Journal of Applied Statistics, 2014
http://dx.doi.org/10.1080/02664763.2014.898136

Taylor & Francis
Taylor & Francis Group

Standard statistical tools for the breed allocation problem

Pablo Martínez-Camblor[a,b*], Carlos Carleos[b], Jesus Á Baro[c] and Javier Cañón[d]

[a]*Oficina de Investigación Biosanitaria (OIB-FICYT), Oviedo, Spain;* [b]*Dpto. de Estadística, IO y DM, Universidad de Oviedo, Spain;* [c]*Dpto. de CC. Agroforestales, Universidad de Valladolid, Spain;* [d]*Lab. de Genética, Facultad de Veterinaria, Universidad Complutense de Madrid, Spain*

(*Received 26 June 2013; accepted 22 February 2014*)

Modern technologies are frequently used in order to deal with new genomic problems. For instance, the STRUCTURE software is usually employed for breed assignment based on genetic information. However, standard statistical techniques offer a number of valuable tools which can be successfully used for dealing with most problems. In this paper, we investigated the capability of microsatellite markers for individual identification and their potential use for breed assignment of individuals in seventy Lidia breed lines and breeders. Traditional binomial logistic regression is applied to each line and used to assign one individual to a particular line. In addition, the area under receiver operating curve (AUC) criterion is used to measure the capability of the microsatellite-based models to separate the groups. This method allows us to identify which microsatellite loci are related to each line. Overall, only one subject was misclassified or a 99.94% correct allocation. The minimum observed AUC was 0.986 with an average of 0.997. These results suggest that our method is competitive for animal allocation and has some interpretative advantages and a strong relationship with methods based on SNPs and related techniques.

Keywords: breed allocation; microsatellite loci; logistic regression; ROC curves

1. Introduction

Frequently, sophisticated methodologies are used in order to handle (sometimes huge) datasets as in -omic studies. These methodologies are commonly based on complex computational algorithms and, sometimes, in some way, there is a loss of the flavor of statistics. Although standard statistical methods are often left aside, they can provide really interesting results in most contexts.

In this paper, we consider the breed allocation problem i.e., the allocation of individual animals to their corresponding lines, by using microsatellite markers. Allocation of individual animals to their corresponding lines may be achieved by analysis of shared genomic patterns. Several methods have been proposed for assigning anonymous samples to reference populations by making

*Corresponding author. Email: pmcamblor@hotmail.com

use of hypervariable genetic markers. For instance, Pritchard *et al.* [14] proposed an unsupervised Bayesian-based method that provides the posterior probability of each individual originating in each of a set of ancestral populations. These methods are frequently based on intensive computations and therefore specific software such as `STRUCTURE` [5], `GeneClass2` [13] or `Genepop` [15] is often used to solve these problems. In this work, we deal with this problem by using standard statistical analysis. This has a connection with methods used in biomedicine to study diagnostic ability, in particular logistic regression (LR) and the receiver operating-characteristic (ROC) curve analysis are used for the breed allocation. In addition, an ROC-surface (average of all involved AUCs) is proposed to measure the global quality of the classification criterion. All analyses were performed by customized but simple, routines in `R`. In Section 2, we introduce the particular problem. In Section 3, the employed methodology is described. Section 4 is devoted to depict the obtained results when these techniques are applied on the current dataset. Finally, Sections 5 and 6 provide further discussion of the work and our main conclusions.

2. The fighting bull breed allocation problem

The Lidia bovine or fighting bull breed is reared in France, Portugal, Spain, and several American countries. It constitutes a racial group that has fragmented into small lines, traditionally called *encastes*, with different levels of gene flow among them and different sets of behavioral traits favored in each one. The more isolated lines may have only a few animals contributing to the gene pool and thus suffer from a loss of genetic variation and inbreeding depression, and others might derive from a mixture of ancestral lines [4].

For each of 70 lines, individuals from the same generation were randomly sampled. Fresh blood was taken from 1811 individuals, with approximately the same number of males and females. They were genotyped for 24 microsatellite loci scattered over most chromosomes. These 24 microsatellite set used in the study consists of those microsatellites that were available at the moment and which informativeness was supported by previous experience in the laboratory. The number is within the usual range for studies based on this kind of genetic markers.

The goal is to use this information to allocate one individual within the adequate line. In addition, the observed association between the lines and the microsatellite loci could associate the genotype with some phenotypic features of the particular line.

