Animal, page 1 of 8 © The Animal Consortium 2017
doi:10.1017/S1751731117001355

animal

Preselection statistics and Random Forest classification identify population informative single nucleotide polymorphisms in cosmopolitan and autochthonous cattle breeds

F. Bertolini[1], G. Galimberti[2], G. Schiavo[1], S. Mastrangelo[3], R. Di Gerlando[3], M. G. Strillacci[4], A. Bagnato[4], B. Portolano[3] and L. Fontanesi[1,†]

[1]*Department of Agricultural and Food Sciences, Division of Animal Sciences, University of Bologna, Viale Fanin 46, 40127 Bologna, Italy;* [2]*Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Via delle Belle Arti 41, 40126 Bologna, Italy;* [3]*Department of Agricultural and Forestry Sciences, University of Palermo, Viale delle Scienze, 90128 Palermo, Italy;* [4]*Department of Veterinary Medicine, Università degli Studi di Milano, Via Celoria 10, 20133 Milano, Italy*

(Received 10 November 2016; Accepted 11 May 2017)

Commercial single nucleotide polymorphism (SNP) arrays have been recently developed for several species and can be used to identify informative markers to differentiate breeds or populations for several downstream applications. To identify the most discriminating genetic markers among thousands of genotyped SNPs, a few statistical approaches have been proposed. In this work, we compared several methods of SNPs preselection (Delta, F_{st} and principal component analyses (PCA)) in addition to Random Forest classifications to analyse SNP data from six dairy cattle breeds, including cosmopolitan (Holstein, Brown and Simmental) and autochthonous Italian breeds raised in two different regions and subjected to limited or no breeding programmes (Cinisara, Modicana, raised only in Sicily and Reggiana, raised only in Emilia Romagna). From these classifications, two panels of 96 and 48 SNPs that contain the most discriminant SNPs were created for each preselection method. These panels were evaluated in terms of the ability to discriminate as a whole and breed-by-breed, as well as linkage disequilibrium within each panel. The obtained results showed that for the 48-SNP panel, the error rate increased mainly for autochthonous breeds, probably as a consequence of their admixed origin lower selection pressure and by ascertaining bias in the construction of the SNP chip. The 96-SNP panels were generally more able to discriminate all breeds. The panel derived by PCA-chrom (obtained by a preselection chromosome by chromosome) could identify informative SNPs that were particularly useful for the assignment of minor breeds that reached the lowest value of Out Of Bag error even in the Cinisara, whose value was quite high in all other panels. Moreover, this panel contained also the lowest number of SNPs in linkage disequilibrium. Several selected SNPs are located nearby genes affecting breed-specific phenotypic traits (coat colour and stature) or associated with production traits. In general, our results demonstrated the usefulness of Random Forest in combination to other reduction techniques to identify population informative SNPs.

Keywords: SNP, breed assignment, Random Forest, *Bos taurus*

Implications

Combining several reduction statistics (Delta, F_{st} and principal component analyses (PCA)) with Random Forest (RF) classification we could select and test population informative single nucleotide polymorphism (SNP) panels containing 96 and 48 SNPs. Single nucleotide polymorphisms were selected among about 50 000 markers analysed in 3304 cattle from six breeds (three cosmopolitan and three autochthonous). The obtained SNP panels might be suitable for breed authentication of cattle-derived products and breed allocation with a low-error rate. The statistical approaches used in this study introduce a useful methodology that can be extended when it is needed to select informative SNPs.

Introduction

The variability within and among livestock populations is the result of natural and artificial selection, genetic drift and admixture events that have contributed to shape the genetic uniqueness and diversity of many different breeds (Notter, 1999; Andersson and Georges, 2004; Decker et al. 2014).

† E-mail: luca.fontanesi@unibo.it

Commercial SNP genotyping tools have been recently developed for several species, including cattle, providing information from many polymorphic sites (Matukumalli et al., 2009). These tools can be used to identify informative markers, creating reduced panels that might be able to differentiate breeds and populations for several downstream applications. Part of these applications could be breed allocation of individuals, breeds of origin of crossbred animals, authentication of mono-breed products and comparative analyses of selection signatures (Wilkinson et al., 2011; Bertolini et al., 2015).

To identify the most discriminating genetic markers among thousands of genotyped SNPs, a few statistical approaches have been proposed. For example, one of the simplest method uses the Delta values that are the absolute allele frequency differences at each polymorphic site in pairwise comparisons between populations (Shriver et al., 1997; Smith et al., 2001). Delta analysis has been explored by Wilkinson et al. (2011) and Hulsegge et al. (2013) to find informative SNPs in several cattle breeds. Another statistic useful for these purposes is Wright's F_{st} analysis that measures the standardized variance in allele frequencies among populations (Wright, 1951). F_{st} has been extensively applied to identify informative genetic markers and population structures in humans and livestock species, including cattle (e.g. Bowcock et al., 1994; Wilkinson et al., 2011; Hulsegge et al., 2013). A third approach is the PCA that is an unsupervised linear technique for dimension reduction and allows to extract axes of maximal variation from data sets (Jolliffe, 2002; Jolliffe and Cadima, 2016). Principal component analysis has been already used in human populations to characterize their structure based on SNP genotyping data (Paschou et al., 2007) and in cattle to reduce dimensionality of large SNP data sets and to identify breed informative SNPs (Lewis et al., 2011; Wilkinson et al., 2011; Bertolini et al., 2015).

