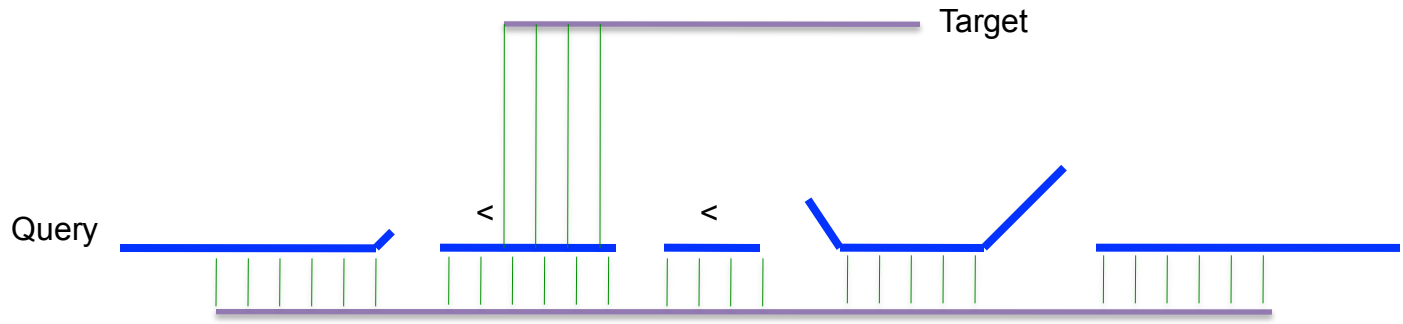


Figure 3. Merge complexity for target and query contigs of independent assemblies.

We will investigate, apply, and assess multiple computational approaches to merging assemblies for the purpose of systematically increasing contig sensitivity, scaffold N50, and long-range accuracy of the genome assembly. Specific approaches to improving assemblies being developed in the Brown lab include digital normalization, read-to-graph alignment for the purpose of error-correcting reads and collapsing heterozygosity prior to assembly, and the extension of long-insert mate-pair constraints to assembly graph analysis [24] [25]. In addition, we will make use of an existing technique for ordering and orienting scaffolds using RNAseq, which should be able to build microchromosomal scaffolds computationally [26] and will of course result in contigs/scaffolds consistent with the RNAseq data thus allowing improved gene prediction.

To evaluate the sensitivity and specificity of these approaches on the chick genome, we will use sequencing information such as k-mer content from the reads included in the genome, concordance between mate-pair ends, and BAC-end pairs, to assess the degree to which each assembly agrees with the underlying data. We will also use RNAseq *de novo* assemblies done with Trinity [27] to identify likely genome mis-assemblies and missing regions. The Cheng lab has a number of extant spleen RNAseq data sets that have been used for assembly evaluation (described above), and the Brown lab has generated a number of HH6-12 embryonic transcriptome samples.

Our primary goal in this objective is to identify data filtering and assembly techniques that perform well on hard to sequence and hard to assemble regions, both to improve the chick genome and also to help serve as a useful technology guide for improving other agricultural genomes that may contain such regions.

2. Objective 2 - Third-generation long read sequencing gap closure.

Once we have exhausted gap closure methods utilizing previous draft assemblies, we will focus on the use of long-read technology from PacBio®. For this objective we envision two phases: first, closure of all possible gaps with random whole genome PacBio® long reads; then, sequencing spanning BACs that target between scaffold gaps. Each PacBio® smart cell produces on average 100 Mb. As a first step only 10x sequence coverage of long reads (>5kb) will be produced to align to an improved galGal4 assembly (Objective 1). This estimate is based on gap closure of 64% with 6.8x mapped-coverage of a primate genome. We will utilize the PBJelly tool

