The contribution of different functional SNP classes to genetic variation in global chicken populations

D.K. Malomane[1], C. Reimer[1], S. Weigend[2], A. Weigend[2] & H. Simianer[1]

[1]University of Goettingen, Department of Animal Sciences, Animal Breeding and Genetics Group, Albrecht-Thaer-Weg 3, 37075 Goettingen, Germany
dmaloma@gwdg.de (Corresponding author)
[2]Friedrich-Loeffler-Institut, Institute of Farm Animal Genetics, Höltystraße 10, 31535 Neustadt, Germany

Summary

To evaluate the contribution of different functional single nucleotide polymorphism (SNP) classes to genetic variation, we used a total 118,676 SNPs genotyped from 18 (3,231 individuals) global chicken breed types. We classified the SNPs into six genomic classes, estimated and compared allele frequency distributions and heterozygosity within breed types for the different SNP classes. There was no difference between the genic and non-genic classes in their contribution to genetic variation. Among the genic regions, allele frequency distributions showed that the missense sites were subjected to selection pressure. Overall the missense sites significantly contributed less to genetic variation than the other regions.

Keywords: Chicken diversity, functional annotation, SNPs

Introduction

Different genomic regions in different breeds might be subject to evolutionary differences resulting in various phenotypes within a species. Downing *et al.* (2010) used derived allele frequency (AF) distributions at different genomic regions to investigate evolutionary differences of gene classes in some breeds of chickens. The differences in allele frequencies at the synonymous and non-synonymous sites could indicate if amino acid changes were neutral or favored under selection. In an event of selective advantage (adaptive Darwinian selection), a non-synonymous mutation will be fixed at a higher rate than a synonymous mutation (Yang, 2002). Defaveri *et al*. (2013) analyzed population genetic parameters including heterozygosity in genic and non-genic regions using SNP and microsatellite data. A number of other genomic studies in different species have explored different contributions of genomic regions to phenotypic (Salinas *et al*., 2016) or genetic variation and prediction of quantitative traits (Abdollahi-Arpanahi *et al*., 2016). However, there have been some contradictions in the research findings regarding the different contribution of genomic regions to genetic variations. We conducted this study to investigate if different genomic regions contribute differently to the variation in global chicken populations in order to help understand the different genomic sources of diversity in the chicken.

Material and Methods

The data used in this study consisted of 173 populations (3,231 individuals), genotyped with Affymetrix 600K SNP array, see Kranis *et al*. (2013) for details. This data set was collected under the umbrella of the SYNBREED project (www.synbreed.tum.de) and combined with samples of 2 Red Jungle fowl populations, *Gallus gallus gallus* and *Gallus gallus spadiceus* taken from previous EU project AVIANDIV, see Lyimo *et al*. (2014). Eighteen groups were established from these populations based on the chicken types.

Data editing and filtering

A total of 580,588 SNPs were obtained from the array. We discarded SNPs which were misplaced at wrong chromosomes and filtered the data for SNP call rates of >99% and animal call rate of >95% using SNP & Variation Suite Version (SVS) 8.1 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com). We retained only SNPs from the 28 autosomal chromosomes. We discarded SNPs that shared the same position on the same chromosome. SNPs were pruned based on linkage disequilibrium (LD) using the following parameters in SVS: *50 SNPs window, 5 SNP steps,* $r^2 > 0.2$. After these filters 123,273 SNPs were remaining for the analysis. LD based pruning of SNPs has been proven to effectively mitigate the effects of ascertainment bias when using SNP genotype data (Malomane *et al*., In Press).

Gene description and annotation

We annotated the SNPs based on Ensembl Gallus_gallus-5.0 variants and SNPs which were not found from the database were discarded, retaining 118,676 SNPs for further analysis. The SNPs were mapped to 6 classes of annotation sets, the intronic (53,070), synonymous (3,389), missense (950), nonsense (10), 3' and 5' untranslated regions (UTR=1,189). These classes together were combined to a genic (58,608) class and the remaining SNPs formed the non-genic (60,068) class. Since only 10 SNPs in our data set were found to be in the 'nonsense' class, we used them only in combination with other genic classes. We used a genic definition consistent with other literature which only considers the region between transcript start and end (Li *et al*., 2012). The downstream and upstream were considered non-genic as they don't fall within the gene boundaries (Aken *et al*., 2016).

