Additional Sequencing of the Chicken Genome

Wesley Warren, LaDeana Hillier, Elaine Mardis & Richard Wilson
Genome Sequencing Center, Washington University School of Medicine

The chicken (*Gallus gallus*) genome represents an important non-mammalian reference sequence that will greatly facilitate the identification of functional sequences within the human genome. Our recent manuscript describing the draft sequence of the chicken genome touched upon several interesting features with direct relevance to vertebrate genome evolution (Hillier et al, 2004). Even in its current draft status, the chicken genome sequence is arguably the most useful species sequenced to date for pair-wise comparisons with the human sequence. Additionally, the chicken is in itself an important model organism, and its genome further serves as a resource for biological research aimed at improving bird health and increasing food production for humans.

We propose here a plan to improve upon the current draft sequence of the chicken genome to address sequence assembly inadequacies and to set the stage for generating a high-quality near-finished genome sequence. One of many important justifications for improving the draft sequence is the fact that the sex chromosomes (Z, W) currently are significantly underrepresented. At present only 30% of Z and 0.5% of W have been sequenced and anchored to those chromosomes, based on their respective estimated sizes. This is not completely unexpected given the paucity of marker information, and that only ~3.3X sequence coverage was produced for those chromosomes since a heterogametic (female) genome was utilized for sequencing. While the draft sequence of this genome has provided an important and useful new resource for investigators studying both mammalian genomes and the biology of birds, a relatively small additional investment will significantly improve the sequence and facilitate the annotation of a more complete gene set. Given the usefulness of the chicken genome sequence for improving the annotation of the human genome sequence, the opportunities for impacting bird health and disease relative to food production, and the current interest in bird pathogens and human health, we believe that this would be a wise investment.

Current status of the chicken sequence

The chicken genome sequence currently exists as a revised version of the original ~6.6X draft assembly that was posted to all of the major genome browsers in March 2004. The sequence was derived using a plasmid-based whole genome shotgun approach, with fosmid and BAC end sequence reads added to improve supercontig construction (all data generated by the WUGSC). In addition to the sequence coverage, a BAC-based physical map was constructed (~20-fold clone coverage) and utilized to further knit together the WGS assembly. The physical map currently consists of 260 contigs, the majority of which are comprised of more than 200 clones each. BAC end sequences generated from a portion of the clones that were fingerprinted were used to anchor the WGS assembly to the physical map. Using a requirement of at least six consistent BAC end pairs, 189 map contigs were integrated with the sequence assembly.

As mentioned above, an improved assembly of the draft sequence recently was generated using a new version of *PCAP* (Huang & Yang, unpublished), and represents a significant incremental upgrade over the earlier assembly. Statistics for the improved assembly are presented in Table 1. Our experience with the mouse genome had suggested that the generation of a reasonably accurate assembly (estimates of 0.3-0.7% chromosomal misassignment and local misordering) using ~6.5X sequence coverage requires integration of the assembly with a physical map and significant manual

correction of any detectable global misassemblies. Therefore, we followed a similar course for the chicken genome and have used the BAC-based physical map as a scaffold to order and orient sequence contigs. In the current iteration, we have been able to construct and localize to chromosomes 84 ultracontigs that represent more than 75% of the genome.

Table 1. PCAP assembly of the chicken genome (2/1/05)

Number of sequence reads: 12,864,915
Number of phred20 bases: 7,480,069,423
Sequence coverage: 7.0X (assuming a genome size of 1.06 Gb)
Number of contigs (>1 kb): 87,112
Total contig sum (>1kb): 1,046,030,915 bp
Average contig length (>1kb): 12,007 bp
Largest contig length: 441,790 bp
Number of supercontigs (>1 kb): 24,931
Total supercontig sum(>1kb): 1,049,863,093 bp
Average supercontig length: 42,110 bp
Largest supercontig length: 50,876,398 bp
Map + sequence assembly integration:
Number of ultracontigs: 84 (772 Mb)
Total bp in ultra & supercontigs >2 kb: 1.055 Gb

Note that USDA and NHGRI recently provided a small amount of funding to add directed sequence reads to the draft chicken genome. This work is currently in progress and should allow substantial improvement of the draft assembly. The expected results of this "pre-finishing" phase have been considered in formulating the plan we describe in this document.