(http://sourceforge.net/projects/pb-jelly/) to integrate long PacBio® aligned reads (>5kb) into the draft assembly consensus base output [28]. Prior to aligning reads with BLASER (alignment software specially adapted for PacBio® reads) reads will be error corrected using an iterative process, implemented in the SMRT® analysis portal. This approach uses the short (200 bp) highly accurate reads to correct single long reads (>5 Kb) from an accuracy of 85% on the seed read to 99% in the base consensus. To provide experimental proof this method is a viable strategy we have utilized a fragmented *Parus major* (great tit bird) genome assembly and selected a 1 Mb scaffold (7 contigs) for this test as this scaffold size and gap structure would reflect what we expect with the high quality chicken reference. After running PBJelly with default parameters we reduced 7 scaffold contigs to 4 with an increase in N50 contig length from 181 to 699 Kb. This 1 Mb scaffold gap alteration is an impressive improvement given our starting long read sequence coverage was estimated to be 1.6x, a significantly lower than optimal coverage input. Published gap closure with mapped PacBio® read-coverage of 24, 4.2, and 6.8x was 69, 20 and 64%, respectively [28]. These encouraging gap-filling results were obtained with vastly different genome architectures, from a primate, bird and fly.

In phase two a modification of the standard approach in use for human and other genomes and already shown to be effective for GGA28 [10] and chromosome Z [6] will be used. Instead of a tile path selection process we will only select gap spanning BACs that are needed to close gaps that were not addressed with mapping and gap assembly of random whole genome PacBio reads in phase one. Once the improved assembly is produced we will map BAC end reads to use as a resource for scaffold gap closure. The primary focus of these directed efforts should be on areas of likely segmental duplication [29] where BACs are critical to allowing creation of accurately assembled sequence and on those microchromosomes that are partially assembled but unfinished.

First we require BACs each have their paired ends (in correct orientation) at least 75 Kb apart and less than 300 Kb apart. It is possible if we had misassemblies, we could then be missing a few BACs, where, for example, the distance in the assembly was only 50 Kb but we had mis-sized the gap. Our somewhat permissive BLAT alignment parameters are that, for both ends independently, at least 200 bp of the BAC end read had to align at >95% identity over the region that aligned and the next best alignment for that BAC end had to have <95% of the number of bases aligned as the "BEST" alignment in order to keep it. Then, of course, the BAC end pairs have to be in the correct orientation and at the correct distance apart. With these parameters, 789 BACs span 191 scaffold gaps and 1020 contig gaps in the current galGal4 assembly. All contig gaps have to be >1 Kb for our consideration as we have plans to close small gaps of this size with our internal gap closure program PyGap. If we increase stringency, requiring the BAC end have at least 500 bases aligning, then for gagGal4 750 BACs span 159 scaffold gaps and 970 contig gaps. Some BACs span multiple gaps, even multiple gap types: 40 scaffold gaps, 24 both a scaffold and contig gap and 686 contig gaps. The above number includes both gaps on both random and ordered scaffolds, but most gaps are on the latter. If we consider microchromosomes as a first priority and look at how many gaps are spanned on the chromosomes with numbers

>20 or chrLGE* (linkage groups that haven't been placed on chromosomes), there are 175 gaps spanned by 74 BACs. Of course this doesn't capture all gaps due to a lack of a spanning BAC meeting our conservative alignment criteria and these galGal4 results will be different when a revised assembly (objective 1 and 2-phase 1) is used as a starting point.

Individual BACs (CHORI-261; http://bacpac.chori.org) that span between scaffold gaps following the completion of objective 2 will be cultured to obtain DNA for long read PacBio® sequencing. For each BAC DNA sample a single SMRT® cell will be sufficient. We plan to use a hierarchical genome assembly process (HGAp) that is described in the PacBio® smart portal system to assemble reads from individual BACs. Our preliminary tests on a single human BAC has resulted in a near complete assembly (2 contigs) of the expected insert size (200 Kb) that when aligned to the same BAC assembly from ABI3730 data yielded no major order or orientation errors and minor base discrepancies (15 within mono- or dinucleotide spans). The HGAp works by first collecting all long reads (>5 Kb), aligns all short and long reads against these long reads, trim and filter as necessary, derive a consensus for long reads, assemble and finally error correct contigs with Quiver. Quiver is a Hidden Markov algorithm that exploits both the base calls and quality values to infer the true underlying DNA sequence and is an open source tool that can be installed from GitHub using these instructions found here: http://git.to.AERIEA. All individual BAC assemblies will be aligned with BLASTZ to the known scaffold gap to evaluate closure. In some cases if a given gap is not closed, i.e. multiple contigs for a BAC assembly, we will attempt to close small gaps (<1Kb) within the BAC assembly using a small gap filling method.

