doi:10.1017/S0043933908000019

Review of the initial validation and characterization of a 3K chicken SNP array

W.M. MUIR[1], G.K. WONG[2,3], Y. ZHANG[3], J. WANG[3], M.A.M. GROENEN[4], R.P.M.A. CROOIJMANS[4], H.-J. MEGENS[4], H.M. ZHANG[5], J.C. MCKAY[6], S. MCLEOD[6], R. OKIMOTO[7], J.E. FULTON[8], P. SETTAR[8], N.P. O'SULLIVAN[8], A. VEREIJKEN[9], A. JUNGERIUS-RATTINK[9], G.A.A. ALBERS[9], C. TAYLOR LAWLEY[10], M.E. DELANY[11] and H.H. CHENG[5]*

[1]Department of Animal Sciences, Purdue University, West Lafayette, IN 47907, USA; [2]University of Alberta, Department of Biological Sciences and Department of Medicine, Edmonton AB, T6G 2E9 Canada; [3]Beijing Institute of Genomics of Chinese Academy of Sciences, Beijing 101300, China; [4]Animal Breeding and Genetics Group, Wageningen University, 6709 PG Wageningen, The Netherlands; [5]USDA-ARS, Avian Disease and Oncology Laboratory, East Lansing, MI 48823, USA; [6]Aviagen, Newbridge, Midlothian, EH28 8SZ, Scotland, UK; [7]Cobb-Vantress, Inc., Siloam Springs, AR 72761, USA; [8]Hy-Line International, Dallas Center, IA 50063, USA; [9]Hendrix Genetics, 5831 CK Boxmeer, The Netherlands; [10]Illumina, Inc., San Diego, CA 92121, USA; [11]Department of Animal Science, University of California, Davis, CA 95616, USA
*Corresponding author: hcheng@msu.edu

In 2004 the chicken genome sequence and more than 2.8 million single nucleotide polymorphisms (SNPs) were reported. This information greatly enhanced the ability of poultry scientists to understand chicken biology, especially with respect to identification of quantitative trait loci (QTL) and genes that control simple and complex traits. To validate and address the quality of the reported SNPs, assays for 3072 SNPS were developed and used to genotype 2576 DNAs isolated from commercial and experimental birds. Over 90% of the SNPs were valid based on the criterion used for segregating, and over 88% had a minor allele frequency of 2% or greater. As the East Lansing (EL) and Wageningen University (WAU) reference panels were genotyped, 1933 SNPs were added to the chicken genetic map, which was used in the second chicken genome sequence assembly. It was also discovered that linkage disequilibrium varied considerably between commercial layers and broilers; with the latter having haplotype blocks averaging 10 to 50 kb in size. Finally, it was estimated that commercial lines have lost 70% or more of their genetic diversity, with the majority of allele loss attributable to the limited number of chicken breeds used.

World's Poultry Science Journal, Vol. 64, June 2008
Received for publication October 14, 2007
Accepted for publication December 12, 2007

Keywords: SNPs; genetics; biodiversity; map

Introduction

The year 2004 was a historic one for biologists, and especially the chicken research community, as the first draft of the chicken genome was published (International Chicken Genome Sequencing Consortium, 2004). The 6.6X coverage of a UCD001 female Red Jungle Fowl (RJF) genome was the first complete description of an avian species genetic blueprint. Simultaneously, and equally important for chicken geneticists, was the identification of more than 2.8 million single nucleotide polymorphisms (SNPs) (International Chicken Polymorphism Map Consortium, 2004). This large dataset was generated by partial shotgun sequencing (1/4X genome coverage) of a Ross broiler male, an experimental White Leghorn female, and a Chinese Silkie female with comparisons of the sequence reads to those from the RJF. This comparative dataset provided a major leap towards understanding how genetic variation contributes and explains phenotypic variation.

Taking advantage of these achievements along with the introduction of high throughput and economical SNP typing platforms, a consortium submitted a USDA grant proposal in 2004 that was subsequently funded in the summer of 2005. This paper briefly documents the goals and current progress of this effort.

