Aus dem Veterinärwissenschaftlichen Department der Tierärztlichen
Fakultät der Ludwig-Maximilians-Universität München
Arbeit angefertigt unter der Leitung von
Univ.-Prof. Dr. Förster

Eine genomweite Populationsstrukturanalyse in Rinderrassen

Inaugural-Dissertation
zur Erlangung der Würde eines Doktor rer. biol. vet.
der Tierärztlichen Fakultät der Ludwig-Maximilians-Universität München

von

Markus Neuditschko

aus Waidhofen/Thaya
Österreich

München 2011

From the Department of Veterinary Sciences
Faculty of Veterinary Medicine
Ludwig-Maximilians-Universität München
Arbeit angefertigt unter der Leitung von
Univ.-Prof. Dr. Förster

A whole-genome population structure analysis within cattle breeds

Inaugural-Dissertation
For the attainment of the title Doctor of Veterinary Biology
From the Faculty of Veterinary Medicine of the
Ludwig-Maximilians-University München

by

Markus Neuditschko

from Waidhofen/Thaya
Austria

München 2011

Gedruckt mit Genehmigung der Tierärztlichen Fakultät der
Ludwig-Maximilians-Universität München

Dekan:	Univ.-Prof. Dr. Braun
Berichterstatter:	Univ.-Prof. Dr. Förster
Korreferent/en:	Univ.-Prof. Dr. Wolf

Tag der Promotion: 30. Juli 2011

This work is dedicated to my family and friends

Borders? I have never seen one.
But I have heard they exist in
the minds of some people.
(Thor Heyerdhal)

TABLE OF CONTENTS

1 INTRODUCTION

Highly informative genetic markers are essential to study the origin, history and evolution of livestock populations (Troy *et al.* 2001; Li *et al.* 2009; Medugorac *et al.* 2009). After a decade of domination by microsatellite markers, recently Single Nucleotide Polymorphisms (SNPs) are becoming more attractive to population genetic studies (Behar *et al.* 2010). High-throughput sequencing (Käller *et al.* 2007), have led that SNPs are the most technologically developed abundant markers in genetic science. Additionally it is becoming increasingly feasible to genotype hundreds or even thousands of individuals for these large numbers of SNPs.

These advantages in genetic science provide new insights into complex population structures (Gompert *et al.* 2010). It simultaneously supports the need of new approaches to study population structure based on whole genome-wide SNP panels spanning hundreds of thousands of loci. The large number of loci and the complexity to assign each individual to exactly one corresponding population poses significant challenges for conventional population structure analyses using commonly applied methods because most of these methods have been focused on small sets of microsatellite loci (Goudet 1995; Pritchard *et al.* 2000) e.g. within cattle most studies have been focused on a small set of microsatellite loci, mostly including the 30 microsatellite markers recommended by ISAG/FAO working group (http://dad.fao.org/) (Cymbron *et al.* 2005; Tapio *et al.* 2006; Li *et al.* 2007). At the same time, it has been predicted that at least tow to six times more SNPs will be necessary to achieve the same resolution as microsatellites when used for individual identification and the study of parentage assessment and relatedness (Morin *et al.* 2004). Furthermore it has been noted that classical population analyses still rely on *a priori* ancestry information that not always respect the natural population structure.

In order to meet these recent issues in population genetics, this study introduces the idea of network analysis into studies of population structures and focuses on a recently developed method (Paschou *et al.* 2007), which identifies highly informative markers applying a method derived from Principal Components Analysis (PCA) on a whole-genome SNP survey. Most informative markers are hereby identified as PCA informative markers (PCAIMs). Given the current interest to detect genome-wide selection signatures in livestock, we have additional investigated the application of PCAIMs to detect breed specific markers, since common

applied methods are computational demanding, while PCAIMs can be detected within seconds.

Hence the three major objectives addressed in this study are:

(i) To ascertain if animals can be allocated accurately to their respective breeds without any *a priori* ancestry information exploiting a whole-genome SNP survey.

(ii) To identify the minimum number of highly informative SNPs needed to guarantee a true assignment of animals to their respective breed/population.

(iii) To examine the utility of PCAIMs to detect putative selection signatures in cattle breeds.

2 LITERATURE REVIEW

2.1 Genetic Markers

To understand the history and evolution of populations it is usually necessary to study a large number of polymorphisms (Cavalli-Sforza 1998). Through molecular revolution over the last few decades, a lot of techniques have been developed to study population structure using genetic markers. At the first stage of research almost all markers identified have been protein and blood group polymorphism, and only a few hundred were previously known (Nei & Roychodhury 1988). These markers are also known as 'classical polymorphisms' to distinguish them from those obtained by direct Deoxyribonucleic acid (DNA) analyses. The use of DNA segments to analyze genetic polymorphisms has resulted in the identification of a great number of markers and genetic polymorphisms, which is of great benefit as a lot of markers are needed to detect population structure and selection signatures within populations, consequently identifying quantitative trait loci (QTL) affecting traits of economic interest (MacNeil & Grosz 2002; Thaller *et al.* 2003), as well as qualitative trait loci (Drögemüller *et al.* 2005; Drögemüller *et al.* 2009; Fontanesi *et al.* 2010). Commonly considered DNA markers are Microsatellites and Single Nucleotide Polymorphisms (SNPs or 'snips').

2.1.1 Microsatellites

Microsatellites are direct tandem repeated sequences of DNA with a repeat size ranging from 1 to 6 base pairs (bp). Hence, a microsatellite is also called a simple sequence repeat (SSR). To identify microsatellite loci suitable for use as genetic markers, different approaches have been developed. One of the prevailing methods to identify microsatellites is the use of primers, which are commonly designed directly from the sequence data. In this context, the primers serve to detect the variation of microsatellites, by flanking the different numbers of repeats provided by Polymerase Chain Reaction (PCR) (Saiki *et al.* 1988; Newton & Graham 1997). Microsatellites are highly polymorphic in many animal and plant species (Morgante & Olivieri 1993; Queller *et al.* 1993; Kim *et al.* 2004). Especially the high variation of this DNA marker have led that during the past decade a large number of population structure studies have extensively used microsatellite loci e.g. within livestock approximately 90% of all projects, microsatellite loci were used as genetic markers, while Single Nucleotide

Polymorphism (SNP) markers were only investigated in about 10% of all projects (Baumung *et al.* 2004). In contrast to the past decade, recent analyses in human (Hannelius *et al.* 2008), cattle (The Bovine HapMap Consortium 2009) and sheep (Kijas *et al.* 2009) show that currently SNPs regaining favour for uncovering genome-wide population structures.

2.1.2 Single Nucleotide Polymorphisms (SNPs)

A SNP is a small (single base pair) genetic change, or variation, that can occur within an individuals DNA sequence. This property of SNPs allows us to make associations between marker alleles with QTL's affecting important economic traits in livestock. It has been hypothesized that these genome-wide studies are our most powerful method identifying genes that are responsible for important traits, e.g. Myostatin (MSTN) for double muscling within animals (Hill *et al.* 2010) as well as for undesirable genes such as bovine leukocyte adhesion deficiency (BLAD), complex vertebral malformation (CVM) and congenital muscular dystony (CMD) (Charlier *et al.* 2008). Single Nucleotide Polymorphisms are a good choice of marker for these studies because of their low mutation rate, high incidence throughout the genome and bi-allelic nature making them amenable to automated detection techniques (Dawson 1999). It has been estimated that a well-designed genome-wide SNP map requires as many as 500,000 SNPs in human (Vega & Kreitman 2000) and likely up to 300,000 SNPs in cattle (Eck *et al.* 2009; Shannon 2010). Hence, the numbers of available SNP chips are increasing annually e.g. within cattle three different SNP chips have been released the last three years namely 10K, 25K and 50K BovineSNPchip. Through the ongoing innovations in high-throughput sequencing and array technology the number of SNPs is rapidly increasing, e.g. The Human International HapMap Consortium is currently using a second generation haplotype map spanning up to 3.1 Million SNPs (www.hapmap.org) (The International HapMap Consortium 2007). This rapid development in the SNP chip technology also provides new insights into the population structure of cattle breeds since previous studies have been focused on small sets of microsatellite loci.

2.2 Genetic Distances in Population genetics

The measure of genetics distance allows one to quantify genetic relationships between two individuals. It is used to describe the proportion of genetic elements (alleles, genes, gametes

and genotypes) that the two individuals do not share. Depending on the kind of similarities (*S*) between individuals most genetic distances vary from 0 to 1 e.g. there is a genetic distance of 1, when two samples have no genetic element in common. Depending on the similarities of individuals (*S*) genetic distances (*D*) can be calculated in three different ways (The International Plant Genetic Resources Institute & Cornell University 2003).

1. D = 1 – *S*, known as linear distance, because it assumes that the relationship with *S* is linear (Figure 1A).

2. D = $\sqrt{(1-S)}$, known as quadratic distances, the similarity relationship follows a quadratic function, so that, to make it linear, the square root must be calculated (Figure 1B).

3. D = $\sqrt{(1-S^2)}$, describes a circular distance (Figure 1C).

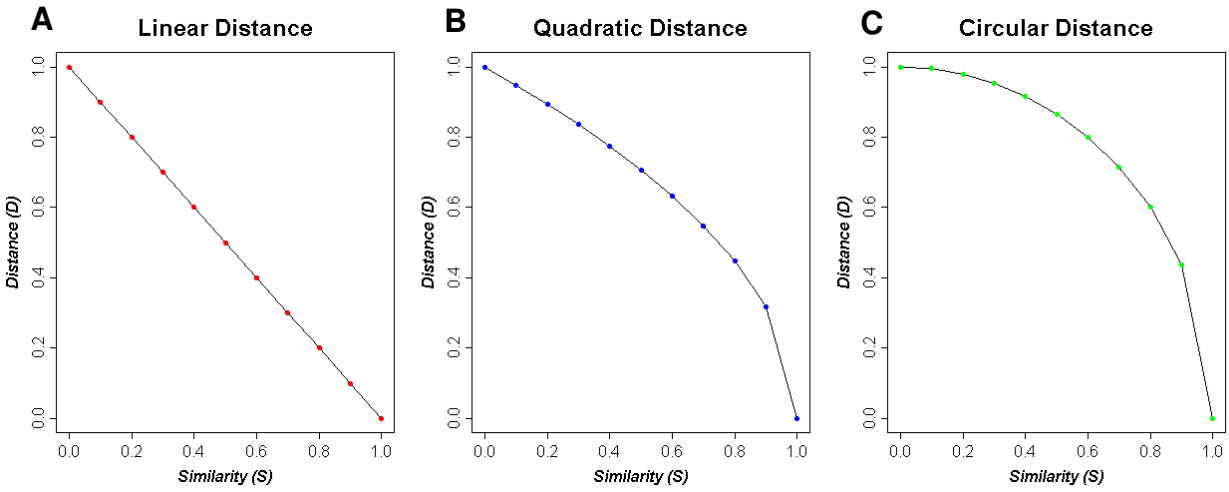


Figure 1 The different kinds of relatedness between Similarity (*S*) and Genetic Distance (*D*). (A) Linear Distance, (B) Quadratic Distance and (C) Circular Distance. **Source:** (The International Plant Genetic Resources Institute & Cornell University 2003)

2.2.1 Genetic distances between individuals using the proportion of shared alleles

The most frequently used measures to determine the proportion of shared alleles between two individuals are the interrelated parameters identical by state (IBS) and identical by descent (IBD), covering very different ranges where IBS describes a stronger genetic relationship

between unrelated individuals compared to IBD. Considering a simple four-allele marker system, where the parents carry the marker alleles (*ab*) and (*cd)*, the two siblings could possibly share zero, one, or two alleles. Within such a simple model the alleles shared by IBS and IBD are identical e.g. the two siblings have no alleles in common (*ac*) and (*bd*), so IBS and IBD is 0, if the siblings share a common allele from the father (*ac*) and (*ad*) IBS and IBD becomes 1 and finally the siblings can possibly share two alleles, where their source is ambiguous, then IBS and IBD is 2. However, in more complex situations as illustrated in Figure 2 the alleles shared by IBS and IBD are different from each other. The main difference between IBS and IBD is that at the calculation of alleles shared by IBD also considers the origins of the alleles, e.g. in Figure 2A, the siblings share the *b* allele by state, but it is evident that these two copies may not have come from the same parent, since both parents carry the *b* allele. In this instance, the siblings do not share an allele identical by descent, because the brothers *b* allele must have come from his father, while the sisters *b* allele originating from her mother. Thus it can also happen that IBD, cannot be directly determined, especially if a parent is homozygous (Figure 2B) and if both parents carry the same pair of alleles (Figure 2C).

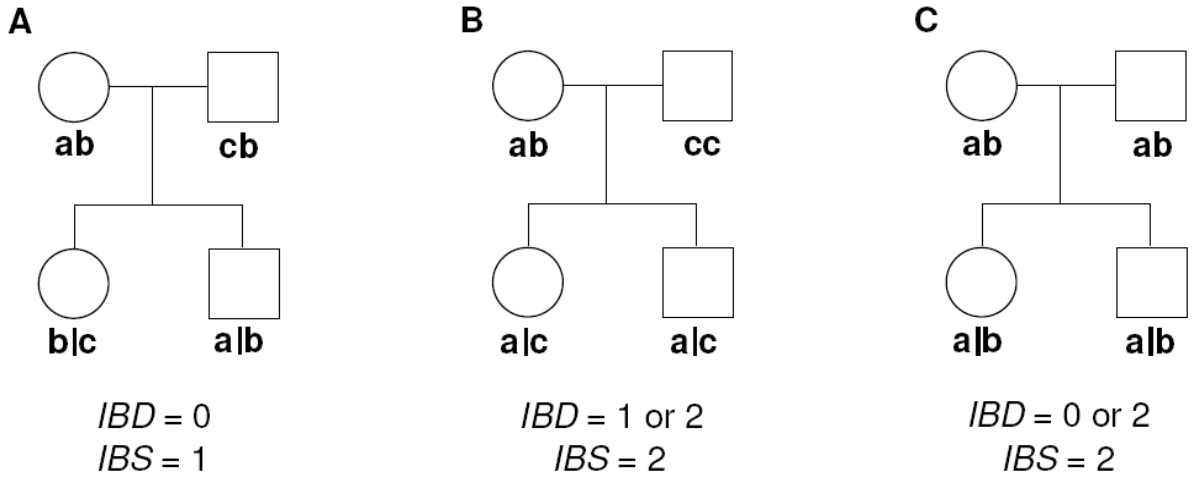


Figure 2 Two generation pedigrees illustrating the difference between IBD and IBS allele sharing considering a simple four-allele marker system. (A) In the first pedigree, the two siblings share allele *b*, which is identical by state (IBS = 1) but not identical by descent (IBD = 0), since the first child received the *b* allele from his father (box), whereas the second received it from his mother (circle). (B) In the second pedigree, the two siblings share two alleles (*ac*), which are identical by state (IBS = 2), in this situation the siblings can share one or two alleles identical by descent (IBD = 1 or 2), since the father carries a copy of one allele (*cc*). (C) The third pedigree demonstrates a similar situation where the two siblings share two alleles (*ab*), which are identical by state (IBS = 2), while the shared alleles identical by descent can be 0 or 2, since both parents carry the same pair of alleles. **Source**: (Sham & Zhao 1998)

In situations as presented in Figure 2 the haplotypes of the F1 generation has to be inferred to reveal alleles shared by IBD, which is carried out by phasing the genotype data (Gusev *et al.* 2009) (Browning & Yu 2009). The examples illustrated in Figure 2 additionally present, that IBS expresses a stronger relationship compared to IBD.

To determine the genome-wide average proportion of shared alleles between two animals we have used PLINK software (Purcell *et al.* 2007), where the genome-wide average proportion of alleles shared identical by state (IBS) between animals *i* and *j* is generally defined as

$$AS = \frac{1}{L} \sum_{l=1}^{L} a_l$$

where $a_l = 0$, if animal *i* and *j* have no allele in common at the l-th locus,
$a_l = 1$, if animal *i* and *j* have only a single allele in common at l-th locus,
$a_l = 2$, if animal *i* and *j* have two alleles in common at the l-th locus,
and L is the number of SNP loci used.

The genome wide proportion of shared alleles generally describes a linear relationship between animals, thus this value can be easily transformed into a so-called allele sharing distance (ASD) by subtracting it from 1 (see Chapter 2.2). Another widely used method in cattle breed analyses is the genomic relationship matrix presented by VanRaden (2008), since this approach additional considers allele frequencies within populations besides the alleles shared by IBS. Another possibility to determine genetic distances are based on genetic models considering allele frequencies e.g. Wright's F statistics (Wright 1931) and Nei`s standard genetic distance (Nei 1972). The major difference between ASD and the classical model based distances is that in the computation of classical ASD allele frequencies are completely ignored (Gao & Starmer 2007).

2.2.2 Genetic distances between subpopulations

One of the classical methods to measure population differences is the aforementioned F statistics. Wright`s F statistics use genotypes of predefined subgroups to calculate variance in allele frequencies, which may also be used to measure the genetic distance of

subpopulations. This concept is based on the idea that those subpopulations that are not intermating will have different allele frequencies to those of the whole population. The genetic distance calculated by F statistics is termed F_{ST}. The F_{ST} values ranges from 0 (no genetic divergence) to 1 (fixation of alternate alleles in different subpopulations), with values approaching 0.25 determining a high degree of genetic differentiation between populations (Mousadik & Petit 1996; Davies *et al.* 1997). The determined F_{ST} values between subpopulations are normally presented by distance matrices and phylogenetic networks using common available software packages e.g. TREEVIEW (Page 1996) and SPLITSTREE (Huson & Bryant 2006). The two additional indexes provided are F_{IS} and F_{IT} which measure the deficiency of average heterozygotes in each population and the excess of average heterozygotes in a group of populations respectively. For a preferable veridical interpretation of the results provided by this analysis it is necessary to separate the individuals into appropriate subgroups. At present, this population separation is mostly based upon geographical origin of samples or phenotypes (Liron *et al.* 2006).

2.3 Methodologies in Population structure analysis

To date, many statistical methods have been proposed to examine the relatedness among populations and to assign individuals to their population of origin. The two major methods currently applied are distance-based clustering methods, which utilize dimensionality reduction techniques e.g. Principal Component Analysis (PCA) in combination with standard clustering tool e.g. k-means (Liu & Zhao 2006; Paschou *et al.* 2007; Lee *et al.* 2009) and model-based methods e.g. STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003).

2.3.1 Distance based Methods

Principal Component Analysis (PCA) is a technique for linear dimension reduction of data complexity, by transforming original vector of correlated variables to the vectors of principal components which are uncorrelated. Through this process a new coordinate system with inherent statistical properties emerges, with the first principal components explaining the main variance in the data (Menozzi *et al.* 1978). In this context, PCA is providing an optimal subspace to investigate data variation in complex population structures and to allocate individuals to their respective population using a common applied clustering algorithm (Patterson *et al.* 2006; Price *et al.* 2006).

The k-means algorithm is an iterative descent algorithm that minimizes the within-cluster sum of squares (WSS) given the number of cluster K (Lloyd 1957; MacQueen 1967).

$$W_K = \sum_{i=1}^{K} \sum_{j \in C_i} \left\| x_j - \mu_i \right\|^2,$$

where x_j is the feature vector representing sample j, μ_i is the center of cluster i, and Ci is the set of samples in cluster i.

Principal component analysis followed by k-means describes a two-stage design that utilizes the proportion of shared alleles between individuals to infer population clusters. The main features of this two-stage design are, that it is computational efficient and assigns individuals to subpopulations without the need of any modelling of data (Liu & Zhao 2006). Principal component analysis is closely related to multivariate technique called Multi Dimensional Scaling (MDS). Indeed, PCA is equivalent to applying MDS on the distance matrix representing the data. Hence, this approach is also termed distance-based clustering method (Gao & Starmer 2007). Both approaches have been recently introduced in population structure analyses (Paschou *et al.* 2007) (Kijas *et al.* 2009). They are providing an optimal subspace to investigate genome-wide data sets including thousands of individuals by maximising variance on the first components (Behar *et al.* 2010). However, the current *Bovine HapMap* study reported that PCA is becoming impractical and less useful for the analysis of larger data sets (The Bovine HapMap Consortium 2009). Furthermore, it has been noted that the first two Principal components (PCs) only accounted for up to 40 % of the variation in this data set.

2.3.2 Model based methods (STRUCTURE)

Model-based methods use standard statistical methods to estimate population parameters, and usually assume Hardy-Weinberg equilibrium for each population. The inference may not be good in the presence of small sample sizes due to the inaccurate estimation of allele frequencies (Gao & Starmer 2007). Model-based inference also depends heavily on the modeling assumptions. The program STRUCTURE is a popular model-based program using Markov chain Monte Carlo (MCMC) within a Bayesian framework (Pritchard *et al.* 2000).

This method has been shown to be an effective and popular procedure in numerous studies (Kim *et al.* 2005) (Lao *et al.* 2006). However, the detailed modelling of the data causes intensive computational demands and associated costs (Liu & Zhao 2006) (Patterson *et al.* 2006) (Price *et al.* 2006), making this algorithm impractical to be applied on high density genome-wide SNP panels. Furthermore, it has been noted that the identification of the true number of clusters (K) in a sample of individuals is not sufficiently accurate by the algorithm applied (Evanno *et al.* 2005). Therefore, Evanno *et al.* (2005) introduced an *ad hoc* statistic ΔK for STRUCTURE to identify stable cluster solutions. However, the final cluster assignments are still very sensitive on the choice of prior distribution parameters of the data used. For example in the study of McKay *et al.* (2008) the *ad hoc* statistic ΔK identified two clusters in a data subset containing eight distinct cattle breeds.

Recently, alternative model-based algorithms as implemented in the programs ADMIXTURE (Alexander *et al.* 2009) and FRAPPE (Tang *et al.* 2005) have been used for uncovering genome-wide population structures. These algorithms are computationally efficient and can be easily applied on genome-wide data sets to infer individual ancestry. However, these methods are still sensitive to the choice of prior distributions of model parameters and rely on *a priori* ancestry information (e.g. ADMIXTURE should only be considered for unrelated individuals).

2.4 Network Theory

Network ideas are attracting the attention of mathematicians, physicians, computer scientists and practitioners of many branches and have already been successfully applied to many different topics as diverse as the Internet and the World Wide Web (Broder *et al.* 2000), epidemiology (May & Lloyd 2001), scientific citation and collaboration (Newman & Girvan 2004), metabolism and ecosystems (Wagner & Fell 2001). A property that seems common to many networks is *community structure*, the division of network nodes into groups based on the density of the connections within and between different sub-graphs (Figure 3). To extract *community structures* within a commonly used distance matrix representing the relationship between each individual, the number of interactions (edges) is strictly minimized according to a mutual neighborhood criterion. The best known criteria's currently applied are the ε neighborhood and *k*-nearest neighborhood criterion.

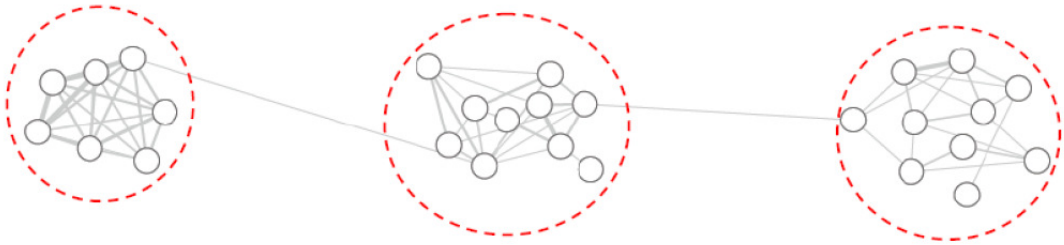


Figure 3 Network representation with three communities. Each individual is presented by a dot. The individuals are connected with an edge, illustrating their connectivity. The distinct communities are denoted by dashed lines, which have dense internal links but between which there is only a lower density of external links. **Source:** (Newman & Girvan 2004)

Given the number of individuals an the respective distances to each other, using the ε neighborhood criterion each individual will be connected to all other individuals which have a distance smaller than ε, whereas at the second criterion each individual will be connected to its *k* nearest neighbors (*k*-NN). The main difference between the two criteria's is, that *k*-NN are determined considering the pair-wise distance between two individuals (X_i,X_j) and the distances of X_i and X_j to all other individuals in the data set, while ε neighbors only depend on the pair-wise distance (X_i,X_j). However, the *k*-NN criterion is the one mostly applied in practice. Once, the *community structures* have been extracted different approaches can be applied to detect and characterize these structures including vertex similarity (Leicht *et al.* 2006), the vertex degree gradient (Bagrow & Bollt 2005), the resistor network (Wu & Huberman 2004) and the Potts Hamiltonian Model (Reichardt & Bernholdt 2004).