3. Methodology

First, suppose that we have N subjects; of which n with the studied characteristic (one particular line/encaste) and $N - n$ without. We also have k variables from which we desire to classify individuals into their appropriate group. This is a familiar problem in biomedicine where the study of diagnostic technique is a topic and where the LR [8] is often used.

The LR is appropriate for finding the best-fitting model to describe the relationship between a dichotomous outcome variable (in the current study, to belong or not to belong to a particular line) and a set of independent variables (microsatellite loci) also called *covariates*. The specific form of the LR model is as follows:

$$\pi(\boldsymbol{x}) = [e^{\boldsymbol{x}\boldsymbol{\beta}^t}] \cdot [1 + e^{\boldsymbol{x}\,\boldsymbol{\beta}^t}]^{-1},$$

where Y is the dependent variable (taken values 0 and 1 in subjects with and without the characteristic, respectively), $\boldsymbol{x} = (1, x_1, \ldots, x_k)$ is a vector with fixed values of the independent variables, $\pi(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{x}]$ stands for the conditional mean of Y given $\boldsymbol{x}$, and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)$

is the vector of parameters. Therefore, applying the *logit transformation*,

$$g(\boldsymbol{x}) = \log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \boldsymbol{x}\cdot\boldsymbol{\beta}^{\boldsymbol{t}}. \tag{1}$$

Of course, the problem is how to estimate $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)$ usually based on the maximum likelihood method. In this case, the estimated coefficient-vector will be the one which maximizes the expression,

$$\begin{aligned} L(\hat{\boldsymbol{\beta}}) &= \log\left(\prod_{i=1}^{N} \pi(\boldsymbol{x}_i)^{y_i}[1-\pi(\boldsymbol{x}_i)]^{1-y_i}\right) \\ &= \sum_{i=1}^{N}\{y_i\cdot\log(\pi(\boldsymbol{x}_i)) + (1-y_i)\cdot\log(1-\pi(\boldsymbol{x}_i))\}, \end{aligned} \tag{2}$$

where y_i's and $\boldsymbol{x}_i$'s are the values of the dependent and independent variables, respectively ($1 \le i \le N$). For each variable in the model, e^{β_j} ($1 \le j \le k$) can be identified with the odd's ratio (OR) associated with this variable and, as usual (the OR is a measure of association commonly used, especially, in epidemiology) it can be interpreted as how much more likely (or unlikely) the outcome is when the variable increases by one unit. For qualitative covariates, a reference group must be fixed (OR = 1) and, rest of ORs (one for group) can be interpreted as how much more likely (or unlikely), with respect to the reference group, the outcome is within the different levels of the covariate. Note e^{β_0}, appeared in the above equation, does not have an obvious interpretation. However, for each subject (which takes the values x for covariates), $g(\boldsymbol{x})$ can be interpreted as a score which measures how likely it is that this particular subject was associated with the studied characteristic. Categorical variables are included by several dummy variables in the same model.

For the current problem, in order to compute how likely is that a subject will be from a particular line, a LR is conducted. The covariates of the LR are related to the microsatellite alleles as follows. Recall that 0, 1 or 2 copies of a certain allele can be carried by an individual in a certain microsatellite locus. For each possible allele of each microsatellite, two covariates are included. The first covariate equals 1 if the individual carries at least one copy of that allele; 0 otherwise, that is, if no such copy is present. The second covariate equals 1 if the individual carries two copies of that allele; 0 otherwise. Seventy different LRs with a number of covariates will be performed and consequently, for each subject, 70 punctuations will be obtained. For more information about the microsatellite loci handling, see, for instance, [3].

Once the K ($K = 70$ in our study) LRs have been fitted, each animal, $T_{j,l}$ (with $j \in 1, \ldots, N_l$, N_l stands for the number of animals in the lth line, and with $l \in 1, \ldots, K$) will be assigned to the line which has the greatest punctuation, i.e. which from the available information, is more plausible. Obviously, if $g_l(\boldsymbol{x}_i) = \max_{1\le l\le K}\{g_1(\boldsymbol{x}_i), \ldots, g_K(\boldsymbol{x}_i)\}$ (g_l stands for the lth punctuation, $1 \le l \le K$), the classification will be correct.

