Linking phenotypic with genomic diversity in the Synbreed Chicken Diversity Panel

H. Simianer[1,], Ngoc-Thuy Ha[1], U. Janßen-Tapken[2], U. Ober[1], M. Erbe[1], A. Weigend[2], and S. Weigend[2]*

[1]Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August-Universität Göttingen, Germany
[2]Institute of Farm Animal Genetics, Friedrich-Loeffler-Institut, 31535 Neustadt-Mariensee, Germany
*hsimian@gwdg.de

Within the framework of the SYNBREED project, a large panel of breeds was characterised, comprising more than 2000 individuals from more than 120 breeds and colour variants, covering a large fraction of breeds kept by fancy breeders as well as most diverse geographical origins, wild ancestors and commercial lines. In this Synbreed Chicken Diversity Panel (SCDP) all animals were genotyped with the newly developed Affymetrix chicken 600k Axiom-SNP-array, and a large proportion of animals was phenotyped according to a standard phenotyping protocol. For the present study we used a subset of the SCDP comprising 1810 adult individuals from 116 populations. After quality control (call rate per SNP >99%, call rate per individual >95%, MAF >5%), 311,006 polymorphic SNPs were used. The objective of this study was to use this diverse set to map SNPs, genes and pathways associated with phenotypic variability in the sample by applying appropriate statistical strategies.

We illustrate our approach with a subset of phenotypes representing the trait complex body size. For all animals, the following phenotypic measures were available: WL: mean of wing length left and right in cm; SL: mean of shank length left and right in cm; ST: mean of shank thickness left and right in mm; KL: keel length in cm; LW: live weight of chicken in g. Since all measured phenotypes reflect some aspects of the size of a bird, we derived a combined measure based on a principle component analysis with all five phenotypic traits and used the first principle component which accounts for 88% of the total variation as additional combined phenotype, which was termed PC.

Genome wide association studies (GWAS) were performed for SL, KL, LW and PC. To account for the population structure of the studied sample, a principal component analysis with the genotype data was performed revealing the first 221 principal components as significant ($p<0.01$) in a Tracy-Widom test. These 221 principal components as well as the sex of the birds were included as covariates in all models. We compared two alternative GWAS strategies:

a) a conventional single marker regression (SMR), in which each of the 311,006 SNPs or 141,425 intragenic SNPs, respectively, was tested;
b) a gene-based score test (GBST) based on a list of all known genes for the species *Gallus gallus* from Ensembl Genes (www.ensembl.org, release 72). SNPs were assigned to genes according to the physical transcription start and end positions. In total, 15,068 genes were available with 11,701 genes containing between one and 252 SNPs. In GBST, a single weighted score statistic for each gene combining all SNPs in that gene was calculated and tested following the approach suggested by Pan (2009).

With both tests, a Bonferroni correction was applied to account for multiple testing. Note that with the SMR each single test is by more than an order of magnitude more conservative than in the GBST, due to the difference in the number of tests (311,006 SNPs *vs* 11,701 genes). We also applied the False Discovery Rate (FDR) approach by Benjamini and Hochberg (1995) which controls the proportion of false positive signals among all positive signals and is known to be less conservative than the Bonferroni correction.

Table 1 shows the number of significant genes that were obtained with the variable combination of approaches. In the SMR, both the total number of SNPs and the subset of SNPs located in known genes were used. The gene counts reflect the number of genes that were found significant with at least one of the analysed phenotypes.

Table 1. Number of significant genes detected with different mapping approaches (SMR *vs* GBST), different testing principles (Bonferroni correction *vs* FDR) and different sets of SNPs and genes used.

	Number of SNPs/genes	$\alpha<0.05$ after Bonferroni-correction	FDR<0.05
SMR	all 311,006 SNPs	2	3
	all 141,425 SNPs in genes	2	12
GBST		7	76

The results show that the GBST unveils a much larger number of genes associated with the analysed genotypes, especially with the less conservative FDR criterion. The advantage of GBST is caused by two mechanisms: (i) it combines all signals within a gene and thus is able to detect a gene in which, say, several SNPs in a SMR are almost significant, while the combined signal exceeds the significance threshold, and (ii) due to the smaller number of genes compared to the number of SNPs, the Bonferroni or FDR correction is less conservative. On the other hand the GBST has the disadvantage that it cannot detect signals in intergenic regions, although it is well known that polymorphisms in regulatory sequences may have a large impact on the functionality of genes. Among the genes found to be significant are CDKAL1 (CDK5 regulatory subunit associated protein 1-like 1, for LW) and UQCC (ubiquinol-cytochrome c reductase complex chaperone), which both can be functionally linked to growth performance in various species.

With the p-values for all genes obtained with the GBST in all four traits (SL, KL, LW and PC), a gene set enrichment analysis was conducted following the idea of Subramanian et al. (2005) which was shown to be efficient compared to other approaches according to results of Hung et al. (2011). This approach is based on a ranked list (ordered by increasing p-value) of all genes from which an enrichment score for a given pathway is calculated. The empirical null distribution of the enrichment score was determined through an extensive permutation of the phenotypes (n=5000 permutations for each of the 141 described KEGG pathways and phenotypes, respectively). For KL, the most significant pathway (p=0.00099) is pathway *gga00040: Pentose and glucuronate interconversions*, which plays a central role in the carbohydrate metabolism where a number of genes at various positions in the pathway (marked in yellow in Figure 1) have contributed to the significance. In general, the majority of identified pathways are linked to the carbohydrate metabolism or to steroid hormone synthesis, both of which have obvious links to growth and body size.

The Synbreed Chicken Diversity Panel represents a valuable resource for the high resolution analysis of phenotypic variability in the entire within species diversity of Gallus gallus. The suggested analytical approach is an efficient way of retrieving relevant genes and pathways in this complex data set.

Acknowledgements

This research was funded by the German Federal Ministry of Education and Research (Bonn, Germany) within the AgroClustEr "Synbreed – Synergistic plant and animal breeding" (Funding identification: 0315526).

References

Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57:289-300.

Hung, J.-H., T.-H. Yang, Z. Hu, Z. Weng, and C. DeLisi. 2011. Gene set enrichment analysis: performance evaluation and usage guideline. *Briefings in Bioinformatics* 13:281-291.

Pan, W. 2009. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology* 33:497–507.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. 2005. Gene set enrichment

analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 102:15545-15550.

Figure 1. Kegg-pathway *gga00040*: *Pentose and glucuronate interconversions*, the most significant obtained for the trait KL, and the genes (marked in yellow) that contribute most to its significance (Pathway downloaded from http://www.genome.jp/kegg/kegg2.html).