Journal of Animal Science, 2025, **103**, skaf083
https://doi.org/10.1093/jas/skaf083
Advance access publication 20 March 2025
Animal Genetics and Genomics

Technical note: anomaly detection for breed purity analysis in pigs using a single breed genotype panel

Xiaohan Jiang,[†] Austin Putz,[†,‡] Wen Huang,[‖] and Juan P Steibel[†,1]

[†]Bioinformatic and Computational Biology Program, Department of Animal Science, Iowa State University, Ames, IA 50011, USA
[‡]Hendrix Genetics, Boxmeer, 5831CK, The Netherlands
[‖]Department of Animal Science, Michigan State University, East Lansing, MI 48824, USA
[1]Correspondence author: jsteibel@iastate.edu

Abstract

Recent advancements in genotyping technologies have revolutionized our ability to estimate breed composition in pigs. The classical compositional regression used for this purpose requires a multibreed panel to detect crossbreeding and its application is limited to breeds in the panel. Some breed entities may not have access to the multibreed panel but may have access to a single breed panel. We present crossbreeding detection methods based on semi-supervised anomaly detection techniques that use a single-breed genotype panel from purebred animals of interest. By utilizing these methods, we identified and assessed breed outliers within large pig genotype panels.

Lay Summary

The genome of an individual contains information about its breed of origin. Classical methods require genomic knowledge of animals from several breeds to estimate the breed composition of an individual. In this paper, we propose methods that can assess the breed purity of an animal when only genotypes of a single breed are known. These methods belong to anomaly detection techniques, which are used to detect abnormal (crossed or admixed individuals) from normal (purebred) ones. We demonstrate that these methods are effective in distinguishing crossbreds from purebreds.

Key words: anomaly detection, pigs, breed

Abbreviations: LW, large white; LR, landrace; DUR, duroc; P, pietrain; MSE, mean square error; LM, linear model; PCA, principal component analysis

Introduction

Genomic data has been broadly applied to estimate the breed composition of domestic animals (Kuehn et al., 2011; Huang et al., 2014). Classical compositional regression models require a multibreed panel to detect crossbreeding (Kuehn et al., 2011; Huang et al., 2014) and they have been proven very effective in detecting crossbreeding of a target breed as well as of other breeds in the reference panel (Funkhouser et al., 2017). In some practical applications, however, entities performing breed purity analysis (e.g., an individual breeders' association) may not have access to a multibreed panel but may have access to a comprehensive reference genotype panel of the breed of interest. Determining the exact breed composition in this situation is impossible, but there is a need to certify whether an unknown animal is purebred or crossbred. We call this breed purity analysis. Thus, statistical methods that can employ a single-breed reference panel to detect crossbreeding should be developed.

This problem is especially well-suited for semi-supervised anomaly detection methods (Chandola et al., 2009). Anomaly detection is a technique for finding unusual datapoints or patterns in a dataset. The term anomaly is also referred to as an outlier (Chandola et al., 2009). In breed purity analyses, crossbred or admixed animals can be considered as genomic outliers when compared to purebreds. Methods for anomaly detection can be supervised, unsupervised, or semi-supervised. If the reference set includes both known 'normal' (purebreds in this case) and 'abnormal' (outliers, i.e., crossbred animals) labels, the method is supervised. If there is no knowledge of the status (purebred or crossbred) of available animals, the method is unsupervised, and it has been used to filter outliers before applying supervised machine-learning models for breed identification (Liu et al., 2022). Finally, if only normal (purebreds) are included in the reference set, the method is semi-supervised. Semi-supervised anomaly detection is more widely applicable and can perform better than supervised techniques in the context of analyzing only one target breed of interest (Chandola et al., 2009).

In this study, we present crossbreeding detection methods based on classic semi-supervised anomaly detection techniques that use a single breed genotype panel from purebred animals and we demonstrate its ability compared to the classical multibreed compositional regression method.

Materials and Methods

Genotype data

The sample for this study was provided by Hypor. It included genotypes from purebred large white (LW; n = 2000), landrace (LR; n = 2000), Duroc (n = 2000), Pietrain (n = 2000),

Received October 15, 2024 Accepted March 15, 2025.

and from crossbred pigs (50% Pietrain, 25% LW, 25% LR, n = 1010). All animals were genotyped using a commercial GeneSeek 50K array that is routinely used by Hypor in its breeding operations. A total of 45,436 SNP genotypes per sample were encoded as 0, 1, and 2 representing the dosage of a reference allele that was held consistent across all genotypes and populations.