The study of *community structure* in networks is closely related to the ideas of graph partitioning in graph theory and computer science, and hierarchical clustering in sociology. Graph partitioning is a problem that arises to allocate a number of individuals or objects into an adequate number of clusters that respects the associations between the individuals. Finding an exact solution for this kind of partitioning task describes an NP (**N**on-deterministic **P**olynomial-time - hard problem), which is extremely difficult to solve for large graphs. To solve such problems a wide variety of heuristic algorithms have been developed that give acceptably good solutions in many cases e.g. the Kernighan-Lin algorithm, which runs in time $O(n^3)$ on sparse graphs (Kernighan & Lin 1973). A solution to the graph partitioning problem is, however not helpful for analyzing and understanding networks. In general networks describe a more complex situation, where besides the number of adequate clusters also the hierarchical structure of cluster within a given network is of main interest. Hence, the aforementioned approaches applied to detect and characterize *community structures* are based

upon the principles of hierarchical clustering. Using these procedures a hierarchical structure of clusters is generated in the form of a tree or *dendrogram*, where the clustering can be halted at any point, and the resulting components in the network are taken to be the communities.

Contrary to traditional methods currently applied in population genetic studies (phylogenetic networks), which primary objective is to study the difference between subpopulations, network analyses additional provide insights into the population structure within the studied subpopulations (see Figure 3). In this context, the network of subpopulations can be further described by network centrality measures. The three centrality measures frequently used in network analyses are *degree centrality*, *closeness centrality* and *betweenness centrality*.

Degree centrality measures the number of direct connections that an individual node has to other nodes within the network (Freeman 1978). For example in a social network it measures the number of friends one has, that is, how many people a person is directly connected to. It has been shown that nodes with higher degree centrality than other in the network are more active and important as members within the community (Frivolt & Bielikova 2005). In this study, we will show that within population analysis this measure can be used to identify important founder individuals within breeds.

Closeness centrality measures how many steps on average it takes for an individual node to reach every other node in the network (Freeman 1978). Thus closeness centrality measures how close, on average, an individual is to other individuals in the network. Closeness centrality has been used to identify important nodes within social networks (Kurida *et al.* 2007).

Betweenness centrality measures the extent to which a node can act as an intermediary or broker to other nodes (Freeman 1978). The more times that a particular node lies on paths that exist between other pairs of nodes in the network, the higher the betweenness centrality is for that node. Nodes that have a high betweenness centrality may act as brokers between subgroups and they may have stronger membership in surrounding communities (Girvan & Newman 2002) (Donetti & Munoz 2004). Betweenness centrality has been already introduced into population genetics analyses by Rozenfeld *et al.* (2008) to study the gene flow in a metapopulation system.

2.5 Resume

The limitations of the current approaches in population structure analyses (see Chapter 2.3) demonstrate that inferring population structure without prior knowledge of individual ancestry, given the amount of data derived from thousands of individuals and thousands of markers, is still a major challenge. We solved these problems by introducing an unsupervised network clustering approach, so-called Super Paramagnetic Clustering (SPC) as implemented in the software package SORTING POINTS INTO NEIGHBORHOOD (SPIN), which uses the Potts Hamiltonian Model to identify community structures in population networks. We evaluate the performance of SPC extensively on a data set of 260 animals representing six cattle breeds genotyped for 54,001 SNPs. Analyzing these six cattle breeds, we demonstrate the utility of SPC to assign individuals to their breeds and to detect fine-scale population structures simultaneously. Additionally, we show that introducing network analysis to population genetics is becoming feasible to identify genetically important founder and recently admixed animals in livestock populations.

Furthermore, we study its application on significant reduced SNP data sets using a novel algorithm developed by Paschou *et al.* (2007) to identify most informative SNP panels within cattle breeds (exact definition see Chapter 3). Contrary to the results in human (Paschou *et al.* 2007), were only 30 SNPs have been used to allocate individuals to the respective subpopulations, we show that in bovine, due to the close relationships of cattle breeds, at least 200 so called PCA-informative SNPs (PCAIMs) are needed to guarantee a true assignment of animals. Furthermore, this study will show that highly informative SNPs can be additionally used to detect selection signatures between breeds.

3 MATERIALS AND METHODS

3.1 Studied Populations

In this study a total of 260 animals from six cattle breeds representing three different selection criteria's (dairy, dual purpose and beef) were analyzed. The breeds included European Braunvieh (BBV) upgraded by US Brown-Swiss, Original Braunvieh (OBV), German Fleckvieh (DFV), Red Holstein (RH), Blue Belgian (BB) and Galloway (GLW). From geographical point of view, the studied breeds can be classified in two groups: **(i)** Alpine breeds covering the three breeds BBV, OBV and DFV, while breeders of BBV population strongly emphasize dairy production. **(ii)** North-western breeds, represented by BB, RH and GLW, where RH represents pure dairy and BB and GLW beef breeds.

Table 1 Summary of sampled subpopulations. The breed origin, country of sampling, breed characteristics and respective subpopulations are indicated, together with the number of sampled animals (n).

Breeds and subpopulations	Breed code	n	Country	Characteristics	Origin
European Braunvieh	BBV	48	Germany and worldwide	medium-large framed cattle, especially selected on high milk yield	Alpine
Original Braunvieh German Original Braunvieh Swiss Original Braunvieh	OBV OBV-G OBV-S	41 8 33	 Germany Switzerland	medium framed, one coloured dual purpose breed form alpine region	Alpine
German Fleckvieh	DFV	42	Germany	medium-large framed dual purpose breed from alpine region	Alpine
Red Holstein	RH	47	Germany and worldwide	selected on high milk yield, large framed Holstein cattle	North-Western
Belgian Blue Belgian Blue Belgian Danish Blue Belgian BB-D with BB-B sire	BB BB-B BB-D BB-BD	45 30 10 5	 Belgium Denmark Denmark	medium-large framed beef cattle, with extreme muscled exterior	North-Western
Galloway Black White White-Black-Pointed Belted Dun	GLW GLW-B GLW-W GLW-WBP GLW-BEL GLW-DUN	38 23 2 2 7 4	Germany and worldwide	small, robust beef cattle, selected for beef production in extensive region.	North-Western

According to sample origin and pedigree information three of these breeds can be subdivided in the following subpopulations or strains: German Original Braunvieh (OBV-G) and Swiss Original Braunvieh (OBV-S), Belgian Blue Belgian (BB-B) and Danish Blue Belgian (BB-D), Black Galloway (GLW-B), Dun Galloway (GLW-DUN), White-Black-Pointed

Galloway (GLW-WBP), White Galloway (GLW-W) and Belted (GLW-BEL). Concerning BB-D it should be noted, that this subpopulation shows evidence of recent extensive use of key BB-B animals in the pedigree, including five animals which have a Belgian sire in the first generation (BB-BD). In this context, we would have preferred animals that did not share a common ancestor for at least 2 generations. Table 1 summarizes the sampled sub-populations.

3.2 SNP Genotypes

The SNP genotypes of all animals were determined by Tierzuchtforschung e.V. München using commercially available service (http://www.illumina.com/; Illumina, San Diego). The returned number of markers was 53,725 SNPs for each animal with an average minor allele frequency (MAF) of 0.25 across all loci. The SNPs were further edited for genotyping errors including only markers with a marker call rate > 95%, MAF < 0.05, departure from Hardy Weinberg equilibrium (HWE) (Wiggintion *et al.* 2005) $P < 0.01$ in at least one population and $P < 0.02$ in at least two populations. This resulted in a total of 46,147 autosomal SNPs that passed the quality control and were used as the final data set for the model and procedure.

3.3 Population Parameters

Genetic diversity within each population was determined by the proportion of polymorphic markers ($\mathbf{P_N}$), allelic richness ($\mathbf{A_R}$), private allelic richness ($\mathbf{pA_R}$) and estimates of gene diversity ($\mathbf{H_E}$).

The proportion of polymorphic markers ($\mathbf{P_N}$) was easily computed by determining the percentage of heterozygous markers within each breed. The $\mathbf{A_R}$ and $\mathbf{pA_R}$ has been computed with the general applied rarefaction method, which trims unequal samples to the same standardized sample *g* (Hurlbert 1971; Petit *et al.* 1998; Kalinowski 2004, 2005) The final estimates of $\mathbf{A_R}$ for each breed are hereby determined by the number of distinct alleles expected in a random subsample of size *g* drawn from the respective breed (Hurlbert 1971; Petit *et al.* 1998), while $\mathbf{pA_R}$ estimates the number of private alleles expected in the breed when random subsamples of size *g* are taken from each of K breeds included in the analysis (Kalinowski 2004).

The gene diversity or expected heterozygosity ($\mathbf{H_E}$) estimates the probability that two randomly chosen alleles from the breed are different. The unbiased estimator of $\mathbf{H_E}$ provided by Weir (1989, 1996) at the l-th locus is hereby defined as

$$H_{El} = \frac{\left(1 - \sum_{l=1}^{k} p_l^2\right)}{1 - \frac{1+f}{n}},$$

where p_l is the frequency of the l-th of k alleles and f is the inbreeding coefficient across animals (n).

It should be noted, that estimates of $\mathbf{A_R}$ and $\mathbf{pA_R}$ were obtained using ADZE-SOFTWARE (Szpiech *et al.* 2008), while $\mathbf{H_E}$ was computed using POWERMARKER (Liu & Muse 2005). To account the existence of subpopulations within the aforementioned breeds, we have calculated the population parameters using two different data sets for these breeds (with and without the respective subpopulations).

3.4 Network clustering (SPC)

Super Paramagnetic Clustering is a hierarchical cluster method that aims to group subjects with similar genetic profiles into stable clusters. The major advantages of this method are the computational efficiency, the robust and reliable clustering results and more importantly the ability to extract community structures without any prior knowledge of data distribution. Based on these features, SPC ascertains population structure without any prior knowledge of individual ancestry given the information of whole-genome SNP panels. Since initial development this method has been successfully applied to genomic analyses as diverse as the analysis of yeast gene expression profiles (Getz *et al.* 2000a), the automatic classification of protein sequences (Tetko *et al.* 2005) and the identification of DNA sequences with active promoter region (Radjiman *et al.* 2006). Full details of the algorithm are described in Blatt *et al.* (1996). Since the method implemented in SPIN differs slightly from the one illustrated, a brief introduction to the method modified and applied in this study is provided.

The input to SPC is a symmetric distance matrix *D* of dimension *n x n*, with the genetic distances for all samples being calculated by subtracting pair-wise identities by state (IBS), as provided by PLINK version 1.07 (Purcell *et al.* 2007), from 1 (see Chapter 12.3). Given *D*, each data point gets associated with a Pott spin variable *s*, which randomly takes on of *q* integer values: $s = 1,2,\ldots.q$ (since the clustering result is insensitive to the choice of *q* (Blatt et al. 1996), we have worked with common applied $q = 20$ (Blatt *et al.* 1996; Getz *et al.* 2000b; Tetko *et al.* 2005). Once the Pott spins have been associated, an initial network is created by connecting neighbour pairs with edges, limiting the interactions to the given number of *k* nearest neighbours (*k*-NN). Since the cluster performance is based on a connected graph (Blatt *et al.* 1997), which links all the existing data, the edges have been superimposed corresponding to a minimal spanning tree associated with the data. The final result is the SPC graph onto the clustering is performed.

Consequently, the cluster performance of SPC is strongly dependent on the number of *k*-NN, e.g. as *k*-NN decreases the number of clusters increases. In previous studies different *k*-NN values have been applied (Blatt *et al.* 1996, 1997), stating that $k\text{-NN} = 10$ works well for most typical data sets (Blatt *et al.* 1997). Here we introduced the modularity (*Q*) (Newman & Girvan 2004) as a quality measure of sub-divided networks to determine optimal *k*-NN in a post-evaluation process (e.g. by calculating *Q* on various SPC graphs and corresponding cluster solutions using different *k*-NN values in the range of 3-105). Modularity ranges from 0 to 1 with values approaching $Q = 1$ indicating better community divisions (Holmström *et al.* 2009). Here it should be noted, that SPC in some situations generates some unclassified data especially at small *k*-NN values. Since *Q* does not allow any unclassified animals, we decided to put these animals in a separate cluster.

To evaluate the clustering performance in a multi-dimensional space, a cost function is used, which is similar to methods used in problems which are hard to optimize (e.g. Traveling Salesman Problem) (Krentel 1988). The cost function applied to SPC is the Hamiltonian of an inhomogeneous ferromagnetic Potts model,

$$H(S) = -\sum_{ij} J_{ij}\delta_{si,sj}$$

where the classification (S) is determined by so-called spin-spin correlation function (G_{ij}) and coupling constant J_{ij}, which is some positive decreasing function of distance. Ferromagnetic Potts models are simulated at a sequence of temperature (T) given a stable ΔT, so that the clustering can be expressed at any level of T. At very low T, all data remain uncorrelated. With increasing temperature the spin-spin correlation between neighbouring points increases and the data points are clustered along the temperature ($0 \leq T < T_{max}$) by calculating G_{ij} using mean field approximation (Barad 2003). Finally the level of T clusters emerging is determined by identifying the edge (ij) with maximal $G_{ij}(T)$ for each data point, which is supplemented by directed growth procedure (Barad 2003). Thus the range of temperatures $\Delta T = T_2 - T_1$ a cluster splits from its parent is used as a measure of the stability and significance of the corresponding data cluster. The more stable the cluster, the larger the range ΔT.

For the final visualization of the SPC graph and corresponding clustering results, open source software CYTOSCAPE (http://www.cytoscape.org) (Shannon *et al.* 2003) was applied. For this purpose we used various additional tools for the analysis of networks as provided by the network analysis tools (*NeAT*) (Brohee *et al.* 2008). In the final network presentation the nodes represent animals and an edge between two nodes represents the relationship of the corresponding animals, which is expressed by the thickness of edges varying in the proportion of genetic distance (1-IBS). To easily detect important founder animals within breeds the node size has been associated with number of direct connections per node (*degree* centrality). To identify key animals that are responsible for the gene flow between subpopulations we additional associated the node size with the *betweenness* centrality. The two different associations of the node size have been determined with *NeAT* and visualized separately from each other.

Within this study the degree-dependent *betweeness*, $bc(k)$ has been determined with the general applied formula (Freeman 1977; Narayanan 2005; Rozenfeld *et al.* 2008),

$$bc(i) = \sum_{i \neq j \neq v} \frac{\sigma_{ij}(v)}{\sigma_{ij}}$$

where σ_{ij} denotes the number of shortest paths connecting nodes i and j and $\sigma_{ij}(v)$ the number of those passing through the node i.

Applying SPC to population genetics leads to an essential difference compared to classical algorithms for construction of phylogenetic trees based on individual distances e.g. Neighbor Joining (NJ) algorithm introduced by Saitou and Nei (1987). NJ algorithm begins with a star tree, which is produced under the assumption that there is no clustering of data points and try to find true neighbours (two data points that are connected by a single node in an unrooted tree) by minimizing the sum of all branch lengths. Therefore, NJ algorithm starts with the closest and finishes with most distant neighbors. Contrary to NJ algorithm, SPC starts at T = 0, where all data points form one cluster. This initial hypothetical taxonomic unit splits successively within a continuous temperature gradient. Thereby, the splitting occurs in an opposite order compared to the NJ algorithm, i.e. SPC algorithm starts to split most distant composed data objects hereby producing a hierarchical structure of clusters.

3.5 Comparative cluster analysis

To evaluate the clustering performance of SPC, the population structure of six cattle breeds was first analysed with PCA coupled with k-means (distance-based clustering approach) and STRUCTURE (model-based clustering approach).

3.5.1 Distance-based clustering (PCA and k-means)

Given the genome-wide proportions of alleles shared by state between individuals (*A*), we directly applied PCA on *A* to compute its singular principle components (PCs). We would like to note that, from mathematical perspective, this procedure is exactly equivalent to applying MDS on the distance matrix (*D*) of *A*. After determining the total genetic variance explained by the first three components, k-means clustering as implemented in *R.2.10* (www.R-project.org) was applied with 10,000 iterations on low-dimensional data to separate the individuals to their respective populations. In order to investigate the genetic relationship between the breeds, different k-means runs have been completed increasing the numbers of clusters (K) from 2 to 12.

3.5.2 Model-based clustering (STRUCTURE)

Contrary to PCA, STRUCTURE v2.3 software could not be applied to the full SNP data set because of software and computational limitations. We applied STRUCTURE on a subset of

4,930 SNPs, constructed by LD-based SNP pruning with $r^2 < 0.06$. The pruned SNP set overcame the computational limitations of this algorithm. Five runs of STRUCTURE were performed with K increasing from 2 to 9. We chose 10,000 iterations of the Gibbs sampler after a burn-in of 10,000 iterations, applying the admixture model. In order to determine the true number of clusters in 260 samples from six cattle breeds, we plotted values of *LnP(D)* (the log probability of data) for each K and estimated the delta K (ΔK) statistics, which is based on the rate of change in *LnP(D)* between successive K values (Evanno *et al.* 2005). Subsequently we compared the numbers of clusters determined by SPC with the *ad hoc* statistic ΔK.

3.6 Phylogenetic Network

To statistically evaluate the general hierarchical structure of SPC analysis, we calculated pair-wise F_{ST} among the six cattle breeds from population allele frequencies across all 46,147 autosomal SNPs (Weir & Cockerham 1984). The resulting F_{ST} distance matrix was plotted using program SPLITSTREE4 (Huson & Bryant 2006). The variation within populations containing subpopulations has been calculated separately.

3.7 PCA informative Markers (PCAIMS)

A detailed description of the algorithm to identify PCA Informative Markers (PCAIMs) used in this study is presented in Paschou *et al*. (2007). The inputs required for the algorithm is a properly encoded SNP genotype data matrix *G* between *m* subjects and *n* SNPs with genotypes encoded -1 for (AA), 0 for (AB) and 1 for (BB) and a positive integer *k* corresponding to the number of principal components. The input data matrix does not allow any missing entries and without rejecting too many informative SNPs, we filled in the missing genotypes by simulating HWE using all available information from respective SNPs in every single breed. To determine the number of *k* principle components we have used the empirical method Horn's parallel analysis as implemented in the statistical programme package *paran* as implemented in R software (http://www.r-project.org). This method employs Monte Carlo estimates to retrain most significant principal components, according to the significance level and number of iterations. Here, we chose a significance level of $p = 0.01$ and 10,000 iterations, which have been suggested in the modified version of parallel analysis (Glorfeld 1995). The fact, that we have applied PCA onto *A* (pair-wise allele sharing proportions

between individuals) to determine the genetic variance explained by the first components and to allocate the individuals to their respective breeds. We found it most attractive to use this matrix to determine the *k* numbers of significant components. Using *k* principle components and input matrix *G* the PCA informative score (p_j) for each SNP has been determined as described in Paschou *et al*. (2007) with

$$p_j = \sum_{i=1}^{k} (v_j^i)^2,$$

where v^i_j are the coefficients of the linear combinations and *k* is the number of significant principal components.

After determining p_j for each SNP, where SNPs with highest scores are most informative to represent the structure of the analysed populations, *A* has been computed along a decreasing number of informative SNPs (using 40,000 to 10 PCAIMs). The final number of PCAIMs that still allows to identification of subpopulations and allocation of individuals has been determined using SPC.

In order to determine the agreement between the SPC solutions on the different number of PCAIMs, we have used the adjusted rand index by Hubert and Arabie (1985). The adjusted rand index ranges between 0 and 1, with values approaching 1 reflect a perfect agreement between two cluster solutions. This method can deal with different numbers of clusters, but does not allow any unclassified data, while SPC in some situations (especially on low density SNP panels) generates some unclassified data; we decided to put these in a separate cluster.

4 RESULTS

4.1 Overall genetic diversity parameters

Genome-wide examination of the variability within breeds was used to compare levels of heterogeneity between breeds and to demonstrate the effect of subpopulations on the interpretation of diversity parameters within breeds. This analysis revealed that RH breed displayed the highest genetic diversity as measured by allelic richness (A_R = 1.855), private allelic richness (pA_R = 0.006) and gene diversity (H_E = 0.357). Conversely the strongly selected alpine breed BBV was ranked lowest using most measures (A_R = 1.800; pA_R = 0.002; H_E = 0.322).

Table 2 Summary of overall genetic diversity parameters in cattle breeds. With abbreviations, (**N**) number of individuals tested per population, (**P_N**) the proportion of SNP which displayed polymorphism (**A_R**) allelic richness, (**pA_R**) private allelic richness and (**H_E**) expected heterozygosity or gene diversity. The maximum and minimum values are expressed in bold, *indicating the main population (OBV-S, BB-B and GLW-B).

			Indices of Genetic Diversity			
Population	**Code**	**N**	**P_N**	**Ar**	**pAr**	**He**
European Braunvieh	BBV	48	**0.915**	**1.800**	**0.002**	**0.322**
Original Braunvieh	OBV	41	0.929	1.840	0.001	0.337
Original Braunvieh*	OBV-S	33	0.881	1.761	0.001	0.320
German Fleckvieh	DFV	42	0.938	1.839	0.005	0.344
Red Holstein	RH	47	**0.944**	**1.855**	**0.006**	**0.357**
Belgian Blue	BB	45	0.930	1,849	0.003	0.348
Belgian Blue*	BB-B	30	0.897	1.760	0.002	0.320
Galloway	GLW	38	**0.915**	**1.800**	0.005	0.325
Galloway*	GLW*	23	0.831	1.610	0.002	0.290

The examination within breeds further revealed that up to 91% of SNPs displayed both alleles within all analyzed breeds with a highest value in RH and lowest values in BBV and GLW respectively (Table 2). Averaged across breeds, 92% indicates that the majority of SNPs display a high degree of polymorphism within all breeds. However, this result also reveals the important position of Holstein Friesian population in the design of the current SNP chip (Bovine SNP50). Concerning the breeds with respective subpopulations namely OBV, BB

and GLW the genetic diversity parameters have been computed a second time excluding animals with different origin and phenotype. The results of this second analysis show that the existence of subpopulations falsifies the aforementioned genetic parameters within these breeds. In this context, the highest differences have been noted within GLW (A_R = -0.190; pA_R = -0.003; H_E = -0.035) followed by BB (A_R = -0.089; pA_R = -0.001; H_E = -0.028) and OBV (A_R = -0.079; pA_R = -0.000; H_E = -0.027).

4.2 Distance based clustering (PCA and k-means)

The results of the PCA analysis are shown in Figure 4. The first dimension (PC1) accounts for 35% of total variance and separates the six cattle breeds into the known geographical regions namely Alpine breeds [BBV-OBV-DFV] and North-Western breeds [BB-RH-GLW]. The second dimension (PC2) summarizes 15% of the variation and clearly distinguishes GLW from two other north-western European breeds (Figure 4A). Additionally, PC2 further distinguishes RH and BB, reveals the existence of substructures within GLW and forebode such in BB breed. The third dimension (PC3), which accounts for 8 %, differentiates the close positioned breeds [BB-RH] and [OBV-DFV] (Figure 4B). Within OBV and BB the respective subpopulations were indistinguishable.