Goals of 2005 USDA consortia

The four goals in the proposal were provided, as shown below:

1. Genotype ~3,000 SNPs that were evenly spaced throughout the entire genome on 60 birds each from 24 elite commercial lines (1,440 individuals), and parents of specific resource populations. This objective was to determine if the SNPs were valid and if yes, then whether they were still segregating (and selectable) within a line or informative in a resource population.
2. Using the same ~3,000 SNPs, genotype the East Lansing (EL) reference panel (4 parents plus 88 progeny) and two families of the Wageningen (WAU) mapping population (4 parents plus 88 progeny). As a result, thousands of genes could be added as genetic markers, to confirm and enhance the integration of the genetic map and the genome sequence, and flag problem areas.
3. Determine the extent of linkage disequilibrium (LD) within commercial populations using a different set of ~1,500 SNPs at 14 genomic locations. It is necessary to know the minimal marker spacing for whole-genome LD QTL (quantitative trait loci) mapping on non-pedigreed populations.
4. From the data collected, it was determined what subset of SNPs would allow traceability of poultry meat or live offspring to its pure line parent, and by using simulations, examined if the usefulness of these markers were adequate for applications with genome-wide marker-assisted selection (GMAS).

As shown below with one example, given the rich dataset, it was possible to extend the analyses to other areas not specifically proposed.

STATUS OF GOALS 1 AND 2

Genetic maps provided the foundation to identifying, tagging, and cloning genes for simple and complex traits. Genetic maps also provided the framework for the whole

genome sequence assembly. The utility of a genetic map is influenced by the number (saturation) and types of marker employed, *i.e.*, genetic maps with many markers that are highly informative are more powerful. The International Chicken Polymorphism Map Consortium (2004) identified over 2.8 million SNPs. However, prior to widespread usage, this *in silico* database needed to be validated, which was the main reason for goals 1 and 2.

To achieve these two goals, a very large collection (2,576) of DNA or blood samples was obtained in the summer of 2005. The samples included:

- Reference panel DNAs (184 in total) – an expanded panel of the East Lansing (EL) reference family (92) and 2 families of the Wageningen (WAU) resource population (92)
- Elite commercial breeding lines (1440 total) – 4 commercial breeding companies each sent 9 lines with 40 individuals per line.
- GMAS panel (116 total) – commercial individuals from several generations with complete pedigree and phenotypic trait measurements.
- Traceability panel (78 total) – commercial individuals from 8 generations with complete pedigree information.
- SNP controls (4) – the original DNAs used in the RJF sequencing or SNP discovery process.
- Lines of special interest (730 total) – a collection of experimental birds from genetic lines of interest, various parents or individuals from experimental and commercial resource populations, etc.
- Individuals contributed from others paid by the US Poultry Genome Coordinators (184 total) – most were parents or progeny of experimental or commercial resource populations.

To obtain SNPs that were evenly-distributed in terms of physical distance) throughout the chicken genome, the sequence (WASHUC1.0) was divided into 3,072 bins, taking into account the recombination rate per chromosome. For each bin, three SNPs from the greater-than 2.8 million SNP dataset were selected. Preference was given to high confidence SNPs in genes, especially those judged to be tolerant coding non-synonymous (cn) SNPs. These cnSNPs result in an amino acid change, and those judged to be 'tolerant' do not deleteriously disrupt the protein based on the SIFT algorithm (Ng and Henikoff, 2001). The reason for selecting cnSNPs was that it was hoped that these polymorphisms might be the cause of phenotypic variation, and not just genetic markers. Furthermore, while it would have been desirable to identify SNPs that influence transcription, these are very difficult to identify based on sequencing information alone. All SNPs were evaluated for the Illumina platform (not every SNP can be converted into an assay suitable for genotyping), and a single suitable SNP was selected for each bin. In addition, 34 SNPs in genes of interest were evaluated.

Following preliminary tests with a limited set of DNA samples, all the samples were delivered to Illumina in bar-coded 96-well plates on August 4, 2005. On October 1, 2005, the results were returned to the investigators via the internet.

The main conclusions of the initial results were:

- All but 14 samples could be genotyped (giving a >99% success rate). To put this into perspective, a number of samples either had no measurable DNA, as determined by both Hoechst and PicoGreen dyes. And unfortunately, at least one foreign shipment was delayed in a US customs warehouse for over 2 weeks. Furthermore, it was interesting to note that the Gentra kits used by some investigators produced single stranded (ss) DNA as judged by the low PicoGreen dye measurements (which binds double stranded (ds) DNA only), yet gave high Hoechst dye readings (which binds

both ssDNA and dsDNA). Nonetheless, the Illumina system was robust enough to amplify almost all the DNAs and provided reliable results.