Data analysis

We calculated the alternative allele frequency globally across genes at each locus for the different genomic regions. We converted the A and B alleles from the array into the reference and alternative allele, 1,659 SNPs did not have the information on whether the alleles were a reference or alternative, and these SNPs were not used in the analysis that involved AF calculations. Heterozygosity was also estimated at each locus and averaged across each genomic region within the different breed types. We used the "stats" package in R v3.2 (R Core Team, 2015) to test differences in proportions of SNPs at the different frequency bins and proportions of heterozygous loci between the various genomic classes.

Results and discussion

Allele frequency distribution in different genomic regions

There were no significant differences in AF between the genic and non-genic regions (Additional figure 1), but differences were observed when the genic regions were compared separately. We present in Figure 1 the comparisons only between missense and synonymous mutations in the exonic regions. The synonymous sites showed AF spectra similar to the overall (shown by genic and non-genic regions in Additional figure 1) spectra in all breed types, with all the local types showing a drop in AF at both edges of the spectra. Unlike the rest of the breed types, the layer lines had a similar distribution of AF between the missense and synonymous sites with only a very slight difference (non-significant, $p>0.05$) at the rare and common SNP bins, suggesting less/no variation at these exonic sites. Both crested types showed a skewed distribution (cf. "Difference" distribution line) at missense sites in comparison to the synonymous sites indicating a more rapid fixation of the alternative allele at the missense sites. Generally the proportion of SNPs on the edges of the AF spectra was higher at missense sites than at synonymous sites. Changes at the synonymous sites don't have functional implications, while they do at missense sites; therefore, this acceleration of missense changes indicates a possible selective pressure (beneficial or harmful). The other genic regions (intronic and UTR) behaved similar to the synonymous sites.