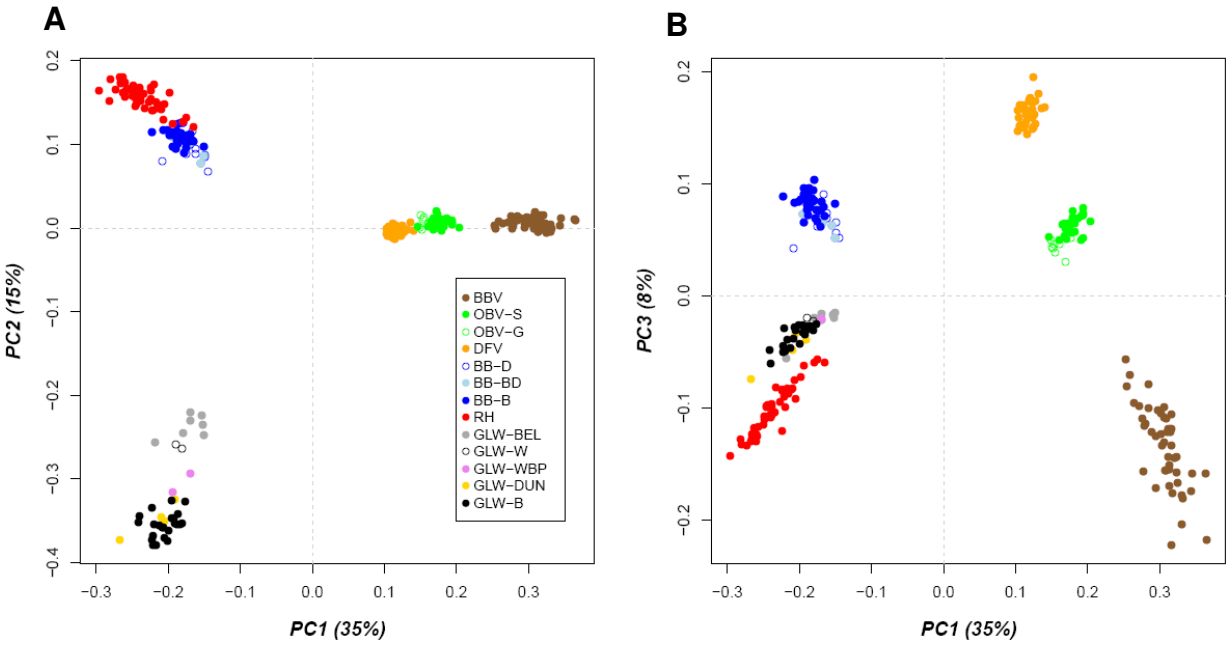
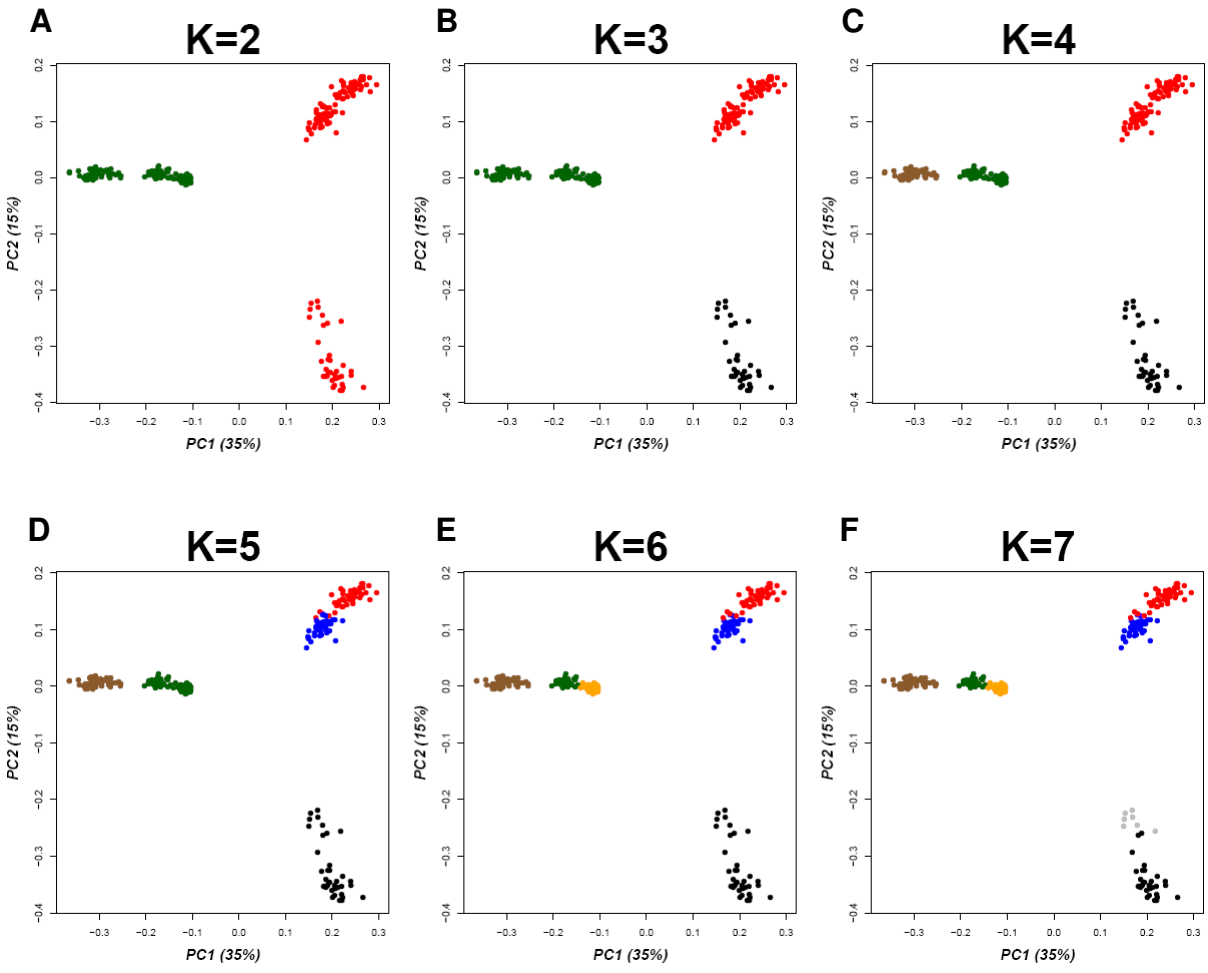


Figure 4 PCA plot of the six cattle breeds respecting the existence of subpopulations. Individual animals are projected onto the subspace of the first three PCs, covering a total variance of 58%, computed using the proportion of allele sharing (IBS) between animals. (A) Scatter-plot using the first and second principal components (PC1 vs. PC2). (B) Scatter-plot using first and third principal components (PC1 vs. PC3).

To investigate the relatedness between breeds and to determine the significance of the existence of subpopulations we did additional k-means clustering runs with 10,000 iterations on low dimensional data increasing the number of clusters (K). Starting with K = 2 (Figure 5A), the six breeds have been separated into the aforementioned geographical regions. Sequentially increasing K to 6, all animals have been ascertained to their respective breeds, while the single breeds have been identified as follows: GLW (Figure 5B), BBV (Figure 5C), RH (Figure 5D), BB (Figure 5D), OBV (Figure 5E) and DFV (Figure 5E). Here it is remarkable that BBV has been separated from the Alpine cluster, while RH and BB have been grouped together. Given K = 7 (Figure 5F) and K = 8 (Figure 5G) the substructures GLW-BEL and BB-D could have been revealed, where one BB-BD animal has been misclassified. Finally, also the subpopulation OBV-G emerged with K = 10 (Figure 5I), but here it should be noted that RH has been separated into two clusters before (Figure 5H). Increasing K to 12 resulted in further differentiation of BBV (Figure 5J) and RH (Figure 5K) but has not revealed the outstanding substructures, namely GLW-W, GLW-WBP and GLW-DUN.



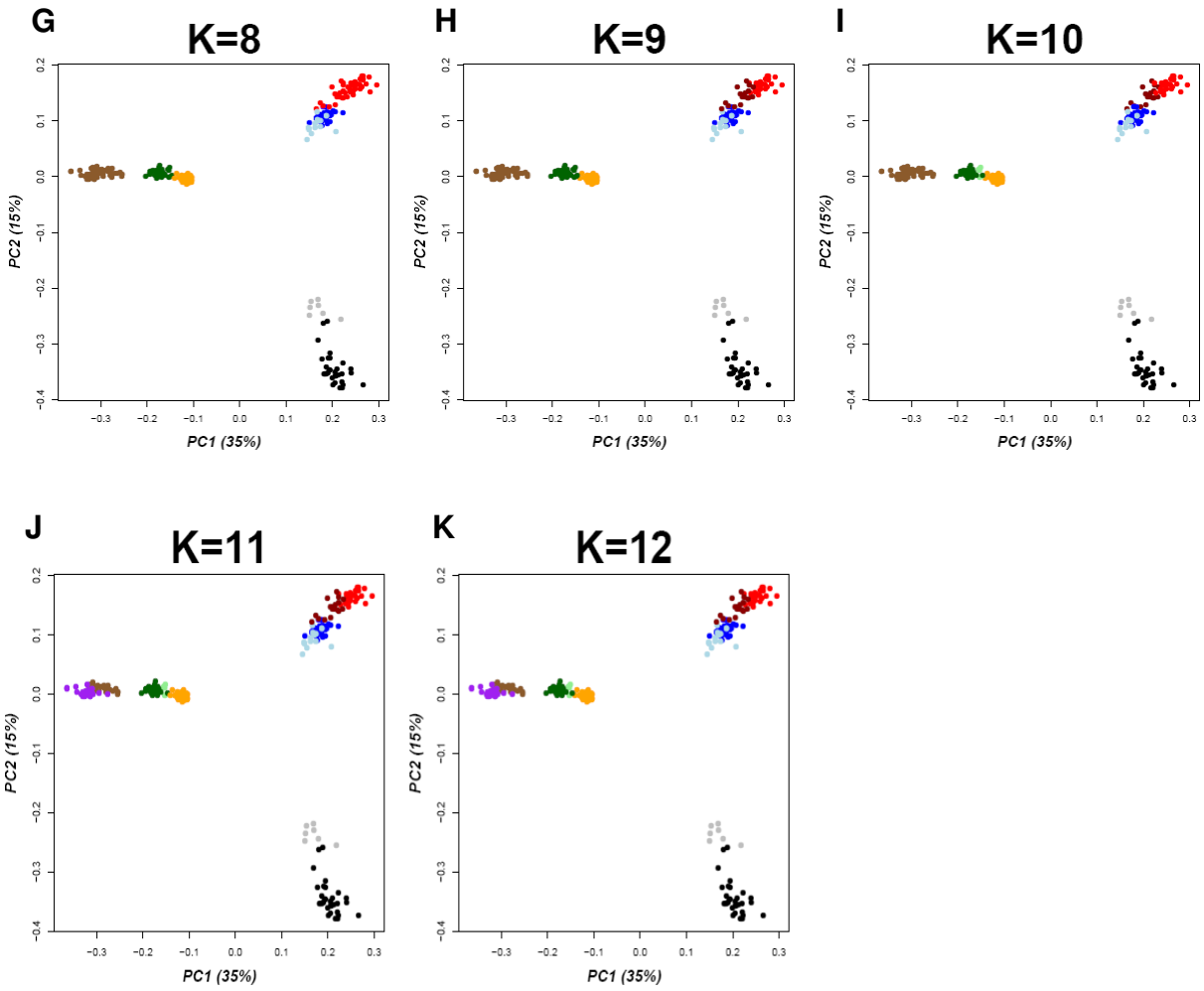


Figure 5 Distance-based clustering result assessed with PCA and k-means. The given number of clusters varied from 2 to 12. Individual animals are projected on the first two PCs, divided into K colours. The panels A to K summarize the different k-means cluster solutions, where each cluster is represented by different colours.

4.3 Model based clustering (STRUCTURE)

To determine the uppermost hierarchical structure within investigated cattle samples, we analyzed the results of five *STRUCTURE* runs increasing K from 2 to 9 (the total number of breeds; respecting the existence of additional subpopulations) with distribution of *LnP(D)* values and *ΔK* statistic presented in (Figure 6). The graph of *LnP(D)* did not show a clear point of change in the slope increasing K from 2 to 7 (Figure 6A). Additional raising K to 9, we observed an increase in *LnP(D)* variance and a significant decrease of mean *LnP(D)*, which suggests an uppermost hierarchical structure at K = 7. The distribution of *ΔK* impressively confirmed this result, with a significant maximum at this K value (Figure 6B).

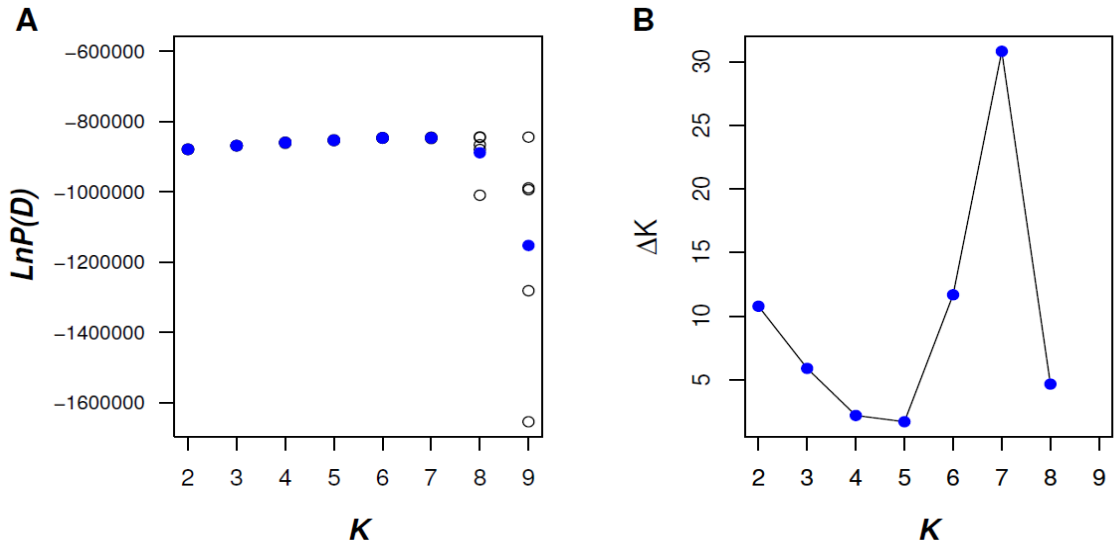


Figure 6 Uppermost hierarchical structure based on *LnP(D)* and *ΔK* values. (A) Estimated likelihood, *LnP(D)* for values of K ranging from 2 to 9. The mean LnP(D) for each K over five runs are represented by blue dots. (B) *ΔK* calculated according to Evanno et al. (2005). The modal value of this distribution corresponds to the true *K*(*) or the uppermost level of structure, here K = 7.

The assignment of individuals to separate clusters with five independent runs per K value revealed that only with K = 2, 3 and 6 consistent cluster solutions could have been achieved for five runs. Given K = 2 and K = 3 *STRUCTURE* confirmed the PCA findings by assigning all Alpine breeds to one and north-western breeds to the other cluster (Figure 7A) and separating GLW from all the other breeds (Figure 7B). At K = 4, three runs resulted in a differentiation of BBV from [OBV-DFV] (Figure 7C) where the remaining breeds cluster like K = 3. In two additional runs, all Alpine breeds stay in a common cluster and North-Western animals split into three distinct clusters according to their breed origin. Assuming K = 5 in three runs [BB- RH] built one cluster while the Alpine breeds have been separated into single clusters (Figure 7D). In another two runs [OBV-DFV] have been assigned to the same cluster, while BB and RH were found in separate clusters. At K = 6 all individual animals have been associated with their breed origin (Figure 7E). Given optimal K = 7, in three runs GLW-BEL has been identified as a subpopulation of GLW with a significant genetic membership proportion of 62% ± 12% (Figure 7F). Referring to the additional subpopulations within Galloway two GLW-W and one GLW-WBP animal have been associated with a membership proportion of 16% ± 3% within the GLW-B cluster, while the other animals (1 GLW-WBP and 4 GLW-DUN) do not differentiate from GLW-B. In the remaining two runs, the subpopulations BB-D and OBV-G have been discovered instead of GLW-BEL with a membership proportion of 39% ± 15% and 52% ± 16% respectively. Under the assumption K = 8, GLW-BEL has been associated into a single cluster at all runs, whilst BB-D has been

represented by a separate cluster in four runs (Figure 7G). In one additional run OBV-G was identified instead of BB-D. With K = 9, one run revealed the simultaneous existence of all three subpopulations (Figure 7H). The remaining four runs have provided new cluster solutions, particularly separating BB and RH population into additional clusters.

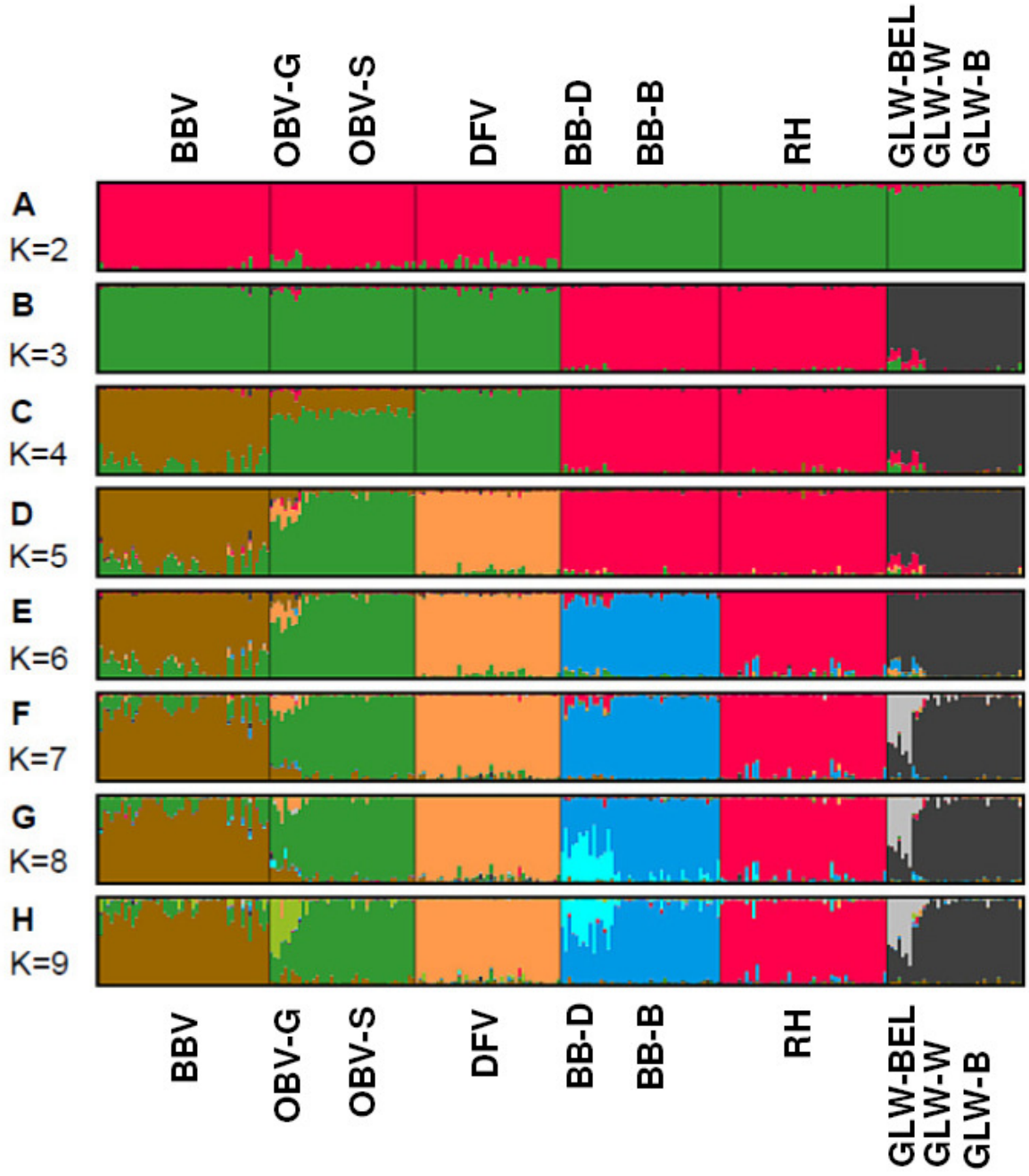
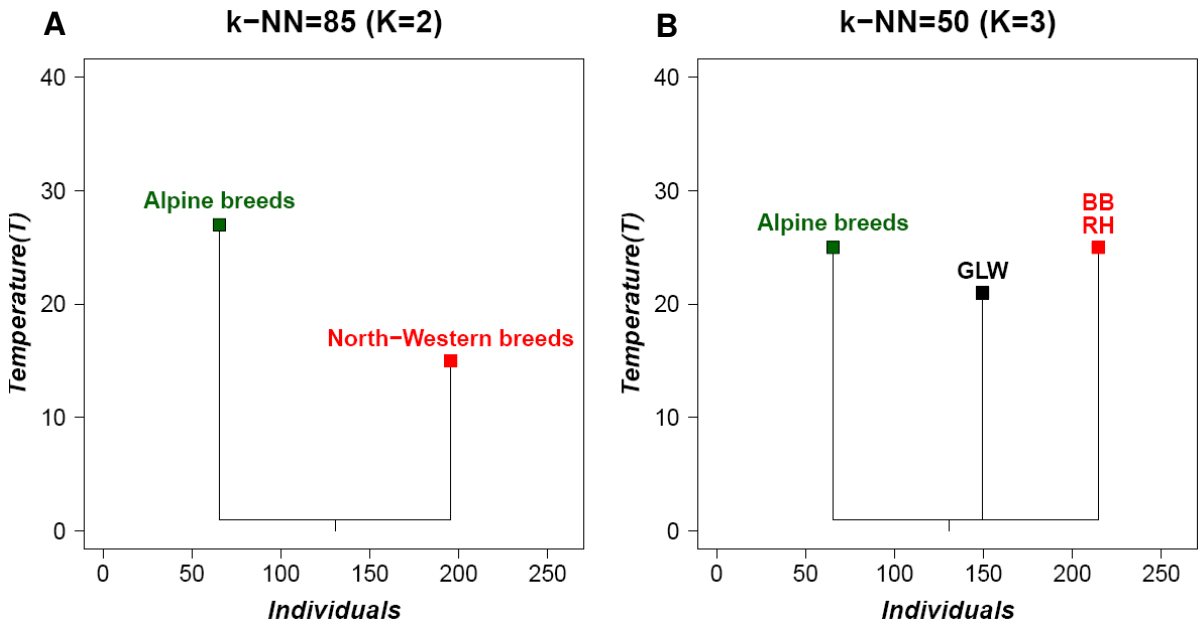


Figure 7 Cluster assignment assessed with STRUCTURE. The given number of clusters (K) varied from 2 to 9. Individual animals are presented by a single vertical column divided into K colours. Each colour represents one cluster, and the length of the coloured segment corresponds to the individuals estimated proportion of membership in that cluster. For each K, five runs were performed. Each run shown here was chosen based on frequency at each run. The panels A to H represent the most frequent patterns at K = 2 to K = 9.

4.4 Network clustering (SPC) of 6 cattle breeds

To determine whether SPC could automatically expose the population structures of cattle breeds without any prior ancestry information, we previously investigated the optimal number of *k*-NN suitable for our data. Consequently, evaluating each clustering at a given number of *k*-NN, starting with *k*-NN = 105, where all animals were assigned to one single cluster, the number was subsequently decreased in steps of 5 until 15, followed by single steps until *k*-NN = 3. Since the number of clusters increases as *k*-NN decreases, the animals have been clustered according to the different levels of *k*-NN, e.g. at *k*-NN = 100-85 the animals have been clustered into Alpine breeds and North-Western breeds (Figure 8A), while at *k*-NN = 80-50 the animals have been separated into 3 clusters, separating GLW from the North-Western breeds (Figure 8B). The additional number of clusters identified were 4 (*k*-NN = 45) with a further separation of North-Western breeds into single breeds (Figure 8C), K = 6 (*k*-NN = 40-30) where all animals have been ascertained to their respective breeds (Figure 8D), K = 7 (*k*-NN = 25) which provides a differentiation between GLW and GLW-BEL (Figure 8E), K = 9 (*k*-NN = 20-6) where animals of the subpopulations GLW-W and OBV-G have been assigned to separate clusters (Figure 8F). Using less than k-NN = 6 particularly BBV breed have been further differentiated (e.g. at *k*-NN = 5 BBV splits into 3 different clusters). It has been further noted that especially with small neighbourhoods (*k*-NN = 5-3) the number of unclassified animals rapidly increased. Nevertheless, unclassified animals have been also noted at runs with *k*-NN = 11 and 12.



