American Genetic Association

Journal of Heredity, 2015, 672–676
doi:10.1093/jhered/esv044
Computer Note
Advance Access publication July 2, 2015

OXFORD

Computer Note

TRES: Identification of Discriminatory and Informative SNPs from Population Genomic Data

Ioannis Kavakiotis, Alexandros Triantafyllidis, Despoina Ntelidou, Panoraia Alexandri, Hendrik-Jan Megens, Richard P. M. A. Crooijmans, Martien A. M. Groenen, Grigorios Tsoumakas and Ioannis Vlahavas

From the Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece (Kavakiotis, Ntelidou, Tsoumakas, and Vlahavas); Department of Genetics, Development & Molecular Biology, School of Biology, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece (Triantafyllidis and Alexandri); and Wageningen University, Animal Breeding and Genomics Centre, Wageningen, The Netherlands (Megens, Crooijmans, and Groenen).

Address correspondence to Ioannis Kavakiotis at the address above, or e-mail: ikavak@csd.auth.gr

Received March 10, 2015; First decision April 12, 2015; Accepted June 17, 2015.

Corresponding editor: Howard Ross

Abstract

The advent of high-throughput genomic technologies is enabling analyses on thousands or even millions of single-nucleotide polymorphisms (SNPs). At the same time, the selection of a minimum number of SNPs with the maximum information content is becoming increasingly problematic. Available locus ranking programs have been accused of providing upwardly biased results (concerning the predicted accuracy of the chosen set of markers for population assignment), cannot handle high-dimensional datasets, and some of them are computationally intensive. The toolbox for ranking and evaluation of SNPs (TRES) is a collection of algorithms built in a user-friendly and computationally efficient software that can manipulate and analyze datasets even in the order of millions of genotypes in a matter of seconds. It offers a variety of established methods for evaluating and ranking SNPs on user defined groups of populations and produces a set of predefined number of top ranked loci. Moreover, dataset manipulation algorithms enable users to convert datasets in different file formats, split the initial datasets into train and test sets, and finally create datasets containing only selected SNPs occurring from the SNP selection analysis for later on evaluation in dedicated software such as GENECLASS. This application can aid biologists to select loci with maximum power for optimization of cost-effective panels with applications related to e.g. species identification, wildlife management, and forensic problems. TRES is available for all operating systems at http://mlkd.csd.auth.gr/bio/tres.

Subject areas: Bioinformatics and computational genetics
Key words: ancestry informative marker, feature selection, marker panel, population genomics, single-nucleotide polymorphism

Single-nucleotide polymorphisms (SNPs) are becoming the marker of choice for a wide range of organisms and applications, such as genetic structure analyses, dispersal studies, wildlife management, forensic problems, and eco-certification (Helyar *et al.* 2011; Ogden 2011; Wilkinson *et al.* 2011). The advent of next-generation sequencing and related platforms has enabled those studies greatly, through the development and genotyping of thousands or even millions of SNPs.

A common aim in many of those studies is to infer individual ancestry or, simply put, to assign individuals to their group of origin (e.g. Paetkau *et al.* 1995; Nielsen *et al.* 2012). For such applications, it is advantageous to select those SNPs, among all genotyped loci, that can best discriminate analyzed individuals. The importance of feature selection for SNP datasets is beyond dispute either for biology or informatics. From a computational view, feature selection, that is, the selection of the most important attributes from a dataset, is an important process with many beneficial effects such as lower computational demands and response times (Guyon and Elisseeff 2003). From a biological point of view, the importance of feature selection, that is, selecting SNPs with maximum information power, has been extensively discussed in the scientific literature (e.g.Wilkinson et al. 2011, 2012; Nielsen *et al.* 2012): although genome-wide data are valuable, they are costly to produce. Smaller panels (based on a subset of SNPs) facilitate the genotyping of several hundreds of individuals with a lower cost and at greater speed than genome-wide genotyping.