3. Objective 3 - Mapping approaches at the missing microchromosome sequence. Many segments of the missing microchromosomes may currently exist in unplaced scaffolds and contigs. An effort was made to capture at least the large contigs on galGal3 chrUn_random by genetic linkage mapping, but this did not expand microchromosome coverage significantly [15]. To fully assemble microchromosomes, it is expected the steps described in Objectives 1-3 will create larger scaffold structures facilitating easier placement by providing adequate sequence tags within unplaced scaffolds and contigs to integrate FISH, linkage, radiation hybrid, optical mapping and physical maps.

As was done with the 60K SNP array [15], sequence scaffolds and contigs that are not placed in the genome assembly will be identified. SNPs will be identified in the East Lansing (EL) and Wageningen (W) reference panels by aligning existing reads to these unplaced contigs. The EL reference mapping population derives from a cross between the inbred UCD001 RJF line, a member of which was used to generate the reference genome sequence, and the UCD003 inbred White Leghorn line. We have access to extensive Illumina UCD003 sequence data from another project that can be aligned with unplaced sequence contigs to identify useful mapping SNPs for the array. With at least 1 SNP per contig, we should be able to generation linkage maps using custom SNP arrays. Currently, Affymetrix will tile probes to query up to 15K SNPs that can generate

data for ~$100 per sample. So it is reasonable to target 15K SNPs and most likely more given the competitive nature of the custom SNP array industry.

As a backup, many labs including the Cheng lab are making custom arrays for genomic selection. It is relatively easy and economical to add 5K or more additional SNPs to these SNP chips. The advantage is that we can leverage existing resources including highly characterized and deep sequence information for other genetic lines that should increase our chances of each SNP being mapped[18].

Although we have not been able to fit additional subcontracts within the budget limits of this proposal, it should be noted that integration of unplaced sequence contigs into the full genome assembly can also utilize radiation hybrid maps, optical mapping and FISH mapping (the latter being used to align scaffolds to specific chromosomes). We have long-standing collaborations with leaders in RH mapping (Alain Vignal, INRA) and FISH mapping (Mary Delany, UC Davis) with whom additional integration efforts can be arranged, once we have contigs and scaffolds of sufficient length and quality. Furthermore, the USDA-NIFA National Animal Genome Research Program coordinators have discussed possible optical mapping of the chicken genome with David Schwartz (U. of Wisconsin, Madison). These approaches extend beyond the scope and budget of this proposal, but they become feasible only with the improvements to the assembly that we propose herein.

4. Objective 4 – Establish the chicken genome resource portal.

In year 1 (see timeline below) while Objectives 1-3 are underway we will mirror the genome reference infrastructure that receives input from the community to further improvements to those regions of biological interest for the current galGal4 reference. This system will be used to post improvements to the reference in quarterly updates (patches) to the galGal4 reference. Over the years this mechanism has been very successful at fixing errors and adding missing genome sequence. Some examples include correcting a base error that affects translation of a PPIP5K2 splice variant on human chromosome 5, addition of whole genome sequence to mouse chromosome 9 that integrated an unlocalized contig that contained gene annotation, and repetitive regions in mouse chromosome 12 have benefited from extensive manual examination and sequence reordering.