These pre-filtering steps normally are coupled with approaches that can classify and assign breeds or individuals. One of these methods is the RF approach, that is an algorithm used for classification and regression and is based on an ensemble of low-correlated decision trees (Breiman, 2001). To guarantee low correlation among decision trees, each tree is built on a different randomly perturbed version of the data set. In a classification context, RF allows to assign an unknown sample to a pre-determined group. This classification technique has been used for genome-wide association studies (GWAS) in case and control analyses for a few human diseases (e.g. Lunetta et al., 2004; Jiang et al., 2009), and more recently it has been tested to perform association analyses on sheep pigmentation, comparing and combining RF with most traditional GWAS methodologies (Kijas et al., 2013). We recently applied RF to predict breed allocation using SNP chip data on four major cattle breeds after a marker preselection obtained by using PCA performed chromosome by chromosome to reduce the dimensionality of the data set (Bertolini et al., 2015). In our previous study, we evaluated the performance of this SNP selection procedure (PCA-chrom + RF) in a few breeds showing that breed classification can be obtained without any error. This strategy could identify informative SNPs that had also been located in genes or close to genes that are known to affect breed-specific traits.

In this work, we extended our previous study by testing and comparing several methods of SNP preselection (Delta, F_{st} and PCA) in addition to RF classifications to analyse thousands of SNP genotypes from several cattle breeds, including cosmopolitan (Holstein, Brown and Simmental) and autochthonous (Cinisara, Modicana and Reggiana) breeds. These latter breeds are raised in the Italian regions of Sicily (Cinisara and Modicana) and Emilia Romagna (Reggiana). They have been adapted for a long time to local environmental conditions and their milk is used to produce mono-breed protected designation of origin cheeses. For each preselection approach (Delta, F_{st} and PCA), reduced panels of 96 and 48 SNPs have been created and evaluated in terms of ability to discriminate both cosmopolitan and autochthonous breeds using RF classification. Linkage disequilibrium (LD) within the selected SNP panels has been also evaluated. Moreover, we analysed the gene content of the regions nearby the selected SNPs, showing that several of them are close to genes that might be responsible for defining breed-specific traits or other economically important traits.

Material and methods

Data set description

A total of 3304 animals belonging to cosmopolitan cattle breeds (2091 Holstein, 738 Brown and 475 Simmental), genotyped with the Illumina BovineSNP50 v1 BeadChip array (Illumina, San Diego, CA, USA) and 311 animals belonging to three Italian local cattle breeds (71 Cinisara, 72 Modicana and 168 Reggiana) genotyped with the Illumina BovineSNP50 v2 BeadChip array (Illumina) were considered for the analysis. Genotyped animals are derived from previous studies that describe in details these data sets (Mastrangelo et al., 2014 and 2016; Bertolini et al., 2015). Reggiana samples represent almost all sires available for this breed (Fontanesi et al., 2015; Mastrangelo et al., 2016). Several steps of filtering were applied to remove SNPs not useful for the analysis: (1) SNPs what were not shared between the two SNP chip versions; (2) SNPs that were mapped on the sex chromosomes or unmapped; (3) SNPs with call rate <0.95 in at least one breed. Minor allele frequency (MAF) was not used to pre-filter SNPs. Missing SNPs were imputed within breed using Beagle 3.3.2 (Browning and Browning, 2007).

Validation step

Each cattle breed was divided into a reference population and a test population. The test population, generated by randomly sampling about 10% of the animals within each breed (209 Holstein, 74 Brown, 47 Simmental, 7 Cinisara, 7 Modicana and 16 Reggiana), was used for the validation of the breed assignment. However, it is worth to mention that RF does not need any cross-validation on a separate test set to get an unbiased estimate of the test set error. Error in the

RF classification is estimated internally, directly during the run. Therefore, the test population could be considered a further validation of the results.

The reference population was composed by the remaining animals and was used for all reduction and allocation analyses. A multidimensional scaling (MDS) plot of the considered breeds was obtained using Plink 1.07 (Purcell et al., 2007).

Single nucleotide polymorphism reduction techniques

A total of four reduction techniques was applied to the reference population of the six investigated breeds: (i) Delta analysis; (ii) F_{st} statistics; (iii) PCA carried out separately on SNPs assigned to each chromosome (chromosome by chromosome PCA or PCA-chrom); and (iv) PCA carried out considering at the same time all SNPs (whole genome PCA or PCA-whole). Each of the four approaches identified a reduced SNP panel including a total of 580 markers, retaining those with the highest values of the corresponding statistic for the whole genome approaches (Delta, F_{st}, and PCA-whole) or the 20 highest ranked SNPs for each chromosome in the chromosome by chromosome approach (PCA-chrom). For this study, no pre-filtering of SNPs in LD was performed as this will be one of the parameters to evaluate the SNP panels.

Average Delta. For a biallelic marker the Delta value is given by $|pAi - pAj|$, where pAi and pAj are the frequencies of allele A in the *i*th and *j*th populations, respectively. Because Delta can be only estimated between pairs of populations, and the data set contained six different populations, values were averaged across all pairwise comparisons to produce an estimated value for each SNP (Wilkinson et al., 2011).