- 2,733 out of the 3,072 SNPs worked giving an 89% success rate. Illumina reported that they have a 30-50% success rate for non-validated SNPs (which fitted the current scenario) and a 90-95% success rate with assays previously genotyped with Illumina; a new mix of SNPs may have 5-10% failure rate due to multiplex amplification issues.
- Based on plate and other controls, the reproducibility rate was 99.996%. In other words, the called genotypes were virtually all correct. This is not the situation with other types of genetic markers, *e.g.*, microsatellites. The highly accurate call rate was a major asset in the generation of the genetic maps as they helped resolve errors with other markers.
- Approximately 90% of the time (2428 of 2706 comparisons), the *in silico* SNPs were verified in the control DNAs. This agreement rate was mainly dependent on the SNP category (tolerant cnSNPs agreed approximately 82% while the others were 95% or higher) and somewhat on the chicken strain originally used to identify the SNP. Interestingly, while it was reported that only one Ross broiler was used, our results suggested this might not be the case, as a high percentage of the disagreements were associated with the broiler control. This suspicion was later confirmed wherein it was determined that the Roslin Institute had sent the Beijing Institute of Genomics two different DNA samples.
- Of the 2,733 working SNPs, 2,416 (>88%) had a minor allele frequency of 2% or greater across all populations, which means that they are more likely to be useful in other populations (only 182 SNPs or 6% were not segregating, which means that these were not SNPs). A highly relevant conclusion is that one can be reasonably confident that the *in silico* SNPs identified by the Beijing Institute of Genomics are real.

With respect to the genetic map, 1933 SNPs were mapped to one or both reference panels. With many of the genetic markers in common between the EL and WAU genetic maps, a single consensus map allowed for the placement of 3850 markers. This consensus map was used in the generation of the second genome sequence assembly. Of the 3,072 SNPs screened in our panel, 233 changed chromosomal positions from the first assembly (WASHUC1.0) and the second (WASHUC2.1). This shows that while the genome sequences are a major advance, they still have a number of problems including gaps and assembly problems.

In conclusion, goals #1 and #2 have been completed and had given very satisfactory results.

STATUS OF GOAL 3

The primary driving force in the first two goals was to characterize and obtain more genetic markers. These markers, in turn, could be applied towards the identification of QTL. The power of genome-wide QTL scans is dependent on the coverage of the genetic markers. More specifically, genetic markers should survey the entire genome and be able to provide adequate density to query the average extent of LD in the resource population of interest. This is why current resource populations are usually derived from F2 or other defined groups that limit the LD, as microsatellite coverage does not normally extend beyond 5-10 cM on average. However, with the large number of SNPs and the extensive coverage that this type of marker provides, it should be possible to screen other populations such as elite commercial lines that likely have much smaller amounts of LD.

Previously, a preliminary study to examine the extent of LD in chicken was undertaken

as part of a collaboration of Wageningen University and Utrecht University in the Netherlands. The design of the experiment to determine LD in commercial lines was by sampling genomic regions with relatively high and low recombination rates, and macrochromosomes 1 and 2 (1 cM ≈ 340 Mb) and microchromosomes 26 and 27 (1 cM ≈ 70 Mb) were selected for this purpose. For each chromosome, two separate regions were interrogated with SNPs at a spacing of 1 per Kb, which we hoped was sufficiently dense to get an accurate determination of LD, with 150-300 Kb total coverage per region. These SNPs were genotyped on a white egg line, a brown egg line, a male broiler line, two female broilers lines (one open, one closed line), two traditional Dutch breeds, and a wild chicken (*G. gallus gallus*) population.

The preliminary conclusions were:

- LD varied considerably between breeds; the white layer has the highest LD (and lowest heterozygosity) and the open broiler line the lowest extent of LD.
- Size of haplotype blocks in broilers were similar to those found in human (10-50 kb)
- LD was higher on macrochromosomes

For more information on a similar study that examines LD in chicken, see Aerts *et al.* (2007).

Based on these results, it was ascertained that dense sampling of at least one SNP every 2 Kb was needed. However, theoretical and empirical studies on other species suggested it was likely that effects of high selection pressure on a genomic region could extend over many cM. To address this possibility, a 'genomic transect' model of SNP selection was devised that allowed combination of high density typing, but still spanning a genomic region of 1 Mb.

Fourteen regions in the chicken genome were selected, 10 of which were selected near genes that were implicated in other studies as candidates for production traits (*e.g.*, fertility, growth, muscle development). Two regions were selected near genes for feather colour, which were renowned for high selection in both commercial and traditional breeds. One region on GGA28 was selected because previous research indicated a possible signature of selective sweep. And finally a 'gene desert' on GGA11 was chosen to serve as a control and also to investigate differences in recombination rate with adjacent gene-rich areas.

