High-density SNP Genotyping Analysis of Broiler Breeding Lines

A.S. Leaflet R2219

Abebe Hassen, associate scientist, animal science;
Jack C. M. Dekkers, professor of animal science;
Susan J. Lamont, distinguished professor of animal science;
Rohan L. Fernando, professor of animal science
Collaborators
Santiago Avendano, senior geneticist, Aviagen Ltd.;
John Ralph, genomics project manager,
Aviagen Ltd.;
Alphons Koerhuis, head of R&D, Aviagen Ltd.;
Jim McKay, science and technology coordinator, EW Group GmbH;
William G. Hill, professor, University of Edinburgh, U.K.

Summary and Implications

The purpose of the genomic initiative project that is described herein is to use high-density single nucleotide polymorphism (SNP) marker genotypes on sires from commercial broiler lines and mean performance of their progeny to identify regions of the genome that are associated with economic traits. This report addresses some of the methodological aspects considered during analysis and interpretation of the data. Considering the large amount of phenotypic and SNP data, the data were analyzed separately for each line using mixed models procedures. Results were then pooled across lines to determine overall significance. Preliminary results showed important association of several genomic regions with mean progeny performance. But, the degree of significance was influenced by the size of heritability value used. Results from this project will be an important step towards implementation of marker-assisted selection in commercial broiler breeding programs.

Introduction

In today's poultry industry, the parents of the next generation are selected using breeding values estimated from mixed model analysis of phenotypic records, along with pedigree information. The fact that selection programs have been able to sustain rapid genetic progress for growth and feed efficiency during the past decades suggests that the traits under selection are affected by many genes. However, many traits of interest in poultry production, e.g. health and survivability, have low heritability or are difficult and/or expensive to measure. For such traits, it is well known that molecular data in the form of genotypes for markers that are linked to so-called quantitative trait loci (QTL) can be an important source of information to bring about favorable genetic changes through marker assisted selection (MAS). By enabling an assessment of the extent to which an individual carries the good or bad genes for a trait based on a DNA sample, MAS can be used to make early selection decisions on traits that are difficult to improve otherwise. This, however, requires information on the location of QTL in the genome and on estimates of their effects.

Despite the amount of selection pressure applied to commercial poultry, high levels of genetic variation still exist at both the trait level and at the DNA level in the form of single nucleotide polymorphisms (SNP), i.e. individuals in a population differ in the nucleotide that they carry at many positions in the DNA sequence. And, indeed, with the genome of the chicken now fully sequenced, a recent study identified around 2.8 million positions across the genome at which modern breeds of chicken differed from their common wild ancestor, the Red Jungle Fowl. It is likely that individuals within modern breeding lines will differ for a good portion of these SNP. Today, genome-wide scans through linkage or association analyses are used to localize genes that have significant contributions to performance differences and/or to estimate the size and direction of such effects. These types of studies are timely and justifiable because current advances in molecular techniques have made routine genotyping of selection candidates for large number of SNP more affordable.

The genomic initiative project is a collaborative effort between Iowa State University and Aviagen, Ltd. with the main objective to identify regions of the chicken genome that influence economic traits in broilers. The project was started in November 2005 and involves several phases, including preliminary evaluation of phenotypic and SNP information, analysis of data, summarizing relevant information, validation of results, and development of strategies to incorporate important markers in routine genetic evaluation and selection programs. The purpose of this report is to present some of the methods that were used in data collection, data analysis, and interpretation of results.

Materials and Methods

Source of data

Data for this study were provided by Aviagen Ltd. and included measurements on economic traits from ten lines that are part of the commercial breeding program of Aviagen Ltd. Phenotypic data used for analysis were the mean progeny performance on 10 traits of up to 200 sires from each line, with the record of each progeny adjusted for fixed and relevant random effects. Traits included growth and survivability traits in high and low hygiene environments, along with breast yield and feed efficiency. The sires were genotyped for up to 6,000 SNP that were chosen to cover the genome from the 2.8 million SNP that were identified in the chicken genome sequencing project.

Pedigree information used in the analysis included relationships between sires spanning four generations.

Analysis of data

The purpose of the statistical analyses of the data was to identify which regions of the genome are associated with performance. This was done by evaluating associations of the genotype of each sire at each of the 6,000 SNP with his progeny mean for each trait in each of the 10 lines. Because of the large number of analyses required (6,000 SNP ∗ 10 traits ∗ 10 lines = 600,000 analyses), some automation of the analysis procedures was necessary.