LW was chosen as the target breed for this study because it exhibits substantial genetic diversity, larger effective population size, and less persistence of linkage disequilibrium compared to other breeds, which makes it more challenging to study breed purity (Badke et al., 2012). SNP with minor allele frequency (MAF) smaller than 0.01 in LW were filtered out, resulting in a 44,616-SNP panel. Genotypes from other breeds were used as test genotypes and to generate in-silico admixed genotypes for more testing (see simulation section).

Simulation of the admixed dataset

Since LW and LR are genomically closer as shown in the principal component analysis (PCA) plot where LW and LR clustered close to each other (Supplementary Figure 1), individual genotypes in the LR panel were used to simulate admixed animals. Four admixed datasets were simulated by randomly mixing LW and LR genotypes at 4 different composition rates following the procedure used by a previous study (Funkhouser et al., 2017).

First, the genomes of 2 animals randomly selected from LW and LR panels respectively were evenly segmented into 24 segments of 1,859 SNP. Then 2, 3, 4, and 6 segments were randomly selected from LR and corresponding segments of LW were replaced to form a synthetic genome. This simulation was repeated 500 times to create a 500-individual 'admixed' pig panel. The genomic composition rate was proportional to the length of the segments of these 2 purebreds, resulting in 75.0%, 83.0%, 87.5%, and 91.6% proportion of Large-White-originated alleles.

Multi-breed composition regression

As a positive control or reference method, the multi-breed compositional regression method (Funkhouser et al., 2017) was implemented. Two reference sets were built: 1) using genotypes from all 4 purebreds and 2) using genotypes from all purebreds except LR. The second reference set was used to assess the ability of this method to detect non-purebred LW animals with ancestors from a breed not included in the reference panel (i.e., LR). The model was (Funkhouser et al., 2017):

$$\mathbf{y} = X\beta + \mathbf{e} \qquad [1]$$

where $\mathbf{y}$ is a vector of length 10,000 containing a single individual's genotypes expressed as proportion of the reference allele (0, 0.5, 1) and $\mathbf{e}$ is a vector of residuals. X is the 10,000 $\times q$ matrix with allele frequencies of randomly selected 10,000 SNP and 100 animals from each reference purebred, and q is the number of purebreds (q = 4 for all purebreds panel and q = 3 for purebreds excluding LR panel). The reduced number of SNP used in this method followed previously published results (Funkhouser et al., 2017) and this SNP density has been shown effective for estimating breed compositions. $\boldsymbol{\beta}$ is a vector of length q representing breed compositions and its solution was constrained by $\sum_{i=1}^{q} \hat{\beta}_{\mathrm{i}} = 1$. Quadratic programming, as implemented in the quadprog R package (Weingessel, 2013), which was used to fit the model in equation [1] subject to this restriction. Moreover, the model was fitted for the genotypes of each individual and R package breedTools (Funkhouser et al., 2017) was used to do the analysis.

Single-breed linear regression model

A total of 1,500 LW pigs were randomly selected to serve as a single-breed reference panel. Test sets included the remaining 500 LW pigs, purebred LR (n = 2,000), Duroc (n = 2,000), Pietrain (n = 2000), crossbred (50% Pietrain, 25% LW, 25% LR, n = 1010), and 4 admixed (each n = 500) pigs. A linear regression model (LM) was used to estimate the coefficients $\boldsymbol{\beta}$ and R^2 of test animals. This model was similar to the one implemented in equation [1], where $\mathbf{y}$ is a vector of length 44,616 containing an individual's genotypes, X is a matrix with 2 columns of length 44,616: the first column is all ones related to the intercept and the second column contains allele frequencies of the single reference breed (LW) to estimate the regression coefficient. $\boldsymbol{\beta}$ is a vector that includes an intercept and a regression coefficient, with no constraints imposed on the estimation of $\boldsymbol{\beta}$. This model was also fitted individually. An anomalous datapoint (genotype of crossbred or admixed animal) should exhibit estimated regression coefficients and R^2 values that depart from the distribution of those statistics computed in purebred LW test animals. The rejection of putative crossbred or admixed genotypes was performed using empirical thresholds described in the section below.