Average F_{st}. Breed allele frequency was calculated for each SNP. Then, for every breed combination (15 in total), F_{st} was calculated for each marker by adapting the formula reported by Karlsson et al. (2007):

$$F_{st_k} = N_k / D_k$$

where k is the SNP marker k, with frequency $p_1^{[k]}$, $p_2^{[k]}$

$$N_k = p_1^{[k]}\left(q_2^{[k]} - q_1^{[k]}\right) + p_2^{[k]}\left(q_1^{[k]} - q_2^{[k]}\right)$$

$$D_k = p_1^{[k]} q_2^{[k]} + q_1^{[k]} p_2^{[k]} = N_k + p_1^{[k]} q_1^{[k]} + p_2^{[k]} q_2^{[k]}$$

The value of each SNP was finally averaged between the 15 breed-by-breed comparisons.

Principal component analysis. Principal component analysis was computed using the prcomp function of the R software 2.12 (http://www.R-project.org) based on SNP allele frequencies of each breed. The principal components that range from PC1 to PC5 explained 100% of the variance (Supplementary Figure S1). Therefore, they were considered for the SNP ranking and selection (see Jolliffe, 2002, for more details about the selection of the relevant principal components). The analysis was carried out autosome by autosome (PCA-chrom) and on the whole SNP list (PCA-whole) on the reference populations. According to Paschou et al. (2007), for each SNP marker, the scores on the five selected Principal Components were squared and summed. These quantities were used to rank the SNP markers.

Breed assignment and further reduction with Random Forest. Random Forest based on the preselected 580 SNPs for each reduction technique (Delta, F_{st}, PCA-chrom and PCA-whole) were built on the reference population using the 'randomForest' package in R (Liaw and Wiener, 2002; www.stat.berkeley.edu/~breiman/RandomForests/). To select the most discriminant SNPs, ranking based on the mean decrease in the Gini index implemented in the function 'importance' of the R package was examined. The mean decrease in the Gini index is a variable importance measure that was specifically devised for ensemble of classification trees, such as RFs. It is based on the contribution of each variable in reducing the within-node heterogeneity of a tree. These contributions are averaged over all the trees that compose the RF (see Hastie et al., 2009 for further details).

For each reduction technique and based on the mean decrease in the Gini index ranking of the 580-SNP panel, two different SNP panels were then created: a 96-SNP panel (a larger panel of 96 SNPs) and 48-SNP panel (a smaller panel of 48 SNPs that included the 48 highest ranked SNPs of the 96-SNP panel). For each of the reduced panel (the 96 and the 48 SNPs panels), a new RF was fitted and the corresponding Out Of Bag (OOB) error rate (a method of measuring the prediction error of the RF classifier) was by default calculated by the R package. The choice of a reduced numbers of SNPs (96 and 48 SNPs) was due to the practical possibilities to develop multiplex SNP panels that might contain a reduced number of SNPs for field applications.

Classification performance of these RFs was also assessed using the test population, not used in determining the SNP panels and for this reason considered as an independent subset of samples. For each breed, LD and MAF were calculated in the 96-SNP panels with the highest mean decrease in the Gini Index value using the software PLINK 1.07 (Purcell et al., 2007). Selected SNPs included in the 96-SNP panel with the lowest OOB were checked in the *Bos taurus* reference genome (UMD 3.1, GCA_000003055.3; http://www.ensembl.org/index.html). The adjacent genes before and after each selected SNP were identified using the software BEDTools (Quinlan and Hall, 2010) and the most recent annotated version of the *Bos taurus* genome was derived by the application Biomart (http://www.ensembl.org).

Results

Population genetics overview

A total of 50 761 SNPs was shared between the two SNP chips and was assigned to an autosomal chromosome.

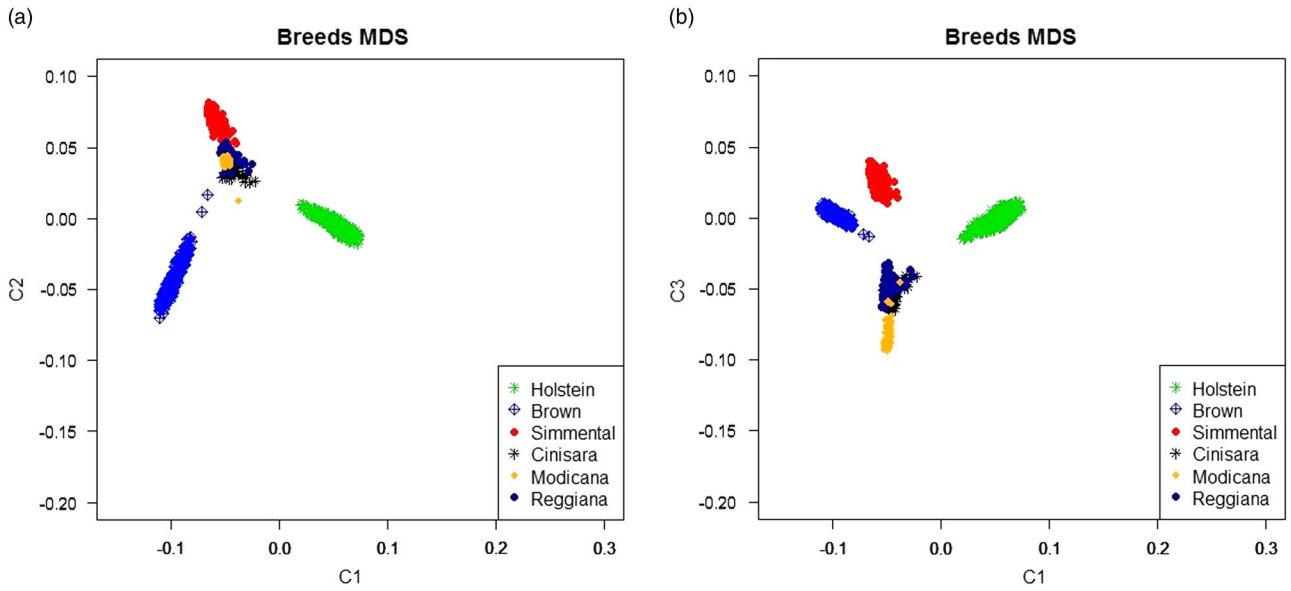

Figure 1 Multidimensional scaling (MDS) of the six analysed breeds considering components 1 and 2 (a) and components 1 and 3 (b).

