# On detection of population stratification in genotype samples using spacial clustering and non-linear optimization

**Vinzent Boerner**

Animal Genetics and Breeding Unit[1], University of New England
Armidale, 2351, NSW, Australia

## Summary

Accounting for population stratification in genotype samples is important to avoid false inference from genome wide association studies. It is usually quantified using model-based ancestry estimation (e.g. ADMIXTURE; Alexander *et al.* (2009)), which has disadvantages with regard to model assumptions and processing time. This article describes a two step procedure for estimating population stratification. In the first step a spacial cluster algorithm is used to detect clusters of genetically homogeneous animals. In a subsequent step genotypes are described as linear functions of within-cluster allele frequencies. The approach was tested on a cattle data set which consisted of 11,639 real genotypes from 11 breeds and 5,000 artificially generated cross-bred genotypes (F1 to F5). It outperformed results obtained from ADMIXTURE in terms of speed and accuracy.

## Introduction

Results from genome wide association studies (GWAS) can be negatively affected by population stratification (Marchini *et al.*, 2004; Price *et al.*, 2010). It is therefore necessary to quantify the latter and expand the fitted model by a related factor. Two major approaches are used for this purpose. The first approach estimates genome proportions of sampled genotypes conditional on a predefined number of ancestral populations, and is embedded in software like STRUCTURE (Pritchard *et al.*, 2000) and ADMIXTURE (Alexander *et al.*, 2009). This approach yields biologically meaningful results at the individual and population level, but results can only be incorporated into GWAS in a subsequent analysis. The second approach performs a singular value decomposition of the matrix of genetic markers and performs GWAS within the Eigen-space. This approach is used in the software EIGENSTRAT (Price *et al.*, 2006). However, it is not obvious how to interpret principal components in terms of ancestral population allele frequencies and individual genome proportions.

This article describes a two step procedure for estimating population stratification which exploits properties of the singular value decomposition, provides biologically interpretable results, and is fast even when the number of markers per genotype is huge. Results were compared to those obtained from ADMIXTURE.

[1]A joint venture of the NSW Department of Primary Industry and the University of New England

## Methods

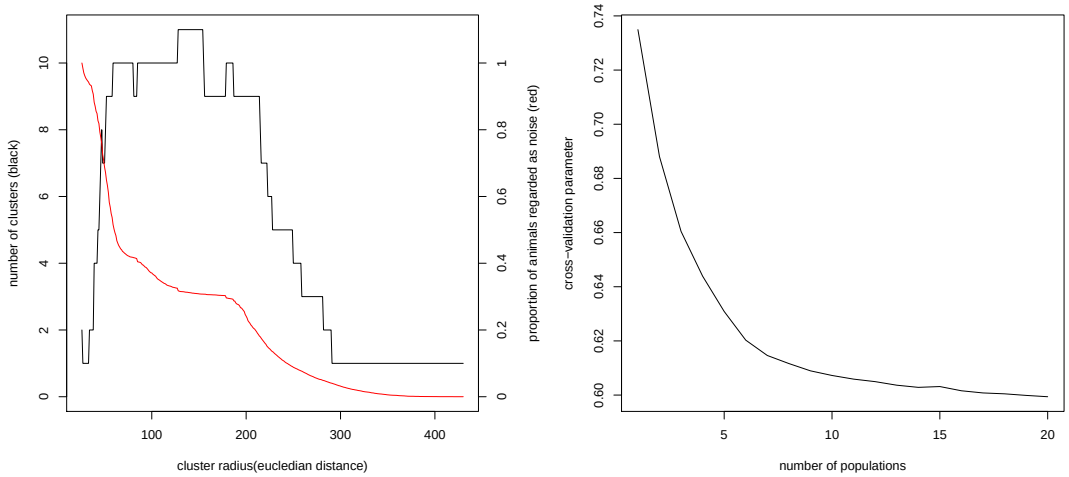
In the first step a spacial cluster algorithm was used to decide whether individual genotypes should belong to a cluster or should be regarded as "noise". In the second step a non-linear optimisation approach (BREEDCOMP; Boerner (2017)) was used to describe an animal's genotype as a linear function of within-cluster allele frequencies.