Initially both phenotypic and genotypic data were evaluated based on simple statistics. Sample means, standard deviations, and scatter plots were used to assess data distributions and to identify possible extreme values. This was followed by evaluation of alternative data analysis procedures, choice of the most appropriate statistical software, development of computer programs in order to implement and automate necessary edits and analyses, and presentation of results. Some of the most important issues considered during development of data analysis protocols were the following:

1.*Fixed vs. mixed models procedures*. Sires used in the study are samples from a commercial population that is under selection. Evaluation of the pedigree structure in the lines showed that some sires belong to groups of several half-sib families. Therefore, the assumption of sires being unrelated was unrealistic and mixed model procedures that account for relationships among sires were used. Thus, the basic model used for analysis of the association of the genotype of a sire at a given SNP with average performance of his progeny for a trait was:

$$y_i = \mu + \beta_j g_{ij} + s_i + e_i$$

where y_i is the adjusted progeny mean of sire i, g_{ij} is the number of "1" alleles that are carried by sire i at SNP j; β_j is the allele substitution effect for SNP j (to be estimated); s_i is the random effect of sire i (with variance-covariance structure $\mathbf{A}\sigma_s^2$ across sires, where **A** is the relationship matrix and σ_s^2 is ¼ of the additive genetic variance of the trait), and e_i is a residual with variance equal to $^1/_n\sigma_e^2$, where n is the number of progeny included in the mean and σ_e^2 is the residual variance. This analysis was conducted for each SNP (6,000), each trait (10) and for each line (10), for a total of 600,000 analyses. Significance of each SNP was tested using a likelihood ratio test, comparing the likelihood of the fitted (full) model to that of a reduced model that did not include the SNP effect.

2.*Single- vs multiple-SNP analyses*. Unlike results from single-SNP analysis that were just described, an analysis that includes the effects of multiple SNP in a region can capture information from the entire region. Furthermore, information on a given region can be compared between different lines; which may not be always possible for a singe-SNP analysis, because a given SNP may be fixed in one or more lines. Furthermore, results from simulation studies in our group have shown similar statistical power for multiple marker and haplotype analyses, suggesting that multiple marker analyses may substitute for the latter. Thus, the following three-SNP analyses were also run for each position in the genome (6,000) and for each trait (10) and each line (10), for another set of 600,000 analyses:

$$y_i = \mu + \beta_{j-1} g_{i,j-1} + \beta_j g_{i,j} + \beta_{j+1} g_{i,j+1} + s_i + e_i$$

where $g_{i,j-1}$ and $g_{i,j+1}$ represent the number of "1" alleles that the sire carries at the SNP up- and down-stream from SNP i.

3.*Source of heritability and standard errors of SNP effects*. The mixed model analyses described above require an estimate of heritability of the trait. Because the sires that were included in the analysis were those that were selected as parents in the breeding program during a given time period, the heritability in our data set does not necessarily represent the heritability in the entire population. Estimation of heritability from the data set was, however, hampered by the limited number of generations included in the data compared to the entire population. Thus, the impact of alternate levels of heritability on results of the analyses was investigated. It was found that the level of heritability used in the analysis has limited impact on estimates of the SNP effects (β) but did affect estimates of standard errors of the β and, thereby, the level of significance (P-values). Results showed that standard errors increased with heritability, resulting in changes in the distribution of P-values, ranging from highly skewed, with many significant P-values, to nearly uniform. There was, however, a strong linear relationship between differences in model likelihood values and heritability. This relationship allowed p-values for alternate heritabilities to be computed by linear approximation, rather than requiring all 1.2 million analyses to be repeated for different levels of heritability.

4.*Combining results across line*. In order to limit the amount of data to a manageable size, data were analyzed separately by line. This analysis approach also allowed analysis of data based on line-specific heritabilities. Results obtained from individual lines were then combined to determine the overall significance of segregating SNP effects. This combining across lines was accomplished by summing the likelihoods of the full model for each line and comparing this to the sum of the likelihoods for the reduced model fitted to each line. This procedure allowed combining results across lines without having to repeat the analyses or to conduct a joint analysis across lines.

5.*Presentation of results.* With so many analyses conducted (greater than 1 million), presentation of results is crucial. To enable interpretation of results, several summary graphs

were created that plotted levels of significance across chromosomes, as well as plots of estimates of effects and allele frequencies. Plots were also created that summarized the most significant results across lines and traits.

Results and Discussion

Data analysis for some of the traits is currently complete. Both single- and three-SNP mixed model analyses were used to evaluate association based different heritability values. The percentage of the 6,000 SNP that were not fixed ranged from 63 to 82% in the ten lines. These results, and the gene frequencies at the segregating loci, confirm the presence of a substantial genetic variation at the molecular level in these lines.

Preliminary results showed promising associations of several regions with mean progeny performance. Figure 1 shows a typical chart of significance by SNP positions. Regions with higher levels of significance indicate associations of SNP genotypes with mean progeny performance. However, due to the large amount of results, several statistical tools and additional visual aids were used to ascertain importance of genomic regions. Although results from the combined analysis across lines provide information on overall significance, association results at each SNP were also assessed in relation to gene frequencies as well as consistency across lines and traits.

Heritabilities estimated from the sample were often smaller than those calculated from the entire population. This is expected considering that data for this analysis were from progeny of selected sires. When the sample heritabilities were used in the analysis, the level of significance was often higher than when the higher heritability estimates from the entire population was used.

The results from these preliminary analyses suggest that through routine genotyping of breeding stock for validated SNP, marker assisted selection can be used to further genetic progress in broilers for traits such as survivability and feed efficiency, which are difficult to improve using standard selection programs.

Figure 1. Levels of significance for the association of sire single nucleotide polymorphism (SNP) genotypes across an example chromosome with mean progeny performance for an example trait. The SNP with high significance indicate regions that contain genes that affect the trait.