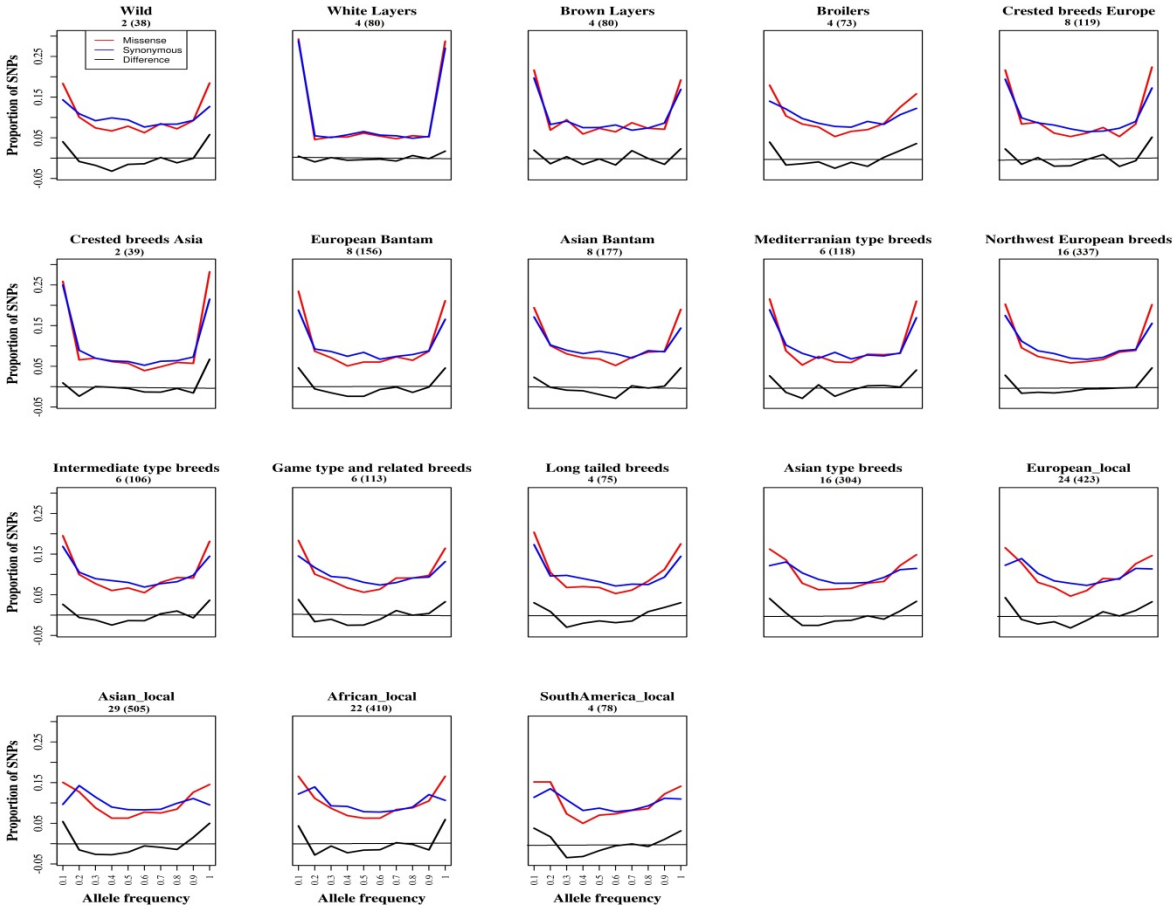


Figure 1. Comparison of the relative proportion of SNPs in allele frequency bins between missense and synonymous classes. Below the breed types' headings are the number of breeds, inside the brackets is number of individuals in each type. The 'Difference' frequency spectra show the difference between the missense and synonymous spectra relative to the zero line.

Proportion of heterozygous SNPs in genomic regions

There was no difference observed between observed heterozygosity (Ho) estimates in the genic and non-genic regions (Additional figure 2). Among the different genic sites (Figure 2), proportion of heterozygous (observed) SNPs was significantly ($p<0.001$) lower in missense sites than in all the other genomic sites except for White Layers which showed no difference in missense (0.123) and synonymous (0.125) sites (similarly to the distribution of AF at these exonic sites).

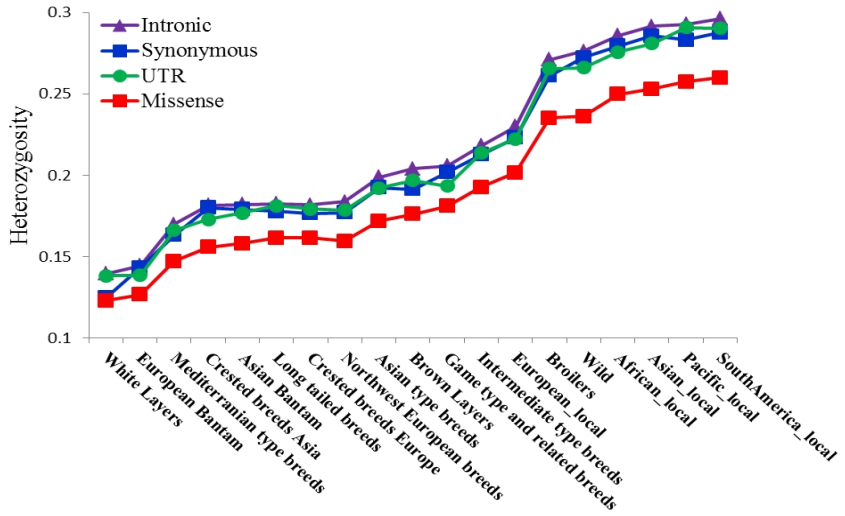


Figure 2. Comparison of observed heterozygosity estimates in the different genic regions.

Overall, different contributions of genic regions are mostly systematic in a sense that missense mutations encompass least genomic variability. With regard to the exonic regions, the results suggest that White Layers experienced evolutionary forces different from the other breed types. Even though they have shown extreme potential selection signals (very high peaks of rare and common SNPs and low Ho), such selection was neutral in the exonic regions. For studying the overall variation in these chickens, both genic and non-genic regions seem to be good representatives while considering only the coding sites or even just the missense regions to study diversity may reflect the amount of genetic variations inappropriately. Therefore, it is advisable to consider the genomic regions together.

Accounting for ascertainment bias

A subset of this data (used in the current study) was used in a previous study where we investigated the strategies to mitigate ascertainment bias when using SNP data. The study has proven that pruning of SNPs based on LD reduces the effects of ascertainment bias and if any, such effects become systematic across the populations and therefore doesn't lead to misinterpretation of differences across the populations (Malomane *et al*., In Press).