A significant limitation for the accurate ordering, orientation and chromosomal placement of the sequence at the current level of contiguity has been the relative paucity of marker information for the chicken genome. Nearly 2,000 markers (about one every 2 cM) have been placed on the consensus linkage map (and integrated into the fingerprint map), some of which are from different chicken strains. As in all linkage maps, there are errors associated with these data that in some cases confound the accurate placement of contigs.

Analysis of the current draft sequence

The amount of finished sequence currently available for the chicken (specifically, jungle fowl #256 which was used for the whole genome shotgun) is limited to 38 BAC clones. A comparison of finished BAC sequences with the draft genome sequence reveals that 98% of finished bases can be aligned with the WGS assembly with an overall substitution rate of 0.02% (phred20 bases only), a deletion rate of 0.01% and an insertion rate of 0.01%. An analysis of WGS supercontigs that matched finished sequence detected no orientation problems, however two order discordances were discovered that would extrapolate to approximately 300 such events in the current draft. Similarly, there were three examples of small supercontigs that were omitted from larger supercontigs, suggesting as many as 500 possible events in the genome. Lastly, we observed four cases where a contig was wrongly inserted into a supercontig, translating to a possible 670 such events in the

current draft sequence. Although the potential numbers of these types of assembly errors are relatively small, they still will disrupt accurate prediction of a significant fraction of genes.

Our placement of human and chicken mRNAs and ESTs onto the draft sequence was quite good overall, in that 94.4% of human mRNAs and 85.7% of human ESTs in Genbank found similarities in the chicken genome. Partial gene prediction could be accomplished for 93% of mRNAs and for almost 90% of available ESTs. However, alignments to mRNAs suggested on the order of 400 rearrangements of supercontigs (primarily moving one supercontig nearer another supercontig within the sequence assembly). Each of these cases required manual review and, in some cases, remains ambiguous. A majority of these suspected rearrangements will require additional sequencing to allow correct ordering and orientation of adjacent contigs.

Comparative analysis

During our recent analysis efforts, specifically when comparing the chicken and human genome sequences, we discovered several examples where additional sequence data will be required before definite evolutionary conclusions can be made. For example, on a large-scale, chicken chromosome 4 shares extensive similarity with human chromosome 4. The first 57 Mb of the human chromosome 4 sequence linearly aligns almost completely with a 30 Mb segment of chicken chromosome 4. However on a smaller scale, a large number of inversions are observed. To determine which inversions are real and which are due to assembly errors, we have reviewed chicken marker and fingerprint map data. Unfortunately, due to the shortage of marker information, many cases are still ambiguous. These cannot be made to appear more human-like in the absence of chicken-specific data. To this point, only one large inversion in the sequence assembly has been identified where additional data indicated that the chicken assembly should have been inverted to agree with the human suggested order. Several other ambiguous examples follow.

a) Bases 76M-90M of human chromosome 16 align with a 14-19 Mb segment of chicken chromosome 11. This particular region contains one of the highest densities of inversions we have found. Here, we have again utilized the underlying sequence assembly, the marker data, the BAC fingerprint data, and alignments to the human sequence in an attempt to refine contig order and orientation. While one supercontig in the region could potentially be flipped, the read pair data are not unambiguous. Likewise, the fingerprint data do not provide additional clues. The sequence assembly is well supported by read pair data, and a misassembly here is not likely. Therefore, without generating additional data specifically aimed at closing the existing gaps, we currently are unable to resolve this region.

b) A region of chicken chromosome 20 contains five supercontigs assembled in the following order:

   s1: 10,051,000-10,060,000
   s2: 10,060,000-10,070,000
   s3: 10,070,000-10,095,000
   s4: 10,095,000-10,106,000
   s5: 10.106.000.10,115,000

   An alignment of this region with human chromosome 20 suggests that the order should be s4, s1, s5, s2 (same orientation, and with the position of s3 undetermined). However, an examination of other data left us unable to accurately place s3, although we did find that

potentially interweaving two supercontigs would result in an order that is more similar (but still not identical) to the human sequence through this region than in our original assembly.

c) Alignments spanning the myosin heavy chain region in human/chicken are also of interest. There is a 1.8 Mb region containing several genes of interest in "sub-regions" 1-5:

