Differentiation power of SNPs for assessing the phylogenetic structure of populations

C. Gärke[*], F. Ytournel[*], B. Bed'hom[¥], I. Gut[‡], M. Lathrop[‡], S. Weigend[§] and H. Simianer[*]

Introduction

The main advantages of single nucleotide polymorphisms (SNPs) compared to microsatellites are the ability for standardization and the low mutation rate (Fries and Durstewitz, 2001). Many studies compared microsatellites and SNPs for whole genome scans in humans, but the situation in livestock data remains to be studied. Because of the differences in population structure, mainly due to much smaller effective population sizes and directional selection in livestock populations, it is difficult to translate the results one to one from human to livestock. Differences in differentiation power of microsatellites and SNPs are thus expected.

In humans the number of SNPs with an equivalent information content of one microsatellite locus varies between 2.7 (Chakraborty et al., 1999), 3.75 (Thalamuthu et al., 2004), 4.29 (Krawczak, 1999), or 5.56 (Glaubitz et al., 2003). In poultry, Schopen et al. (2008) found, that the number of needed SNPs to compensate every microsatellite was not constant. With 12 microsatellites they needed 2.3 SNPs for each microsatellite. They ascertained that more SNPs would be required for an increasing number of microsatellites. Herráez et al. (2005) found that 2.65 SNPs matched one microsatellite in Galloway cattle.

The aim of this study was to assess the number of SNPs needed to reach the same differentiation power than 29 microsatellites. For that purpose, we first realized a Principal Component Analysis (PCA) and plotted each individual in a frame defined by the two first components. We then used Euclidean distances obtained from the first two coordinates of the individuals in the PCA-analysis to compare the differentiation power for both marker types.

Material and methods

Biological material: Eight different chicken breeds, Broiler Dam Line (BDL), Brown Egg Layer (BEL), Green legged Partidge (GLP), Godollo Nhx (GOD), Orlov (ORL), Padova (PAD), Rhode Island Red (RIR) and White Egg Layer (WL), were used to compare the application of microsatellites and single nucleotide polymorphisms (SNPs). Eight unrelated animals were chosen for each breed. Genotypes were available for 29 microsatellites and 2,931 SNPs. The chicken microsatellites are from the FAO-panel for poultry. All selected

[*] Animal Breeding and Genetics Group, Georg-August-University Göttingen, 37075 Göttingen, Germany
[¥] INRA, AgroParisTech, UMR1313 Animal Genetics and Integrative Biology, Jouy-en-Josas, France
[‡] Centre National de Genotypage, 91057 Evry, France
[§] Institute of Animal Genetics, Friedrich-Loeffler-Institute, Mariensee, 31535 Neustadt, Germany

SNPs were genotyped using the Illumina GoldenGate array. SNPs were distributed at random over the whole genome.

Statistical analysis: Principal components analysis (PCA) was used to reduce the number of variables to a small number of principal components. We used the statistic program R version 2.9.1 and the package ade4 (Chessel et al., 2004) to implement the PCA for different subsets: all microsatellite markers, all SNPs, or various subsets of SNPs. These different subsets were obtained in the following way. We chose at random 29, 100, 150, 200, 300, 400, 500, 1,000, 1,500, 2,000 or 2,500 of the 2,931 SNPs and repeated this step 100 times for each number of SNPs except the complete set. For each of the subsets, we calculated the pairwise Euclidean distances based on the first two principle components, which reflects the genetic difference between all animals. The distance between animal's *j* and *j'* were:

$$d_{j,j'} = \sqrt{(x_j - x_{j'})^2 + (y_j - y_{j'})^2}$$

Then, the distance of all animals within a breed *i* was:

$$d_{i.} = \sum_{j=1}^{7} \sum_{j'>j} d_{ij,ij'}$$

The distance between the animals of two breeds *i* and *i'*:

$$d_{ii'} = \sum_{j=1}^{8} \sum_{j'=1}^{8} d_{ij,i'j'}$$

Finally, the proportion of the total distance contained over all breeds (PtDoB) was:

$$PtDoB = \frac{\sum_{i=1}^{8} d_{i.}}{\sum_{i=1}^{8} d_{i.} + \sum_{i=1}^{7} \sum_{i'>i} d_{ii'}}$$

This parameter reflects the level of differentiation: the smaller it is, the better is the differentiation. This value was also computed for a particular subset of SNPs containing the 50 SNPs flanking the 25 microsatellites from which the position on the genome was known.