Principal component analysis and reconstruction

Principal component analysis is a method of dimensionality reduction (Jolliffe, 2002). An application of this method in anomaly detection is to reconstruct the original data using the PCA components and find anomalous items using reconstruction error (Jablonski et al., 2015).

First, the reference genotype matrix $Y_{reference}$ (1,500 × 44,616) was centered by subtracting matrix $1_{1500}u'$, where 1_{1500} is a 1,500 × 1 column vector of ones, and u' (1 × 44,616) is the row vector of column means of $Y_{reference}$. The centered reference genotype matrix $centered\_Y_{reference}$ (1,500 × 44,616) was given by:

$$centered\_Y_{reference} = Y_{reference} - 1_{1500}u' \qquad [2]$$

After centering, PCA was performed on $centered\_Y_{reference}$, and the eigenvectors matrix V (44,616 × 1,500) was obtained using singular value decomposition (SVD) by solving equation [3] as implemented in R function prcomp():

$$centered\_Y_{reference} = U\Sigma V^T \qquad [3]$$

where U (1,500 × 1,500) is the matrix of left singular vectors, Σ (1,500 × 1,500) is the diagonal matrix containing the singular values, and V^T (1,500 × 44,616) is the transpose matrix of V, which contains the 1,500 principal components for the reference data.

Eigenvectors V were then used to project the genotypes of a test individual onto all principal components and to reconstruct the genotypes. In this process, the first step was to center the genotypes of a test individual y_{test} (a row vector of length 44,616) by subtracting u' in equation [4].

$$centered\_y_{test} = y_{test} - u' \qquad [4]$$

Second, the $centered\_y_{test}$ was projected into the 1,500 principal components by multiplying it with V, producing the PCA score vector pca_y_{test} of length 1,500:

$$pca\_y_{test} = centered\_y_{test} \cdot V \qquad [5]$$

Third, the genotype of the test individual was reconstructed from the PCA scores. The reconstruction process reprojected the scores pca_y_{test} back to the original centered feature space by post-multiplying to V^T (1,500 × 44,616). Thus, these 2 steps were equal to: $centered\_y_{test} \cdot V \cdot V^T$, where $V \cdot V^T$ acted as a projection matrix as V has orthogonal columns, and $centered\_y_{test} \cdot V \cdot V^T$ was the orthogonal projection of $centered\_y_{test}$ onto the column space of V. If test genotypes lie closer to the column space of V (the PCA space of reference genotypes), their reconstruction will be more accurate. Finally, u' was added back to restore the original scale of genotypes. Thus, the reconstructed $\hat{y}_{test}$ (a vector of length 44,616) was represented by:

$$\hat{y}_{test} = pca\_y_{test} \cdot V^T + u' = centered\_y_{test} \cdot V \cdot V^T + u' \qquad [6]$$

The code for the whole process is available at: https://github.com/xiaohanj-isu/pca_reconstruction_SNPchip. Reconstruction correlation and mean square error (MSE) of reconstruction were used as statistical criteria for detecting crossbreeding. The correlation was calculated as:

$$\operatorname{cov}(y_{test},\ \hat{y}_{test})/(\sqrt{\operatorname{var}(y_{test})} \cdot \sqrt{\operatorname{var}(\hat{y}_{test})}) \qquad [7]$$

and MSE was calculated as:

$$\frac{1}{n}\,(y_{test} - \hat{y}_{test})(y_{test} - \hat{y}_{test})^T \qquad [8]$$

where n = 44,616 for each individual. An empirical threshold was computed from the distribution of those statistics in purebred LW test genotypes.

Phasing and imputation

Another way to perform anomaly detection is through genotype imputation. Genotype imputation has been developed for the prediction of missing genetic variants from known genotypes (Yun et al., 2009). Imputation accuracy is higher when imputation is performed within a breed compared to using another breed's reference population (Friedenberg and Meurs, 2016). This can be exploited to detect admixture and crossbreeding with a single-breed panel.

SHAPEIT 5 (Hofmeister et al., 2023) and IMPUTE 5 (Rubinacci et al., 2020) were utilized for genotype phasing and imputation. In the test sets, 14,305 SNP were masked as missing for imputation. The masked SNP corresponded to those that exhibited a correlation ≥0.85 in the PCA reconstruction of test LW genotypes, i.e., SNP that were easier to be imputed accurately in the target breed.