Note that each model could be used in order to classify an individual as with or without the characteristic. Consequently, the lth model ($1 \le l \le K$) allows one to classify the individuals in the lth line. The ROC curves, $\mathcal{R}(\cdot)$, are a valuable tool in order to describing the intrinsic accuracy of a model. It is a plot of the *sensitivity* (S_E) of the classification model (i.e. the ability of the model to detect the condition of interest, the line) versus the complementary of the *specificity* ($1 - S_P$) of this model (i.e. the inability of the test to recognize individual from different lines). In addition, the AUC (area under ROC curve, $\mathcal{A} = \int_0^1 \mathcal{R}(t)\,\mathrm{d}t$) is often used in order to measure the quality of binary classifications [6]. There exists a rich literature on both the ROC curves and the AUC (see, for instance, [20] and references therein) and they have a wide range of applications

(see [10,11] for recent reviews). In addition, the AUC has also been used as goodness-of-fit criterion in stepwise model selection (see [12]). In this setting, each particular AUC, $\mathcal{A}_l$ ($1 \leq l \leq K$), can be understood as the global capacity of the lth LR to identify the animals of the lth line but not like a general line identification. However, the AUC shows the model capacity to distinguish the origin of the animals. The ROC-surface volume (joint K-ROC curves) can be seen as a measure (ranges between 0.5 and 1) of this capability. In the current problem, due to the great number of possible covariates, a number of them are previously selected. This selection is based on the ORs out of which the univariate LRs is made. Once fixed the number of covariates to be selected, V, this procedure is equivalent to computing the P-values of the Fisher's exact test between the dependent variable and each independent covariate and then to select the V-variables with the smallest P-values. Hence, the V most significant covariates are initially included within the model. New covariates are included in the model only if the increment in the ROC-surface volume is larger than 0.05.

4. Results

Twenty-four loci on 1811 animals from 70 different lines were collected. The variability of the loci alleles ranged between 7 and 22. The Shannon information index [19] ($\text{SII} = -\sum_{i=1}^{n}[f_i \cdot \log_2(f_i)]$, where f_i is the frequency of the ith allele and n is the number of different alleles observed) ranged between 1.30 and 3.27. Figure 1 depicts the different values for both different number of alleles and the SII order by SII.

The total number of possible (dummy) covariates were 540, which are the results of asking to the dataset about the presence of a particular allele in each particular loci. The 40 covariates with the smallest P-values were finally used in order to build the models. Figure 2 depicts the minimum P-value of each loci in each model. Lines in the extremes of the plot (those labeled as '15', '11','13', '42', '12', '47', '46' and '55') had a weaker relationship with the Loci information (green zones). In general, each Loci had a strong relationship with several lines. Both the loci and the lines are ordered by clustering based on the Euclidean distance of the original dataset. Note that, in this context, the Euclidean distance between two individuals will depend on the number of shared alleles.

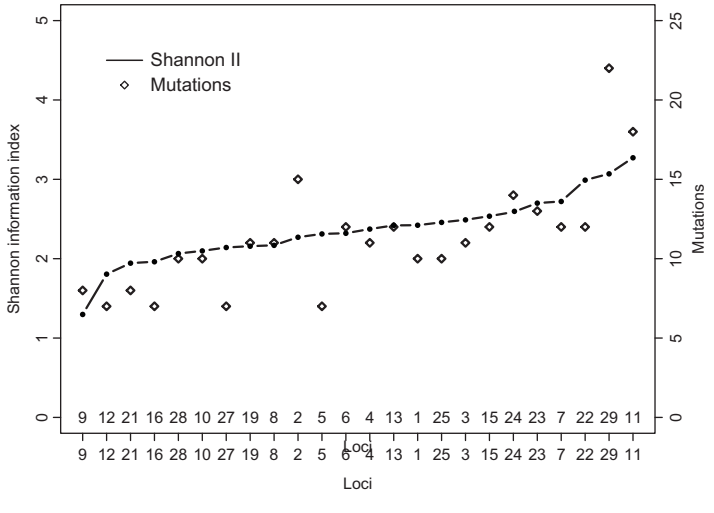

Figure 1. The Shannon information index and number of alleles for different collected loci.