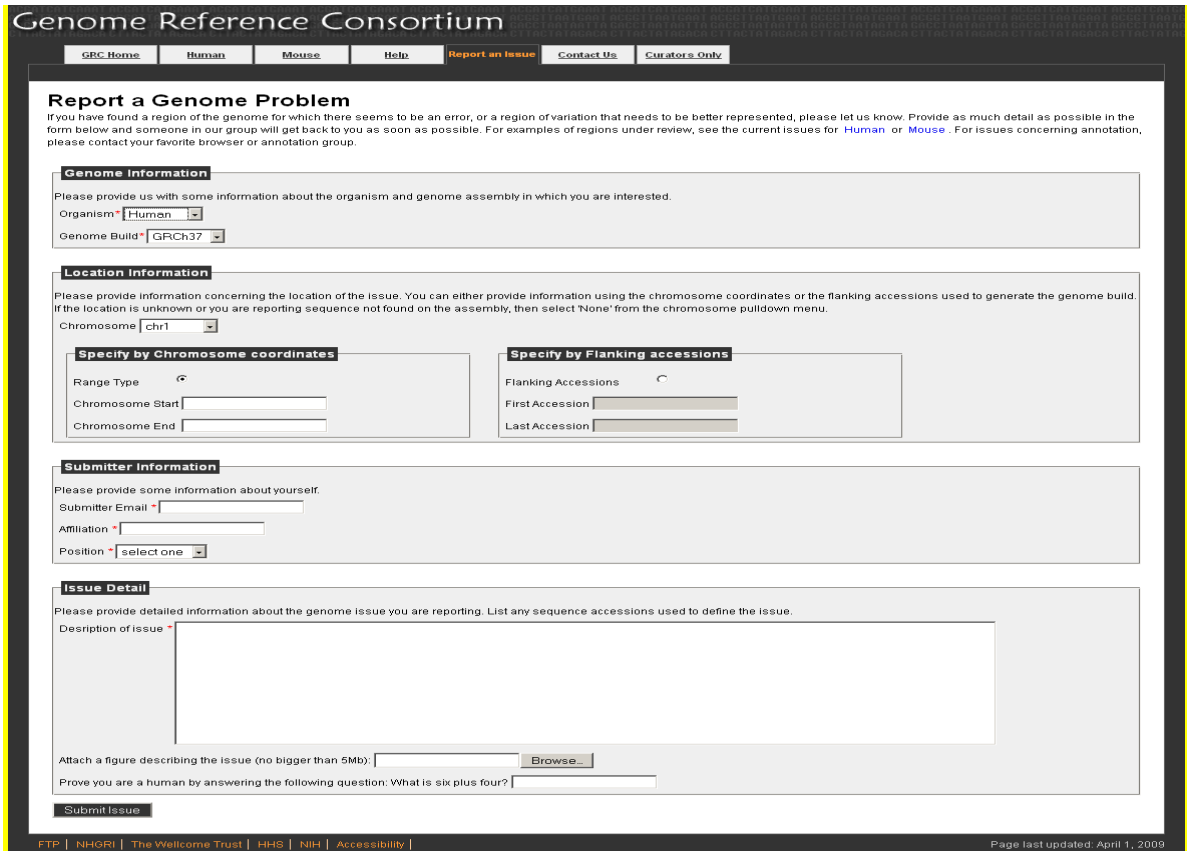