Imputation correlation and mean square error of imputation were used as statistical criteria for detecting crossbreeding. Correlation and mean square error were calculated using equations [7] and [8] (n = 14,305) in the previous section, but y_{test} and $\hat{y}_{test}$ are vectors of length 14,305 representing observed genotypes and re-imputed genotypes for masked SNP per individual. An empirical threshold was computed from the distribution of those statistics in purebred LW test genotypes.

Performance evaluation

Each applied method used specific metrics to detect crossbred or admixed animals among purebred individuals. In all cases, the distribution of one or more statistics was characterized for purebred LW test genotypes and it was used as an empirical reference distribution to compare individual test statistics from putative crossbred individuals.

For single-breed linear regression, the test statistics used were the linear coefficient and R^2. For PCA and Imputation, the test statistics were the correlation and MSE between the observed and the reconstructed/imputed multi-SNP genotypes for each animal.

In all cases, the threshold used to call crossbred animals was the upper 5th percentile (for metric MSE), or lower 5th percentile (for metrics correlation, linear coefficient, and R^2) of the distribution of the test statistic for purebred LW animals. Specifically, any tested animals with a metric value larger (for MSE) or smaller (for correlation coefficient, linear coefficient, and R^2) than the threshold were rejected as non-purebreds. This rule would produce 5% of false rejections for LW and the proportion of true rejection in test sets was used as a measure of the performance of each method.

Results and Discussion

The multibreed compositional regression method coupled with a reference panel of LW, LR, Duroc, and Pietrain, estimated breed compositions that closely matched true compositions (Table 1). All LW pigs had a LW coefficient that was larger than 0.97. Average LR proportion for LW animals was estimated at 1.2%, which was low per-se, yet the highest among all other estimated breed coefficients. This result was consistent with previous observations (Funkhouser et al., 2017). Also, the LW coefficient in simulated admixed genotypes were very close to true values, and even animals with LW segments as high as 91.6% were clearly separated from 100% LW purebreds. Simulated admixed genotypes were also tested against a compositional regression panel that did not contain LR in the panel, to emulate the case where the 'foreign breed' was not part of the reference. When LR was omitted from the reference panel, LW proportion in admixed animals was overestimated to be close to 0.9, which did not accurately reflect their breed composition (Table 2).

Table 3 shows the proportion of rejections for genomic datasets from purebreds, crossbreds, and in-silico admixed panels using LM, PCA, and Imputation methods. All these 3 methods can achieve a 100% rejection rate in all other purebred sets, as well as for the experimental cross of 50% Pietrain, 25% LW, and 25% LR. This demonstrated the potential of semi-supervised anomaly detection to discover purebred animals from different breeds or crosses.

In the context of in-silico admixed genotypes, imputation-based and PCA-based anomaly detection showed similar performance for the purpose of detecting non-purebred animals. They both achieved almost a 100% rejection rate of admixed genotypes with LW proportion ≤83%. Imputation had the highest rejection rate for the 91% LW admixed set, even detecting 80% of the admixed samples, while PCA had the highest rejection rate of approximately 95% for the 87.5% LW admixed set. The single breed LM method performed

Table 1. Breed composition validation using multi-breed regression with 4 purebreds reference panel (foreign breed LR included; Mean regression coefficients, SD in parentheses)

	Predicted				
Dataset	LW	LR	DUR	*P*	R^2
LW (100% LW)[1]	0.973 (0.042)	0.012 (0.039)	0.006 (0.014)	0.009 (0.0184)	0.394 (0.024)
LR (100% LR)[1]	0.015 (0.044)	0.971 (0.042)	0.006 (0.013)	0.008 (0.020)	0.408 (0.023)
Duroc (100% DUR)[1]	0.004 (0.045)	0.005 (0.041)	0.987 (0.015)	0.003 (0.021)	0.641 (0.023)
Pietrain (100% P)[1]	0.008 (0.045)	0.008 (0.042)	0.005 (0.015)	0.978 (0.022)	0.561(0.02)
50% P, 25% LW, 25% LR[2]	0.230 (0.033)	0.249 (0.022)	0.020 (0.013)	0.501 (0.018)	0.357 (0.025)
91.6% LW, 8.4% LR[3]	0.897 (0.028)	0.086 (0.037)	0.007 (0.013)	0.010 (0.016)	0.363 (0.023)
87.5% LW, 12.5% LR[3]	0.859 (0.01)	0.123 (0.01)	0.007 (0.016)	0.011 (0.009)	0.352 (0.019)
83.0% LW, 17.0% LR[3]	0.815 (0.017)	0.167 (0.017)	0.008 (0.011)	0.010 (0.03)	0.341 (0.028)
75.0% LW, 25.0% LR[3]	0.736 (0.061)	0.245 (0.061)	0.008 (0.018)	0.011 (0.034)	0.322 (0.021)