However, the identification of these (minimum) SNPs can be problematic. When dealing with large numbers of SNP markers, automated methods for selecting loci with the most power across a range of application scenarios are required (Helyar *et al.* 2011). There are 2 fundamentally different approaches. One way is to use available software to estimate the assignment power of individual loci (WHICHLOCI, Banks *et al.* 2003) or combinations of those (i.e. GAFS, Topchy *et al.* 2004 and BELS, Bromaghin 2008). These software have their own advantages and disadvantages (reviewed in Helyar *et al.* 2011), but their most important caveat is that they have computational limitations. As Helyar *et al.* (2011) mentioned, although, presumably, there are no constraints on the number of loci or individuals that can be analyzed, the analysis may be prohibitive on a desktop computer. More specifically, applications that search for the best combination of loci (GAFS and BELS) can be extremely computationally intensive for high-dimensional datasets, such as SNP datasets. For backward elimination approaches (such as in BELS), evaluation time can sometimes be proportional to m^2 (where m is the number of features, i.e. SNPs). For more sophisticated, exhaustive, searches (such as in GAFS) up to 2^m possible subsets have to be examined (Witten *et al.*, 2011).WHICHLOCI is probably the only available fast software. Finally, as reported extensively in Anderson (2010), those applications have been implemented in a way that leads to a systematic upward bias in the predicted accuracy of the chosen set of markers for population assignment, since the same set of individuals is used to initially train and to subsequently estimate the accuracy of their model.

A second completely different approach is to rank loci solely according to their informativeness, that is, the marker information content, which is the amount of information that a locus holds regarding the ancestry of an individual. The use of markers with high informativeness reduces the number of markers needed for correct assignment (Rosenberg et al 2003). Several measures/criteria of marker informativeness have been proposed (Rosenberg *et al.* 2003; Ding *et al.* 2011 and references therein), such as Delta (Shriver *et al.* 1997), pairwise Wright's F_{ST} given by Wright (1951), global Wright's F_{ST} by Wright (1951), pairwise Weir and Cockerham F_{ST} by Weir and Cockerham (1984), global pairwise Weir and Cockerham F_{ST} by Weir and Cockerham (1984), and informativeness for assignment (I_n) (Rosenberg *et al.* 2003). Principal component analysis has also been used for the same purpose (Paschou *et al.* 2007). Wilkinson *et al.* (2011) and Ding *et al.* (2011) have performed comparisons between some of these approaches, concluding, respectively, that pairwise Wright's F_{ST} and I_n were 2 of the best evaluators/criteria, whereas global statistics did not return satisfactory results. The available literature, however, is sometimes contradictory in which method is the best (see Paschou *et al.* 2007 vs. Wilkinson *et al.* 2011), often stating that differences in assignment power are marginal, while agreeing that no single method outperforms the rest in all circumstances, but that the power of assignment success, and the required number of SNPs, is dependent on the studied species, the levels of genetic heterogeneity and the pool of samples considered, as well as the desired stringency of the assignment (Ding *et al.* 2011; Wilkinson *et al.* 2011).

It must be stressed that all the above methods and software from both approaches produce a limited set of markers appropriate for assignment purposes. They should not be used in downstream estimation of general population genetics parameters (e.g. in larger sample sizes) since this is a biased subset of the total number of markers.

The application of different methods for marker prioritization and decision making in the construction of SNP panels is becoming more important as large high-throughput assays become readily available. Though new computer programs exist for elementary analysis of large SNP datasets (e.g. Xu *et al.* 2010), a gap of tools is apparent regarding the evaluation of large numbers of SNP loci for individual assignment. Recent scientific efforts to use SNP data for assignment purposes (Paschou *et al.* 2007; Wilkinson *et al.* 2011; Nielsen *et al.* 2012) have relied on bioinformatics methods of each separate laboratory, and no software solutions are currently available to meet the needs of genetisists (working with nonmodel species in most cases).

In order to address these setbacks, we developed the *Toolbox for Ranking and Evaluation of SNPs (TRES)*, a collection of algorithms built in a user-friendly and computationally efficient software that can manipulate and analyze datasets even in the order of millions of genotypes in a matter of seconds.