Among these SNPs, 39 878 passed the final filter step of the call rate and were considered for the subsequent analyses. A multi-dimensional scaling plot considering the components 1 and 2 (Figure 1a) shows a clear separation of the three commercial breeds and a clustering of the autochthonous breeds, which is confirmed also considering the components 1 and 3 (Figure 1b). This clustering of the autochthonous breeds included not only Cinisara and Modicana breeds, that are raised in the same area, but also Reggiana breed (raised in Emilia Romagna region, Province of Reggio Emilia – North of Italy), which is located about 1300 km far away from the region of sampling of Cinisara and Modicana. Reggiana is also historically and geographically separated from the two Sicilian breeds.

Description of the 580-single nucleotide polymorphism panels

PCA-chrom selected a 580-SNP panel including 20 SNPs for each autosome, independently from the across chromosome ranking of SNPs located on each chromosome. All the three whole genome SNP preselection approaches (Delta, F_{st} and PCA-whole), selected SNPs located in all autosomes but with different density according to the applied method (Table 1). The lowest number of retained SNPs for F_{st} was on bovine chromosome (BTA) 27 ($n = 2$), BTA23, BTA25, and BTA28 ($n = 7$). Five SNPs (on BTA28) was the lowest number of preselected SNPs with the PCA-whole method. The highest numbers of preselected SNPs for both Delta and PCA-whole approaches were on BTA6 (89 and 82 SNPs, respectively). F_{st} preselected the highest number of SNPs (39) on BTA2. A total of 51 SNPs was shared among the four 580-SNP panels (i.e. Delta, F_{st}, PCA-whole and PCA-chrom). Delta and PCA-whole shared the highest number of preselected SNPs (340; 58.6% of the panels). Delta and PCA-chrom shared 268 common SNPs (46.2%). The two PCA approaches (PCA-whole and PCA-chrom) had 295 common SNPs (50.9%). The most divergent 580 panels were obtained by Delta and F_{st} that shared only 15.7% of SNPs. The Venn diagram with all the combinations of shared SNPs is reported in Supplementary Figure S2a.

Characteristics of the reduced SNP panels

Two reduced SNP panels (96 and 48 SNPs) were selected according to the mean decrease in Gini coefficient, where a high value means a high contribution of the SNP in shaping the structure of the trees that compose the RF, and hence in determining category assignments. These lists of SNPs are reported in the Supplementary Table S1. The 96-SNP panels contained SNPs on many different chromosomes (except: BTA17 and BTA25 for Delta; BTA21, BTA24 and BTA25 for F_{st}; BTA12, BTA15, BTA17 and BTA27 for PCA-whole; BTA14 and BTA15 for PCA-chrom). BTA6 still contained the highest number of SNPs in Delta, F_{st} and PCA-whole (16, 16 and 15, respectively), while the highest number of SNPs for PCA-chrom was located on BTA10, with 15 SNPs (Table 1). The reduced 48-SNP panel contained more chromosomes without SNPs: eight chromosomes for Delta; nine for PCA-chrom; 10 for F_{st} and PCA-whole (Table 1). Again, the chromosomes with the highest number of SNPs were BTA6 for Delta, F_{st} and PCA-whole and BTA10 for PCA-chrom (Table 1). No SNPs were common by all the four 96-SNP panels. A total of six SNPs was shared among three panels: Delta, F_{st} and PCA-whole that shared one SNP on BTA6 (6 : 38576012); Delta, F_{st} and PCA-chrom that shared one SNP on BTA8 (8 : 96594716) and BTA11 (11 : 67559090); Delta, PCA-whole and PCA-chrom that shared one SNP on BTA10 (10 : 19366262), BTA11 (11 : 9851948) and BTA20 (20 : 30398960). Pairs of 96-SNP panels shared from 7 to 10 SNPs (Supplementary Figure S2b), whereas pairs of 48-SNP panels only had in common two to four SNPs (Supplementary Figure S2c).