[1] real genotype of purebreds, [2] real genotypes of crossbreds, [3] simulated admixed genotypes.
Abbreviations: DUR, Duroc; LW, large white; LR, landrace; P, Pietrain; SD, standard deviation.

Table 2. Breed composition using multi-breed regression with 3 purebreds reference panel (foreign breed LR excluded; mean regression coefficients, SD in parentheses)

	Predicted			
Dataset	LW	DUR	P	R^2
91.6% LW, 8.4% LR[1]	0.950 (0.027)	0.022 (0.016)	0.028 (0.023)	0.361 (0.02)
87.5% LW, 12.5% LR[1]	0.934 (0.027)	0.029 (0.017)	0.038 (0.024)	0.348 (0.019)
83.0% LW, 17.0% LR[1]	0.916 (0.03)	0.037 (0.018)	0.047 (0.025)	0.333 (0.019)
75.0% LW, 25.0% LR[1]	0.881 (0.03)	0.052 (0.018)	0.067 (0.026)	0.304 (0.017)

[1] simulated admixed genotype.
Abbreviations: LW, large white; LR, landrace; SD, standard deviation.

Table 3. Comparison of proportion of rejection among LM, PCA, and imputation using single breed anomaly detection

	Proportion of rejection, %					
	LM	LM	PCA	PCA	Imputation	Imputation
Dataset	coefficient	R^2	correlation	MSE	correlation	MSE
LW[1]	5	5	5	5	5	5
LR[1]	100	100	100	100	100	100
DUR[1]	100	100	100	100	100	100
P[1]	100	100	100	100	100	100
50% P, 25% LW, 25% LR[2]	100	100	100	100	100	100
91.6% LW, 8.4% LR[3]	41.4	49	65.4	75.4	80.6	79
87.5% LW, 12.5% LR[3]	70.4	78.4	94.2	95.6	92.2	91.8
83.0% LW, 17.0% LR[3]	93.4	96.6	98.4	98.8	99.4	99.4
75.0% LW, 25.0% LR[3]	100	100	100	100	100	100

[1] real genotype of purebreds, [2] real genotypes of crossbreds, [3] simulated admixed genotypes.
Abbreviations: DUR, Duroc; LW, large white; LR, Landrace; P, Pietrain; LM, linear model; PCA, reconstruction-based principal component analysis; MSE, mean square error.
The proportion of rejection was obtained with respect to the distribution of metrics in reference breed LW.

relatively worse in the admixed set with a high proportion of LW (91.6% and 87.5% LW), with a rejection rate below 50% in genotypes with 91.6% LW genomic segments. However, it could detect over 90% of admixed animals with a proportion of LW ≤83%, very close to that of the Imputation and PCA methods. In general, the higher the proportion of LW, the more difficult it is to distinguish an admixed genotype from purebred genotypes.

Although the outlier detection rate is a main factor evaluated when comparing these anomaly detection methods, there are other considerations for their practical implementation. For instance, PCA and LM are linear methods, which makes them computationally simpler and easier to implement in breeding programs. In contrast, imputation-based anomaly detection can utilize linkage disequilibrium to capture subtler differences in complex genotypic patterns (Rubinacci et al., 2020), which increases the computational cost of detection. However, some breeding programs include imputation in their usual data analysis pipeline (Dassonneville et al., 2011), making its implementation for anomaly detection readily accessible for those cases.