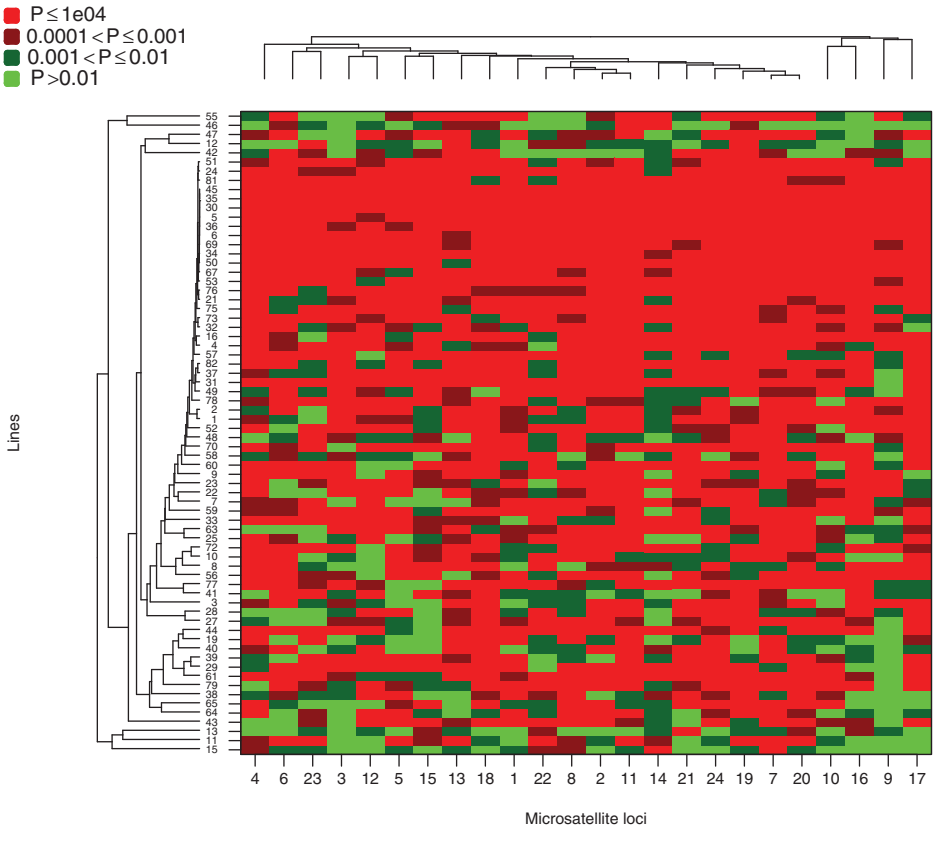

Figure 2. Heat map for the relationship between loci and line. The clusters are based on the Euclidean distances on the original data.

The allocation capacity for the (complete) model was really good. Only one subject was incorrectly classified (99.94% were correctly allocated). In particular, one subject which belong to the line labeled as '32' was allocated within the line '61'. These lines are far from each other in the cluster analysis. The minimum observed AUCs were 0.986 for the line '32' and 0.993 for the line '61'. The remaining AUCs were 1. The ROC surface (average of these AUCs) was 0.997 (95% bootstrap confidence interval of (0.995–1.00)).

The prediction capacity decreases when a leave-one-out cross-validation method is applied. By using the 40 most significant covariates (from the possible 540), the observed true classification rate was 0.791 (0.772–0.810). However, this proportion can be improved by increasing the number of selected covariates. In particular, when we take the 100 most significant ones (instead of the 40 ones previously taken), the true classification rate was 0.866 (0.850–0.882). Finally, when the 250 most significant variables are used, the value of the true-classification rate achieved 0.879 (0.864–0.894). In the last case, the particular discriminant capacity (to belong to or not to belong to a particular line) was really good for most of the lines. The minimum observed AUC was 0.870 (for the line labeled as '15'). Fifteen lines achieved an absolutely correct separation (AUC = 1). The volume under ROC surface was 0.987 (95% bootstrap confidence interval of (0.983–0.991)). Figure 3 depicts the 70 ROC curves. In addition, a box-plot for the AUC distribution and the 3D ROC surface is also shown. The 3D ROC surface stands for a three-dimensional figure where all the ROC curves (sorted by the AUC value) are plotted sequentially on the z-axis. Because the number of curves is high enough, a surface is obtained.