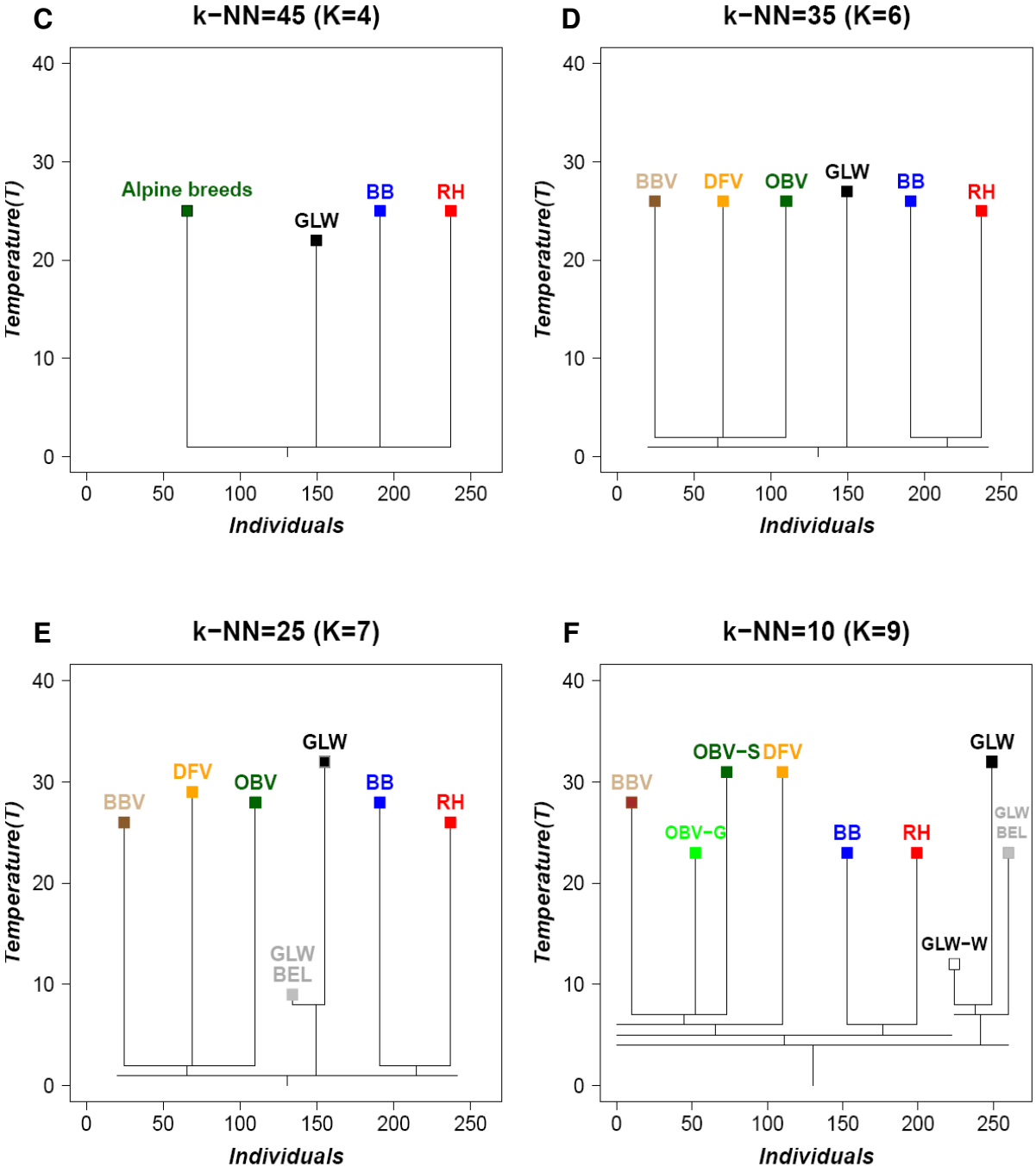


Figure 8 SPC cluster solutions at various numbers of *k*-NN. Presented are the cluster solutions from K = 2 to 9 (panels A to F), given the respective number of k-NN. Each identified cluster is symbolized by a box; with vertical positions determining the significance of each cluster, while horizontal positions are indicating the proximity between clusters. The clusters are presented in different colours, where next to the identified clusters the respective breed abbreviations are given.

Hence, runs at *k*-NN < 5 have been removed from further analysis. The optimal choice of *k*-NN was determined by plotting *Q* values for the range of *k*-NN 105-5 as is shown in Figure 9. Although no significant differences were detected in the range of *k*-NN 6-40, a local

maximum was reached at *k*-NN = 10, with $Q = 0.835$ (Figure 9). The final settings for SPC algorithm were this taken at 20 components (q) Potts spins, and 10 nearest neighbours (*k*-NN) for all further analyses.

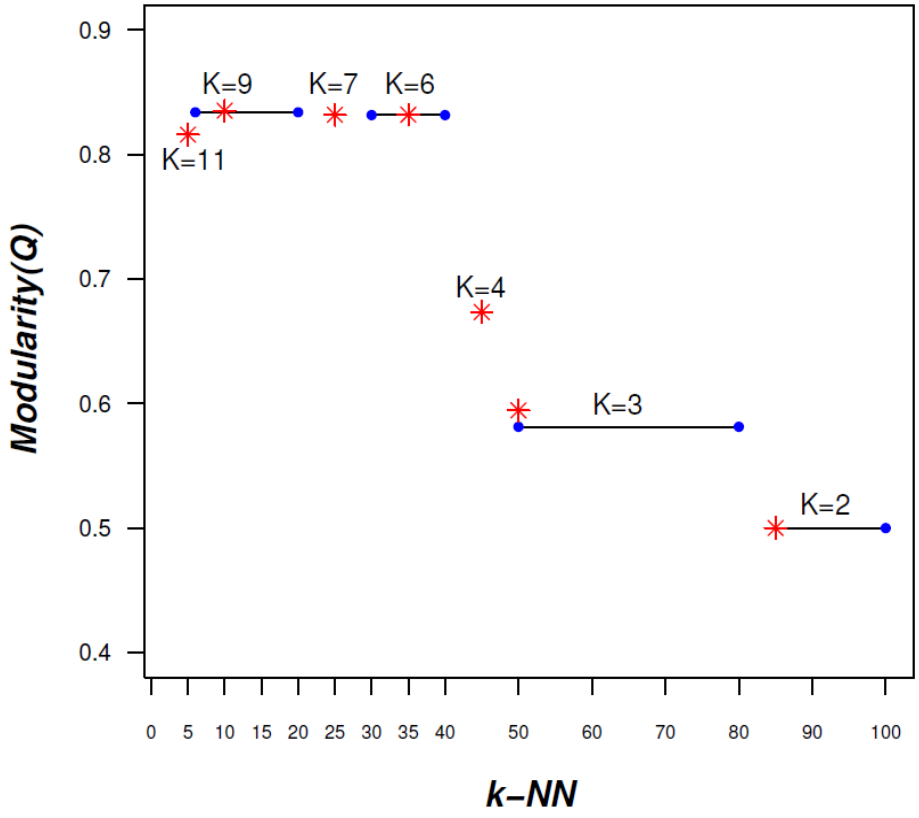


Figure 9 Determining the optimal number of *k*-NN. Modularity measures (Q) for each SPC run reducing the number of *k*-NN form 105 to 5. As modularity is strongly influenced by the number of determined clusters at respective *k*-NN, the detected numbers of clusters is also given. Since some runs reproduce the same cluster solutions with small variation in Q, the average Q values for these runs have been plotted, with horizontal lines indicating the range of the determined cluster result, while the run with the local maximum for each number of clusters identified is recognized by (*).

A typical hierarchical tree of clusters achieved from this analysis is presented in Figure 10. As shown in the figure, the tree starts to split into three major groups and ends up with a final organization of clusters, where the animals have been successively allocated to single breeds and subpopulations along a continuous temperature gradient (T).

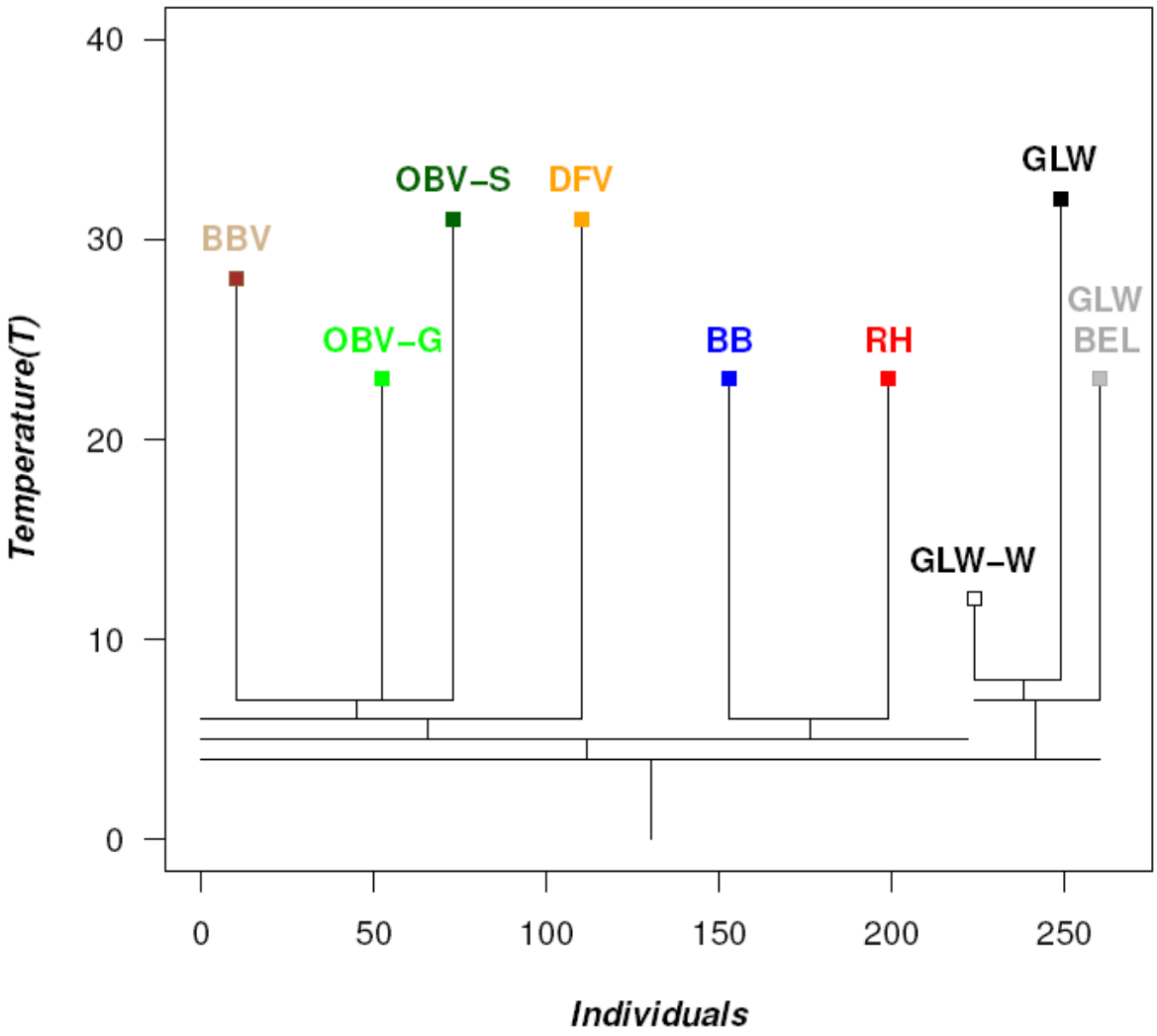


Figure 10 Super Paramagnetic Clustering of cattle breeds including 260 animals. Dendrogram representing the clustering of animals with optimal *k*-NN = 10. The individual animals have been separated into nine clusters, representing the six cattle breeds and the additional existence of three subpopulations. Each cluster is represented by a box; with vertical positions determining the stability of each cluster, while horizontal positions are indicating the proximity between clusters.

The shape of the hierarchical tree observed perfectly reflects the geographical origin of the breeds by separating the breeds into Alpine region [BBV-OBV-DFV] including a further differentiation into Braunvieh populations [BBV-OBV] and [DFV], Western Europe [BB-RH] and North Europe [GLW]. To visualize the genetic relationships between the determined subpopulations more detailed, the SPC cluster solution has been further presented in a heat map (Figure 11). From this reordered distance matrix *D* one can easily infer the shapes of the identified subpopulations and the genetic relationship between them, with red indicating small genetic distances and white/yellow suggesting large genetic distances between the individuals and subpopulations respectively.

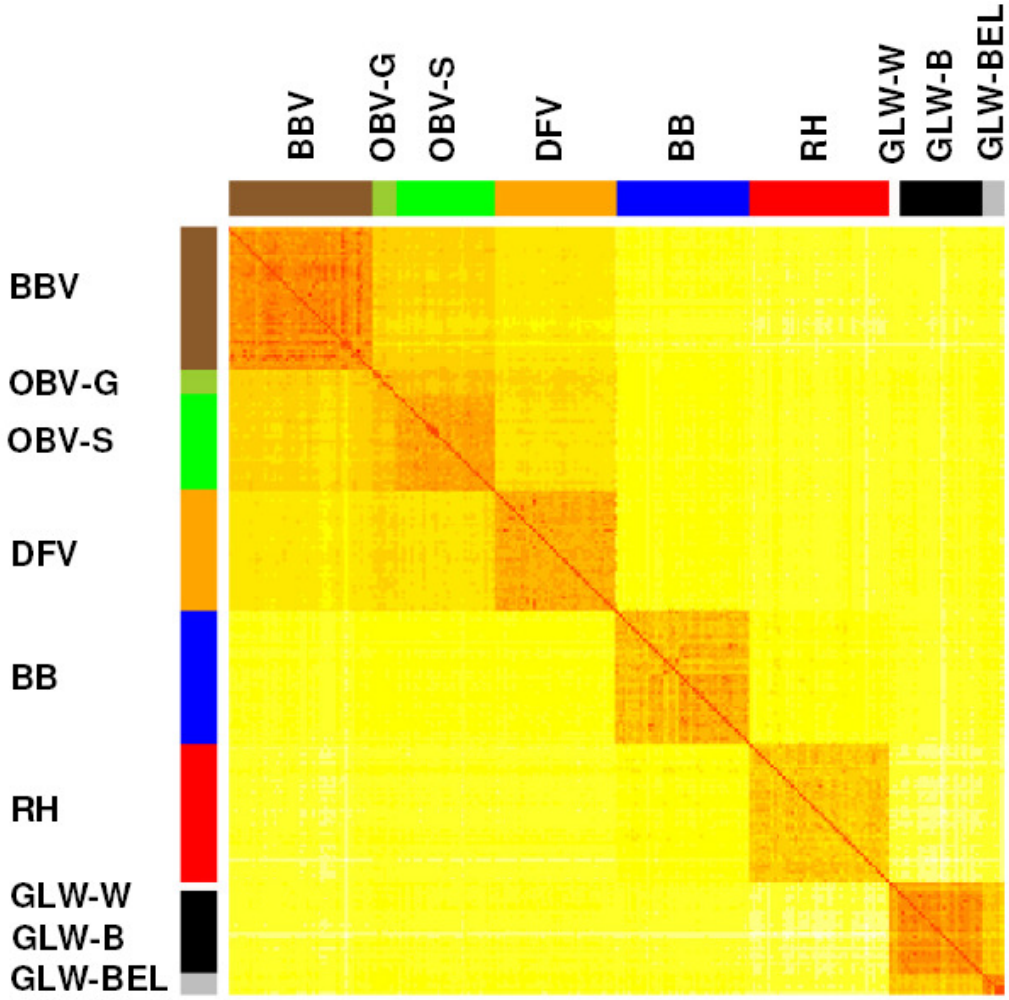


Figure 11 SPC reordered distance matrix (1-IBS). From this organized matrix one can easily infer the shapes of the identified breeds and the relationship between them, with white (red) indicating large (small) distances between individuals. Each of the colored bars along the top horizontal and left vertical axes represents the determined breeds and sub-populations. With the determined clusters colored brown (BBV), yellow-green (OBV-G), green (OBV-S), orange (DFV), blue (BB), white (GLW-W), black (GLW-B) and gray (GLW-BEL).

The only exception of a clear cluster solution concerns GLW strains with a small number of analysed samples, where GLW-DUN and one white-black-pointed animal has been assigned to the GLW-B, whilst the other white-black-pointed sample has been allocated to the white grouping. Within BB population SPC fails to cluster the animals into Belgian (BB-B) and Danish (BB-D) origin (Figure 10 and Figure 11). However, the comparable low temperature level is indicating that substructure exists within this breed because a high temperature level is reflecting the robustness of the clusters obtained, i.e., if it contains a fair number of closely connected animals. Within RH the low temperature level directly refers to high diversity level determined for this breed (see Chapter 4.1), which causes that RH animals are comparable less related than animals from the other breeds.

In order to investigate individual relationships within breeds in more detail, we additional visualized the extracted SPC graph used for the final clustering solution (Figure 12). To determine close related animals within this graph, the edge line width varies in proportion of the genetic distances, with thick edges representing small genetic distances. Furthermore, the node size has been associated with the degree centrality, hereby detecting important founder individuals

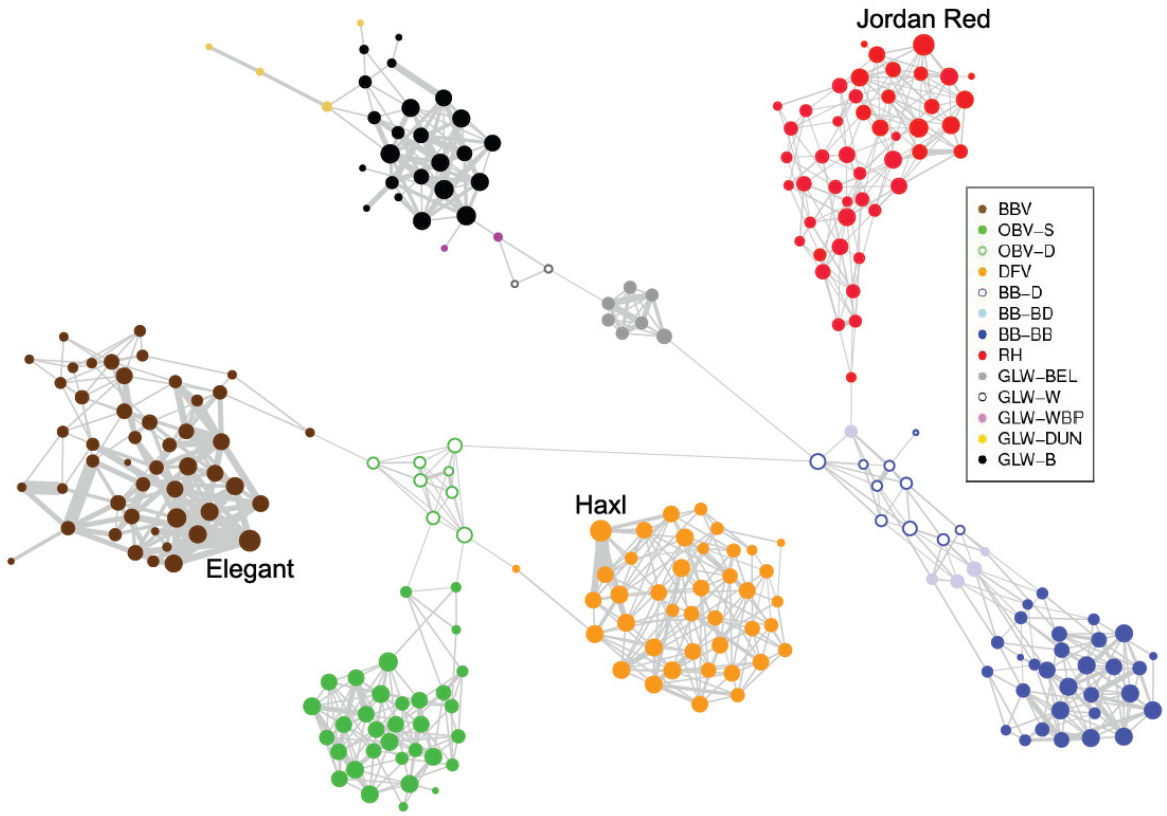


Figure 12 The network of interactions between cattle breeds as provided by the SPC graph. Each animal is represented by a node; with the different shades denote the sample origin. The network has been drawn with longer edges between vertices in different breeds than between those in the same community, to make community groupings clearer. The thickness of edges, which varies in the proportion to the genetic distance, has been used to visualize individual relationships within breeds. The node size, which varies in proportion of the number of edges per node (degree centrality), illustrates how well each individual is connected within the breed, hereby detecting important founder animals namely Elegant within BBV, Haxl within DFV and Jordan Red within RH.

The community structure as presented in the figure clearly reflects the aforementioned cluster solution (Figure 10), concurrently providing illuminating information about each animal, simultaneously revealing important founder individuals in the respective populations, namely

Jordan Red within RH as well as ***Elegant*** and ***Haxl*** within BBV and DFV respectively. Additionally, the topology of this network highlights the population structure within breeds and the transitional positions of animals acting as hubs between subpopulations (Figure 12). In this context, it is very interesting to see that Danish bred Belgian Blue animals with Belgian Blue sire (BB-BD) (except one animal) strongly link BB-D to BB-B. Concerning the animal, that does not connect to BB-B animals although it stems from a BB-B sire (Figure 12), it has to be noted that this bull has been directly exported to Denmark for cross-breeding and has not been used as an artificial insemination sire in Belgium. Hence, the only common ancestor this animal shares with the other BB-B animals has been found in the fifth generation of his pedigree. To further investigate the complex population structure within BB we have determined the betweenness centrality of the individuals (Figure 13). Given this network structure particularly the aforementioned BB-BD animals and BB animals that directly connect to these transitional animals have been associated with high betweenness values. However, the highest betweenness value has been associated to a BB-D animal that directly connects to three BB-BD animals. Screening the pedigree of this animal has shown that this animal has a BB-B sire in second generation. Concerning the BB-D samples it has been further noticed that more than half of the samples do share a common BB sire namely ***Opticen d`au Chêne***, which is simultaneously an important founder within the whole BB population.

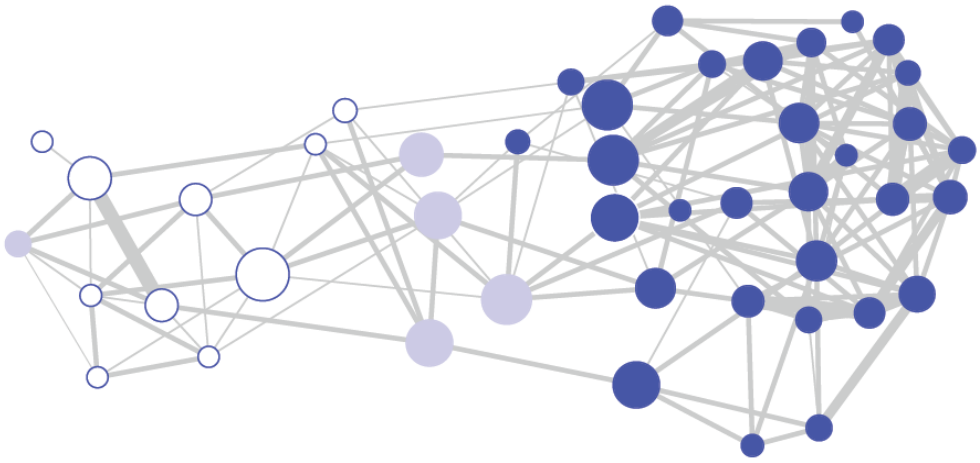


Figure 13 Network of Belgian Blue (BB) individuals from two subpopulations of Danish (BB-D) and Belgian (B-BB) ancestry. The edges between individuals have been associated with the genetic distance, while the size of each node indicates its betweenness centrality (i.e., the proportion of all shortest paths getting through the node). The different node shades identify BB-B individuals (dark blue), BB-BD individuals (light blue) and BB-D individuals (white with blue borders).

To investigate, if the Danish subpopulation (BB-D) can be separated from the main Belgian population a reduced data set, without the five BB-BD animals have been designed. According to the modularity measures SPC has been applied with an optimal k-NN = 7 (Q = 0.837). Excluding the five BB-BD animals and re-running SPC on this data, BB-D individuals have been ascertained to a single cluster, simultaneously increasing the significance of the whole BB cluster (Figure 14). However, the high temperature level of the differentiation and the fact, that the individuals have been separated with a low number of *k*-NN, indicate that these two breeds are genetically strongly related.