As no pre-filtering step was performed to remove SNPs in high LD, we evaluated if part of the SNP data set could

Table 1 *Chromosome distribution of single nucleotide polymorphisms (SNPs) of the 580-SNP panel selected with the whole genome-based approaches: Delta, F_{st}, PCA-whole (PCA-chrom has a fixed number of 20 SNPs each chromosome) and chromosome distribution of the 96-SNP panels and 48-SNP panels considering the reduction techniques from which they derived*

	580-SNP panel				96-SNP panel				48-SNP panel			
Chromosome	Delta	F_{st}	PCA-whole	PCA-chrom	Delta	F_{st}	PCA-whole	PCA-chrom	Delta	F_{st}	PCA-whole	PCA-chrom
1	21	24	19	20	2	4	4	2	1	2	2	0
2	16	39	18	20	2	3	6	1	0	3	2	0
3	11	28	13	20	4	3	4	1	1	2	3	1
4	32	24	36	20	4	1	6	5	1	0	3	5
5	57	38	55	20	7	13	8	1	4	8	3	0
6	89	29	82	20	16	16	15	2	8	9	7	0
7	35	27	23	20	4	4	6	3	2	1	2	1
8	21	17	18	20	5	1	1	5	2	1	1	0
9	11	16	10	20	1	1	2	3	1	0	1	1
10	19	27	28	20	1	4	5	15	0	1	3	8
11	26	20	25	20	8	6	7	7	6	1	3	3
12	10	18	12	20	1	3	0	3	0	0	0	2
13	31	25	23	20	5	3	1	2	3	3	1	1
14	19	15	23	20	4	4	7	0	2	2	5	0
15	9	15	6	20	3	1	0	0	1	1	0	0
16	29	24	25	20	3	5	5	2	1	1	2	0
17	6	31	10	20	0	6	0	2	0	2	0	2
18	22	22	27	20	3	3	1	5	2	3	0	4
19	16	24	14	20	2	1	3	1	1	0	1	1
20	18	16	18	20	3	2	2	6	2	2	2	4
21	9	8	7	20	1	0	1	2	1	0	1	1
22	13	13	11	20	5	1	2	3	2	1	1	2
23	8	7	6	20	2	0	0	6	0	0	0	2
24	13	15	15	20	4	0	2	2	2	0	2	0
25	6	7	13	20	0	0	2	4	0	0	0	3
26	19	17	20	20	2	7	3	3	2	3	3	1
27	2	11	6	20	1	1	0	2	1	0	0	1
28	5	7	5	20	1	1	1	4	1	0	0	2
29	7	16	12	20	2	2	2	4	1	2	0	3

PCA = principal component analyses.

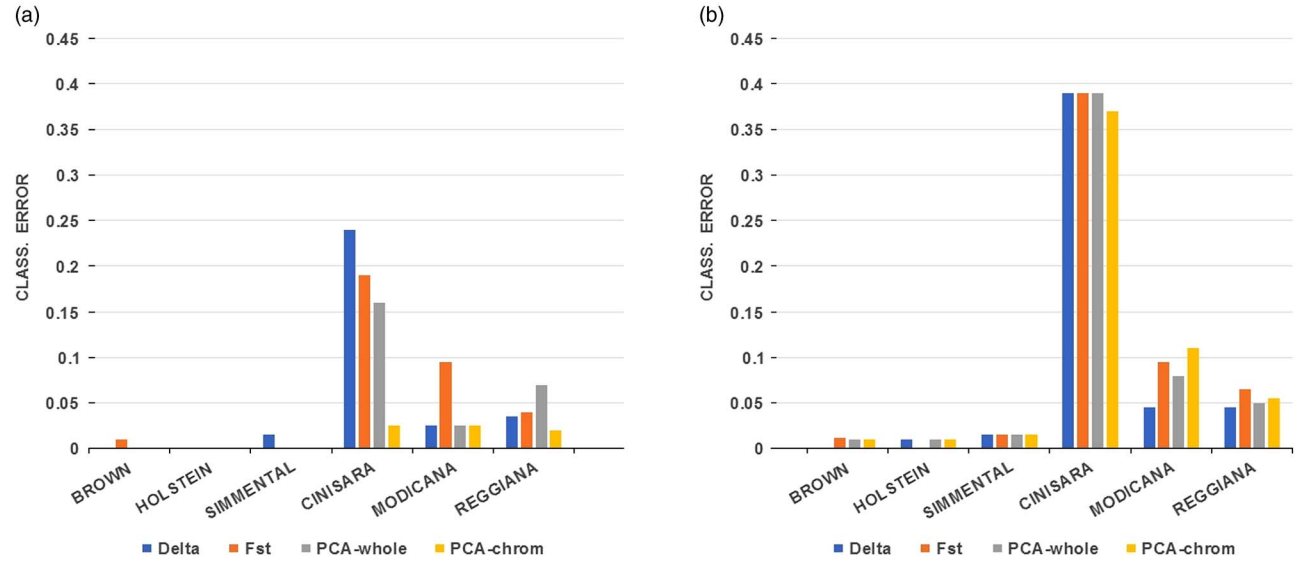

Figure 2 Classification error (from 0 to 1.0) of the single breeds (Brown, Holstein, Simmental, Cinisara, Modicana, Reggiana) using (a) the 96-SNP panels (b) the 48-SNP panels. SNP = single nucleotide polymorphism; PCA = principal component analyses.

include highly correlated SNPs. Considering $r^2 = 0.6$ as a threshold, the LD calculated considering all markers of the 96-SNP panels identified different numbers of SNPs in LD with $r^2 > 0.6$ among three groups, while no SNP in LD were identified in the PCA-chrom panel (Supplementary Table S1). The PCA-whole 96-SNP panel includes two SNPs on BTA25 that were in LD in Holstein and Modicana, two SNPs on BTA24 and two SNPs on BTA6 that were in LD only in Simmental. Two additional SNPs located on BTA6 (positions: 71421017 and 71452210 on UMD3.1) were in LD in Simmental, Cinisara and Reggiana. The Delta 96-SNP panel included a few SNPs in LD on BTA3 (two SNPs in Modicana), on BTA6 (two in Holstein, Brown, Modicana and Reggiana; two SNPs in Brown) and BTA22 (two SNPs in Holstein, Brown and Simmental). The largest number of SNPs in LD ($n = 20$) in different breeds was in the list of the F_{st} 96-SNP panel. BTA6 was the chromosome with the highest number of SNP in LD (two in Holstein, 10 in Brown, four in Cinisara, 12 in Modicana, four in Reggiana). Two of these SNPs located on BTA6 (positions: 39257620 and 39346170 in UMD3.1) and other two SNPs were in LD across all breeds (Supplementary Table S1).