Conclusion

The presented results clearly demonstrate the effectiveness and limitations of semi-supervised anomaly detection techniques for breed purity analyses when using a single breed reference genotype panel. Anomaly detection enables the identification of admixed individuals compared to purebreds, but only when the proportion of the foreign breed is relatively large (in the present results ¼ or more). Genotype imputation and PCA reconstruction showed similar performance and can be applied based on breed entities' detection needs and conditions. Notably, this study illustrates a challenging scenario of anomaly detection in the admixture of a large proportion of genomic segments from a highly variable target breed with a small proportion of segments from a genomically close foreign breed.

Supplementary Data

Supplementary data are available at *Journal of Animal Science* online.

Acknowledgments

This is a collaborative research between the Departments of Animal Sciences at Iowa State University, Ames, IA 50011, USA, and Hendrix Genetics, Boxmeer, 5831 CK, The Netherlands.

Conflict of interest statement

All authors have no conflicts of interest.

Author Contributions

Xiaohan Jiang (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Austin Putz (Data curation, Formal analysis, Investigation, Methodology, Resources, Writing—review & editing), Juan Steibel (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing—original draft, Writing—review & editing), and Wen Huang (Formal analysis, Investigation, Methodology, Software, Writing—review & editing)

Literature Cited

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, and J. P. Steibel. 2012. Estimation of linkage disequilibrium in four US pig breeds. BMC Genomics 13:24. doi:10.1186/1471-2164-13-24

Chandola, V., A. Banerjee, and V. Kumar. 2009. Anomaly detection: a survey. ACM Comput. Surv. 41:1–58. doi:10.1145/1541880.1541882

Dassonneville, R., R. F. Brøndum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M. S. Lund, V. Ducrocq, and G. Su. 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. J. Dairy Sci. 94:3679–3686. doi:10.3168/jds.2011-4299. Available from:https://www.sciencedirect.com/science/article/pii/S0022030211003651

Friedenberg, S. G., and K. M. Meurs. 2016. Genotype imputation in the domestic dog. Mammalian Genome 27:485–494. doi:10.1007/s00335-016-9636-9

Funkhouser, S. A., R. O. Bates, C. W. Ernst, D. Newcom, and J. P. Steibel. 2017. Estimation of genome-wide and locus-specific breed composition in pigs. Transl. Anim. Sci. 1:36–44. doi:10.2527/tas2016.0003

Hofmeister, R. J., D. M. Ribeiro, S. Rubinacci, and O. Delaneau. 2023. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. Nat. Genet. 55:1243–1249. doi:10.1038/s41588-023-01415-w

Huang, Y., R. O. Bates, C. W. Ernst, J. S. Fix, and J. P. Steibel. 2014. Estimation of U.S. Yorkshire breed composition using genomic data1. J. Anim. Sci. 92:1395–1404. doi:10.2527/jas.2013-6907

Jablonski, J. A., T. J. Bihl, and K. W. Bauer. 2015. Principal component reconstruction error for hyperspectral anomaly detection. IEEE Geosci. Remote Sens. Lett. 12:1725–1729. doi:10.1109/lgrs.2015.2421813

Jolliffe. 2002. Principal component analysis for special types of data. In: Principal Component Analysis. New York, NY: Springer. p. 338–372. doi:10.1007/0-387-22440-8_13

Kuehn, L. A., J. W. Keele, G. L. Bennett, T. G. McDaneld, T. P. L. Smith, W. M. Snelling, T. S. Sonstegard, and R. M. Thallman. 2011. Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 Bull Project1,2. J. Anim. Sci. 89:1742–1750. doi:10.2527/jas.2010-3530

Liu, R., Z. Xu, J. Teng, X. Pan, Q. Lin, X. Cai, S. Diao, X. Feng, X. Yuan, J. Li, et al. 2022. Evaluation of six machine learning classification algorithms in pig breed identification using SNP array data. Anim. Genet. 54:113–122. doi:10.1111/age.13279

Rubinacci, S., O. Delaneau, and J. Marchini. 2020. Genotype imputation using the positional burrows wheeler transform. PLoS Genet. 16:e1009049. doi:10.1371/journal.pgen.1009049

Weingessel, A. 2013. Quadprog: Functions to solve Quadratic Programming Problems. R package version 1.5-5. http://CRAN.R-project.org/package=quadprog

Yun, L., C. Willer, S. Sanna, and G. Abecasis. 2009. Genotype imputation. Annu. Rev. Genomics Hum. Genet. 10:387–406. doi:10.1146/annurev.genom.9.081307.164242