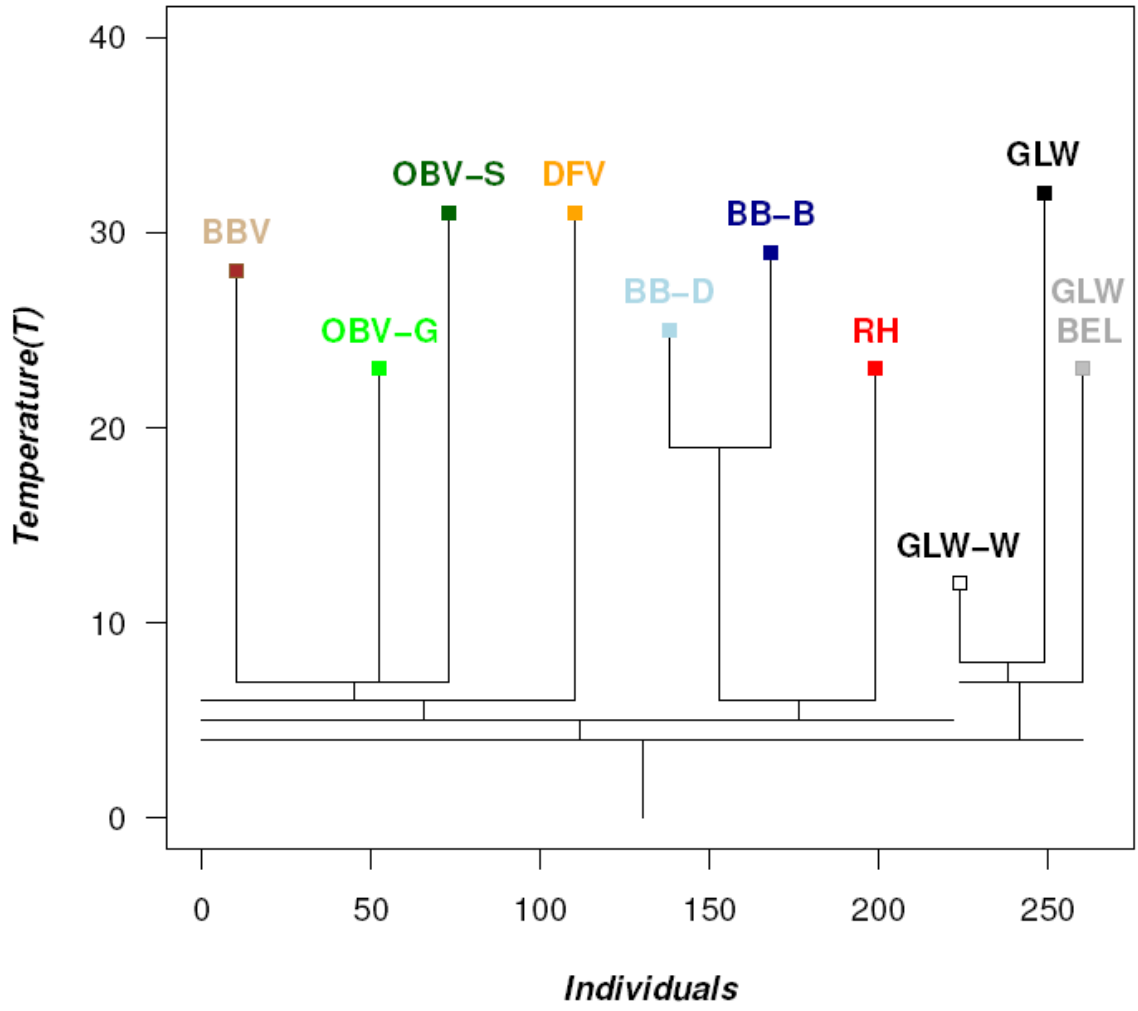


Figure 14 Super Paramagnetic Clustering excluding BB-BD individuals. Within this analysis the five BB-D animals with a Belgian sire in the first generation (BB-BD) have been excluded. Dendrogram representing the clustering of animals with optimal *k*-NN = 7. At this SPC run the animals have been separated into 10 clusters, respecting the differentiation of BB-D and BB-B into single clusters. Each cluster is represented by a box; with vertical positions determining the stability of each cluster, while horizontal positions are indicating the proximity between clusters.

Within GLW a similar situation was noticed, where one GLW-WBP animal has been assigned to the white population, whilst another animal has been loosely connected to GLW-B (Figure 12). According to GLW-DUN it has been observed, that these animals have been peripherally connected to GLW-B, where animals linked to GLW-B do show a high admixture with GLW-B, according to pedigree information. However, for more significant results in this case, additional animals have to be analyzed.

4.5 Phylogenetic network

The pair-wise F_{ST} estimates between the six breeds (upper triangle) and between nine subpopulations (lower triangle) based on a total of 46,147 markers are illustrated in Table 3.

Table 3 The pair-wise F_{ST} values between breeds and subpopulations. The upper triangle shows pair-wise F_{ST} values between six generally recognized breeds. The lower triangle recognizes the *post hoc* information on the substructure within GLW, BB and OBV. Minimum and maximum values are expressed in bold.

Breed				OBV	DFV	BB	RH	GLW	**Breed**
BBV				0.068	0.089	0.116	0.119	**0.142**	BBV
OBV-S	0.076				**0.059**	0.090	0.093	0.116	OBV
OBV-G	0.064	0.032				0.093	0.095	0.118	DFV
DFV	0.089	0.066	0.053				0.073	0.113	BB
BB-B	0.123	0.104	0.091	0.100				0.118	RH
BB-D	0.119	0.097	0.083	0.092	**0.023**				
RH	0.119	0.099	0.084	0.095	0.079	0.071			
GLW-B	**0.152**	0.133	0.122	0.127	0.130	0.123	0.127		
GLW-BEL	**0.152**	0.130	0.119	0.124	0.125	0.119	0.119	0.076	
Breed	BBV	OBV-S	OBV-G	DFV	BB-B	BB-D	RH	GLW-B	

The additional subpopulations GLW-W, GLW-WBP and GLW-DUN have been analyzed separately see results below. Considering the design with six breeds the lowest F_{ST} value (0.059) was estimated for [OBV-DFV] and highest (0.142) for [BBV-GLW], where the average over all pairs is 0.100. Regarding the existence of three additional subpopulations namely OBV-G, BB-D and GLW-BEL the lowest F_{ST} value (0.023) was estimated for [BB-B-BB-D] and highest (0.152) for [BBV-GLW-B] and [BBV-GLW-BEL], whereas an average over all pairs amounts 0.101. Comparing the variation of respective subpopulations within OBV ($F_{ST\,[OBV\text{-}S\text{-}OBV\text{-}G]}$ = 0.032) and BB ($F_{ST\,[BB\text{-}B\text{-}BB\text{-}D]}$ = 0.023), suggests that OBV-S

and OBV-G are genetically more differentiated from each other compared to BB-B and BB-D, hereby reflecting more diffuse barrier between BB strains. For instance, removing the five BB-BD animals has slightly increased the breed differentiations between the two subpopulations (F_{ST} = 0.025), which is still smaller compared to OBV subpopulations. The analysis of variation within GLW identified and showed that GLW-BEL is most distinct from GLW-B ($F_{ST\ [GLW\text{-}B\text{-}GLW\text{-}BEL]}$ = 0.0763), followed by GLW-W ($F_{ST\ [GLW\text{-}B\text{-}GLW\text{-}W]}$ = 0.0401) and GLW-DUN ($F_{ST\ [GLW\text{-}B\text{-}GLW\text{-}DUN]}$ = 0.0238).

The topology of the phylogenetic network constructed with F_{ST} distances between the six cattle breeds perfectly coincides with SPC analysis (Figure 15), hereby confirming the grouping of the Alpine cluster into [BBV-OBV] and [DFV].

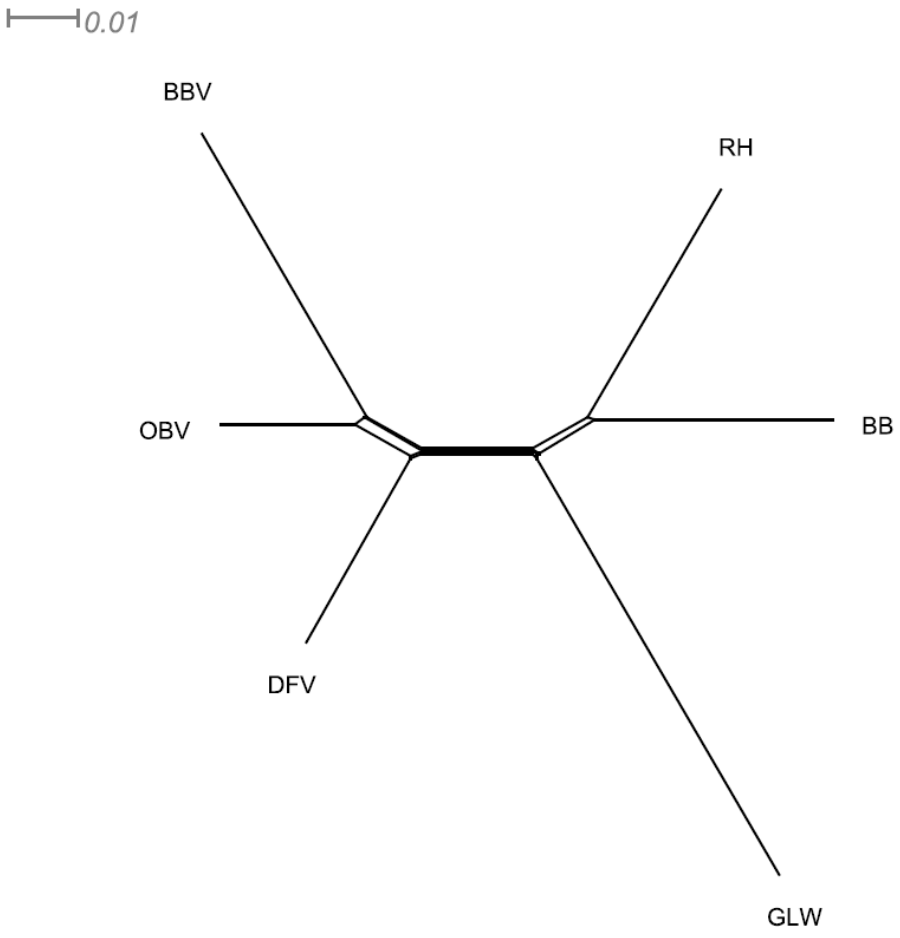


Figure 15 Phylogenetic Network. Phylogenetic network constructed from F_{ST} distances among six cattle breeds using all autosomal SNPs.

4.6 Network clustering (SPC) on low density SNP panels

Finally we investigated the performance of SPC on a low density SNP panel, therefore we applied SPC on subsets of *D*, subsequently decreasing the number of high informative SNPs (in steps of 10,000 until 5,000, followed by steps of 1,000 until 1,000 and finally in steps of 100, 50 and 10 until a set of 10 markers). Concerning the calculation of high informative SNPs it should be noted, that we have used $k = 5$ principal components according to Horn's parallel analysis. The cluster solutions provided by SPC on the reduced SNP data sets were highly accurate in accordance with the population structures extracted using all autosomal markers (46,147 SNPs) down to 200 PCAIMS. Concerning the number of clusters it has been noted that the subpopulations GLW-W and OBV-G could not have been identified as a single cluster with less than 20,000 PCAIMs. However, using 200 and down to 50 PCAIMs we still have recognized seven distinct clusters but this was not possible with less then 50 PCAIMs. These results are illustrated in Figure 16.

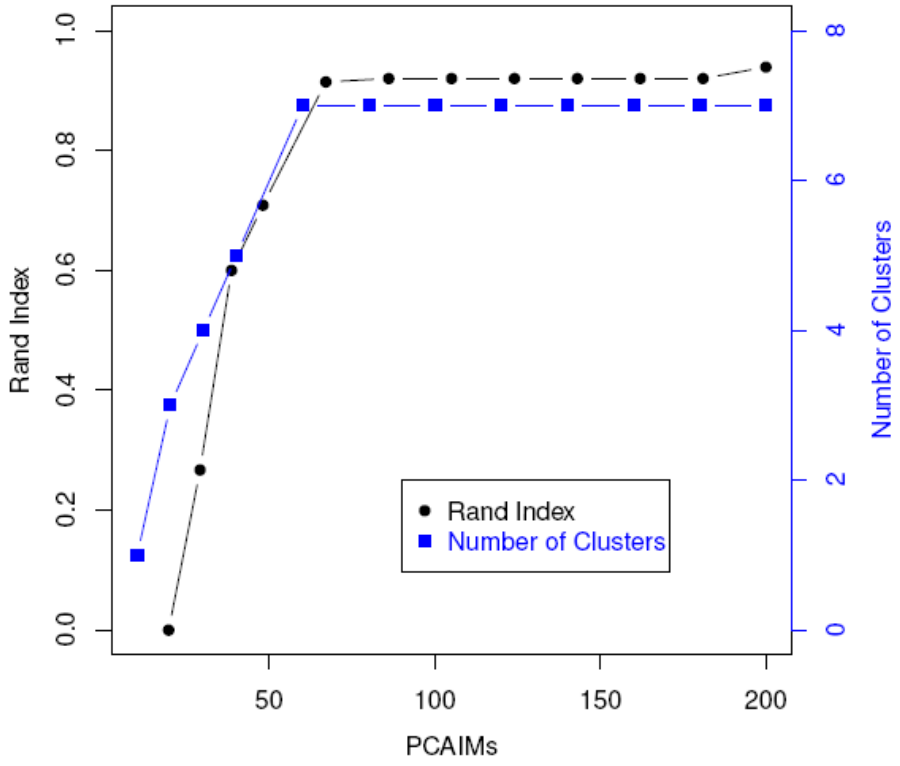


Figure 16 Adjusted Rand Index results between the cluster solutions generated with SPC on sets of 10 to 200 PCAIMs. The Rand index attains its maximum 1 only if the determined numbers of clusters and predicted membership of animals coincide with the original cluster solution. As this index is influenced by the number of clusters, the detected number of clusters is also given at the decreased number of PCAIMs.

Decreasing the PCAIMS from to 200 to 50 it was very interesting to see, that the subpopulation GLW-BEL still could have been identified as a single cluster, while the cluster solutions between BBV and OBV become more diffuse. In this context, we additional noted that animals from GLW-W clusters have been allocated to GLW-BEL cluster. The fact, that the population structure from some breeds can be still detected with small sets of PCAIMs (e.g. with the top 30 PCAIMs GLW and DFV animals have been correctly clustered, while OBV and BBV as well as BB and RH animals have been assigned to single clusters) shows that especially for close related breeds more PCAIMs are required to separate them into single clusters compared to less related ones. Furthermore, this cluster solutions lead to the assumption that the algorithm introduced by Paschou *et al.* (2007) supports SNPs that are strongly selected in the breeds analysed. Therefore we plotted the PCA scores of each SNP and tagged the regions, which have been previously associated with important phenotypes in our samples namely the absence of horns (polled locus) which is selected in GLW and mapped on proximal part of BTA1 (Drögemüller *et al.* 2005), belt pattern on BTA3 selected in GLW-BEL (Drögemüller *et al.* 2009), double muscling on BTA2 selected in BB (Grobet *et al.* 1997) and spotted loci on BTA6 selected in DFV (Fontanesi *et al.* 2010) (Figure 17).

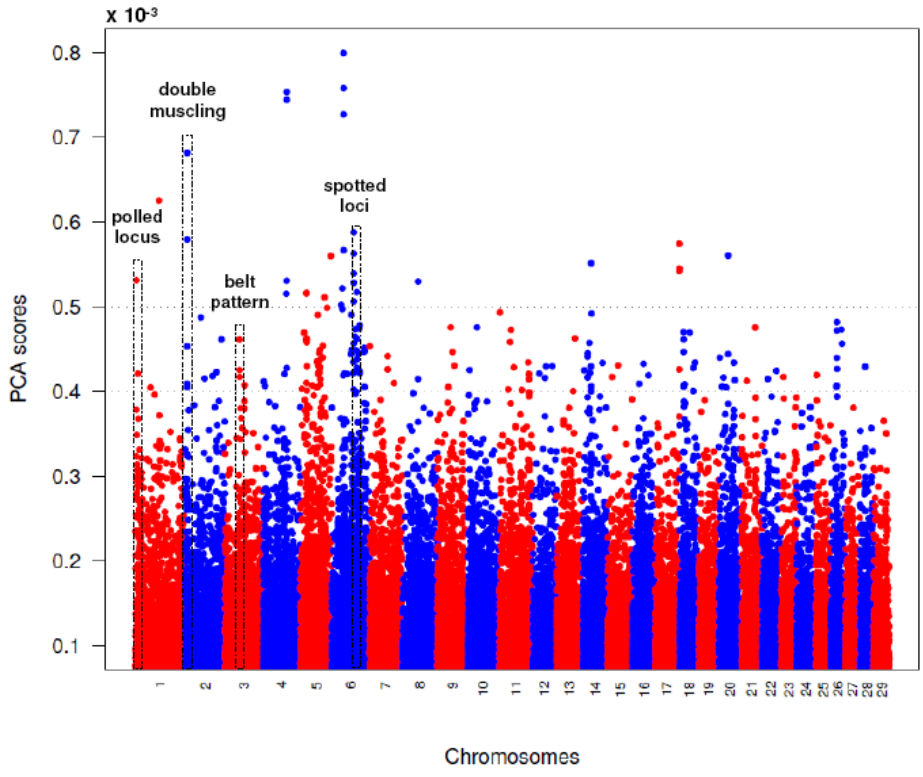


Figure 17 PCA scores for each SNP differentiating the six cattle breeds. PCA scores p_j for each SNP, with SNPs sorted by chromosome and physical position along chromosome. Presented p_j were multiplied by 10^{-3}. The dashed boxes indicate the position of the known phenotypes, while the horizontal dotted line represents the threshold for the top 30 PCAIMs.

As shown in the figure three out of the four aforementioned phenotypes have been associated with relatively high PCA scores (p_j), while SNPs associated with highest scores are located on BTA4 and BTA6.

According to Paschou *et al*. (2007) and applied method for the estimation of PCA-correlated SNPs (PCAIMs; Chapter 3.7) this approach should detect breed specific regions, i.e. markers under differential selection in applied design. Table 4 presents 30 most informative PCAIMs in the data design including six cattle breeds, i.e. PCAIMs over horizontal dotted line in Figure 17. These most probably differentially selected markers are sorted by chromosome and physical position along the chromosome. The known and putative selection signatures are listed along chromosomes.

Table 4 Top 30 PCAIMs and known selection signatures. The comments hereby summarize the current information about the listed SNPs according to present differentiation analyses and literature. SNPs with known function are deposited with the colour of the respective breed, while SNPs with unknown function stay blank. Presented p_j were multiplied by 10^{-3}.

BTA	SNPs	Position (bp)	PCA-scores	Comments
1	ARS-BFGL-BAC-4401	1547169	0.531	polled locus (GLW) (Drögemüller et al. 2005)
1	BTA-95639-no-rs	77706623	0.623	
2	BFGL-NGS-112454	7025774	0.579	double muscling (BB) (Grobet et al 1997)
2	Hapmap54028-rs29023584	7136739	0.682	double muscling (BB) (Grobet et al 1997)
4	ARS-BFGL-NGS-102407	78358392	0.515	selection signature in RH (Medugorac 2010)
4	ARS-BFGL-NGS-16805	79217610	0.531	selection signature in RH (Medugorac 2010)
4	Hapmap53144-ss46525999	79623055	0.744	selection signature in RH (Medugorac 2010)
4	BFGL-NGS-116590	79701138	0.754	selection signature in RH (Medugorac 2010)
5	ARS-BFGL-NGS-10032	19603246	0.515	Selection signature in BBV (Medugorac 2010)
5	UA-IFASA-5221	21082215	0.516	Selection signature in BBV (Medugorac 2010)
5	BTA-37834-no-rs	99617785	0.511	
5	ARS-BFGL-NGS-7711	122241829	0.560	
6	Hapmap52018-BTA-75646	29803112	0.502	
6	Hapmap33117-BTC-032493	33761326	0.522	
6	Hapmap26308-BTC-057761	37963148	0.799	selection signature in OBV (Medugorac 2010)
6	ARS-BFGL-NGS-45457	38102328	0.727	selection signature in OBV (Medugorac 2010)
6	Hapmap31285-BTC-041097	38256890	0.567	selection signature in OBV (Medugorac 2010)
6	Hapmap33628-BTC-041023	38326148	0.758	selection signature in OBV (Medugorac 2010)
6	Hapmap33128-BTC-041916	72346412	0.588	selection signature in DFV (Fontanesi et al 2010)

BTA	SNPs	Position (bp)	PCA-scores	Comments
6	Hapmap26269-BTC-041695	72377595	0.506	selection signature in DFV (Fontanesi et al 2010)
6	ARS-BFGL-NGS-38827	72401391	0.539	selection signature in DFV (Fontanesi et al 2010)
6	Hapmap27692-BTC-042876	72445235	0.528	selection signature in DFV (Fontanesi et al 2010)
6	Hapmap32220-BTC-042831	72478577	0.562	selection signature in DFV (Fontanesi et al 2010)
6	Hapmap56688-rs29025335	83365546	0.517	
8	BTB-00348223	54529421	0.530	
14	BTB-01532239	22634363	0.551	
17	ARS-BFGL-NGS-24012	76186394	0.542	QTL affecting live weight (MacNeil and Grosz 2002)
17	ARS-BFGL-NGS-108169	76208699	0.574	QTL affecting live weight (MacNeil and Grosz 2002)
17	ARS-BFGL-NGS-18349	76454249	0.545	QTL affecting live weight (MacNeil and Grosz 2002)
20	BTA-103550-no-rs	32979501	0.561	

To further investigate and illustrate the ability of the PCAIM approach to detect breed specific markers or selection signatures, we analyzed a dataset including only GLW-B and GLW-BEL samples (30 animals, 41,174 SNPs). Within this data set two principal components were found significant. Screening the 30 top PCAIMs especially SNPs located on BTA3 and BTA21 have been associated with highest PCA scores (Figure 18).

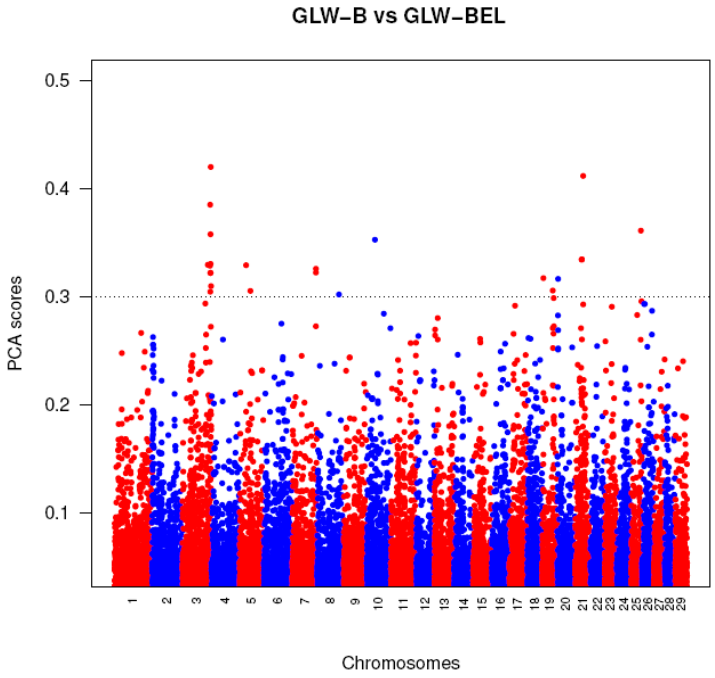


Figure 18 PCA scores for each SNP differentiating GLW-B from GLW-BEL. PCA scores p_j for each SNP, with SNPs sorted by chromosome and physical position along chromosome. Presented p_j were multiplied by 10^{-3}.The SNPs with highest PCA scores are hereby most informative to separate these two breeds. The horizontal dotted line represents the threshold for the top 30 PCAIMs.

