SpinNet: A New Tool To Study The Population Structure With A Genome-Wide SNP Survey

M. Neuditschko[*], J. Maxa[†], I. Russ[*], J. Schär[†] and I. Medugorac[†]

Introduction

Recent technical advances in single nucleotide polymorphism (SNP) chip technology have led to SNPs becoming the most developed and abundant markers in livestock science. However, current available applications of algorithms in population genetics turn out to be impractical due to intensive computational demand (e.g. STRUCTURE) (Price *et al.* (2006)) given the vast amount of data derived from thousands of individuals and thousands of markers. This study addresses this problem and introduces the idea of network analysis into the field of studies on population structure. Network theory describes the ability to sub-divide a network of nodes into community structures, which provides help in understanding and visualizing the structure of the respective network. To identify these community structures, many different approaches have been developed, including vertex similarity, the vertex degree gradient, the resistor network and the Potts Hamiltonian model. In this study, we have used an unsupervised clustering approach, so-called Super Paramagnetic Clustering (SPC) (Blatt *et al.* (1996)), which uses the Potts Hamiltonian model (Reichardt and Bornholdt 2004)) to identify community structures in population networks. Since SPC uses a so-called spin-spin correlation function to extract the community structure in networks, we call this method SpinNet. The objective of this study was to test a new method, which automatically identifies population structures without any prior ancestry information in panel bovine samples genotyped with a high density whole-genome SNP panel.

Material and methods

SNP genotyping. A total of 260 individuals representing six cattle populations were analyzed. The breeds included were European Braunvieh upgraded by US Brown-Swiss (BBV, Germany, [48]), Original Braunvieh (OBV, Germany [7], Switzerland [34]), German Fleckvieh (DFV, Germany [41]), Red Holstein (RH, Germany [47]), Blue Belgian (BBB, Belgium [30], Denmark [15]) and Galloway (GLW, sampled in Germany but originating from Scotland [38]). According to pedigree information and phenotypes the GLW can be sub-divided in following sub-populations or strains: black [23], dun [4], white-black-pointed [2], white [2] and belted [7]. In this context, we preferred individuals that did not share a common ancestor for at least 2 generations. The SNP genotypes were determined by a commercially available service (http://www.illumina.com/; Illumina, San Diego). The

[*] Tierzuchtforschung e.V. Munich, 85586 Grub, Germany

[†] LMU Munich, Chair of Animal Genetics and Husbandry, 80539 Munich, Germany

returned number of SNPs was 53,725 SNPs for each individual with an average minor allele frequency (MAF) of 0.25 across all loci. The SNPs were further been edited for genotyping errors, MAF < 0.05, HWE, P < 0.01 in at least one population and P < 0.02 in at least two populations respectively. This resulted in a total of 46,147 autosomal SNPs that passed the quality control and were used for the final course of the model/procedure.

Algorithm to identify community structures. The input for the Super Paramagnetic Clustering (SPC) algorithm represents a symmetric distance matrix D of dimension n x n, with the genetic distances for all samples being calculated by easily subtracting pair-wise identities by state (IBS) from 1. Additional inputs to SPC are the number of *k*-nearest neighbors (*k*-NN), a spin variable *s*, which takes on of *q* integer values: $s = 1,2,\ldots q$, a stable delta T and the minimal cluster size. To evaluate the clustering performance in high-dimensional space, a cost function is used, which is similar to methods used in hard-optimization problems (e.g. Traveling Salesman Problem). The cost function applied to SPC is the Hamiltonian of an inhomogeneous ferromagnetic Potts model,

$$H[\{s\}] = \sum_{\langle i,j\rangle} J_{ij}\left[1-(\delta s_i, s_j)\right]$$

where the classification {s} is determined by a so-called spin-spin correlation function ($\delta s_i, s_j$) and the nearest neighbor interaction J_{ij}, which is some positive decreasing function of the increasing distance between neighboring points *i* and *j*. Ferromagnetic Potts models are simulated at a sequence of temperatures (T), so that the clustering can be expressed at any level of T. At very low T, all data points remain uncorrelated. With increasing temperature the spin-spin correlation between neighboring points increases and the data points are clustered along the temperature by measuring the correlation of the nearest-neighbor spins. Consequently, the clustering result of the Potts model is strongly dependent on the number of *k*-NN, e.g. as *k*-NN decreases the number of clusters increases. Here we introduce the modularity (Q) (Newman (2006)) as a quality measure of sub-divided networks to determine optimal *k*-NN.

Results and discussion

To determine whether SPC could automatically expose the population structures of cattle breeds without any previous information, the algorithm was used with *q*=20 component Potts spins, each interacting with *k*-NN = 10, with respect to modularity measures and a minimal cluster size of two. Applying SPC to the network of cattle breeds, 9 communities have been extracted (Figure 1), which perfectly corresponds to our previous knowledge (breed and geographical origin) (Medugorac *et al.* (2009)) and investigations with the benefit of hindsight. As the figure shows, the network starts to split into three major groups, each represented by the dashed circles, and ends up with a fine-scale community structure, in which the relationship between breeds and animals is expressed by the thickness of edges varying in the proportion of genetic distance. This nature of network theory perfectly reflects the geographical origin of the breeds (Alpine region [BBV-OBV-DFV], Western Europe [BBB-RH] and North-Europe [GLW]). Additionally, it expresses the relationships of individuals within breeds and reveals the existence of sub-populations that are German Original Braunvieh, Belted Galloway and White Galloway. The only exception of a clear cluster solution concerns two Galloway animals, where one white-black-pointed animal has been assigned to the white, while the other one has been loosely connected to the black ones. This result possibly indicates the different levels of White/Black within these animals. However, for more significant results in this case, additional animals have to be analyzed. Within the Blue Belgian population the algorithm fails to cluster the animals into Belgian and Danish origin, since this data set contains five Danish animals which have a Belgian sire in the first generation. Excluding these "crossbred" animals in a second data set, the Danish subpopulation could be extracted (result not shown). To determine optimal *k*-NN in different data sets we have introduced Q to SPC, e.g. with and without crossed Danish-Belgian animals. Optimal *k*-NN = 10 for complete data set and *k*-NN = 7 for data set without crossbred animals. This result indicates that optimal *k*-NN varies with input data, hence should be determined for each data set separately.