In order to check for the presence of a structure in the fixed data sets (all microsatellites, 50 flanked SNPs and all SNPs), we did a permutation test (Doerge and Churchill, 1996) with 10,000 replicates by assigning randomly each individual to a population and deduced the significance of the PtDoB value corresponding to a type I error rate of 5%.

Results and discussion

The first two principal components for the microsatellites described 24.2% of the variance (14.2% and 12.0%) for the microsatellites and 23.6% (15.4% and 8.2%) for the complete set of the SNPs (2,931 SNPs). Figure 1 illustrates the results for these two subsets (1A and 1D) as well as for one of the replicates with 29 SNPs (1B) and one of the replicates with 500 SNPs (1C). The more SNPs were used in the PCA, the better was the differentiation (Figure 1B, 1C, 1D): increasing the number of SNPs reduced the distances of the animals within the breeds. The PCA with the SNPs provided here a better differentiation of the eight breeds than the microsatellites, even with the same number of markers (Figure 1A *vs*. Figure 1B).

The mean PtDoB values ± the standard error from 100 replicates for the different subsets are plotted in Figure 2. The highest (worst) PtDoB-value was achieved with the microsatellites. Thus the observed values were lower with all sets of SNPs even with the same number of SNPs (SNP29). The more SNPs were used, the smaller are the PtDoB values and thus the better the observed differentiation. The value obtained with the microsatellites exceeds the upper boundary of the confidence intervals (under the hypothesis of a normal distribution,

i.e. two standard deviations) of all replicated data. Figure 2 also shows no difference between the 50 SNPs flanking the microsatellites and the subsets of 50 randomly chosen SNPs, so an impact of the chromosomal position of the microsatellite loci can be excluded.