5 DISCUSSION

5.1 General Features of network clustering (SPC)

Genotypic data is now widely used to infer population structures of a wide range of living organisms (Baumung *et al.* 2004; Li *et al.* 2007; Li *et al.* 2008; Li *et al.* 2009). The two main tools currently applied for identification of population structure and subdivision are based on parametric e.g. STRUCTURE software (Pritchard *et al.* 2000) and non-parametric approaches e.g. PCA (Menozzi *et al.* 1978; Pritchard *et al.* 2000). Consequently recent studies have started to use both approaches to ascertain population structure in livestock populations (Salmela *et al.* 2008; Kijas *et al.* 2009; The Bovine HapMap Consortium 2009). However, with current genotyping technologies, the vast amount of data derived from thousands of individuals and thousands of markers, both approaches are becoming operationally impractical to provide an optimal solution to assess population structure. STRUCTURE can only be applied on a reduced SNP data sets within a reasonable amount of computational time, and requires *a priori* complicated linkage disequilibrium (LD) modelling, under assumptions that do not hold when SNPs are densely genotyped (Gao & Starmer 2007). PCA is limited to the utilization of a clustering algorithm (e.g. k-means) which can not give the optimal number of clusters objectively (Gao & Starmer 2007) furthermore visualization of a large number of PCAs simultaneously to reveal population structure is not possible, limiting complex interpretation of relationships to a series of two dimensional visual expressions., To over come these limitations, we have introduced a hierarchical clustering method (SPC) that can take advantage of the vast amount of SNP markers available, simultaneously revealing the degree of population structure without any *a priori* ancestry information.

In order to evaluate the efficiency to detect population structure without any ancestry information, we have compared the findings of SPC with common applied methods namely PCA and STRUCTURE. The latter two methods are performing well to separate the animals into the single breeds. However, these methods could neither reproduce the phylogenetic relationships between breeds, nor reveal the correct clustering of subpopulations that would have suggested recent admixture in some breeds. They were even less suited to highlight the central position of important founder animals within breeds. By contrast SPC analyses of these breeds (Figure 10) revealed a hierarchical structure between breeds that perfectly corresponds to the known phylogenetic relationship between breeds, hereby suggesting the

existence of subpopulations within breeds. Given the network structure (Figure 12) which arises trough the clustering procedure, additionally highlighted important founder individuals within breeds and the important roles of admixed animals, which account for the clustering solution. The combination of all these findings implies that cattle breeds can be structured into different subpopulation respective to coat-colour strains and geographic origin and reveal degrees of admixture on a very fine scale. These results are therefore in agreement with the genetic structure revealed with classical population genetic analysis (Wrights F_{ST}) (Weir & Cockerham 1984) revealing shortest F_{ST} distances between GLW-DUN and GLW-B and recently admixed subpopulation BB-D and BB-B. We are aware that these breed differences are not significant, due to the small sample sizes. However, in this case, they highlight the advantages of the network clustering approach since the findings provided by SPC perfectly reflect the known breed history.

A major advantage of SPC is that it has a single global optimum that can be obtained analytically, and so multiple runs of SPC unambiguously produces the same results. By contrast admixture-based models can have multiple local optima, and the computational algorithms used can produce different results depending on their starting point. The program STRUCTURE detected the existence of subpopulations at an optimal K = 7, but only in different runs and on average with lower *LnP(D)*. Super Paramagnetic Clustering supports the results of STRUCTURE at optimal K = 7 but additionally provides essential add-on information about population structure and suggests an optimal number of clusters K = 9 at *k*-NN = 10 chosen at maximal modularity (*Q*). The admixture-based models like STRUCTURE are more effective in discrete settings while PCA performs better on continuous settings (Engelhardt & Stephens 2010).

The analysis of population structure and individual relationships in subdivided populations of domesticated farm animals as well as in subpopulations of most wild species presents the combination of some discrete and continuous subpopulations. According to our experience the admixture-based model is strongly affected by the presence of isolated and effectively small subpopulations within a continuous setting (Medugorac *et al.* 2009; Ramljak *et al.* 2011). The delta *K* approach (Evanno *et al.* 2005) applied to such data sets underestimates the true number of clusters (McKay *et al.* 2008; Medugorac *et al.* 2009; Ramljak *et al.* 2011). The results presented here demonstrate the superiority of SPC to handle the combination of discrete and continuous subpopulations by not imposing a partition on BB population when

there are no natural subpopulations present (Blatt *et al.* 1996). In this context, we have demonstrated the ability to identify key animals and to maximise population differences within breeds, which is exceedingly useful for the design of genome-wide association analyses (Matukumalli *et al.* 2009).

There are further developments of the model-based algorithms as implemented in the program STRUCTURE (ADMIXTURE and FRAPPE). These programs are faster than STRUCTURE and consequently are amenable to the analysis of larger data sets in reasonable computing time. Nonetheless, all other attributes of the program STRUCTURE are kept, e.g. multiple local optima during multiple runs and sensibility to presence of isolates within a continuous setting.

Besides the self-organized determination of the number of clusters, SPC additionally determines the robustness of each cluster identified in a hierarchical tree of clusters. As a consequence, clusters identified at lower hierarchical level are less stable. Concerning our results the low robustness of some clusters identified can be interpreted as follows:

(i) Within BBV it reveals a close relationship between BBV and OBV, since BBV contains admixed animals showing an OBV contribution from up to 50% according to pedigree informations. In this context, we have demonstrated that removing admixed animals from the data the stability of the cluster increases and that clusters detected at low temperatures might contain putative substructures (Figure 14).

(ii) At RH the low cluster stability directly relates to the reported ascertainment bias introduced in the construction of the BovineSNP50 chip assay (Gautier *et al.* 2009; Matukumalli *et al.* 2009; Gautier *et al.* 2010; Neto & Barendse 2010), since SNPs are enriched and thus biased with sequences derived from Holstein Friesian populations, which is resulting in less related animals compared to the other breeds analysed (Figure 10). According this result the determined genetic parameters can be taken as evidence, with RH showing highest diversity values at all parameters (Table 2).

(iii) Within the determined subpopulations (OBV-G, GLW-W and GLW-BEL) the low robustness reveals the close relationships of these animals to the main population. However these results are not representative due to the small sample sizes within these groups.

Our results demonstrate that network analyses provide a holistic and powerful approach to detect fine-resolution population structure without any *a priori* ancestry information. In this context, network tools can be of major importance for association and phylogeny studies. Furthermore, our results suggest that network analyses can be especially useful to study the population structures of indigenous breeds with unknown or considerably reduced population information, hereby providing essential information for conservation and genetic resource management decisions in domesticated species (Ruane 1999). At the same time, it is becoming feasible to detect putative important founders within breeds and to accurately study gene flow between and within populations (Rozenfeld *et al.* 2008) by applying specific network properties such as the degree centrality and betweenness centrality. These properties emphasize again the advantages of network illustrations and analysis in the characterization of livestock populations. In essence, this study shows, in contrast to previous assertions (Turakulov & Easteal 2003), that it is possible to detect fine-scale population structures only from genetic distances.

5.2 Specific Features of network clustering (SPC)

The main feature of SPC is that the multi-dimensional interactions of individuals are limited to the mutual neighbourhood criterion (*k*-NN). With this criterion, computational efficiency is increasing and running time is greatly reduced. For example, a complete run given a whole-genome SNP panel has not taken longer than two minutes on an Intel Core2, 3 GHz CPU with 3.48 GB of RAM. Another feature of our approach is the lack of manually-tuneable parameters other than the *k*-NN which can be automatically chosen by maximisation of modularity (*Q*).

However, it has been shown that the clustering is strongly dependent on the given numbers of *k*-NN. In order to determine optimal *k*-NN we have introduced *Q* to SPC. Maximizing *Q* in different subsets of data, e.g. with (260 samples) and without (255 samples) crossed Danish-Belgian sire we have consistently received higher *Q* values and a community structure which is in an agreement with the expected genetic structure according to pedigree data and the known breed history. However, determination of an optimal *k*-NN still remains challenging, since *Q* shows a flat distribution around *k*-NN = 6-40. Our results demonstrate that a step-wise reduction of the numbers of *k*-NN offer additional insight into underlying population structure, hereby concentrating on *k*-NN = 10, since that number has been found to be appropriate for most data sets (Blatt *et al.* 1997). However, in near future better methods should be available on the optimal choice of *k*-NN for such partitioning problems (Maier *et al.* 2009). At present the four steps involved to detect fine-scale population structures are:

(i) Estimate allele sharing distances from genome-wide SNP survey for all pair-wise combinations of individuals;

(ii) Determine optimal *k*-NN in a post-evaluation process using modularity (*Q)*;

(iii) Create a network clustering using the SPC algorithm;

(iv) Visualize the SPC graph with software tool like CYTOSCAPE.

5.3 Performance on low density SNP Panels (PCAIMs)

Microsatellites and SNPs are used for a diversity of scientific studies and commercial applications in cattle, such as linkage mapping (Grosse *et al.* 1999), genetic diversity and differentiation (Ibeagha-Awemu & Erhardt 2005), individual identification and kinship investigations (Heaton *et al.* 2002). Simulation-based studies show that genetic analysis requires a large number of SNP markers compared to microsatellite markers (reviewed by Morin *et al.*, 2004) e.g. previous studies have been focused only on a small set of microsatellite loci, mostly considering the 30 microsatellite markers recommended by ISAG/FAO working group (http://dad.fao.org) (Kantanen *et al.* 2000; Wiener *et al.* 2004; Cymbron *et al.* 2005; Tapio *et al.* 2006; Li *et al.* 2007). Recent developments since this recommendation has been the interest to investigate genome wide informative (population specific) SNPs. To obtain informative SNP marker sets within populations, various parameters have been suggested (e.g. estimations of paternity exclusion (PE), cumulative paternal exclusion probability (PEP) and polymorphism information content (PIC). These criteria's are based on the allelic frequencies of each marker within the studied population. Hence, these measures are more powerful to determine high informative microsatellite marker sets, due to the higher information content (i.e. heterozygosity, allele frequencies) of multi-allelic microsatellites over that of the bi-allelic SNPs (Herraez *et al.* 2005).

For SNP marker sets the currently applied methods to determine informative marker sets are the informativeness for assignment (I_n) as defined by Rosenberg *et al.* (2003) (Kijas *et al.* 2009) and the method introduced by Paschou *et al.* (2007) that describes so called PCA-correlated SNPs (PCAIMs) (Lewis *et al.* 2010). Informativeness for assignment is an F_{ST}-correlated measure that computes the mutual information on allele frequencies, while PCAIMs are determined by the correlation within the subspace spanned by the number of significant principal components. The major advantage of this method is, that compared to I_n, the informative SNPs are not determined based on allele frequencies, where the knowledge of individual membership to a studied population is a prerequisite. Furthermore, Paschou *et al.* (2007) demonstrated on a data set covering 10,805 SNPs genotyped in 11 populations, that PCAIMs performs slightly better than I_n-SNPs. Since the final number of SNPs needed to describe population structure is not directly provided by the method, we found it most attractive to couple this approach with SPC. In previous studies (Paschou *et al.* 2007; Lewis *et al.* 2010) particularly k-means clustering have been used to determine the number of

minimum SNPs needed to describe population structure, due to the simplicity of this algorithm. The fact that k-means can not give the optimal number of clusters objectively (Gao & Starmer 2007), we suggest SPC is a more suitable method to determine the optimal number of PCAIMs due to its superiority by not imposing a partition on the data when there are no natural subdivisions present (Blatt *et al.* 1997). Hence, in this study, we have attempted to determine a minimum number of SNPs that guarantee a true assignment of animals respecting the existence of subpopulations by applying SPC on subsets of PCAIMs.

Comparing the results achieved in this study using cattle with those presented in human (Paschou *et al.* 2007), we needed slightly more PCAIMs to guarantee a true assignment of the animals. The underlying population structure of these two species as well as applied data sets is fairly different and different numbers are to be expected. However, the reason that obviously more markers are needed to investigate the population structure of cattle breeds may reflect the relative high relationships among investigated cattle breeds compared to human samples and likely differences in effective population size (*Ne*). The relationships between livestock samples are generally expected to be higher than in humans, where artificial insemination and cross-breeding is widespread. As our results demonstrate, especially closely related populations namely the subpopulations GLW-W, GLW-WBP and GLW-W, which could be detected using all autosomal SNPs (46,147), were not detectable with even hundreds of PCAIMs. Considering only the main populations we showed that using 200 PCA-correlated SNPs, we were able to assign the studied animals with 100% accuracy to their population of origin. However, a final advice on an optimal number of PCAIMs can not be provided at this stage. As our results show, the identification of the optimal number of PCAIMs is highly variable and strongly relies on the structure of the investigated data (e.g. Lewis *et al.* 2010 suggested an optimal number of 2,000 PCAIMs analysing 13 different cattle breeds as presented in the BovineHapMap project).

Concerning the general application of the determined PCAIMs in this study, the screening of the top PCAIMs shows, that SNPs identified as population structure informative in this study are likely to refer to selection signatures within the breeds analyzed. This suggests that population structure in the formation of breeds has influenced different parts of the genome. Hence, to get a similar recommendation comparable with that presented by ISAG/FAO working group for microsatellite markers a large number of studies excluding ascertainment bias at population and marker level is required to develop a SNP marker set suitable for use in

different cattle breeds and populations. Concerning the number of markers, we expect that the number of SNPs has to be at least four times more compared to microsatellites to achieve the same resolution as suggested by Morin et al. (2004). In this context it should be noted that Paschou *et al.* (2008) introduced an additional method, a so-called Rank Revealing Decomposition, which corrects PCAIMs for linkage disequilibrium (LD). Through this process the number of PCAIMs can be further reduced, simultaneously supporting a true clustering of animals. In this study, this approach has not been examined due to low density of SNPs in bovine compared to human samples. However, since the number of SNPs is steadily increasing in bovine, this approach should be also implemented in future studies.

5.4 Performance of PCAIMs to detect selection signatures in cattle breeds

Besides the possibility to investigate the population structure on low informative SNP panels, we have also demonstrated that PCAIMs could possibly be used to detect population specific markers (Table 4). As shown in this table, the derived set of PCAIMs identify well known breed specific regions namely polled locus on BTA1 (Drögemüller *et al.* 2009) (GLW), double muscling on BTA2 (Grobet *et al.* 1997) and a QTL region on BTA17 affecting live weight (MacNeil & Grosz 2002) (BB) as well as the recently published spotted loci on BTA6 (DFV) (Fontanesi *et al.* 2010). The high informative markers detected on BTA4 and BTA5 show high evidence with selection signatures associated with RH and Braunvieh population (BBV and OBV) (Medugorac *et al.* 2010) using the commonly applied association mapping approach as suggested by Charlier *et al.* (2008) and the Cross Population Extended Haplotype Homozygosity (XP-EHH) (Sabeti *et al.* 2007). As our results demonstrate PCAIMs are especially useful in a variety of different research scenarios, such as association mapping, conservation studies and population genetics (Pritchard & Donnelly 2001; McKeigue 2005; Wan *et al.* 2010).

The PCAIMs could also be potentially useful for investigators conducting genome-wide association studies considering a pair-wise breed comparison e.g. comparing GLW-BEL with GLW-B. In particular we detected a SNP with genome-wide highest PCA-Score on BTA3 matching exactly the belted locus (Drögemüller *et al.* 2009). The SNP associated with the second highest value is located on BTA21, where *Bos Taurus* Syntaxin binding protein 6 (STXBP6) have been mapped (Zimin *et al.* 2009). Hence, we suggest implementing this approach in a pair-wise breed comparison as suggested by Matukumalli *et al.* (2009) to

effectively detect breed specific markers and possible selection signatures related to breed formation.

The SPC method is simple and computationally fast (e.g. the genome-wide analysis undertaken in this study required less than one minute) and thus allows the analysis of very large genome-wide data sets with thousands of individuals. Furthermore, this method does not rely on any assumptions (e.g. inheritance pattern, ancestral alleles and selection variations) or modeling of the data. Therefore it is conceivable to combine PCA-scores with commonly applied test statistics e.g. XP-EHH (Sabeti *et al.* 2007) and integrated Haplotype Score (iHS) (Voight *et al.* 2006) as suggested by Grossman *et al.* (2010) to improve genome-wide detection and characterization of positive selection in livestock populations.

6 CONCLUSION

6.1 General conclusion on the Utility of network clustering

In this study, we have described and demonstrated that network-based analysis can be applied successfully to infer the genetic structure in subdivided populations at low levels of genetic differentiation without any *a priori* knowledge of clustering and stratification. This tool provides new insights into the history of domesticated breeds of livestock because of its ability to reduce the level of complexity of large-scale data sets. It would be of valuable help to extract the population structure and stratification within single populations with appropriate relevance to association studies. Hence, this new approach described here is a very valuable complement and alternative to the time-consuming clustering programs because this algorithm detects fine-scale population structures with a remarkable reduction of computing time and effort, as it works within seconds exploiting a whole-genome SNP survey.

6.2 General conclusion on the Utility of PCAIMs

Applying SPC on small sets of PCAIMs shows that it becomes feasible to uncover the structure of cattle breeds without knowing *a priori* the origin of individuals. In comparison to some studies of human subpopulations more SNPs are needed to guarantee a true clustering of cattle individuals in here investigated samples. Furthermore, we have shown that for fine-scale population structures considerable more markers are needed. However, our results suggest that PCAIMs maybe a valuable help for geneticists in analyzing complex population structures using ever-increasing genome-wide data sets. Furthermore, it has been demonstrated that PCAIMs are potentially useful to determine population specific SNPs hereby identifying putative selection signatures within breeds. The preliminary results of this study demonstrate that PCAIMs could be very powerful to detect these signatures carrying out a pair-wise breed comparison. The results from this study suggest that PCAIMs derived from SPC as a possible independent selection signature test.

Hence the three stated objectives considering the application of network theory in population genetics, to identify the minimum number of SNPs and the application of PCAIMs to detect selection signatures could have been met.

7 SUMMARY

The aim of this study was to develop and approve a new procedure for determining genome-wide population structures with a high density SNP chip in cattle (BovineSNP50 BeadChip). Furthermore, this study investigates the application of population informative SNPs regarding the allocation of animals to their respective breeds with low density SNP panels and regarding the detection of breed specific markers or selection signatures within cattle breeds.

The identification of groups with similar genetic profiles is one of the major challenges in population genetics. Since most existing methods still rely on *a priori* information of individual ancestry and underlying assumptions that not always respect the natural population structure. Based on network theory and recent advances in single nucleotide polymorphism (SNP) chip technology, we investigated an unsupervised network clustering approach called Super Paramagnetic Clustering (SPC) that when applied on genome-wide SNP data identifies the natural divisions of individuals into densely connected subgroups without use of any ancestry information. This methodology allows a straightforward characterization of hierarchical population structure and the detection of ascertainment bias in population structures. At the same time we investigated the application of SPC on low-density SNP panels by determining population specific SNPs so called PCA informative SNPs (PCAIMs) as introduced by Paschou *et al.*(2007). Screening the top 30 PCAIMs we also observed that these SNPs simultaneously detect regions under selection within the breeds analyzed.

We first applied SPC procedure on a whole genome-wide SNP panel of 46,147 SNPs genotyped in 260 bovine samples containing six outbred cattle breeds. Performing SPC on our bovine data, it was possible to effectively assign animals to their breed simultaneously revealing the degree of population structure within and between breeds. Analyzing closely related breeds namely Black Galloway, White Galloway and Belted Galloway, as well as German- and Swiss Original Braunvieh we show that SPC can be used to detected fine-scale subpopulation structures. Furthermore, it is becoming feasible to reveal genetically important founder individuals within breeds (e.g. Haxl in German Fleckvieh). To evaluate the performance of SPC, population structure was further examined using Principal Components Analysis (PCA) coupled with a standard clustering tool (k-means) and software STRUCTURE. Applying SPC on low density SNP panels, it was possible to effectively assign animals to their breeds down to 200 informative SNPs. However, with this minimum SNP number a

fine-scale population structure could not have been reproduced and especially the cluster solutions concerning close related breeds namely between Braunvieh and Original Braunvieh became more diffuse. Comparing these results with studies in human, where meaningful cluster solutions could have been achieved with just 30 informative SNPs, obviously more SNPs are needed in cattle to guarantee a true assignment of animals. Plotting the genome wide PCA-Score of each SNPs we have noticed that the top PCAIMs indicate regions of strong selection e.g. from the top 30 PCAIMs, 21 PCAIMs (70%) could have been associated to regions with known selection signatures in the breeds analyzed. In this context, we have also demonstrated that PCAIMs effectively identify selection signatures by performing a pair-wise breed comparison. For example by comparing belted Galloways (GLW-BEL) with black Galloways (GLW-B), SNPs covering the region causing the belt pattern were strongly associated with highest PCA scores.

In summary the methods presented in this thesis addresses consequential questions in ecology, evolution, behavior or conservation and simultaneously describe new approaches to study genome-wide population structure moving from high to low density SNP panels. The major advantages of SPC and PCAIMs presented in this study are that they are simple, computationally extremely efficient and less dependent on *a priori* assumptions and thus allow the analysis of very large genome-wide datasets with thousands of animals without prior breed and pedigree information. Furthermore, here introduced methods can be very useful for the detection of fine stratification in associations or mapping populations.

8 ZUSAMMENFASSUNG

Das Ziel dieser Arbeit war es, ein neues Verfahren für die Erfassung von Populationsstrukturen in Rinderrassen, basierend auf genomweiten SNP-Genotypen (BovineSNP50 BeadChip), zu entwickeln. Des Weiteren wurde in dieser Studie die Anwendung von hoch informativen SNPs bezüglich der Erfassung von Populationsstrukturen anhand weniger informativen SNP-Genotypen und der Identifikation von Selektion Signaturen in Rinderrassen, untersucht.

In der Populationsgenetik ist die Identifikation von genetisch homogenen Gruppen noch immer eines der am intensivsten diskutierten Themen, da die meisten Methoden für die Erfassung von Populationsstrukturen Informationen über die Abstammung benötigen. Da diese Informationen in manchen Situationen nicht zur Verfügung stehen oder die natürliche Populationsstruktur nicht korrekt wiedergeben, haben wir uns entschieden die Anwendungspotentiale der sozialen Netzwerktheorie für die Erfassung von Populationsstrukturen basierend auf genomweiten SNP-Genotypen zu analysieren. In dieser Studie wurde für die Analyse von genomweiten Populationsstrukturen die hierarchische Clustermethode Super Paramagnetic Clustering (SPC) verwendet. Bei dieser Methode werden für das Clustern, nur Beziehungen zwischen Individuen in nächster Nachbarschaft herangezogen, wodurch sich ein Netzwerk von Individuen ergibt. Die Anwendung dieser Methode ermöglicht eine rasche Charakterisierung der hierarchischen Populationsstruktur ohne jegliche Abstammungsinformationen und das Auffinden von Erfassungsverzerrungen in Populationsstrukturen. Gleichzeitig untersuchten wir die Anwendung von SPC für die Rekonstruktion von Populationsstrukturen anhand weniger informativen SNPs, so genannten PCA informativen SNPs (PCAIMS) (Paschou *et al.* 2008). Eine detaillierte Darstellung dieser SNP-Genotypen zeigte, dass diese SNPs gleichzeitig Regionen unter starker Selektion, in den untersuchten Rinderrassen, kennzeichnen.

Für die ersten vorliegenden Untersuchungen haben wir den SPC Algorithmus an einem genomweiten SNP-Genotypen Datensatz angewendet, welcher 46.147 SNPs und 260 Tiere aus sechs verschiedenen Rinderrassen umfasste. Unter der Anwendung von SPC konnten die einzelnen Tiere den entsprechenden Rassen zugeordnet und die hierarchische Populationsstruktur zwischen den Rassen aufgezeigt werden. Durch die Analyse von eng verwandten Populationen (z.B.: Deutsches Original Braunvieh und Schweizer Original

Braunvieh, sowie Schwarzen Galloways und Belted Galloways) konnte gezeigt werden, dass diese Methode in der Lage ist Subpopulationen zu erkennen und auszuweisen. Mit Hilfe von weiteren Netzwerkanalysen konnten wir zusätzlich den Einfluss von wichtigen Gründertieren (z.B. Haxl bei Deutschen Fleckvieh) relativ einfach nachweisen. Zur Evaluierung der SPC Methode wurden vergleichende Ergebnisse mit der Hauptkomponentenanalyse (PCA) und der Computersoftware STRUCTURE aufgestellt.