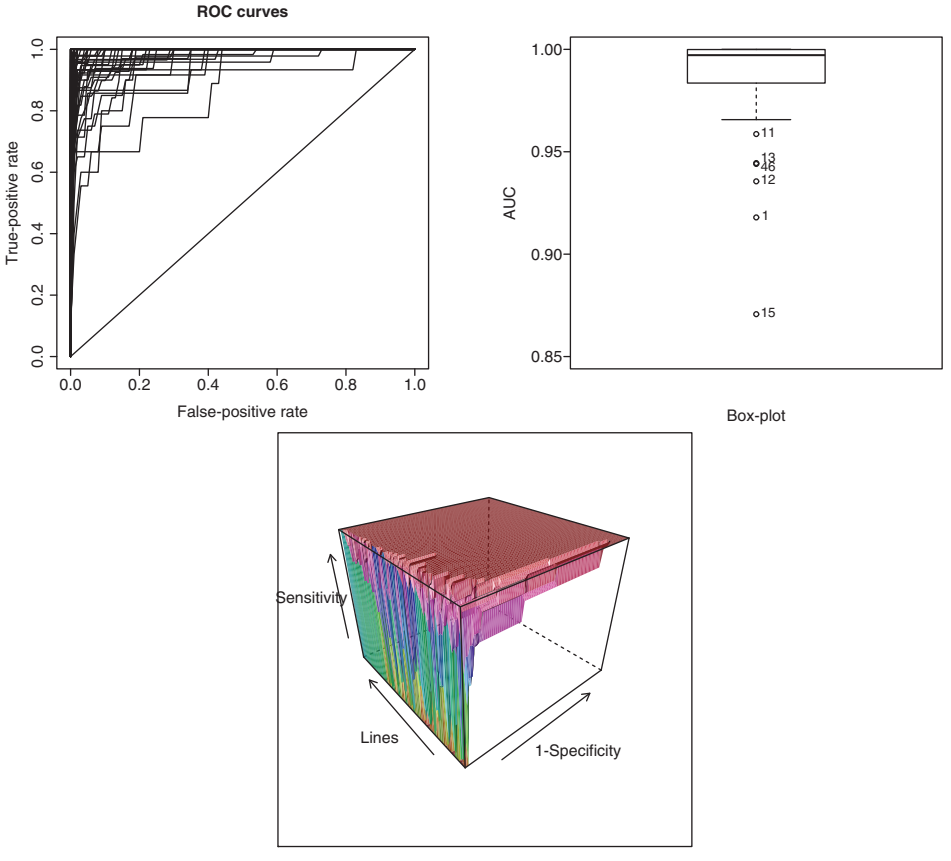

Figure 3. Upper, 70 ROC curves (left) and box-plot for the AUC distribution (right). Lower, ROC-surface.

When all models are considered jointly and each subject is assigned to its most plausible line, the observed overall percentage of true classification was 87.9% (95% confidence interval of (86.4–89.4)). The percentages of true-classification vary greatly with the line. The method classifies poorly the lines labeled as '15' with only the 22.2% of true classification percentages, '12' with the 41.7%. Both lineages are characterized by a high degree of admixture (data not shown). The remaining lines have, at less, the 50% of true classification. Sixteen lines were classified correctly. Figure 4 depicts the punctuation of each animal for each line. In color, the subjects which really are for the line. Note that, in some lines, subjects can be seen with a really low punctuation in the model (line) (ojo que quito!) they belong to. Obviously, these subjects are finally incorrectly classified. In Figure 4, the true-classification rate and a 95% confidence interval are also shown.

There are subjects which obtain poor punctuations for all the lines i.e., there is no strong evidence that they were from any of the considered lines. It is really difficult to assign, correctly, these subjects to a particular line. When we only classify the subjects whose maximum punctuations are larger than a fixed value (rest of the subjects would be labeled as *undefined*), for instance 0.9 (a total of 1491, 82.3%), the true classification rate was 0.945 (0.933–0.957). This percentage can be improved by taking a cut-off of 0.99 until the 0.950 (0.938–0.962). In the last case, the total number of classified subjects was 1415.