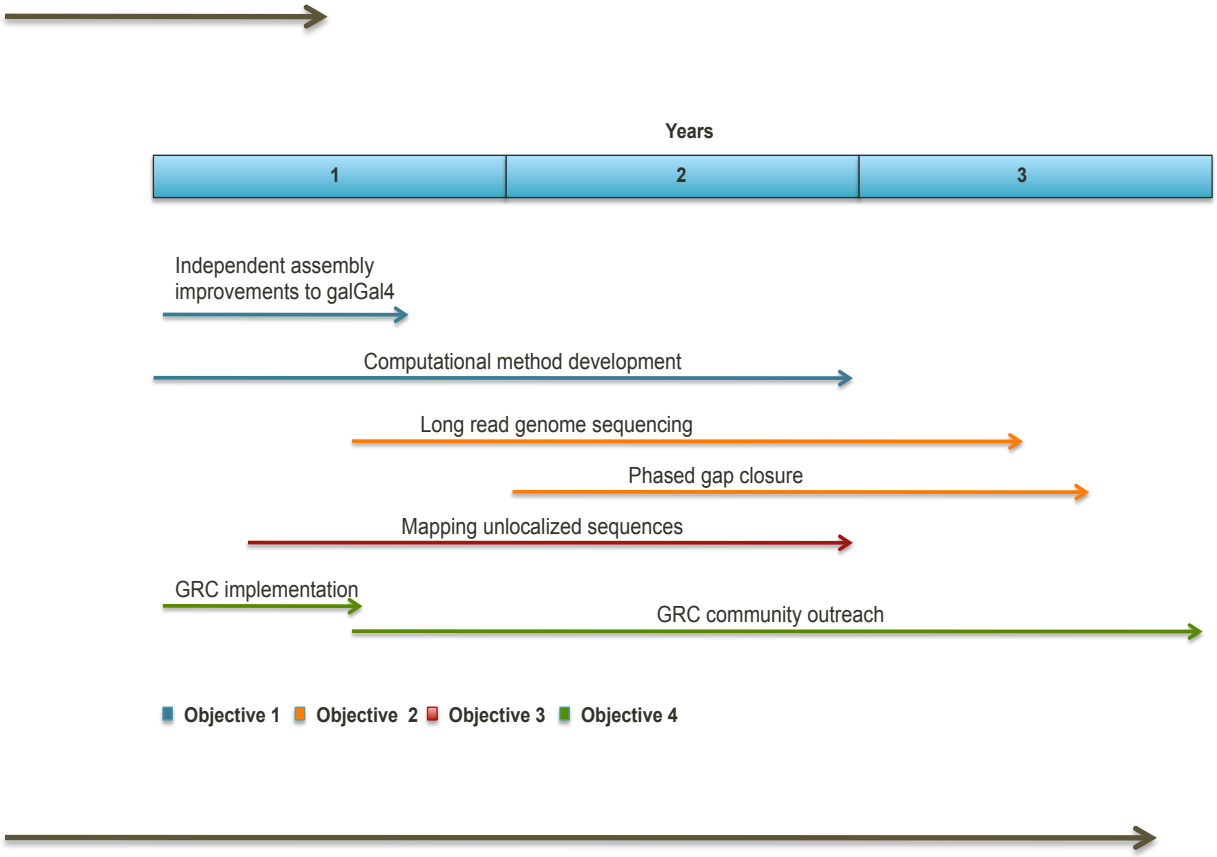