Die Ergebnisse der hoch informativen SNP Genotypen (PCAIMs) zeigten, dass es möglich ist mit einer signifikant reduzierten Anzahl von Markern (200 PCAIMs) die einzelnen Populationen zu unterscheiden und die Herkunft der Tiere zu bestimmen. Eine detaillierte Strukturierung innerhalb der Populationen ist bei dieser reduzierten Anzahl von SNPs jedoch nicht mehr möglich. Vergleicht man dieses Ergebnis mit aktuellen Resultaten bei Menschen (30 PCAIMs), werden bei Rindern offenbar mehr informative SNPs benötigt um eine 100%ige Herkunft der Tiere zu garantieren. Durch die detaillierte Darstellung der genomweiten PCAIMs haben wir festgestellt, dass diese SNPs gleichzeitig Regionen unter starker Selektion kennzeichnen z.B.: Von den 30 informativsten SNPs konnten 21 (70%) bekannten Selektions- Signaturen zugeordnet werden. In diesem Zusammenhang haben wir gezeigt, dass die Identifikation von Selektions-Signaturen mit Hilfe von PCAIMs am effektivsten in einem paarweisen Rassenvergleich angewendet werden kann (z.B.: In einem Vergleich von Schwarzen Galloways und Belted Galloways, sind vor allem SNPs welche die bekannte Belted Region kennzeichnen mit hohen PCA Werten assoziiert worden.

Die, in dieser Arbeit, präsentierten Methoden beziehen sich auf essentielle Fragen in der Populationsgenetik und beschreiben gleichzeitig neue Ansätze für die Erfassung von Populationsstrukturen unter der Anwendung von genomweiten und informativen SNP-Genotypen. Die wesentlichen Vorteile der hier vorgestellten Methoden sind, dass sie einfach, rechnerisch effizient und ohne individuelle Abstammungsinformationen angewendet werden können. Daher können relativ einfach große Datensätze mit tausenden von Tieren ohne die Kenntnis von Abstammungen objektiv analysiert werden und gleichzeitig Erhebungsfehler aufgezeigt werden. Darüber hinaus liefern diese Methode wichtige Erkenntnisse über die Populationsstruktur, welche ein wesentlicher Bestandteil von Assoziationsstudien ist.

9 LIST OF FIGURES

10 LIST OF TABELS

11 GLOSSARY

Allele	One form of a gene or a genetic marker.
BLAD	Bovine leukocyte adhesion deficiency
CentiMorgan (cM)	A unit for measuring genetic distance; A Morgan is 100 cMs. A cM is approximately Equivalent to a 1 % recombination value if (double) high levels of crossover are ignored.
Deoxyribonucleic Acid (DNA)	The genetic material and main focus of molecular genetics and genomics.
DNA marker	A small piece of DNA that can be identified and Used as genetic marker
Dominant markers	Genetic markers showing dominant inheritance. In genomics, most PCR-based markers are dominant markers, e.g., presence or absence of a PCR product.
Chromosome	(chroma = color, soma = body), part of the nucleus and carrier of DNA.
CVM	Complex vertebral malformation
Genetic distance (D)	A measure to quantify genetic relationship between individuals, e.g. F_{st} values
Genome	The complete set of DNA carried by a gamete.
Genotype	Genetic constitution of an individual.

Haplotype	Set of closely linked genetic markers present on one chromosome, which trend to be inherited together.
Hardy-Weinberg Equilibrium (HWE)	A state in which gene and genotypic frequencies Remain constant from generation to generation in a large random mating population without mutation, selection or migration.
Heterozygosity	State of an individual having 2 different alleles of a gene.
Identical by Descent (IBD)	Two alleles are identical by descent (IBD) if they are identical copies of the same ancestral allele.
Identical by State (IBS)	Two alleles are identical by state (IBS) if they share the same mutational expression.
Linkage disequilibrium (LD)	Gametic disequilibrium
Microsatellites	Repetitive DNA with repeats ranging in size from 1 to 6 bp. It is also referred to as simple sequence Repeat (SSR).
Minor allele frequency (MAF)	The lowest allele frequency at a locus that is observed in a population.
Network theory	The study of graphs representing the relationships of individuals.
Polymerase chain reaction (PCR)	A technique for amplifying a target portion of a DNA molecule. Some genetic markers used in genomic analyses, such as microsatellites are PCR-based markers.

PCA informative Markers (PCAIMs)	Population informative markers, e.g. SNPs which are strongly selected in one population.
Principal Components Analysis (PCA)	Linear dimensionality reduction technique that Seeks to identify a small number of "dimensions" or "components" that capture most of the relevant structure in the data.
Quantitative trait loci (QTL)	Genes controlling quantitative traits, e.g. milk yield (MY).
QTL mapping	A set of procedures for detecting genes controlling quantitative traits (QTL) and estimating their genetic effects and genome locations.
Single Nucleotide Polymorphisms (SNPs)	Bi-allelic co-dominant genetic markers see dominant markers.
Super Paramagnetic Clustering (SPC)	Algorithm that identifies robust clusters based upon network structures.

12 APPENDIX

12.1 Article for the 9th World Congress on Genetics applied to Livestock Production (2010)

SpinNet: A New Tool To Study The Genetic Structure Of Cattle Breeds With A Genome-Wide SNP Survey

M. Neuditschko[1], J. Maxa[2], I. Russ[1], J. Schär[2] and I. Medugorac[2]

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 bovine 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 a 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, [47]), Original Braunvieh (OBV, Germany [7], Switzerland [38]), German Fleckvieh (DFV, Germany [41]), Red Holstein (RH, Germany [47]), Blue Belgian (BBB, Belgium [30], Denmark [15]) and Galloway containing two sub-populations namely White- and Belted Galloway (GLW, sampled in Germany but originating from Scotland, with the color variations black [28], dun [4], white black pointed [2], white [2] and belted [7]). As pedigree information was available, 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 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 important inputs to SPC are the number of *k*-nearest neighbors (*k*-NN), 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,

[1] Tierzuchtforschung e.V. Munich, 85586 Grub, Germany
[2] LMU Munich, Chair of Animal Genetics and Husbandry, 80539 Munich, Germany

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

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.

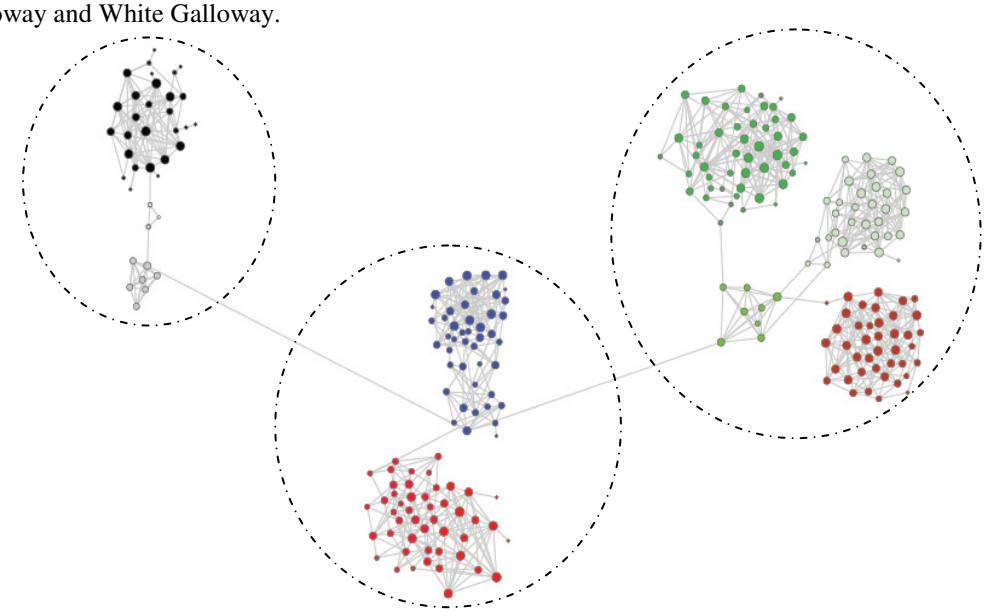


Figure 1: Community structure in cattle breeds extracted with the SPC algorithm. The breeds have firstly been separated into three major groups, denoted by the dashed circles, before they have been separated into communities. The 9 communities shown are Galloway in black dots (covering the colour variations black, dun and white black pointed), Belgian Blue (blue dots), Red Hostein (red dots), Deutsches Fleckvieh] (brown dots), European Braunvieh (dark green dots), Original Braunvieh with an separation into Swiss (cyan dots) and German (green dots) origin coloured 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.

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 modularity (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.

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 time and effort, as it works within seconds exploiting a whole-genome SNP survey.

Acknowledgements

Researchers are grateful for funding form Deutsche Forschungsgemeinschaft (DFG) through project number (ME3404/2-1) and Tierzuchforschung e.V. Munich. Furthermore we thank the numerous breeding associations who sent us samples free of charge to support this study.

References

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (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., Veit-Kensch, C.E., Tarberlet, P., Luntz, B., Mix, H.M., and Förster, M., (2009). *Mol. Ecol.,* 18:3394-3410

12.2 Poster for the 9th World Congress on Genetics applied to Livestock Production (2010)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Tierzuchtforschung e.V. Munich
Chair of Animal Genetics and Husbandry,
Faculty of Veterinary Medicine

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

1. Introduction

- Given the vast amount of SNPs, currently applied methods are becoming impractical to ascertain population structure.
- Since STRUCTURE can only be applied on LD-corrected SNP data sets, while PCA is limited to the utilization of a clustering algorithm.
- To meet these problems, we have introduced Super Paramagnetic Clustering (SPC) into population structure analysis.

2. An overview of the method

- SPC is a hierachical cluster method that aims to cluster subjects with similar genetic profiles into stable clusters along a continuous temperature gradient (T) (Fig. 1A).
- The inputs to SPC is a distance matrix (D), with the genetic distances being calculated as 1-IBS (identities by state) (Fig. 1B) and optimal number of k-nearest neighbours (k-NN).
- The major advantages of this method are the computional efficiency and the ability to extract community structures without use of any *a priori* ancestry information.

3. Results and discussion

- Applying SPC to the network of cattle breeds*, 9 communities have been extracted (Fig.2), which perfectly corresponds to classical population differentiation analysis (AMOVA).
- Given the network of cattle breeds, we were additional able to demonstrate recent admixture in BB population and to identify important founders within breeds.
- Considering the structures within BB and GLW populations, network illustrations and analysis can be exceedingly useful for association and phylogeny studies.

*__breed abbreviations:__ European Braunvieh (BBV), Original Braunvieh sampled in Switzerland (OBV-S) and Germany (OBV-G), German Fleckvieh (DFV), Blue Belgian (BB) originating form Belgian (BB-B) and Denmark (BB-D), Red Holstein (RH) and GLW with the color variations: black (GLW-B), white (GLW-W), white-black-pointed (GLW-WBP), dun (GLW-DUN) and belted (GLW-BEL)

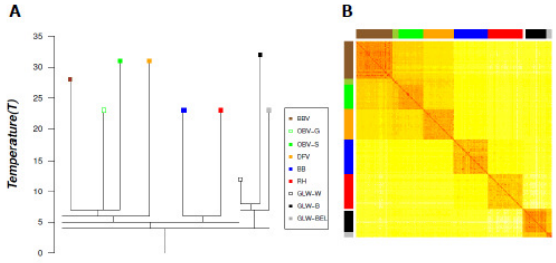


Figure 1 SPC of cattle breeds. (A) Dendrogram representing the clustering of individuals given optimal k-NN = 10. Each cluster is represented by a box, with vertical positions determining the stability of each cluster, while horizontal positions are indicating the proximity between two clusters. (B) Corresponding ordered distance matrix (D), with colored bars along the top horizontal and left vertical axes identifying the determined breeds and sub-populations. Full names of the breeds and sub-populations are mentioned in the footnote.

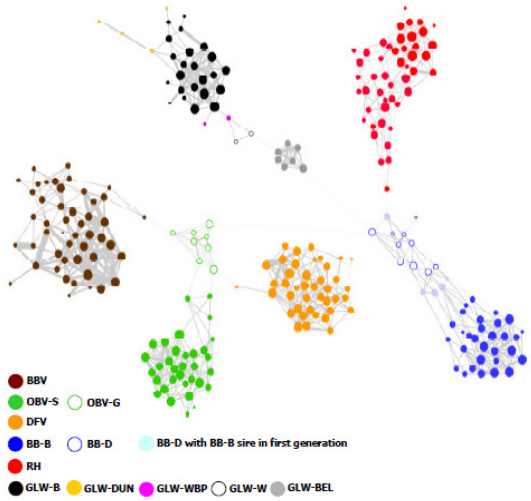


Figure 2 Community structure in cattle breeds extracted with the SPC algorithm. The network has been drawn with longer edges between vertices in different breeds than between those in the same community, to make community groupings clearer. The thickness of edges, which varies in the proportion to the genetic distance, has been used to visualize individuals relationship 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.

9th WCGALP 5th August 2010

12.3 Program files

IBS similarity matrix (PLINK)

```
1:  plink --cow --file test--cluster --matrix
```

IBS similarity matrix (VanRaden) Herman Schwarzenbacher 9.6.2010 (R.2.10)

```
#p   = vector of allele frequencies
#M = genotype matrix: rows = animals, col = SNP coded 0,1,2
#G = IBS matrix

1:  M = M-1
2:  P = 2*(p-.5)
3:  Z = t(t(M) -P)
4:  scl = (2*sum(p*(1-p)))
5:  G = Z %*%(Z)/scl
```

PCA analysis and plot (R.2.10)

```
1:  a1 = read.table("ibs_260animals.txt")
2:  a= a1
3:  an_pca = prcomp(a)
4:  summary(an_pca)
5:  plot(an_pca) #identifying the variance explained by the PCAs

#calculation of PCAs

6:  center = an_pca$center
7:  scale = an_pca$scale
8:  hm = as.matrix(a)

9:  pca1 = drop (scale(hm, center = center, scale = scale ) %*%
        an_pca$rotation[,1])
10: pca2 = drop (scale(hm, center = center, scale = scale ) %*%
        an_pca$rotation[,2])
11: pca3 = drop (scale(hm, center = center, scale = scale ) %*%
```

```
        an_pca$rotation[,3])
12: x = cbind(pca1,pca2)
13: dbb =    x[1:45,1:2]
14: dbbb =    x[11:15,1:2]
15: bbb =    x[16:45,1:2]
16: bbv =    x[46:93,1:2]
17: dfv =    x[94:134,1:2]
18. belglw = x[135:141,1:2]
19: wglw   = x[142:143,1:2]
20: wpglw  = x[144:145,1:2]
21: dglw   = x[146:149,1:2]
22: bglw   = x[150:172,1:2]
23: dobv   = x[173:180,1:2]
24: sobv   = x[181:213,1:2]
25: rh     = x[214:260,1:2]
26: plot(pca1,pca2, type="n", xlab="", ylab="")
27: title(xlab="PC1 (35%)", ylab="PC2 (15%)", font.lab=4, cex.lab
     =1.25)
28: abline(h=0, v=0, col = "lightgray", lty=2)
29: points(dbb,pch=21,col="blue")
30: points(dbbb,pch=19, col="lightblue")
31: points(bbb,pch=19,col="blue")
32: points(bbv,pch=19,col="tan4")
33: points(dfv,pch=19,col="orange")
34: points(belglw,pch=19,col="darkgray")
35: points(wglw,pch=21,col= "black")
36: points(wpglw,pch=19,col="violet")
37: points(dglw,pch=19,col="gold")
38: points(bglw,pch=19,col="black")
39: points(dobv,pch=21,col="green")
40: points(sobv,pch=19,col="green")
41: points(rh,pch=19,col="red")
42: legend(0.205, -0.06, c("BBV","OBV-S","OBV-D","DFV","BB-D","BB-
    BD", "BB-B", "RH", "GLW-BEL", "GLW-W", "GLW-WBP", "GLW-DUN",
    "GLW-B"), col =
    c("tan4","green","green","orange","blue","lightblue","blue","red
    ","darkgray","black","violet","gold","black"), pch =
    c(19,19,21,19,21,19,19,19,19,21,19,19,19), cex=.9)
```

Modularity (Q) (R.2.10)

```
1:  a = read.table("wmatrix_10edges.txt")
2:  A = as.matrix(a)
3:  sum_A = sum(A)
4:  x = 1/sum_A
5:  y = p_m%*%A%*%p_t
6:  e = x*y
7:  z = sum(diag(e))
8:  s = sum(e^2)
9:  Q = z-s
10: Q
```

Modularity plot

```
1:  a1 = read.table("input_modularity_mean.txt", header=TRUE)
2:  b1 = read.table("input_modularity_limit.txt", header=TRUE)
3:  c1 = read.table("input_modularity_avg.txt", header=TRUE)
4:  x = as.vector(a1$knn)
5:  y = as.vector(a1$modularity)
6:  k2  = a1[1:4,1:3]
7:  k2x = as.vector(k2$knn)
8:  k2y = as.vector(k2$modularity)
9:  k3  = a1[5:11,1:3]
10: k3x = as.vector(k3$knn)
11: k3y = as.vector(k3$modularity)
12: k6  = a1[13:15,1:3]
13: k6x = as.vector(k6$knn)
14: k6y = as.vector(k6$modularity)
15: k9  = a1[17:27,1:3]
16: k9x = as.vector(k9$knn)
17: k9y = as.vector(k9$modularity)
18: limx = as.vector(b1$knn)
19: limy = as.vector(b1$modularity)

20: avgx = as.vector(c1$knn)
21: avgy = as.vector(c1$modularity)
```

```
22: library(grid)
23: plot(x,y, type="n", ylim=c(0.4,0.9), xlab="", ylab="",
    axes=FALSE)
24: points(k2x,k2y, type="l")
25: points(k3x,k3y, type="l")
26: points(k3x,k3y, type="l")
27: points(k6x,k6y, type="l")
28: points(k9x,k9y, type="l")
29: points(limx,limy, pch=21, col="blue", bg="blue", cex=0.7)
30: points(avgx,avgy, pch=8, col="red", cex=1.2)
31: grid.text("K=2", vjust=10.5, hjust=-7.5,gp=gpar(fontsize=12,
    fontface=1))
32: grid.text("K=3", vjust=4, hjust=-2.5,gp=gpar(fontsize=12,
    fontface=1))
33: grid.text("K=4", vjust=-3.3, hjust= 1,gp=gpar(fontsize=12,
    fontface=1))
34: grid.text("K=6", vjust=-16.3, hjust=2.8,gp=gpar(fontsize=12,
    fontface=1))
35: grid.text("K=7", vjust=-16.3, hjust=4.6,gp=gpar(fontsize=12,
    fontface=1))
36: grid.text("K=9", vjust=-16.3, hjust=6.9,gp=gpar(fontsize=12,
    fontface=2))
37: grid.text("K=11", vjust=-12.5, hjust=6.4,gp=gpar(fontsize=12,
    fontface=1))
38: axis(1,
    c(0,5,10,15,20,25,30,35,40,50,60,70,80,90,100),cex.axis=0.7)
39: axis(2, las=2,c(0.4,0.5,0.6,0.7,0.8,0.9),cex.axis=0.8)
40: box()
41: title(xlab="k-NN", ylab="Modularity (Q)", font.lab=4,
    cex.lab=1.25)
```

Heatmap (R.2.10)

```
1:  require(plotrix)
2:  library(grid)

3:  a1 = read.table("heatmap_spc_sorted.txt")
4:  a = a1[,-1]
```

```
5:  mat = as.matrix(a)
6:  b1 = read.table("colorbar.txt", header=TRUE)
7:  col = as.vector(b1$col)
8:  rlab = col
9:  clab = col

10: library(gplots)

11: heatmap.2(mat, Rowv=FALSE, Colv=FALSE, dendrogram="none",^
    density.info="none", key=FALSE,trace="none",
    labRow=FALSE,labCol=FALSE,symm=TRUE, margins=c(7, 7),
    ColSideColors=clab, RowSideColors=rlab)
```

Determining the Number of Significant Principal Components (R.2.10)

```
1:  require(paran)
2:  a1 = read.table("ibs_matrix.txt")
3:  paran(a1)
```

Selecting PCA-Correlated (SNPs) Paschou et al. (2008) Matlab

```
1:  Load test.txt -ASCII
2:  [scores] = PCAscores(test,k);
```

Genome-wide PCAIMs plot (R.2.10)

```
1:  a1 = read.table("pcaim_scores.txt")
2:  x = 1:46147
3:  v = a1*1000
4:  z = cbind(x,v)

5:  chr1 =  z[1:3023,1:2]
6:  chr2 =  z[3024:5476,1:2]
7:  chr3 =  z[5477:7810,1:2]
8:  chr4 =  z[7811:10090,1:2]
9:  chr5 =  z[10091:12026,1:2]
10: chr6 =  z[12027:14332,1:2]
```

```
chr7 =  z[14333:16341,1:2]
chr8 =  z[16342:18484,1:2]
chr9 =  z[18485:20328,1:2]
chr10 = z[20329:22284,1:2]
chr11 = z[22285:24325,1:2]
chr12 = z[24326:25796,1:2]
chr13 = z[25797:27381,1:2]
chr14 = z[27382:28909,1:2]
chr15 = z[28910:30417,1:2]
chr16 = z[30418:31838,1:2]
chr17 = z[31839:33285,1:2]
chr18 = z[33286:34481,1:2]
chr19 = z[34482:35703,1:2]
chr20 = z[35704:37143,1:2]
chr21 = z[37144:38371,1:2]
chr22 = z[38372:39525,1:2]
chr23 = z[39526:40494,1:2]
chr24 = z[40495:41651,1:2]
chr25 = z[41652:42526,1:2]
chr26 = z[42527:43486,1:2]
chr27 = z[43487:44363,1:2]
chr28 = z[44364:45215,1:2]
chr29 = z[45216:46147,1:2]

library(grid)
plot(z, las=1, xlim = c(0, 46147),ylim = c(0.1,0.8), type="n",
    axes=F,     xlab="Chromosomes", ylab="PCA scores")
axis(2, ylim = c(0.1,0.8), col="black", las=2)

abline(h=0.5, v=NULL, col = "black", lty=3)
abline(h=0.465, v=NULL, col = "gray",  lty=3)

points(chr1, pch=19, col="red", cex =.6)
mtext("1",side=1,line=0.5,at= 1512, cex =.7, las=2)
points(chr2, pch=19, col="blue", cex =.6)
mtext("2",side=1,line=0.5,at= 4250,cex =.7, las=2)
points(chr3, pch=19, col="red", cex=.6)
mtext("3",side=1,line=0.5,at= 6644, cex =.7, las=2)
```

```
45: points(chr4, pch=19, col="blue", cex=.6)
46: mtext("4",side=1,line=0.5,at= 8951, cex =.7, las=2)
47: points(chr5, pch=19, col="red", cex=.6)
48: mtext("5",side=1,line=0.5,at= 11059, cex =.7, las=2)
49: points(chr6, pch=19, col="blue", cex=.6)
50: mtext("6",side=1,line=0.5,at= 13180, cex =.7, las=2)
51: points(chr7, pch=19, col="red", cex=.6)
52: mtext("7",side=1,line=0.5,at= 15337, cex =.7, las=2)
53: points(chr8, pch=19, col="blue", cex=.6)
54: mtext("8",side=1,line=0.5,at= 17413, cex =.7, las=2)
55: points(chr9, pch=19, col="red", cex=.6)
56: mtext("9",side=1,line=0.5,at= 19407, cex =.7, las=2)
57: points(chr10, pch=19, col="blue", cex=.6)
58: mtext("10",side=1,line=0.5,at= 21307, cex =.7, las=2)
59: points(chr11, pch=19, col="red", cex=.6)
60: mtext("11",side=1,line=0.5,at= 23305, cex =.7, las=2)
61: points(chr12, pch=19, col="blue", cex=.6)
62: mtext("12",side=1,line=0.5,at= 25061, cex =.7, las=2)
63: points(chr13, pch=19, col="red", cex=.6)
64: mtext("13",side=1,line=0.5,at= 26589, cex =.7, las=2)
65: points(chr14, pch=19, col="blue", cex=.6)
66: mtext("14",side=1,line=0.5,at= 28146, cex =.7, las=2)
67: points(chr15, pch=19, col="red", cex=.6)
68: mtext("15",side=1,line=0.5,at= 29664, cex =.7, las=2)
69: points(chr16, pch=19, col="blue", cex=.6)
70: mtext("16",side=1,line=0.5,at= 31128, cex =.7, las=2)
71: points(chr17, pch=19, col="red", cex=.6)
72: mtext("17",side=1,line=0.5,at= 32562, cex =.7, las=2)
73: points(chr18, pch=19, col="blue", cex=.6)
74: mtext("18",side=1,line=0.5,at= 33884, cex =.7, las=2)
75: points(chr19, pch=19, col="red", cex=.6)
76: mtext("19",side=1,line=0.5,at= 35093, cex =.7, las=2)
77: points(chr20, pch=19, col="blue", cex=.6)
78: mtext("20",side=1,line=0.5,at= 36424, cex =.7, las=2)
79: points(chr21, pch=19, col="red", cex=.6)
80: mtext("21",side=1,line=0.5,at= 37758, cex =.7, las=2)
81: points(chr22, pch=19, col="blue", cex=.6)
82: mtext("22",side=1,line=0.5,at= 38949, cex =.7, las=2)
```

```
83: points(chr23, pch=19, col="red", cex=.6)
84: mtext("23",side=1,line=0.5,at= 40010, cex =.7, las=2)
85: points(chr24, pch=19, col="blue", cex=.6)
86: mtext("24",side=1,line=0.5,at= 41073, cex =.7, las=2)
87: points(chr25, pch=19, col="red", cex=.6)
88: mtext("25",side=1,line=0.5,at= 42089, cex =.7, las=2)
89: points(chr26, pch=19, col="blue", cex=.6)
90: mtext("26",side=1,line=0.5,at= 43007, cex =.7, las=2)
91: points(chr27, pch=19, col="red", cex=.6)
92: mtext("27",side=1,line=0.5,at= 43925, cex =.7, las=2)
93: points(chr28, pch=19, col="blue", cex=.6)
94: mtext("28",side=1,line=0.5,at= 44790, cex =.7, las=2)
95: points(chr29, pch=19, col="red", cex=.6)
96: mtext("29",side=1,line=0.5,at= 45682, cex =.7, las=2)
97: box()
```

Scatter plot smoothing (R.2.10)

```
a1 <- read.table("SSP_LD.txt", header=TRUE)

y <- a1$Xdf
x <- a1$Dist
a <- as.data.frame(cbind(x,y))
low <- lowess(a)
x <- low$x
y <- low$y
plot(x,y, type="l")
z <- data.frame(low)

write.table(z, file="SSP_LD_rg.txt", sep="\t")
```

13 REFERENCES

Alexander D, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19, 1655-1664.