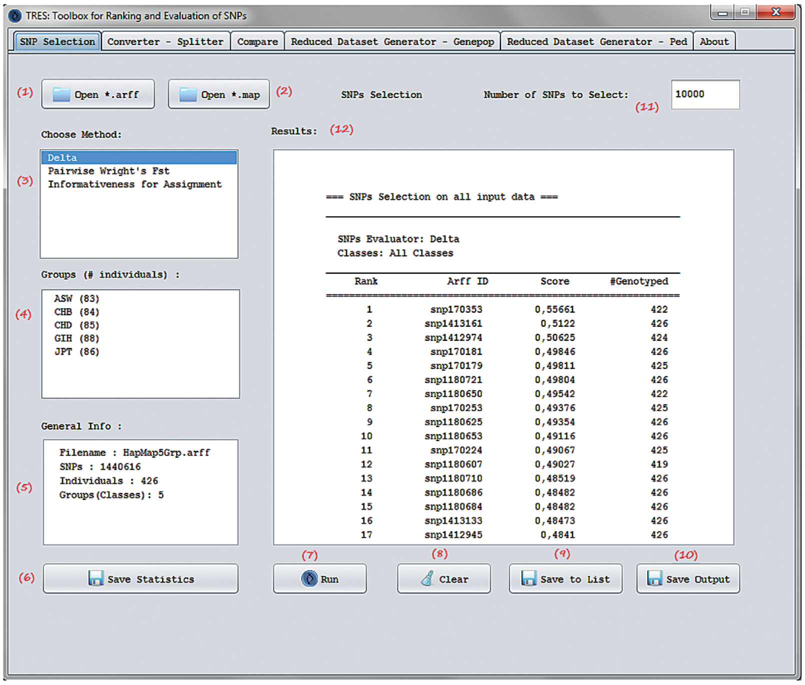
Application Features and Functionalities

TRES is a user-friendly application for selecting the most informative SNP markers for assignment purposes from a SNP dataset. Moreover it offers a collection of algorithms for data manipulation.

SNP Marker Evaluation

The first and main tab of the application is the "SNP Selection" tab where the SNP evaluation is performed. Three methods (presented in detail below) are provided for ranking and later on evaluating the discriminating power of each SNP (Figure 1, spot 3). The number of top-ranked SNPs to be returned is user defined, according to the needs of individual genotyping assays that a researcher needs to prepare (Figure 1, spot 11). Users can also choose specific subgroups (populations) to be used for the evaluation (Figure 1, spot 4), that is, users can evaluate the SNPs that better distinguish specific subsets of samples or all samples together.

Another functionality of TRES, offered in the Compare tab, compares the lists of top-ranked predefined number of loci produced



Number of SNPs to Select: 10000

Choose Method:
- Delta
- Pairwise Wright's Fst
- Informativeness for Assignment

Groups (# individuals) :

ASW (83)
CHB (84)
CHD (85)
GIH (88)
JPT (86)

General Info :

Filename : HapMap5Grp.arff
SNPs : 1440616
Individuals : 426
Groups(Classes): 5

```
=== SNPs Selection on all input data ===
______________________________________________________________

 SNPs Evaluator: Delta
 Classes: All Classes
______________________________________________________________
     Rank        Arff ID        Score      #Genotyped
==============================================================
        1      snp170353      0,55661             422
        2     snp1413161       0,5122             426
        3     snp1412974      0,50625             424
        4      snp170181      0,49846             426
        5      snp170179      0,49811             425
        6     snp1180721      0,49804             426
        7     snp1180650      0,49542             422
        8      snp170253      0,49376             425
        9     snp1180625      0,49354             426
       10     snp1180653      0,49116             426
       11      snp170224      0,49067             425
       12     snp1180607      0,49027             419
       13     snp1180710      0,48519             426
       14     snp1180686      0,48482             426
       15     snp1180684      0,48482             426
       16     snp1413133      0,48473             426
       17     snp1412945       0,4841             426
```

Figure 1. TRES application window. Numbers on gray background indicate features described in the text, except for numbers 5 (area for presenting general information of the dataset), 7 (begins execution), and 8 (clears results' area).

with any combination of the 3 different methods offered by TRES (i.e. between Delta, pairwise Wright's F_{ST}, and informativeness for assignment, see below) and returns the common loci; Ding *et al.* (2011) have suggested that in some cases union of the most informative SNPs among different methods could also be considered as the best panel, though researchers need to be also aware of the differences between the various methods for evaluating ancestry informativeness of SNP markers.