1. 1-650,000	MyHC
2. 650,000-740,000	MAP2K4
3. 740,000-990,000	MYCD
4. 990,000-1,180,000	DNAH9
5. 1,180,000-1,700,000	AK127379

Alignment with the human genome sequence predicts the order 1 5' 4 2 3' (with primes denoting reversed orientation). At a finer scale, alignments predict the reversal of a fragment of approximately 50 kb with sub-region 3. In this case, after review of the underlying data, the physical map strongly supported our initial assembly through the region. There are some remaining questions as to the precise order of marker data, however the overall placement is not far out of range. This particular region in our whole genome assembly is spanned by a single supercontig that has good supporting read pair information. Thus, by our usual criteria, the current assembly is acceptable. However, we are currently experimenting with other assembly algorithms to determine whether alternative paths exist for this region. As previously stated, the availability of additional linkage markers and/or sequencing data would help to resolve the order and orientation of this region in the current chicken genome sequence.

Gene detection and annotation

Analysis of the current assembly with the goal of creating an index of all chicken genes has been difficult. The overwhelming consensus among those focused on this task (Hillier, Birney, Miller, Ponting, Bork, et al.) is that much of the sequence is still simply too discontinuous to accurately predict genes. For example, in the current draft for any particular gene, one encounters the following scenarios:

1) "Complete salad." (descriptive term courtesy of E. Birney) For example, the gene P17482 currently has three of its exons on two contigs localized to chromosome 2, another exon on a chromosome 27 contig, with at least two other exons missing. Obviously, this type of example will result in genes being missed or at least incomplete.

2) Stretched/long introns. For example, the gene Q8N6G6 is mainly on chromosome Z, with three "tight" islands of exon structure (one of two exons, one of three exons, one of four exons) separated by potential introns of 10,000 bp and 5,000 bp. These long "introns" contain predicted genes on the opposite strand. The 5' end of the gene appears to be copied on chromosome 10 (and/or there is a complicated paralog). Large introns are not unheard of in the chicken genome, however this sort of "island" arrangement is more likely due to assembly issues. This type of problem likely will lead to the islands being annotated as separate genes.

3) Complex duplications/paralogous structure. For example, it appears that the gene P50238 may be present in three copies within the same region on chromosome 8. Within this region, exons are either missing or misplaced, leading to an inability to accurately reconstruct the intron-exon

structure of this gene.

These examples represent the predominant error types. Largely, they are due to the heuristics of the software tools we have available for gene prediction, most of which were designed to work with more complete genomes and to avoid including pseudogenes within the predicted gene index. For a ~6X draft with a significant number of remaining sequence gaps, this leads to "drop outs" that must be identified by other methods and manually parsed.

A number of interesting examples of specific genes that are missing or incomplete in the current draft sequence have been detected. All of these provide good evidence that improved sequence continuity will be necessary before a stable gene index can be produced for the chicken genome. One such example is the VKORC1 gene, which encodes the vitamin K epoxide reductase protein that recycles vitamin K. The VKORC1 gene is present in all available mammalian genomes, as well as the *Fugu* and zebrafish genomes, yet was apparently absent from the assembled chicken genome sequence. A TBLASTN search of the 440,000 unassembled reads revealed a single read that contains most of the second of three expected exons. A closer look suggests that the VKORC1 gene lies in a region of the genome that is underrepresented in the current assembly, perhaps due to cloning problems or simply for statistical reasons. In contrast, the paralog of this gene - VKORC1L1 - is present in the assembled genome.

Many of the problems described above will only be resolved by additional (preferably targeted) sequencing of the chicken genome. Specifically, we would advocate increasing the sequence coverage - at least through difficult and low coverage regions - by sequencing BAC clones picked based on the physical map. Additionally, oligonucleotide-directed sequencing aimed at closing the existing ~100,000 gaps would greatly facilitate gene prediction. While the current draft sequence is useful for initial global studies of the chicken genome, the comprehensive comparative analyses envisioned by many in the scientific community simply are not possible.