Out Of Bag values in the different single nucleotide polymorphism panels. After the preselection of the 580-SNP panels, and the reduction of this list to 96 and 48-SNP panels based on their ranking in terms of informativeness, RF analyses were applied separately to the four panels with the purpose of learning a classification rule to assign animals to the six pre-defined groups (the six cattle breeds considered in this study) using information based on the three different SNP levels. Mean OOB rates using the different 580-SNP, 96-SNP and 48-SNP panels of the four reduction techniques are reported in Table 2. The highest OOB rate for the 580-SNP panel was reached by the F_{st} approach (0.60%) whereas the lowest OOB rate was observed for the PCA-chrom approach (0.03%). PCA-whole and Delta reached 0.06% and 0.08% OOB rates, respectively.

The OOB error rates of the different preselection methods for the 96-SNP panels increased according to the following order (Table 2): PCA-chrom (0.12%), PCA-whole (0.71%), F_{st} (0.77%) and then Delta (0.83%). The 48-SNP panel showed an expected but not linear increase of the OOB. The lowest rate was for the PCA-whole method (0.89%), followed by Delta (1.17%), PCA-chrom (1.35%) and then F_{st} (1.38%). In Figure 2 it is shown the classification errors for each breed using the 96-SNP panels (a) and the 48-SNP panels (b). The local breeds showed the highest error rates for all preselection methods. Cinisara showed the highest number of miss-assigned animals for all approaches and for both reduced panels, reaching about 0.40 for all methods in the 48-SNP panel. The Cinisara only reached almost 100% of correct allocation using PCA-chrom with the 96-SNP panel. The Modicana showed the highest error rate using F_{st}. The Reggiana was the local breed with the lowest error rates for both 96 and 48-SNP panels. The cosmopolitan breeds were almost always correctly assigned with the 96-SNP panels. Few errors rate values (<0.006) were observed for the 48-SNP panels while the 96-SNP panels show only a very limited number of not correctly assigned animals (error rate around 0.004). The analyses of the test populations confirmed the general low error rate of assignment, with the minimum value of 0.53% reached by PCA-chrom with the 96-SNP panel and the maximum value of 1.86% reached by Delta with the 48-SNP panel (Supplementary Table S2).

Table 2 *Out Of Bag (OOB) error rate (%) of the whole training population using the single nucleotide polymorphism (SNP) panels (580, 96 and 48 SNPs) derived from the four reduction techniques*

Reduction approach	OOB-580 (%)	OOB-96 (%)	OOB-48 (%)
Delta	0.08	0.77	1.17
F_{st}	0.6	0.83	1.38
PCA-whole	0.05	0.71	0.89
PCA-chrom	0.03	0.12	1.35

PCA = principal component analyses.
For details breed-by-breed see Figure 2.

Gene annotation of the 96-SNP panels. Each SNP of the 96-SNP panel (obtained with the different approaches) was investigated to detect genes within or nearby the selected SNPs (Supplementary Table S1). A total of 39 SNPs on the Delta panel are within annotated genes, 37 SNPs for the F_{st} panel, 45 SNPs for the PCA-whole panel, and 38 SNPs for PCA-chrom. Considering the whole list of genes, whose SNPs are within or nearby, several of those have already been shown to be associated with cattle production traits or coat colour such as the KIT proto-oncogene receptor tyrosine kinase (*KIT*) gene detected by PCA-whole, the kirre like nephrin family adhesion molecule 3 (*KIRREL3*) gene), the platelet derived growth factor receptor alpha (*PDGFRA*), detected by PCA-whole and PCA-chrom, the secreted phosphoprotein 1 (*SPP1*) detected by PCA-whole a*nd* F_{st}, and ligand dependent nuclear receptor corepressor like (*LCORL*), detected by the four approaches.

Discussion

In this study, we tested several SNP reduction techniques (Delta, F_{st} and PCA-based) combined with RF for breed assignment in a large cohort of dairy and dual purpose cattle that belong not only to cosmopolitan breeds but also to local breeds that might have been influenced by cosmopolitan breeds. Thousands of SNPs provided by a high-throughput genotyping platform were analysed in a scalable manner and their number was reduced to define sets of population informative markers useful for breed allocation. Similar reduction techniques (Delta, Wright's F_{st}, Weir & Cockerham's F_{st} and PCA-whole) were already tested in cattle, using different commercial breeds, but each represented by <30 animals (Wilkinson et al., 2011). In that study, the reduction method that required the lowest number of SNP markers to verify the animal's breed origin was the Wright's F_{st} approach, while PCA showed a lower individual assignment power. Wilkinson et al.

(2011) did not report the level of LD among selected SNPs and how this parameter could be affected by the different statistical methods of SNP preselection. Hulsegge et al. (2013), who followed the process described by Wilkinson et al. (2011), did not obtain relevant differences for most statistical measures in the number of SNPs and informativeness if LD restrictions were or were not applied in selecting SNPs.