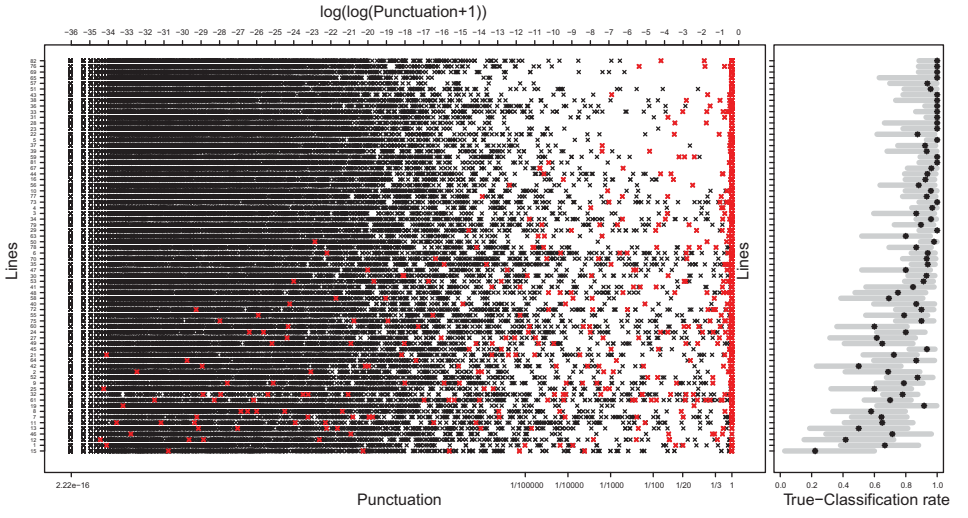

Figure 4. Punctuations of each bullfigther for each line. In color, the subject from the line. Also, true-classification rate and 95% confidence interval for each line.

5. Discussion

Although the true classification rate of the method performed well (probability = 0.879), it is still a bit worse than that obtained by the method based on the maximum-likelihood (0.910). This method described in Cañón *et al.* [2, Section 2.5.3], correctly classified 30 lines. Besides overlaps on classification, the two methods are based on different approaches of classification, and obtain different results. For instance, the proposed method obtained better true classification rates for the lines labeled as '61', '11' or '32'. On the other hand, results obtained by the LR procedure were quite similar to those obtained from other methods based on more computing intensive procedures, particularly based on data-mining and machine-learning approaches [7]. Of course, it is possible to improve its performance by making use of information on Mendelian sampling and relationship between the alleles within the same microsatellite loci [3]. Moreover, the merits of LR has already been studied. Setakis *et al.* [18], via simulations, show that, methods based on LR (i) are flexible, (ii) computationally fast, (iii) easy to implement and (iv) provide good protection against the effects of cryptic relatedness, even though they do not explicitly model the population structure.

One of the main issues, as a most interesting topic in the area, is the number of variables to be used. Similarly to other models, the absolute quality (true classification rate) of each particular model does not decrease with the number of covariates (measured via pseudo R^2). However, the overall quality of the procedure can decrease when too much covariates are included. We have observed that each models (70, in this case) must be fitted separately. When there are few *positives* (animals from a particular line) in the model, if too many variables are included, the model turns unstable and the classifications are somewhat spurious. For instance, the true classification rate for the line labeled as '15' was 0.333 when 100 variables were used and decreased to 0.222 for 250 variables. We think that this is an interesting open problem which requires particular and more research.

6. Main conclusions

Problems related with the -omic sciences often involve a huge quantity of information. Frequently, the solutions to the main questions are tackled with algorithms specifically developed for

this goal. However, standard statistical methods can prove useful for dealing with most problems. In this paper, well-known multivariate models were used to deal with the allocation problem by using microsatellite loci information. AUC was also used in order to measure the quality of the performed classification and the ROC surface was employed as stopping rule for introducing new variables within the models. The marker information is included in the form of dummy variables. In particular, for specific loci, each allele is represented by two variables, the response to the questions: *Does almost one parent have this allele?* (yes/no) and *Do two parents have this allele?* (yes/no). Note that, due to this way of introducing the variables, the relationship with problems based on SNPs information (usually a huge number of SNPs is available) and other statistical techniques such as the logic regression (see, for instance [16,17]) or the random forest (see, for instance, [1] or [9] for software tools) is direct.

The main advantage of the current method lies perhaps in its natural interpretation for non-geneticists (in particular, statisticians). In addition, each model allows for discriminating each line and identifying which microsatellite loci are specifically involved. Moreover, multivariate methods take into account the internal data structures (correlations matrix) therefore they do not need the independent (between loci) assumption (linkage equilibrium) no independent alelles within the same locus (Hardy–Weiberg equilibrium).