The cluster analysis step used the OPTICS algorithm for spacial clustering (Ankerst *et al.*, 1999) and was carried out on a Euclidean distance matrix $K$ of dimension $N_a \times N_a$ where $N_a$ is the total number of genotyped animals. $K$ was constructed from a matrix $P$ of dimension $N_a \times 50$. $P$ was obtained from $DR[1 : N_a, 1 : 50]$. $D$ and $R$ were matrices obtained from the singular valued decomposition $Z = LDR$, where $Z$ is an matrix of marker genotypes of dimension $N_a \times N_m$ where $N_m$ is the number of markers. OPTICS requires no further input parameter and yields a list of ordered data points. This list is subsequently cut into cluster and "noise" where the cutting process requires a single external input parameter (MPT). MPT is the minimum number of points to regard a point aggregation as a cluster and is evaluated along the radius around data points. As a result of an increasing radius the number of clusters increases and the "noise" decreases. However, beyond a certain radius the number of clusters will shrink again, thus forming a distinctive curve with a clear maximum. In the second step genome proportions were estimated using the approach of Boerner (2017). The problem to solve can be written as $Z' = FQ' + E$. $F$ is an $N_m \times N_k$ matrix of within-cluster allele frequencies where $N_k$ is the number of clusters, $Q$ is a $N_a \times N_k$ matrix of the animals' genome proportions, and $E$ is an $N_m \times N_a$ matrix of non-explainable residuals. Assuming that animals are independent the above equation may be solved by minimising $trace(E'E)$. Thus, for animal $i$ it becomes: $\arg\min_{Q_{i,:}} f(Q_{i,:}) = E'_{:,i}E_{:,i} - 2Z_{i,:}FQ'_{i,:} + Q_{i,:}F'FQ'_{i,:}$. To obtain meaningful results the parameter space of values in $Q_{i,:}$ must be constrained to $Q_{i,:} \geq 0$ $\{i = 1, .., N_k\}$ and $\sum_j^{N_k} Q_{i,j} = 1$.

The above combination of algorithms was tested on a cattle data set which consisted of 11,639 individuals from 11 different breeds (Brahman (1,492), Angus (1,473), Murray Grey (316), Limousin (1,395), Charolais (899), Hereford (1,500), Simmental (337), Shorthorn (1,126), Wagyu (1,497), Santa Gertrudis (1,474) and Droughtmaster (130)). Because genotypes of these animals were from various SNP panels, the analyses were based on a set of 4,022 SNP in common across all panels. Five generations of cross-bred animals were generated from these pure-bred animals (F1 to F5). To generate the F1 individuals, 2,000 pure-bred parents were randomly chosen to form 1000 sire-dam pairs where their haplotypes were obtained by random phasing, and gametes were formed using 25 randomly located cross-overs. Haplotypes and genotypes of offspring were generated by gamete union. To generate the F2 to F5 individuals, the 2,000 parent genotypes were selected from the previous 1,000 offspring implying more than one offspring per parent. Thus, the total number of artificially admixed offspring was 5,000 and the total data set size was 16,639. All computations were carried out on an desktop computer with an Intel(R) Core(TM) i7-3770 processor and 32GB of memory.

## Results and Discussion

The first step for both ADMIXTURE as well as the OPTICS - BREEDCOMP cascade was the correct determination of the number of founder populations. When OPTICS was




(a) Number of clusters (black) and the propotion of animals regarded as crosses (red) when using the OPTICS cluster algorithm.

(b) ADMIXTURE cross-validation parameter as a measure to find the optimum number of populations

Figure 1: Population detection characteristics of OPTICS and ADMIXTURE

used the maximum number of populations could be easily inferred because the number of clusters as a function of the cluster radius formed a distinct curve (see Figure 1). Contrarily, no inference could be made from the ADMIXTURE cross-validation approach because the relevant parameter was still decreasing even with the number of populations set to 20. With the MPT parameter set to 100, OPTICS generated 405 cluster solutions of which 27 had 11 clusters. However, the composition of these 27 solutions differed only slightly with correlations between within-cluster allele frequencies >0.999. Since ADMIXTURE could not detect an optimum number of populations, subsequently the relevant parameter was regarded as prior knowledge and set to 11. But even then the true population allele frequencies correlated less to those estimated by ADMIXTURE than to the within-cluster allele frequencies generated from OPTICS results with 11 clusters (see Table 1). Individual genome composition results from ADMIXTURE were compared to BREEDCOMP results where the latter used an arbitrary OPTICS solution with 11 clusters to calculate within-cluster allele frequencies. Genome proportions estimated by BREEDCOMP correlated to the true genome proportions with 0.99, whereas genome proportions estimated by ADMIXTURE reached a correlation of 0.84 only. This is also reflected in the maximum difference between the estimated and true genome proportions which was 0.4 and 0.99 for BREEDCOMP and ADMIXTURE, respectively.