Acknowledgement

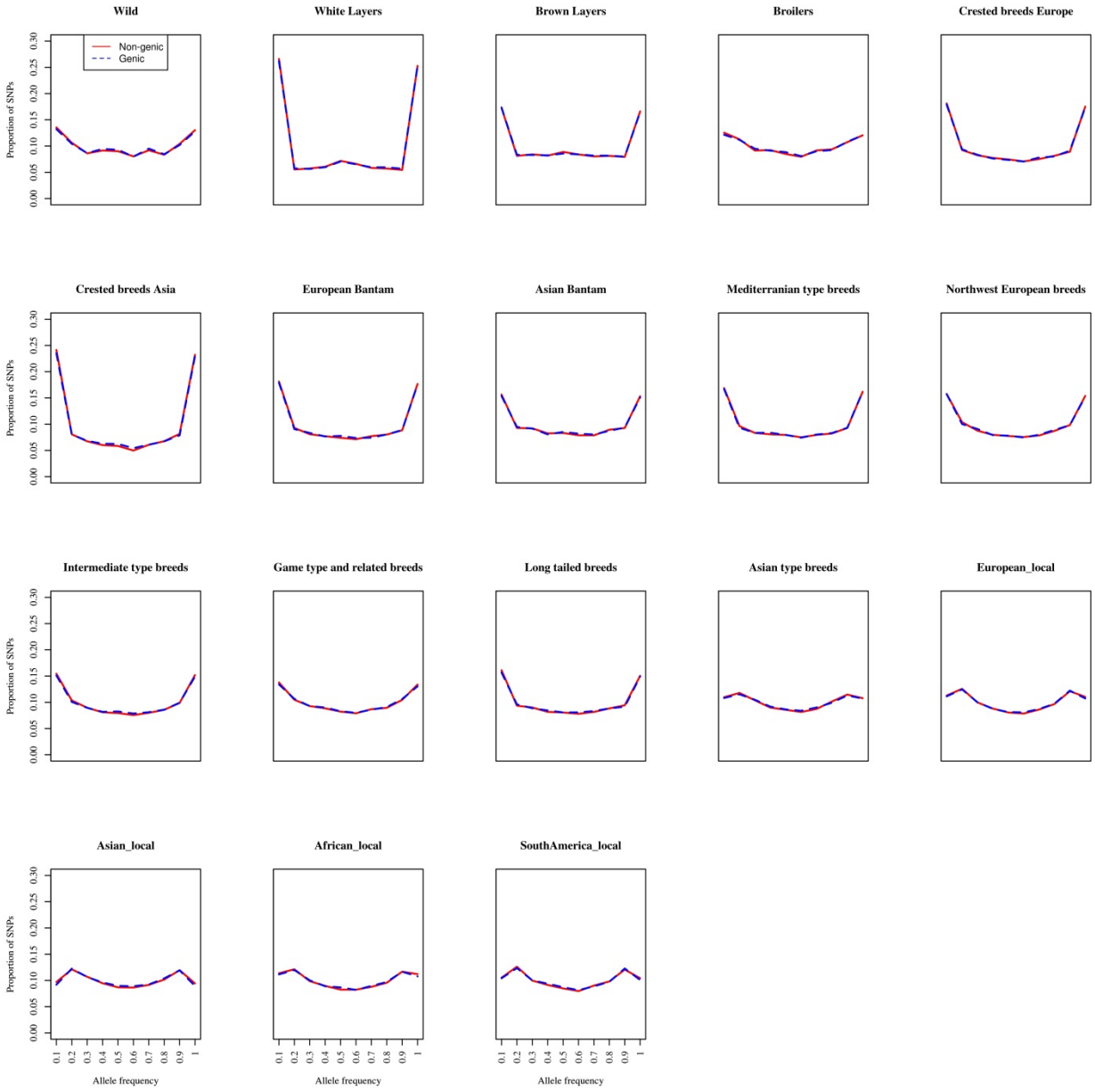
We acknowledge the German Federal Ministry of Education and Research (FKZ 0315528E) for funding the SYNBREED project. We are very grateful to all the breeders for their assistance in sampling. This work is part of a PhD project supported financially by the Erasmus Mundus.