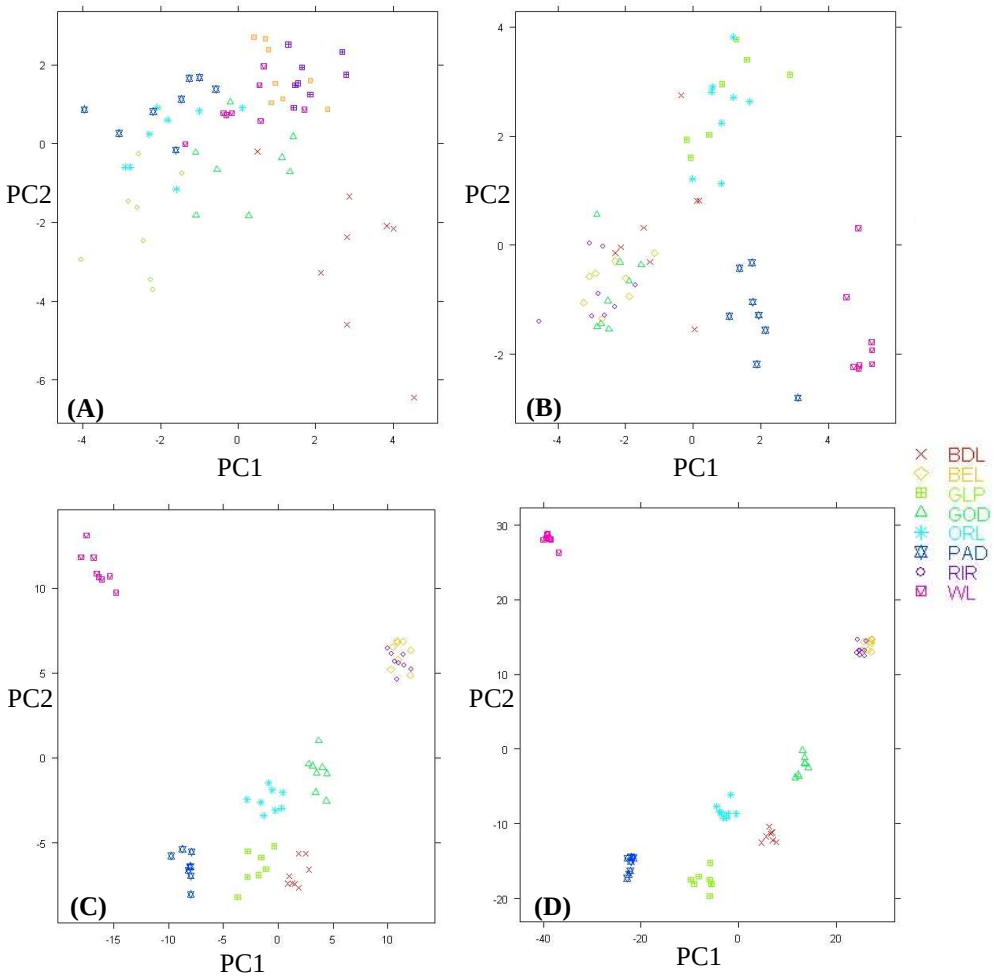

Figure 1: Plot of the first two components (PC1, PC2) of the PCA analysis for different subsets, using (A) the complete set of 29 microsatellites, (B) 29 SNPs at random, (C) 500 SNPs at random and (D) the complete set of 2,931 SNPs.

Table 1: Results of the permutation tests for the non-random marker sets.

Data set	Observed PtDoB	5% threshold	Mean PtDoB value
29 microsatellites	0.0554	0.1061	0.1111
50 SNPs	0.0330	0.1055	0.1112
2,931 SNPs	0.0062	0.1043	0.1111

Results from the permutation tests indicated that there was a structure in all data sets, even in those where it was not obvious when looking at the plots of the PCA analysis (e.g. for the 29 microsatellites, i.e. Figure 1A). Results are presented in Table 1.

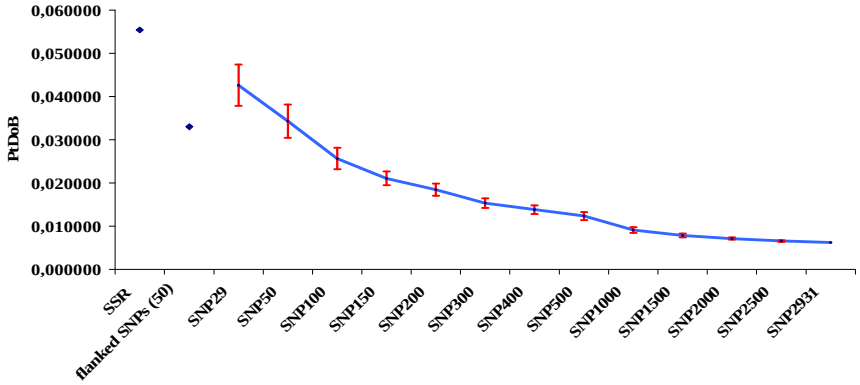

Figure 2: Calculated Part of the total distance contained over all breeds (PtDoB) for different subsets (Mean ± standard deviations for replicated samples)

Conclusion

In this dataset the 29 microsatellites did not provide a better differentiation than the same number of SNPs. The differentiation between breeds improved massively with an increasing number of SNPs. The results indicate that studies based on high throughput SNP genotyping will substantially improve the breed differentiation even with moderate numbers of SNPs.

Acknowledgment

We are grateful to Michele Tixier-Boichard and Leif Andersson for providing SNP data. DNA samples were taken from the chicken DNA bank established during the EC AVIANDIV project.

References

Chakraborty, R., Stivers, D., Su, B. et al. (1999). Electrophoresis, 20:1682-1696.

Chessel, D., Dufour, A.B. and Thioulouse J. (2004). R News, 4:5-10.

Doerge, R.W. and Churchill, G.A. (1996). Genetics, 142:285-294.

Fries, R. and Durstewitz, G. (2001). Nature Biotechn., 19:508.

Glaubitz, J.C., Rhodes, E. and Dewoody, J.A. (2003). Mol. Ecology, 12:1039-1047.

Herráez, D.L., Schäfer, H., Mosner, J. et al. (2005). Z. f. Naturforschung, 60c:637-643.

Krawczak, M. (1999). Electrophoresis, 20:1676-1681.

Schopen, G.C.B., Bovenhuis, H., Visker, M.H.P. et al. (2008). Anim. Genet., 39:451-453.

Thalamuthu, A., Mukhopadhyay, I., Ray, A. et al. (2004). BMC Genet., 6:27-31.