Input Files

TRES receives as input PED (Pedigree) and ARFF (Attribute–Relation File Format) files (Figure 1, spot 1). PED files (Purcell *et al.* 2007) are Pedigree files used by various software such as PLINK, for reconstructing a full pedigree based on genotypes. Currently, except from PED files, there are many other file formats that contain SNP datasets such as HapMap, VCF, GenePop, and others. Fortunately, there are many reliable converters, for instance, PGDSpider (Lischer & Excoffier, 2012) that can convert any well-known file format to PED file. An ARFF file (Hall *et al.* 2009) is an ASCII text file that describes a list of instances (i.e. individuals) sharing a set of features (i.e. SNPs). It is a popular file format in the field of data mining and machine learning, as it is used by the Weka machine learning library (Hall *et al.* 2009). ARFF files are used only in the SNP evaluation process. The use of ARFF was essential in order to benefit from implemented classes from WEKA that made the whole analysis process more efficient which is a most desired feature when dealing with very high-dimensional datasets. Finally, TRES can also accept complementarily a MAP file (Figure 1, spot 2) that contains information regarding the genomic position of the SNPs.

Results—Output

Results are presented in 2 ways, depending on the existence of the MAP file. All results and information are presented within an application window (Figure 1, spot 12), but TRES can also export results to text files (Figure 1, spot 10). TRES can also save the list of selected SNPs (Figure 1, spot 9), without additional information in order to be used at the construction of datasets with selected SNPs. Descriptive statistics are also given (Figure 1, spot 6) including allele and genotype frequencies for each group and for the total population.

Dataset Manipulation Algorithms

The first dataset manipulation algorithm is a converter. Although the initial genetic datasets are offered in PED files, as mentioned before, ARFF files are used in the analysis process. In order to increase the usability, the application offers an algorithm that converts PED to ARFF files.

The second available algorithm splits a dataset in 2 parts. Anderson (2010) and Witten *et al.* (2011) stressed the fact that "statistical classification procedures should be assessed using data that are separate from those used to train the classifier" and proposed "assessing the power of the resulting locus panels using a separate holdout set of individuals that was not used in any way to choose the set of loci for the marker panel." Following comments by Anderson (2010), our program avoids biased estimates of accuracy by providing an algorithm for splitting the initial PED file into 2 new ones based on a user given (x) value. The first file (training dataset) contains x percentage of the initial data samples and is used to train the model (i.e. to run one of the available ranking algorithms and obtain the ranked list of SNPs). The second file (test dataset) contains the remaining (1 – x) percentage and can be used to evaluate the model (evaluate the SNP ranking through assignment of these individuals in GENECLASS). It must be stressed that the split is applied separately and equally within each subpopulation to assure that no subpopulation will be under-represented in either of the newly produced PED files.

Finally, users can also construct GENEPOP (Rousset 2008) files as input datasets for GENECLASS 2.0 for a desired number of top-ranked SNPs which occurred from the analysis step. GENECLASS2.0 (Piry *et al.* 2004) is a software that offers an extensive list of genetic assignment methods. Such methods assign an individual to the population in which the individual's genotype is most likely to occur. Through the "Reduced Dataset Generator - GENEPOP" tab, the application can produce a file as reference population for GENECLASS2.0 (based on the training dataset) and a file with individuals to be assigned with GENECLASS2.0 (based on the test dataset) containing only a user defined number of top-ranked SNPs which is the SNP list produced from the SNP evaluation analysis. In this way users can easily evaluate the assignment accuracy of subsets of SNPs offered by GENECLASS2.0 and eventually compare their datasets and different markers of informativeness. The same functionality is offered for reduced ped and map file construction through the "Reduced Dataset Generator – PED."

A more detailed description of the application can be found at TRES's user guide.

Implementation

The application is implemented in Java and therefore can be executed in all operating systems. Java Swing library was used in order to build the graphical user interface (GUI). Parts of programming code from WEKA machine learning library were used (Hall *et al.* 2009), mainly for data handling. TRES has no intrinsic limits on the number of SNPs and individuals that it can analyze. This is possible, due to the dynamic nature of the data structures that have been used in its implementation.