Proposed next steps

We propose the following steps to improve the current draft chicken genome sequence:

1. Shotgun sequencing of targeted BAC clones. Using the current WGS sequence assembly, sequence-mapped BAC clones, the recently-available chicken RH panel and the physical map, we propose to select a tiling path of BAC clones across the chicken genome. This path would serve as a clone resource for the chicken research community, and as a BAC-based representation of the genome on microarrays for comparative hybridization. Additionally, the path would allow the use of restriction digests to confirm local sequence assembly, and the ability to select and sequence specific regions of the genome to improve the assembly or to analyze regions of interest to the research community (e.g., the Z and W sex chromosomes). To improve problematic regions of the current assembly and to enhance the quality and representation of the genome, approximately 3,500 BAC clones will be selected and shotgun sequenced at a low (~3X) level of coverage. This approach will permit local assembly of difficult regions of the genome and will be helpful in discerning unique copies of segmental duplications often associated with diseases and evolutionary mechanisms. In addition, ~1,500 BAC clones will be chosen from regions of the genome with very poor WGS sequence coverage (including the sex chromosomes) and shotgun sequenced at a higher (~6X) coverage level.

To further address the lack of sequence coverage on the sex chromosomes, we will collaborate with other laboratories (e.g. D. Page, Whitehead Institute and D. Griffin, Kent University) to include Z-

and W-specific BAC clones among those targeted for 6X coverage. In addition, we propose that a homogametic (ZZ) BAC library be constructed at a minimum of 6X genome equivalents, and all clones (~50,000) fingerprinted. This resource will allow us to 1) significantly increase clone coverage of the Z chromosome, and 2) aid in the placement of BAC clones and contigs not associated with the W chromosome. Also, we anticipate that several currently unanchored fingerprint contigs can be localized on the Z and W chromosomes by utilizing the chicken RH panel or FISH techniques. Lastly, in order to provide additional sex chromosome-specific marker sequences, we will generate and sequence plasmid subclone libraries from flow-sorted Z and W chromosome preparations (D. Griffin, Kent University).

2. Initial genome sequence finishing. Upon completion of the clone-based sequencing, a manual review of the genome sequence will be performed. This process will aim to resolve misassemblies, to make joins that were missed by the relatively stringent assembly algorithm, and to edit sequence regions not properly represented by the resulting consensus. Using the resulting improved genome sequence, primer-directed sequence reads will be performed to further close gaps and resolve low quality regions. Based on past experience, we expect that ~100,000 directed reads will be required.

3. Additional genome assembly and closure. Upon completion of the efforts described above, some additional but minimal work to close gaps and improve local assembly will be warranted. For example, we will aim for a final average contig length of ~200 kb. To this end, we will first ensure correct assembly, order and orientation of all contigs and BAC clones, expecting that most regions of the genome will be represented by high quality contiguous sequence. A few regions however, will require additional manual curation to meet the order and orientation criteria, wherein very minimal additional laboratory work (mainly PCR and directed sequencing) would be necessary. This level of “finishing” is dramatically less laborious and expensive than those applied for the human and mouse genomes, but will provide a product of the highest quality.

Proposed budget

We estimate that the cost of additional chicken genome sequencing will be $6.95M. (see Table 2). The proposed work would build upon a new PCAP assembly that includes ~198,000 directed pre-finishing reads that utilize fosmid clones as templates (in progress).

Table 2. Budget estimates for finishing the chicken genome

Description	Units	Cost/Unit	Component Cost
3X BACs[a] (3,500)	2,688,000 reads	$0.60/read	$1,612,000
6X BACs[b] (1,500)	2,304,000 reads	$0.60/read	$1,382,400
BAC subclone libraries	5,000	$100/library	$ 500,000
ZZ BAC library			$ 30,000
Fingerprinting	50,000 clones	$2.50/fp	$ 125,000
Pre-finishing	1.1 Gb	$0.001/bp	$1,100,000
Finishing	1.1 Gb	$0.002/bp	$2,200,000
Total costs			**$6,949,400**

a – 768 reads/BAC
b – 1536 reads/BAC

Our estimates of success will be measured in terms of 1) reduction in the total number of supercontigs, 2) increase in supercontig length, overall sequence continuity and gene prediction gains.

Timeline

Project initiation can begin immediately upon funding approval. The expected time to completion is 12-18 months. Some of the proposed tasks will be performed in parallel.