Comparing these previous findings with the approaches used in our study, the power of assignment is comparable in all the strategies, but the 96-SNP panel derived by the PCA selected chromosome by chromosome (PCA-chrom), derived and modified by Bertolini et al. (2015), reached the lowest OOB rate. The F_{st}-based approach using a formula that slightly differed from Wilkinson et al. (2011) was the method with the highest number of total miss-assigned animals, and one of the highest considering the autochthonous breeds. Moreover, PCA-chrom reduced the number of SNPs with high LD limiting the need of a preselection step based on this measure. This is probably because working chromosome by chromosome reduces the risk to select multiple SNPs because they are in LD, rather than for their importance, thus reducing the risk of bias. The 96-SNP panel derived by PCA-chrom could select informative SNPs that were particularly useful for the assignment of minor breeds that reached the lowest value of OOB error even in the Cinisara, whose value was quite high in all the other panels. While Holstein, Simmental and Brown are worldwide distributed with specific breeding plans, Cinisara and Modicana are not subject to breeding programmes, whereas Reggiana is characterized by limited selection programme that has been developed during the last 20 years (Mastrangelo et al., 2016; http://www.razzareggiana.it). This selection plan could have influenced the differentiation of the Reggiana breed among the others. In fact, the Reggiana was the local breed with the lowest error rates for both 96 and 48-SNP panels, closer to the cosmopolitan breeds than the local breeds. Despite this, the MDS plot does not show a clear separation among these minor breeds that clustered with each other and partially overlap the Simmental breed. This is also confirmed by the RF analysis that underlines the lowest efficiency in breed discrimination for the local breeds. Therefore, despite the reduced genetic variability in these breeds, the lack of specific breeding programme (Cinisara and Modicana) and external influences before the recent development of breed-specific consortia, might have facilitated their admixture with cosmopolitan breeds. This might point to the need of specific breeding plans that on one hand can protect genetic variability within each local breed, but on the other hand can emphasize breed-specific characteristics, increasing the value of breed-labelled products that are linked to them.

Random Forest has been recently proposed as an alternative approach to GWAS studies for simple and complex traits. In our previous work (Bertolini et al., 2015), RF selected SNPs that were close to genes related to production traits. Even if the current study applied several different reduction approaches to select SNPs, they provide a confirmation of the observations of our previous work. In both studies, several genes located nearby the selected SNPs are associated with production traits and coat colour suggesting a potential role of these markers to capture phenotypic differences among the investigated breeds. The *KIT* gene is well known to be associated with spotted phenotypes (Reinsch et al., 1999; Fontanesi et al., 2010b) and markers within this gene have already been proposed for breed traceability including the Reggiana breed (Fontanesi et al., 2010a and 2015). The *KIRREL3* gene was recently located in a breed-specific large-effect pleiotropic QTL, after analyzing 10 different US breeds (Saatchi et al., 2014). The *PDGFRA* gene was associated with milk composition (Cole et al., 2011) and the *SPP1* gene was associated with growth-related traits (Cohen-Zinder et al., 2005; Allan et al., 2007). The *LCORL* gene, which was found using all the reduction approaches, is strongly related to body weight/height in several species including cattle (Takasuga, 2016).

The results obtained showed that when increasing the number of breeds the error rate increases mainly for local breeds, probably as a consequence of their admixed origin or lower selection pressure that did not fix many SNPs. The 96-SNP panel constituted by the approach of applying PCA chromosome by chromosome, derived by our previous study (Bertolini et al., 2015) and implemented here is the panel with the lowest error rate and the lowest number of SNPs in LD with each other. The applicability of reduced SNP panels with low classification error rate is therefore still possible also for autochthonous breeds in which the total or partial lack of selection programmes have not shaped the genome as it might be the case for cosmopolite breeds. Many selected SNPs are located close to genes that are important for several economic and breed-specific traits, confirming the possibility of using RF as an alternative and complementary approach to identify indirectly putative selection signature regions in the investigated populations.

In conclusion, our study tested different solutions for SNP preselection and identified which one of those, in combination with RF, allowed the definition of the most breed informative 96 and 48-SNP panels. PCA-chrom preselection method, previously applied on cosmopolitan breeds only, was confirmed to be the best preselection strategy to be combined with RF, with no SNPs in LD, and therefore would not need any previous LD-filtering steps, and its application is potentially suitable also for local admixed breeds.

Acknowledgements

This work was funded by Innovagen MiPAAF and by PON01_02249 –R& C 2007–2013 – MIUR projects.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/S1751731117001355

References

Allan MF, Thallman RM, Cushman RA, Echternkamp SE, White SN, Kuehn LA, Casas E and Smith TP 2007. Association of a single nucleotide polymorphism in

SPP1 with growth traits and twinning in a cattle population selected for twinning rate. Journal of Animal Science 85, 341–347.

Andersson L and Georges M 2004. Domestic-animal genomics: deciphering the genetics of complex traits. Nature Review Genetics 5, 202–212.

Bertolini F, Galimberti G, Calò DG, Schiavo G, Matassino D and Fontanesi L 2015. Combined use of principal component analysis and Random Forests identify population-informative single nucleotide polymorphisms: application in cattle breeds. Journal of Animal Breeding Genetics 132, 346–356.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR and Cavalli-Sforza LL 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368, 455–457.

Browning SR and Browning BL 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81, 1084–1097.