List of References

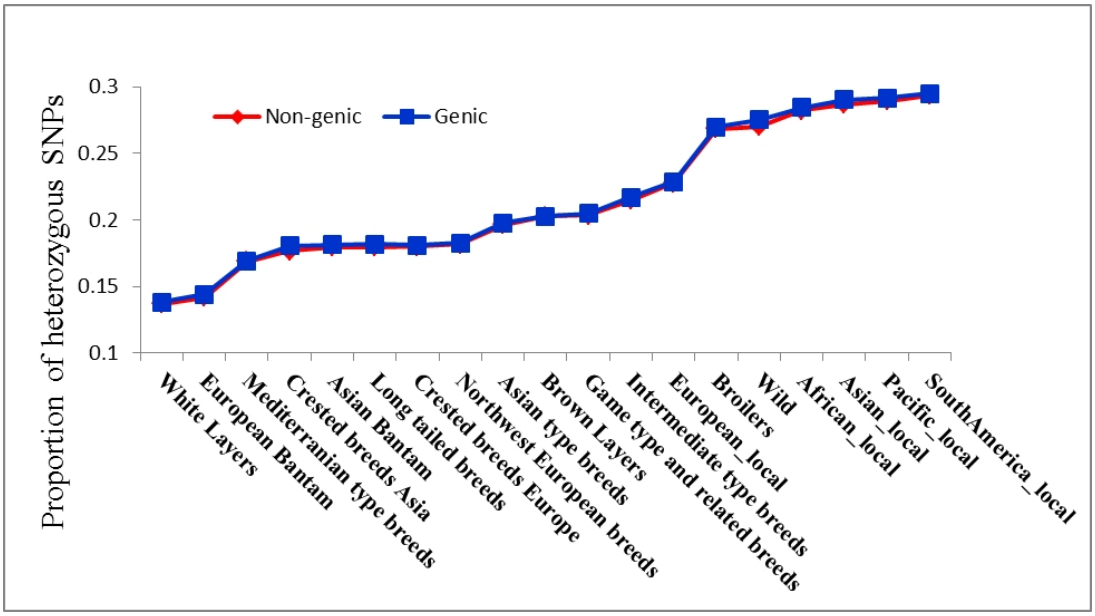
Abdollahi-Arpanahi, R., G. Morota, B. D. Valente, A. Kranis, G. J. Rosa, & D. Gianola, 2016. Differential contribution of genomic regions to marked genetic variation and prediction of quantitative traits in broiler chickens. Genet Sel Evo. 48(1): 10.

Aken, B. L., S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, et al., 2016. The Ensembl gene annotation system. Database. 2016.
Defaveri, J., H. Viitaniemi, E. Leder & J. Merilä, 2013. Characterizing genic and nongenic molecular markers: Comparison of microsatellites and SNPs. Mol Ecol Res. 13(3): 377-392.
Downing, T., A.T. Lloyd, C. O'Farrelly & D.G. Bradley, 2010. The differential evolutionary dynamics of avian cytokine and TLR gene classes. J. Immunol. 184(12): 6993-7000.
Kranis, A., A.A. Gheyas, C. Boschiero, F. Turner, L. Yu, S. Smith, et al., 2013. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 14(1): 59.
Li, X., C. Zhu, C. Yeh, W. Wu, E.M. Takacs, K.A. Petsch, et al., 2012. Genic and nongenic contributions to natural variation of quantitative traits in maize. Genome Res. 22: 2436-2444.
Lyimo, C. M., A. Weigend, P.L. Msoffe, H. Eding, H. Simianer, & S. Weigend, 2014. Global diversity and genetic contributions of chicken populations from African, Asian and European regions. AnimGenet. 45(6): 836-848.
Malomane, D. K., C. Reimer, S. Weigend, A. Weigend, A. R. Sharifi, & H. Simianer, In Press. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. BMC Genomics.
R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Salinas, F., C.G. de Boer, V. Abarca, V. García, M. Cuevas, S. Araos, et al., 2016. Natural variation in non-coding regions underlying phenotypic diversity in budding yeast. Sci Rep. 6(1).
SNP & Variation Suite ™ (Version 8.1). Bozeman, MT: Golden Helix, Inc. Available from *http://www.goldenhelix.com*.
Yang, Z., 2002. Inference of selection from multiple species alignments. Curr Opin Genet Dev. 12(6): 688-694.

Additional figures



Additional figure 1. Comparison of the relative proportion of SNPs in allele frequency bins between genic and non-genic classes.



Additional figure 2. Observed heterozygosity estimates in genic and non-genic regions.