Acknowledgements

This work was supported by the Grant MTM2011-23204 of the Spanish Ministry of Science and Innovation (FEDER support included).

Samples were provided by UCTL (Union de Criadores de Toros de Lidia) within the frame of a research project founded by the INIA and the European Regional Development Fund no: RZ2008-00005-C02-02. This study also received financial support of INIA, RTA2011-00060-C02-O2.

References

[1] L. Breiman, *Random forests*, Mach. Learn. 45(1) (2001), pp. 5–32.
[2] J. Cañón, P. Alexandrino, I. Bessa, C. Carleos, Y. Carretero, S. Dunner, N. Ferran, D. Garcia, J. Jordana, D. Laloë , A. Pereira, A. Sanchez, and K. Moazami-Goudarzi, *Genetic diversity measures of local European beef cattle breeds for conservation purposes*, Genet. Select. Evol. 33 (2001), pp. 311–332.
[3] J. Cañón, M.L. Checa, C. Carleos, J.L. Vega-Pla, M. Vallejo, and S. Dunner, *The genetic structure of Spanish Celtic horse breeds inferred from microsatellite data*, Anim. Genet. 31 (2000), pp. 39–48.
[4] J. Cañón, I. Tupac-Yupanqui, M.A. Garcia-Atance, M. Cortes, D. Garcia, J. Fernandéz, and S. Dunner, *Genetic variation within the Lidia bovine lineage*, Anim. Genet. 39 (2008), pp. 439–445.
[5] G. Evanno, S. Regnaut, and J. Goudet, *Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study*, Mol. Ecol. 14(8) (2005), pp. 2611–2620.
[6] R. Fluss, D. Faraggi, and B. Reiser, *Estimation of the Youden Index and its associated cut point*, Biomet. J. 47(4) (2005), pp. 458–472.
[7] B. Guinand, A. Topchy, P.S. Page, M.K. Burnham-Curtis, P.K. Punch, and K.T. Cribner, *Comparisons of likelihood and machine learning methods of individual classification*, J. Heredity 93(4) (2002), pp. 260–269.
[8] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed., New York, Wiley, 2000.
[9] A. Liaw and M. Wiener, *Classification and regression by random forest*, R News 2/3 (2002), pp. 18–22.
[10] P. Martínez-Camblor, C. Carleos, and N. Corral, *Powerful nonparametric statistics to compare k independent ROC curves*, J. Appl. Statist. 38(7) (2011), pp. 1317–1332.
[11] P. Martínez-Camblor, C. Carleos, and N. Corral, *General nonparametric ROC curve comparison*, J. Korean Statist. Soc. 42(1) (2013), pp. 71–81.
[12] P. Martínez-Camblor, J. de Uña-Álvarez, and C. Díaz-Cote, *Expanded renal transplantation: A multi-state approach*, (2013), Unpublished manuscript.
[13] S. Piry, A. Alapetite, J.M. Cornuet, D. Paetkau, L. Baudouin, and A. Estoup, *GeneClass2: A software for genetic assignment and first-generation migrant detection*, J. Heredity 95 (2004), pp. 536–539.
[14] J.K. Pritchard, M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data*, Genetics 155 (2000), pp. 945–959.

[15] F. Rousset, *Genepop'007: A complete reimplementation of the Genepop software for Windows and Linux*, Mol. Ecol. Resources 8 (2008), pp. 103–106.
[16] I. Ruczinski, C. Kooperberg, and M. LeBlanc, *Logic regression*, J. Comput. Graph. Statist. 12(3) (2003), pp. 475–511.
[17] I. Ruczinski, C. Kooperberg, and M LeBlanc, *Exploring interactions in high-dimensional genomic data: An overview of logic regression, with applications*, J. Multivar. Anal. 90 (2004), pp. 178–195.
[18] E. Setakis, H. Stirnadel, and D.J. Balding, *Logistic regression protects against population structure in genetic association studies*, Genome Res. 16 (2006), pp. 290–296.
[19] C.E. Shannon, *A mathematical theory of communication*, Bell Syst. Tech. J. 27 (1948), pp. 379–423.
[20] X.H. Zhou, N.A. Obuchowski, and D.K. McClish, *Statistical Methods in Diagnostic Medicine*, New York, Wiley, 2002.