Breiman L 2001. Random Forests. Machine Learning 45, 5–32.

Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, Lee JH, Drackley JK, Band MR, Hernandez AG, Shani M, Lewin HA, Weller JI and Ron M 2005. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Research 15, 936–944.

Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ Jr, Crooker BA, Van Tassell CP, Yang J, Wang S, Matukumalli LK and Da Y 2011. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genomics 12, 408.

Decker JE, McKay SD, Rolf MM, Kim J, Molina Alcalá A, Sonstegard TS, Hanotte O, Götherström A, Seabury CM, Praharani L, Babar ME, Correia de Almeida Regitano L, Yildiz MA, Heaton MP, Liu WS, Lei CZ, Reecy JM, Saif-Ur-Rehman M, Schnabel RD and Taylor JF 2014. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. PLoS Genetics 10, e1004254.

Fontanesi L, Scotti E and Russo V 2010a. Analysis of SNPs in the KIT gene of cattle with different coat colour patterns and perspectives to use these markers for breed traceability and authentication of beef and dairy products. Italian Journal of Animal Science 9, e42.

Fontanesi L, Scotti E, Samorè AB, Bagnato A and Russo V 2015. Association of 20 candidate gene markers with milk production and composition traits in sires of Reggiana breed, a local dairy cattle population. Livestock Science 176, 14–21.

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

Hastie T, Tibshirani R and Friedman JH 2009. The elements of statistical learning, 2nd edition. Springer, New York.

Hulsegge B, Calus MP, Windig JJ, Hoving-Bolink AH, Maurice-van Eijndhoven MH and Hiemstra SJ 2013. Selection of SNP from 50K and 777K arrays to predict breed of origin in cattle. Journal of Animal Science 91, 5128–5134.

Jiang R, Tang W, Wu X and Fu W 2009. A Random Forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinformatics 10, S65.

Jolliffe IT 2002. Principal component analysis. 2nd edn. Springer-Verlag, New York, NY, USA.

Jolliffe IT and Cadima J 2016. Principal component analysis: a review and recent developments. . Philosophical Transactions of the Royal Society. Series A, Mathematical, Physical, and Engineering Sciences 374, 20150202.

Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NH, Zody MC, Anderson N, Biagi TM, Patterson N, Pielberg GR, Kulbokas EJ 3rd, Comstock KE, Keller ET, Mesirov JP, von Euler H, Kämpe O, Hedhammar A, Lander ES, Andersson G, Andersson L and Lindblad-Toh K 2007. Efficient mapping of mendelian traits in dogs through genome-wide association. Nature Genetics 39, 1321–1328.

Kijas JW, Serrano M, McCulloch R, Li Y, Salces Ortiz J, Calvo JH and Pérez-Guzmán MD, International Sheep Genomics Consortium 2013. Genome wide association for a dominant pigmentation gene in sheep. Journal of Animal Breeding and Genetics 130, 468–475.

Lewis J, Abas Z, Dadousis C, Lykidis D, Paschou P and Drineas P 2011. Tracing cattle breeds with principal components analysis ancestry informative SNPs. PLoS One 6, e18007.

Liaw A and Wiener M 2002. Classification and regression by Random Forest. R News 2, 18–22.

Lunetta KL, Hayward LB, Segal J and Van Eerdewegh P 2004. Screening large-scale association study data: exploiting interactions using Random Forests. BMC Genetics 5, 32.

Mastrangelo S, Saura M, Tolone M, Salces-Ortiz J, Di Gerlando R, Bertolini F, Fontanesi L, Sardina MT, Serrano M and Portolano B 2014. The genome-wide structure of two economically important indigenous Sicilian cattle breeds. Journal of Animal Science 92, 4833–4842.

Mastrangelo S, Tolone M, Di Gerlando R, Fontanesi L, Sardina MT and Portolano B 2016. Genomic inbreeding estimation in small populations: evaluation of runs of homozygosity in three local dairy cattle breeds. Animal 10, 746–754.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS and Van Tassell CP 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One 4, e5350.

Notter DR 1999. The importance of genetic diversity in livestock populations of the future. Journal of Animal Science 77, 61–69.

Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW and Drineas P 2007. PCA-correlated SNPs for structure identification in world-wide human populations. PLoS Genetics 3, 1672–1686.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and Sham PC 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81, 559–575.

Quinlan AR and Hall IM 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.

Reinsch N, Thomsen H, Xu N, Brink M, Looft C, Kalm E, Brockmann GA, Grupe S, Kühn C, Schwerin M, Leyhe B, Hiendleder S, Erhardt G, Medjugorac I, Russ I, Förster M, Reents R and Averdunk G 1999. A QTL for the degree of spotting in cattle shows synteny with the KIT locus on chromosome 6. Journal of Heredity 90, 629–634.

Saatchi M, Schnabel RD, Taylor JF and Garrick DJ 2014. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. BMC Genomics 15, 442.

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R and Ferrell RE 1997. Ethnic-affiliation estimation by use of population-specific DNA markers. American Journal of Human Genetics 60, 957–964.

Smith MW, Lautenberger JA, Shin HD, Chretien JP, Shrestha S, Gilbert DA and O'Brien SJ 2001. Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. American Journal of Human Genetics 69, 1080–1094.

Takasuga A 2016. PLAG1 and NCAPG-LCORL in livestock. Animal Science Journal 87, 159–167.

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

Wright S 1951. The genetical structure of populations. Annals of Human Genetics 15, 323–354.