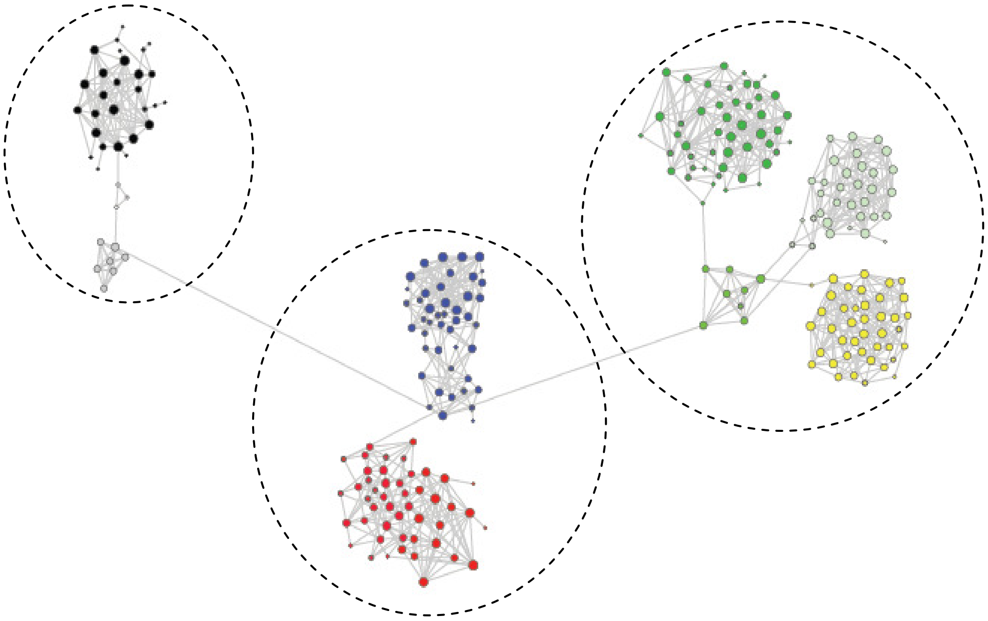

Figure 1: Community structure in cattle breeds extracted with the SPC algorithm. The breeds have firstly been separated into three major groups, denoted by the dashed circles, before they have been separated into communities. The 9 communities shown are Galloway in black dots (covering the colour variations black, dun and white black pointed), Belgian Blue (blue dots), Red Hostein (red dots), Deutsches Fleckvieh (yellow dots), European Braunvieh (dark green dots), Original Braunvieh with an separation into Swiss (cyan dots) and German (green dots) origin and the sub-populations White Galloway (white dots) and Belted Galloway (grey dots). The network has been drawn with longer edges between vertices in different breeds than between those in the same community, to make the community groupings clearer. The thickness of edges, which varies in proportion to the genetic distance, has been used to visualize individual's relationships within breeds. The node size, which varies in proportion of the number of edges per node (degree), illustrates how well each individual is connected within the breed.

Conclusion

These results clearly show that network clustering can be applied successfully to study the genetic structure of domesticated subpopulations without any *a priori* knowledge of clustering and stratification. This tool provides new insights into the history of domesticated breeds because of its ability to reduce the level of complexity of large-scale data sets. It will be of invaluable help to extract the population structure and stratification within single breeds with appropriate relevance to association studies. Hence this new approach described here is a very valuable alternative to the time-consuming clustering programs, e.g. STRUCTURE program, because this algorithm detects fine-scale population structure with a remarkable reduction in computing time and effort, as it works within seconds exploiting a whole-genome SNP survey.

Acknowledgements

Researchers are grateful for funding from Deutsche Forschungsgemeinschaft (DFG) through project number ME3404/2-1 and Tierzuchforschung e.V. Munich. Furthermore we thank the numerous breeders and breeding associations who sent us samples free of charge to support this study. In particular, we thank Swissgenetics (CH-3052 Zollikofen) for organizing parts of OBV samples.

References

Price, A.L., Patterson, N.J., Plenge, R.M., *et al.* (2006). *Nat. Genet.,* 38:904-909.

Blatt, M., Wiseman, S., and Domany, E. (1996). *Phys. Rev. Lett.*, 76:3251–3255.

Reichardt, J., and Bornholdt, S. (2004). *Phys. Rev. Lett.*, 93:218701–218704.

Newman, M.E.J., (2006). *PNAS.*, 103:8577–8582.

Medugorac, I., Medugorac, A., Russ, I., *et al.* (2009). *Mol. Ecol.,* 18:3394-3410.