Bagrow J, Bollt E (2005) Local method for detecting communities. *Physical Review E* 72, doi: 10.1103/PhysRevE.1172.046108.

Barad O (2003) *Advanced Clustering Algorithm for Gene Expression Analysis using Statistical Methods*, Weizmann Institute.

Baumung R, Simianer H, Hoffmann I (2004) Genetic diversity studies in farm animals - a survey. *Journal of Animal Breeding and Genetics* 121, 361-373.

Behar D, Yunusbayev B, Metspalu M*, et al.* (2010) The genome-wide structure of the Jewish people. *Nature* 466, 238-242.

Blatt M, Wiseman S, Domany E (1996) Superparamagnetic Clustering of Data. *Physical Review Letters* 76, 3251-3255.

Blatt M, Wiseman S, Domany E (1997) Data Clustering Using a Model of Granular Magnet. *Neural Computation* 9, 1805-1842.

Broder A, Kumar R, Maghoul F*, et al.* (2000) Graph structure in the web. *Computer Networks* 33, 309-320.

Brohee S, Faust K, Lima-Mendez G*, et al.* (2008) NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Research* 36, doi:10.1093/nar/gkn1336.

Browning B, Yu Z (2009) Simultaneous genotype calling and haplotype phase inference improbes genotype accuracy and reduces false and positive associations for genome-wide association studies. *The American Journal of Human Genetics* 85, 847-861.

Cavalli-Sforza L (1998) The DNA revolution in population genetics. ***Science*** **14, 60-65.**

Charlier C, Coppieters W, Rollin F*, et al.* (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. ***Nature*** **40, 449-454.**

Cymbron T, Freeman A, Malheiro M, Vigne J, Bradley D (2005) Microsatellite diversity suggests different histories for Mediterranean and Northern European cattle populations. ***Proceedings of the Royal Society of London, Series B: Biological Sciences*** **272, 1837-1843.**

Davies C, Simovich M, Hathaway S (1997) Population genetic structure of a California endemic branchiopod. ***Hydrobiologia*** **359, 149-158.**

Dawson E (1999) SNP maps: more markers needed? ***Molecular Medicine Today*** **4, 419-420.**

Donetti L, Munoz M (2004) Detecting network communities: A new systematic and efficient algorithm. ***Journal of Statistical Mechanics: Theory and Experiment*** **10, P10012.**

Drögemüller C, Demmel S, Engensteiner M, Rieder S, Leeb T (2009) A shared 336 kb haplotype associated with the belt pattern in three divergent cattle breeds. ***Animal Genetics*** **41, 304-307.**

Drögemüller C, Wöhlke A, Mömke S, Distl O (2005) Fine mapping of the polled locus to a 1-Mb region on bovine chromosome 1q12. ***Mammalian Genome*** **16, 613-620.**

Eck S, Benet-Pages A, Flisikowski K*, et al.* (2009) Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. ***Genome Biology*** **10, R82.**

Engelhardt B, Stephens M (2010) Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis. ***PLOS Genetics*** **6, doi:10.1371/journal.pgen.1001117.**

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14, 2611-2620.

Falush D, Stephens M, Pritchard J (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 164, 1567-1587.

Fontanesi L, Tazzoli M, Russo V, Beever J (2010) Genetic heterogeneity at the bovine KIT gene in cattle breeds carrying different putative alleles at the spotting locus. *Animal Genetics* 41, 295-303.

Freeman L (1977) Set of measures of centrality based on betweenness. *Sociometry* 40, 35-41.

Freeman L (1978) Centrality in social networks: Conceptual clarification. *Social Networks* 1, 215-239.

Frivolt G, Bielikova M (2005) The anatomy of large-scal hypertextual web search engine. *C Proceedings of the 1 st International Workshop on Representation Analysis of Web Space*.

Gao X, Starmer J (2007) Human population structure detection via multilocus genotype clustering. *BMC Genetics* 8, doi:10.1186/1471-2156-1188-1134.

Gautier M, Flori L, Riebler A*, et al.* (2009) A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genomics* 10, doi:10.1186/1471-2164-1110-1550.

Gautier M, Laoe D, Moazami-Goudarzi K (2010) Insights into the Genetic History of French Cattle from Dense SNP Data on 47 Worldwide Breeds. *PLOS One* 5, e13038. doi:13010.11371/journal.pone.0013038.

Getz G, Levine E, Domany E (2000a) Coupled two-way clustering analysis of gene microarray data. ***Proceedings of the National Academy of Sciences of the United States of America*** **97, 12079-12084.**

Getz G, Levine E, Domany E, Zhang M (2000b) Super-paramagnetic clustering of yeast gene expression profiles. ***Physica A*** **279, 457-464.**

Girvan M, Newman M (2002) Community structures in social and biological networks. ***PNAS*** **99, 7821-7826.**

Glorfeld L (1995) An Improvement on Horn's Parallel Analysis Methodology for Selecting the Correct Number of Factors to Retain. ***Educational and Psychological Measurement*** **55, 377-393.**

Gompert Z, Forister M, Fordyce J*, et al.* (2010) Bayesian analysis of molecular variance in pyrosequence quantifies population genetic structure across the genome of Lycaeides butterflies. ***Molecular Ecology*** **19, 2455-2473.**

Goudet J (1995) FSTAT(Version 1.2): A Computer Program to Calculate F-Statistics. ***Journal of Heredity*** **86, 485-486.**

Grobet L, Martin L, Poncelet D*, et al.* (1997) A deletion in the bovine myostatin gene causes the double-musceld phenotype in cattle. ***Nature Genetics*** **17, 71-74.**

Grosse W, Kappes S, Laegrid W*, et al.* (1999) Single nucleotide polymorphism (SNP) discovery and linkage mapping of bovine cytokine genes. ***Mammalian Genome*** **10, 1062-1069.**

Grossman S, Shylakhter I, Karlsson E*, et al.* (2010) A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positve Selection. ***Sciene*** **327, 883-886.**

Gusev A, Lowe J, Stoffel M*, et al.* (2009) Whole population, genome-wide mapping of hidden relatedness. ***Genome Research*** **19, 318-326.**

Hannelius U, Salmela E, Lappalainen T*, et al.* (2008) Population substructure in Finland and Sweden revealed by the use of spatial coordinates and a small number of unlinked autosomal SNPs. *BMC Genetics* 9, 1471-2156.

Heaton M, Harhay G, Benett G*, et al.* (2002) Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. *Mammalian Genome* 13, 272-281.

Herraez DL, Schäfer H, Mosner J, Fries H, Wink M (2005) Comparison of Microsatellite and Single Nucleotide Polymorphism Markers for the Genetic Analysis of a Galloway Cattle Population. *Zeitschrift für Naturforschung* 60c, 637-643.

Hill E, McGivney B, Gu J, Whiston R, MacHugh D (2010) A genome-wide SNP-association study confirms a sequence variant (g.66493737C>T) in the equine myostatin (MSTN) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses. *BMC Genomics*, doi:10.1186/1471-2164-1111-1552.

Holmström E, Bock N, Brännlund J (2009) Modularity density of network community divisions. *Physica D* 238, 1161-1167.

Hubert L, Arabie P (1985) Comparing partitions. *Journal of Classification*, 193-218.

Hurlbert S (1971) Nonconcept of species diversity: A critique and alternative parameters. *Ecology* 4, 577-586.

Huson D, Bryant D (2006) Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* 23, 254-267.

Ibeagha-Awemu E, Erhardt G (2005) Genetic structure and differentation of 12 African Bos indicus and Bos taurus cattle breeds, inferred from protein and microsatellite polymorphisms. *Journal of Animal Breeding and Genetics* 122, 12-20.

Kalinowski S (2004) Counting alleles with rarefaction: private alleles and hierarchical sampling designs. ***Conservation Genetics*** **5, 539-543.**

Kalinowski S (2005) HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. ***Molecular Ecology Notes*** **5, 187-189.**

Käller M, Lundeberg J, Ahmadian A (2007) Arrayed identification of DNA signatures. ***Expert Review of Molecular Diagnostics*** **7, 65-76.**

Kantanen J, Olsaker I, Holm L*, et al.* (2000) Genetic diversity and population structure of 20 north European cattle breeds. ***Journal of Heredity*** **91, 446-457.**

Kernighan B, Lin S (1973) An Effective Heurisitc Algorithm for the Traveling-Salesman Problem. ***Operations Research*** **21, 498-516.**

Kijas J, Townley D, Dalrymple B*, et al.* (2009) A Genome Wide Survey of SNP Variation reveals the Genetic Structure of Sheep Breeds. ***PLOS One*** **4, e4668 doi:4610.1371/journal.pone.0004668.**

Kim J, Verdu P, Pakstis A*, et al.* (2005) Use of autosomal loci for clustering individuals and populations of East Asian origin. ***Human Genetics*** **117, 511-519.**

Kim K, Min M, An J, Lee H (2004) Cross-Species Amplification of Bovidae Microsatellites and Low Diversity of the Endangered Korean Goral. ***Journal of Heredity*** **95, 521-525.**

Krentel M (1988) The Complexity of Optimization Problems. ***Journal of Computer and System Sciences*** **36, 490-509.**

Kurida A, Daescu O, Ammann L, Kakhniashvili D, Goodman S (2007) Centrality measures for the human red blood cell interactome. ***Engineering in Medicine and Biology Workshop*****, 98-101.**

Lao O, Duijn Kv, Kersbergen P, Knijff Pd, Kayser M (2006) Proportioning whole-genome single-nucleotide-ploymorphism diversity for the identification of geographic population structure and genetic ancestry. ***American Journal of Human Genetics*** **78, 680-690.**

Lee C, Abdool A, Huang C (2009) PCA-based population structure inference with generic clustering algorithms. ***BMC Bioinformatics*** **10, doi:10.1186/1471-2105-1110-S1181-S1173.**

Leicht E, Holme P, Newman M (2006) Vertex similarity in networks. ***Physical Review E*** **73, doi: 10.1103/PhysRevE.1173.026120.**

Lewis J, Abdas C, Dadousis C*, et al.* (2010) Tracing the Origin of Cattle Breeds with PCA-based Ancestry Informative SNPs. ***Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*****, 348p.**

Li H, Han W, Zhu Y*, et al.* (2009) Analysis of genetic structure and relationship among nine indigenous Chinese chicken populations by the Structure program. ***Journal of Genetics*** **88, 197-203.**

Li J, Absher D, Tang H*, et al.* (2008) Worldwide Human Relationships inferred from Genome-Wide Patterns of Variation. ***Science*** **319, 1100-1104.**

Li M, Vilkki J, Ivanova Z*, et al.* (2007) The genetic structure of cattle populations (Bos taurus) in northern Eurasia and the neighbouring near Eastern regions: implications for breeding strategies and conservation. ***Molecular Ecology*** **16, 3839-3853.**

Liron J, Peral-Garcia P, Giovambattista G (2006) Genetic Characterization of Argentine and Bolivian Creole Cattle Breeds Assessed through Microsatellites. ***Journal of Heredity*** **97, 331-339.**

Liu K, Muse S (2005) PowerMarker: An integrated analysis environment for genetic marker analysis. ***Bioinformatics*** **21, 2128-2129.**

Liu N, Zhao H (2006) A non-parametric approach to population structure inference using multilocus genotypes. ***Human Genomics*** **2, 353-364.**

Lloyd S (1957) Least squares quantization in PCM. ***IEEE Transactions on Information Theory*** **28, 128-137.**

MacNeil MD, Grosz MD (2002) Genome-wide scans for QTL affecting carcass traits in Hereford x composite double backcross populations. ***J. Anim Sci.*** **80, 2316-2324.**

MacQueen J (1967) Some Methods for classification and Analysis of Multivariate Observations. ***Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*****, 281-297.**

Maier M, Hein M, Luxburg Uv (2009) Optimal construction of *k*-nearest neighbor graphs for identifiying noisy clusters. ***Theoretical Computer Science*** **410, 1749-1764.**

Matukumalli L, Lawley C, Schnabel R*, et al.* (2009) Development and Characterization of a High Density SNP Genotyping Assay for Cattle. ***PLOS One*** **4, e5350. doi:5310.1371/journal.pone.0005350.**

May R, Lloyd A (2001) Infection dynamics on scale-free networks. ***Physical Review E*** **64, doi: 10.1103/PhysRevE.1164.066112.**

McKay S, Schnabel R, Murdoch B*, et al.* (2008) An assessment of population structure in eight breeds of cattle using whole genome SNP panel. ***BMC Genetics*** **9, 1471-2156.**

McKeigue P (2005) Prospects for admixture mapping of complex traits. ***American Journal of Human Genetics*** **76, 1-7.**

Medugorac I, Medugorac A, Russ I*, et al.* (2009) Genetic diversity of European cattle breeds highlights the conservation value of traditional unselected breeds with high effective population size. ***Molecular Ecology*** **18, 3394-3410.**

Medugorac I, Rothammer S, Russ I, Förster M (2010) Genome-wide Mapping of Breed Differences - A reciprocal Case-Control Design. ***Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*****, 348p.**

Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthenic maps of human gene frequencies in Europeans. ***Science*** **201, 786-792.**

Morgante M, Olivieri A (1993) PCR-amplified microsatellites as markers in plant genetics. ***The Plant Journal*** **3, 175-182.**

Morin P, Luikart G, Wayne R, group atSw (2004) SNPs in ecology, evolution and conservation. ***Trends in Ecology and Evolution*** **19, 208-216.**

Mousadik A, Petit R (1996) High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. ***Theoretical and Applied Genetics*** **92, 832-839.**

Narayanan S (2005) ***The Betweenness Centrality of Biological Networks*****, Virginia Polytechnic Institute and State University.**

Nei M (1972) Genetic distance between populations. ***American Naturalist*** **106, 283-292.**

Nei M, Roychodhury A (1988) Human Polymorphyic Genes: World Distribution. ***Oxford University Press*****.**

Neto LP, Barendse W (2010) Effect of SNP origin on analyses of genetic diversity in cattle. ***Animal Production Science*** **50, 792-800.**

Newman M, Girvan M (2004) Finding and evaluating community structure in networks. ***Physical Review E*** **69.**

Newton C, Graham A (1997) PCR. Introduction to Scientific Techniques. ***BIOS Scientific Publisher Ltd.*** **2.**

Page R (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. ***Computer Application in the Biosciences*** **12, 357-358.**

Paschou P, Drineas P, Lewis J, ***et al.*** **(2008) Tracing Sub-Structure in the European American Population with PCA-Informative Markers.** ***PLOS Genetics*** **4, e1000114. doi:1000110.1001371/journal.pgen.1000114.**

Paschou P, Ziv E, Burchard E, ***et al.*** **(2007) PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations.** ***PLOS Genetics*** **3, e160. doi:110.1371/journal.pgen.0030160.**

Patterson N, Price A, Reich D (2006) Population Structure and Eigenanalysis. ***PLOS Genetics*** **Genet 2, e190. doi:110.1371/journal.pgen.0020190.**

Petit R, Mousadik A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. ***Conservation Biology*** **12, 844-855.**

Price A, Patterson N, Plenge R, ***et al.*** **(2006) Principle components analysis corrects for stratification in genome-wide association studies.** ***Nature Genetics*** **38, 904-909.**

Pritchard J, Donnelly P (2001) Case-control studies of association in structured or admixed populations. ***Theoretical Population Biology*** **60, 227-237.**

Pritchard J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. ***Genetics*** **155, 945-959.**

Purcell S, Neale B, Todd-Brown K, ***et al.*** **(2007) PLINK: a tool set for whole-genome association and population-based linkage analyses.** ***American Journal of Human Genetics*** **81, 559-575.**

Queller D, Strassmann J, Hughes C (1993) Microsatellites and Kinship. ***Trends in Ecology and Evolution*** **8, 285-288.**

Radjiman S, Lianyi H, Jian-Sheng W, Zong CY (2006) Super Paramagnetic Clustering of DNA sequences. ***Journal of Biological Physics*** **32, 11-25.**

Ramljak J, Ivankovic A, Veit-Kensch C, Förster M, Medugorac I (2011) Analysis of genetic and cultural conservation value of three indigenous Croatian cattle breeds in a local and global context ***Journal of Animal Breeding and Genetics*** **128, 73-84.**

Reichardt J, Bernholdt S (2004) Detecting Fuzzy Community Structures in Complex Networks with a Potts Model. ***Physical Review Letters*** **93, doi: 10.1103/PhysRevLett.1193.218701.**

Rosenberg N, Li L, Ward R, Pritchard J (2003) Informativeness of Genetic Markers for Inference of Ancestry. ***American Journal of Human Genetics*** **73, 1402-1422.**

Rozenfeld A, Arnaud-Haond S, Hernandez-Garcia E, ***et al.*** **(2008) Network analysis identifies weak and strong links in a metapopulation system.** ***PNAS*** **105, 18824-18829.**

Ruane J (1999) A critical review of the value of genetic distance studies in conservation of animal genetic resources. ***Journal of Animal Breeding and Genetics*** **116, 317-323.**

Sabeti P, Varilly P, Fry B, ***et al.*** **(2007) Genome-wide detection and characterization of positve selection in human populations.** ***Nature*** **449, 913-918.**

Saiki R, Gelfand D, Stoffel S, ***et al.*** **(1988) Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase.** ***Science*** **239, 487-491.**

Saitou N, Nei M (1987) The Neighbor-joining Method: A New Method for Reconstructin Phylogenetic Trees. ***Molecular Biology and Evolution*** **4, 406-425.**

Salmela E, Lappalainen T, Fransson I*, et al.* (2008) Genome-Wide Analysis of Single Nucleotide Polymorphisms Uncovers Population Structure in Northern Europe. *PLOS One* 3, e3519. doi:3510.1371/journal.pone.0003519.

Sham S, Zhao J (1998) Linkage analysis using affected sib pairs. *Guide to Human Genome Computing* 2nd ed.

Shannon J (2010) Understanding Genomics: An Introduction to the Terminology. *WCDS Advances in Dairy Technology* 22, 327-338.

Shannon P, Markiel A, Ozier O*, et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 13, 2498-2504.

Szpiech Z, Jakobsson M, Rosenberg N (2008) ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* 24, 2498-2504.

Tang H, Peng J, Wang P, Risch N (2005) Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology* 28, 289-301.

Tapio I, Värv S, Bennewitz J*, et al.* (2006) Prioritization for conservation of northern european cattle breeds based on analysis of microstellite data. *Conservation Biology* 20, 1768-1779.

Tetko I, Facius A, Ruepp A, Mewes H (2005) Super paramagnetic clustering of protein sequences. *BMC Bioinformatics* 6, 1471-2105.

Thaller G, Kramer W, Winter A*, et al.* (2003) Effects of DGAT1 variants on milk production traits in German cattle breeds. *J. Anim Sci.* 81, 1911-1918.

The Bovine HapMap Consortium ea (2009) Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science* 324, 528-532.

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. ***Nature*** **449, 851-861.**

The International Plant Genetic Resources Institute, Cornell University (2003) Measures of Genetic diversity. 1-71.

Troy C, MacHugh D, Bailey J*, et al.* (2001) Genetic evidence for Near-Eastern origins of European cattle. ***Nature*** **410, 1088-1091.**

Turakulov R, Easteal S (2003) Number of SNPs loci needed to detect population structure. ***Human Heredity*** **55, 37-45.**

VanRaden P (2008) Efficient Methods to Compute Genomic Predictions. ***Journal of Dairy Science*** **91, 4414-4423.**

Vega FDL, Kreitman M (2000) Human Genome Variation: Analysis, Management and Application of SNP Data. ***Pacific Symposium on Biocomputing*** **5, 633-635.**

Voight B, Kudaravalli S, Wen Y, Pritchard J (2006) A Map of Recent Positive Selection in the Human Genome. ***PLOS Biology*** **4, e72 doi:10.1371/journal.pbio.0040072.**

Wagner A, Fell D (2001) The small world insied large metabolic networks. ***Proceedings of the Royal Society London Series B*** **280, 1803-1810.**

Wan X, Yang C, Yang Q*, et al.* (2010) Detecting two-locus associations allowing for interactions in genome-wide association studies. ***Bioinformatics*** **26, 2517-2525.**

Weir B (1989) Sampling properties of gene diversity. ***Plant population genetics, breeding and genetic resources*****, 23-42.**

Weir B (1996) Genetic data analysis II. ***Sunderland(MA)*** **Sinauer Associates.**

Weir B, Cockerham C (1984) Estimating f-statistics for the analysis of population structure. ***Evolution*** **38, 1358-1370.**

Wiener P, Burton D, Williams J (2004) Breed relationships and definition in British cattle: a genetic analysis. ***Heredity*** **93, 597-602.**

Wiggintion J, Cutler D, Abecasis G (2005) A Note on exact tests of Hardy-Weinberg equilibrium. ***American Journal of Human Genetics*** **76, 887-893.**

Wright S (1931) Evolution in Mendelian populations. ***Genetics*** **16, 97-159.**

Wu F, Huberman B (2004) Finding communities in linear time: a physics approach. ***European Physical Journal B*** **38, 331-338.**

Zimin A, Delcher A, Florea L*, et al.* (2009) A whole-genome assembly of the domestic cow, Bos taurus. ***Genome Biology*** **10, R42 (doi:10.1186/gb-2009-1110-1184-r1142).**

ACKNOWLEDMENTS

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. I want to thank the Ludwig Maximilian University (LMU) for giving me the permission and providing the resources to under take this project. Furthermore I thank Tierzuchtforschung München e.V. for providing financial support and data to work with.

In particular I would like to thank Herman Raadsma and Mehar Khatkar (University of Sydney), without whose help this project would not have been achieved.

Thanks a lot to Martin Dinkel, Renate Damian, Tatiana Schröter and Gisela Andorfer for good cooperation and technical work.

I would also like to thank Doctor Jennifer Schär, Janey Heine, Sophie Rothammer, Monika Höpfl and Ashraf Awad for all help in the work but also for being wonderful colleges and supportive friends.

Special thanks also to Doctor Jan Maxa for his continuing support, advice and help throughout the time in Munich.

Special mention must go to Doctor Ivica Medugorac for his supervision throughout the whole project. His patience, time, guidance and effort were unsurpassed and are greatly appreciated.

Finally, thanks must go to my family and friends for their encouragement and continuing support.