Marker Evaluation Methods

TRES offers the following metrics to evaluate SNPs: (1) Delta, probably the most commonly used measure of marker informativeness for human populations (Shriver *et al.* 1997). (2) Pairwise Wright's F_{ST}, and (3) informativeness for assignment measure (Rosenberg *et al.* 2003) which, as stated, have both proved to be highly informative. In all cases, loci are scored based on the above criteria and later on ranked from highest to lowest value.

Delta

For a biallelic marker, the delta value is given by the following equation:

$$\delta = \left| p_A^i - p_A^j \right|$$

where $p_A{}^i$ is the frequency of the allele A in the i^{th} population and $p_A{}^j$ is the frequency of the same allele in the j^{th} population. It is important to mention that delta is calculated only between 2 populations, so if there are more than 2 populations, the delta value is computed for each one of every possible combination between existing populations and their average is subsequently calculated in order to produce a value for each SNP marker.

Pairwise Wright's F_{ST}

Pairwise Wright's F_{ST} (Wright 1951) for more than 2 populations is computed with the same approach as outlined for delta. For a biallelic marker, the F_{ST} value is given by the following equation:

$$F_{ST} = \frac{H_t - H_s}{H_t}$$

where H_s is the average expected heterogygosity across subpopulations and H_t is the expected heterozygosity of the total population (Beebee & Rowe 2004), and they are given by the following equations:

$$H_t = 2 * p_A * p_B \quad H_s = p_A^i * p_A^j + p_B^i * p_B^j$$

where $p_A{}^i$ is the frequency of the allele A in the i^{th} population, $p_A{}^j$ is the frequency of the same allele A in the j^{th} population and p_A is the frequency of allele A in all populations. Notations for allele *B* are defined similarly.

Informativeness for Assignment (I_n)

I_n is a mutual information-based statistics that takes into account self-reported ancestry information from sampled individuals (Rosenberg *et al.* 2003, Ding *et al.* 2011).

$$In = \sum_{j=0}^{N} (-\mathrm{p_j} \log_2 \mathrm{p_j} + \sum_{i=0}^{K} (\mathrm{p_{ij}} \log_2 \mathrm{p_{ij}}) / \mathrm{K})$$

Where $i = 1,2,\ldots,K$ are the populations with $K \geq 2$ and N are the loci, p_{ij} denotes the frequency of allele j in population *I* and p_j denotes the average frequency of allele j over *K* populations.

In all cases, allele frequencies are calculated as in Ding et al (2011).

Computational Performance

In order to measure the computational performance, we used 2 different SNP datasets to calculate the response time of the application. The first analysis was conducted on the dataset of Wilkinson *et al.* (2012). This dataset comprised of a comprehensive coverage of pig breed types present in Britain; it consisted of 446 pigs from 7 traditional British breeds, 5 commercial purebreds, 1 imported European breed and 1 imported Asian breed that were genotyped with the PorcineSNP60 BeadChip (59436 SNPs, Ramos *et al.* 2009). A standard laptop computer with an Intel Core™ Duo CPU T9600 at 2.80GHz processor with 4 GB main memory was adequate for the application to convert the PED file into ARFF in less than 60 s and to select the top-100 SNPs using the Delta metric for the comparison of all populations or just 2 populations in 28 and 15 s, respectively. It should be stated that data could not be loaded to BELS (inadequate memory allocation) and WHICHLOCI or GAFS (both programmes crashed).

The second analysis was performed on data originating from the HapMap project (International HapMap Consortium 2003). Our dataset consisted of 426 individuals that belonged to 5 different population groups genotyped at 1440616 SNPs (Figure 1), that is, more than half a billion genotypes. Since the size of the dataset was very big (>5 GB), we conducted the analysis in a more powerful laptop with an Intel Core™ i7 4500u processor at 3.4GHz and 16 GB RAM. We ran the application allocating 12 GB RAM. The dataset was fully loaded into the application in 155 s. SNP selection analysis based on the Delta metric was completed in 230 s. Consequently, our application has no intrinsic limits on the number of SNPs and individuals that can be analyzed and restrictions can only arise from user's hardware.