Genotyping was conducted on 18 commercial chicken lines. These lines, presumed to be under very high selection pressure for production traits, were compared to a number of chicken populations that were under no or mild selection during the same period. These included two wild Red Jungle Fowl populations, and six old Dutch (traditional) breeds. Furthermore, two experimental lines selected for negative growth were included. A selection of 11 smaller population samples from a few additional Dutch breeds and breeds from Uganda, China, Pakistan, and Bangladesh were also genotyped to ascertain geographic distribution of SNPs and evaluation of SNPs deposited in dbSNP. Finally, twelve Ceylon Jungle Fowl (*G. lafayettei*) were included for out-group comparison and ascertainment of the ancestral state of the SNPs.

Initial results indicated that selective sweeps can be identified on several genes. However, only for one gene does this appear unambiguously in commercial breeds vs. non-selected breeds. In other genomic regions, the pattern of heterozygosity or LD could be interpreted as resulting from a selective sweep but are difficult to discern from stochastic effects such as population demography. We are currently working on more detailed statistical analysis that should allow us to better interpret effects of selection.

Biodiversity

Due to the large database of information that was collected, it was possible to address a number of other interesting questions. Among these, the most interesting was whether, following decades of intense selection primarily for meat or egg production, and competition that has reduced the industry to a few multi-national companies, does sufficient genetic diversity exist as 'insurance' to address potential future needs, such as new or emerging diseases? To address this question, allele loss was assessed by determining the relationship with inbreeding, or through the use of SNP "weights" that correct for ascertainment bias and avoided the need to extrapolate from the inbreeding coefficient. This type of analysis was not possible previously with microsatellite markers.

Results from using both methods indicated that individual commercial breeding lines have lost 70% or more genetic diversity of which only 25% of this loss can be recovered by combining all stocks of commercial poultry (Muir and Cheng, 2007). However, and interestingly, this did not mean that modern agricultural practices were the primary source of this allele loss as, in fact, the majority of the alleles were lost prior to the formation of the current industry. These results emphasize the need for concerted national and international efforts to preserve chicken biodiversity.

Conclusions and future direction

'Disruptive' technologies have and will continue to play an important role in the technology-driven field of genomics. The impact has been previously seen with PCR-based genetic markers, automated DNA sequencers, DNA microarrays, etc. and now high throughput SNP typing. Furthermore, the accuracy and throughput of these systems is advancing greatly. For our study, in a 2 month period, more than 2,500 DNA samples were genotyped with more than 3,000 SNPs resulting in >7.5 million data points with near 100% accuracy at a cost of US 4.2 cents per data point. While impressive, improvements have already been implemented to make it possible to run 60,000 or more SNPs at 1 cent or less per data point on a single DNA sample.

The implication of this advancement alone means that genotyping is probably not the rate limiting step. Rather it is in the generation of sufficient number of animals with accurate trait measurements. Breeding companies are well positioned in this aspect, and it will be interesting in the coming years to observe how they incorporate this new wealth of genetic data in practical poultry breeding situations.

Besides genotyping, advancements in DNA sequencing and transcript profiling are also being made. Certainly there is a place for these technologies to be incorporated and particularly in the identification of the causative genes for QTL. The main challenge will be when and how to properly analyze all the information, incorporate the results, and integrate all these methods.

Acknowledgements

We would like to thank Laurie Molitor, Tom Goodwill, and a large number of individuals that made this project possible by providing their technical and logistical support. This work was supported in part by USDA NRICGP #2004-05434.

References

AERTS, J., MEGENS, H.J., VEENENDAAL, T., OVCHARENKO, I., CROOIJMANS, R. GORDON, L., STUBBS, L. and GROENEN, M. (2007) Extent of linkage disequilibrium in chicken. *Cytogenetics and Genome Research* **117**: 338-345.

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

INTERNATIONAL CHICKEN POLYMORPHISM MAP CONSORTIUM (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* **432**: 717-722.

MUIR, W. and CHENG, H. (2007) A world wide and genome wide assessment of biodiversity in commercial poultry populations. Proceedings of the National Breeders Roundtable. *In press*.

NG, P.C. and HENIKOFF, S. (2001) Predicting deleterious amino acid substitutions. *Genome Research* **11**: 863-874.