Figure 4. A screen image of the GRC problem reporting page.

All work being done as part of the GRC is viewable to the public through a browser interface, (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/). Investigators can identify and report problems through an interface on the website, as well as see the most current issues that have been resolved (Figure 4). A ticketing system, supported by NCBI, has been established to track the various issues reported and to document standard operating procedures employed in resolving these issues. Human and Mouse chromosomes are divided between WTSI and The Genome Institute, for all experimentalist work, while EBI and NCBI provide the informatics infrastructure and support for the project. A similar system would be used for our project. This portal and accompanying information will be disseminated by the US Poultry Coordinator, via AnGenMap, and poultry meetings (e.g., Poultry Workshop at Plant and Animal Genome).

Project timeline for proposal objectives.


Years
1
2
3
Independent assembly improvements to galGal4
Computational method development
Long read genome sequencing
Phased gap closure
Mapping unlocalized sequences
GRC implementation
GRC community outreach
Objective 1
Objective 2
Objective 3
Objective 4

References:

1. Rosegrant, M.W. and X. Cai, *Water scarcity and food security: alternative futures for the 21st century.* Water Sci Technol, 2001. **43**(4): p. 61-70.
2. Consortium, I.C.G.S., *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.* Nature, 2004. **432**(7018): p. 695-716.
3. Wallis, J.W., et al., *A physical map of the chicken genome.* Nature, 2004. **432**(7018): p. 761-4.
4. Groenen, M.A., et al., *A consensus linkage map of the chicken genome.* Genome Res, 2000. **10**(1): p. 137-47.
5. Wong, G.K., et al., *A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms.* Nature, 2004. **432**(7018): p. 717-22.
6. Bellott, D.W., et al., *Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition.* Nature, 2010. **466**(7306): p. 612-6.
7. Ye, L., et al., *A vertebrate case study of the quality of assemblies derived from next-generation sequences.* Genome Biol, 2011. **12**(3): p. R31.
8. Shiina, T., et al., *Extended gene map reveals tripartite motif, C-type lectin, and Ig superfamily type genes within a subregion of the chicken MHC-B affecting infectious disease.* J Immunol, 2007. **178**(11): p. 7162-72.
9. Alkan, C., S. Sajjadian, and E.E. Eichler, *Limitations of next-generation genome sequence assembly.* Nat Methods, 2010. **8**: p. 61-65.
10. Gordon, L., et al., *Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions.* Genome Res, 2007. **17**(11): p. 1603-13.
11. Dalloul, R.A., et al., *Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis.* PLoS Biol, 2010. **8**(9).
12. Khatib, H., et al., *Sequence-tagged microsatellite sites as markers in chicken reference and resource populations.* Anim Genet, 1993. **24**(5): p. 355-62.
13. Bumstead, N. and J. Palyga, *A preliminary linkage map of the chicken genome.* Genomics, 1992. **13**(3): p. 690-7.
14. Muir, W.M., et al., *Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds.* Proc Natl Acad Sci U S A, 2008. **105**(45): p. 17312-7.
15. Groenen, M.A., et al., *The development and characterization of a 60K SNP chip for chicken.* BMC Genomics, 2011. **12**(1): p. 274.
16. Dodgson, J.B., M.E. Delany, and H.H. Cheng, *Poultry Genome Sequences: Progress and Outstanding Challenges.* Cytogenet Genome Res, 2011.
17. Rubin, C.J., et al., *Whole-genome resequencing reveals loci under selection during chicken domestication.* Nature, 2010. **464**(7288): p. 587-91.
18. Elferink, M.G., et al., *Signatures of selection in the genomes of commercial and non-commercial chicken breeds.* PLoS One, 2012. **7**(2): p. e32720.
19. Warren, W.C., et al., *The genome of a songbird.* Nature, 2010. **464**(7289): p. 757-62.

20. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-8.
21. Huang, X. and A. Madan, *CAP3: A DNA sequence assembly program.* Genome Res, 1999. **9**(9): p. 868-77.
22. Sommer, D.D., et al., *Minimus: a fast, lightweight genome assembler.* BMC Bioinformatics, 2007. **8**: p. 64.
23. Yao, G., et al., *Graph accordance of next-generation sequence assemblies.* Bioinformatics, 2012. **28**(1): p. 13-6.
24. Pell, J., et al., *Scaling metagenome sequence assembly with probabilistic de Bruijn graphs.* Proc Natl Acad Sci U S A, 2012. **109**(33): p. 13272-7.
25. Kelley, D.R., M.C. Schatz, and S.L. Salzberg, *Quake: quality-aware detection and correction of sequencing errors.* Genome Biol, 2010. **11**(11): p. R116.
26. Mortazavi, A., et al., *Scaffolding a Caenorhabditis nematode genome with RNA-seq.* Genome Res, 2010. **20**(12): p. 1740-7.
27. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotechnol, 2011. **29**(7): p. 644-52.
28. English, A.C., et al., *Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology.* PLoS One, 2012. **7**(11): p. e47768.
29. Volker, M., et al., *Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution.* Genome Res, 2010. **20**(4): p. 503-11.