Availability

TRES is a Java application and therefore can be executed in all operating systems (tested on Windows, Linux, and MacOs). The user guide contains information about: the installation process; the requirements of the application; the exact structure of the input data files; the software's features and functionalities; and finally a step-by-step complete analysis scenario using TRES. The executable, user manual and example datasets are freely available at http://mlkd.csd.auth.gr/bio/tres/.

Acknowledgments

The authors would like to thank Dr Efstratios Kontopoulos for his valuable comments and suggestions and Dr Rob Ogden and Dr Pamela Wiener for providing SNP datasets for the evaluation of TRES. The authors would also like to thank 12 postgraduate students from the School of Biology of the Aristotle University of Thessaloniki, who took part in their task-oriented user evaluation and helped them improve the GUI based on their comments, suggestions, and errors throughout the evaluation.

References

Anderson EC. 2010. Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources*. 10:701–710.

Banks MA, Eichert W, Olsen JB. 2003. Which genetic loci have greater population assignment power? *Bioinformatics*. 19:1436–1438.

Beebee T, Rowe G. 2004. *An introduction to molecular ecology*. Oxford, UK: Oxford University Press.

Bromaghin JF. 2008. BELS: backward elimination locus selection for studies of mixture composition or individual assignment. *Molecular Ecology Resources*. 8:568–571.

Ding L, Wiener H, Abebe T, Altaye M, Go CP, Kercsmar C, Grabowski G, Martin LJ, Hershey GKK, Chakorborty R and Baye TM. 2011.Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics*. 12:622.

Guyon I, Elisseeff A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*. 3:1157–1182.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The WEKA data mining software: an update. *SIGKDD Explorations*. 11:10–18.

Helyar SJ, Hemmer-Hansen J, Bekkevold D *et al*. 2011. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*. 11:123–136.

International HapMap Consortium. 2003. The International HapMap Project. *Nature*. 426:789–796.

Lischer HE, Excoffier L. 2012. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*. 28:298–299.

Nielsen EE, Cariani A, Mac Aoidh E, Maes GE, Milano I, Ogden R, Taylor M, Hemmer-Hansen J, Babbucci M, Bargelloni L, *et al*.; FishPopTrace consortium. 2012. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nature Communications*. 3:851.

Ogden R. 2011. Unlocking the potential of genomic technologies for wildlife forensics. *Molecular Ecology Resources*. 11(Suppl 1):109–116.

Paetkau D, Calvert W, Stirling I, Strobeck C. 1995. Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*. 4:347–354.

Paschou P, Ziv E, Burchard EG *et al*. 2007. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics*. 3:1672–1686.

Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A. 2004. GENECLASS2: a software for genetic assignment and first-generation migrant detection. *Journal of Heredity*. 95:536–539.

Purcell, S. *et al*. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*. 81:559–575.

Ramos AM, Crooijmans R, Affara NA *et al*. 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS One*. 4:8.

Rosenberg NA, Li LM, Ward R, Pritchard JK. 2003. Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics*. 73:1402–1422.

Rousset F. 2008. GENEPOP'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*. 8:103–106.

Shriver MD, Smith MW, Jin L *et al*. 1997. Ethnic affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics*. 60:957–964.

Topchy A, Scribner K, Punch W. 2004. Accuracy-driven loci selection and assignment of individuals. *Molecular Ecology Notes*. 4:798–800.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution*. 38:1358–1370.

Wilkinson S, Wiener P, Archibald AL, Law A, Schnabel RD, McKay SD, Taylor JF, Ogden R. 2011. Evaluation of approaches for identifying population informative markers from high density SNP chips. *BMC Genetics*. 12:45.

Wilkinson S, Archibald AL, Haley CS, Megens HJ, Crooijmans RP, Groenen MA, Wiener P, Ogden R. 2012. Development of a genetic tool for product regulation in the diverse British pig breed market. *BMC Genomics*. 13:580.

Witten IH, Frank E, Hall MA. 2011. *Data mining: practical machine learning tools and techniques. 3rd ed*. Burlington, MA: Morgan Kaufmann.

Wright S. 1951. The genetical structure of populations. *Annals of Eugenics*. 15:323–354.

Xu S, Gupta S, Jin L. 2010. PEAS V1.0: a package for elementary analysis of SNP data. *Molecular Ecology Resources*. 10:1085–1088.