The OPTICS - BREEDCOMP cascade also outperformed the ADMIXTURE unsupervised mode in terms of speed. An unsupervised ADMIXTURE run with the number of populations equal to 11 needed 45 real time minutes. In comparison, OPTICS provided all 405 solutions given in Figure 1a in 22 real time seconds. A BREEDCOMP run for a single OPTICS solution needed about 30 real time seconds to estimate the genome composition of all 16,639 animals. Thus, it is possible to run OPTICS and evaluate approximately 90 cluster solutions in the time necessary for a single ADMIXTURE run. In addition, it is worthwhile noting that this difference in speed is accentuated as the number of markers per genotype and the number of genotyped animals increases. For an unbiased compari-

Table 1: Correlations between the allele frequencies of the 11 populations estimated by OPTICS and ADMIXTURE and the allele frequencies of the true populations.

| | true population | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| OPTICS | 1 | 1 | 1 | 1 | 1 | 1 | 0.999 | 1 | 1 | 1 | 1 |
| ADMIXTURE | 0.999 | 0.982 | 0.994 | 0.999 | 0.991 | 0.964 | 0.83 | 0.99 | 0.954 | 0.995 | 0.984 |
| OPTICS | 0.314 | 0.639 | 0.394 | 0.844 | 0.675 | 0.639 | 0.839 | 0.754 | 0.556 | 0.675 | 0.599 |
| ADMIXTURE | 0.25 | 0.529 | 0.318 | 0.533 | 0.628 | 0.539 | 0.7 | 0.572 | 0.869 | 0.596 | 0.52 |
| true | 0.315 | 0.639 | 0.394 | 0.839 | 0.675 | 0.639 | 0.839 | 0.75 | 0.556 | 0.675 | 0.599 |

Upper part: maximum of the correlations between the allele frequencies of the true population and the allele frequencies of all the suggested populations. Middle part: Second largest correlation between the allele frequencies of the true population and the allele frequencies of all the suggested populations. Lower part: maximum off-diagonal values from a correlation matrix between true population allele frequencies.

son one would also have to account for the run time of the ADMIXTURE cross-validation procedure, which was 18 hours for this data set.

Beside speed and accuracy differences, the key difference between ADMIXTURE and the OPTICS - BREEDCOMP cascade is that OPTICS tries to detect point aggregation in $R^n$ whereas ADMIXTURE moves vectors of allele frequencies through $R^n$ until all data points are best explained. ADMIXTURE appears to be useful even for samples containing cross-bred animals only. However, a stabilised cross will occur as point aggregation as well and will attract ADMIXTURE to place a population allele frequency vector within or at least very close to it leading to false inference. By contrast OPTICS can only be applied if it can be assumed that at least a part of the genotype sample is from pure-breed animals.

## REFERENCES


Alexander, D. H., J. Novembre & K. Lange, 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19(9): 1655–1664.

Ankerst, M., M. M. Breunig, H.-P. Kriegel & J. Sander, 1999. OPTICS: ordering points to identify the clustering structure. In: Sigmod Rec, volume 28, pp. 49–60. ACM.

Boerner, V., 2017. On breed composition estimation of cross-bred animals using non-linear optimisation. In: Proc. Assoc. Advmt. Anim. Breed. Genet. 22, Townsville, Australia, 2017.

Marchini, J., L. R. Cardon, M. S. Phillips & P. Donnelly, 2004. The effects of human population structure on large genetic association studies. Nat. Genet. 36(5): 512–517.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick & D. Reich, 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38(8): 904–909.

Price, A. L., N. A. Zaitlen, D. Reich & N. Patterson, 2010. New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11(7): 459–463.

Pritchard, J. K., M. Stephens & P. Donnelly, 2000. Inference of population structure using multilocus genotype data. Genetics 155(2): 945–959.