Integrating Population Genomics and Genomic Selection

***P. Ajmone-Marsan*[1]*, E. Nicolazzi*[1]*, R. Negrini*[1]*, N. Macciotta*[2]*, L. Fontanesi*[3]*, V. Russo*[3]*, A. Bagnato*[4]*, E. Santus*[3]*, D. Vicario*[4]*, J.B.C.H.M. van Kaam*[5]*, A. Albera*[6]*, F. Filippini*[7]*, C. Marchitelli*[8]*, G. Mancini*[8]*, A. Nardone*[8]*, A. Valentini*[8]**

[1]*Università Cattolica sel S. Cuore, Piacenza, Italy;* [2]*University of Sassari, Italy;* [3]*DIPROVAL, University of Bologna, Italy,* [4]*University of Milano, Italy,* [5]*ANARB, Italy;* [6]*ANAPRI, Italy;* [7]*ANAFI, Italy;* [8]*ANABORAPI, Italy;* [9]*ANABIC, Italy;* [10]*Tuscia University, Viterbo, Italy*

Abstract

The availability of panels of several thousand SNPs ordered on the genome has initiated the era of population genomics, that is the application of genomic approaches to population genetics. One application of population genomics is the investigation of patterns of diversity along the chromosomes in search for signatures left by past and recent selection. These signatures are locus specific and can be identified and distinguished from the genome wide effects caused by genetic drift and demographic events. In this paper we searched for outlier behaviour within the 54,001 SNPs of the Illumina Beadchip Array assayed on 2682 bulls belonging to Italian Brown and four other Italian breeds, one dairy (Italian Friesian), one dual purpose (Italian Simmental), and two beef (Marchigiana and Piedmontese) investigated within the Italian SELMOL project on molecular genetics applied to animal breeding. Outlier values of the F_{st} genetic differentiation index averaged along 9 markers sliding windows were searched in pairwise breed comparisons by a permutation strategy. A total of 8944 sliding windows were significant in at least one of the four comparisons that included Italian Brown. Among these, 869 SNPs were significant in all three comparisons vs dual purpose and beef breeds. These two subsets of 8944 and 869 SNPs were used in a genomic prediction exercise. The 749 Italian Brown genotyped bulls were divided in training (the 600 older bulls) and prediction (the 149 younger bulls) populations. In all cases DGVs, Bayes-A estimates of Milk Yield, Protein Percent, Udder Score and Total Economic Index, were not significantly higher than those obtained with a random marker subset of the same size. Selection signatures likely identify genomic regions subjected to historical selection that do not match with those in which genes controlling economic traits are segregating in modern populations. This hampers the use of the selection signature approach for identifying marker subsets useful in genomic selection.

1. Introduction

Population genomics is a term first proposed in a publication on human genetic diseases (Gulcher and Stefansson, 1998) to indicate the use of genomic technologies in population genetics studies. The availability of panels of many thousand and sometimes many hundred thousand SNPs ordered along the genome has recently marked a paradigm shift in the way populations can be investigated. One major advance is the ability to identify genomic regions that are under selection. These can be detected by comparing the distribution of allele frequencies at marker loci within or between populations or groups of populations, in search for markers significantly departing from neutral behaviour. The comparison of the distribution of allele frequencies can be either direct or through different statistics, function of allelic or genotypic frequencies, as F_{st} (e.g. The Bovine HapMap Consortium *et al.*, 2009) and linkage disequilibrium (e.g. Ennis, 2007). In addition, specific tests for detecting significant effects have been developed (e.g. Voight *et al.,* 2006). The objectives of this study are i) to detect selection signatures in the Italian Brown cattle breed and ii) to evaluate the performance of markers under selection in the genomic prediction of genetic values of young bulls. To reach these goals we used SNP data produced within the Italian SELMOL project on the application of molecular genetics to animal breeding.

2. Materials and Methods

2.1 Animals

A total of 2295 animals from 5 breeds were genotyped with the 54001 SNP markers included in the Illumina BovineSNP50 BeadChip: 775 Italian Brown (BRW), 419 Italian Friesian (FRI), 379 Piedmontese (PIM), 229 Marchigiana (MCG) and 493 Italian Simmental (SIM).

2.2 Genomic data

Following clean up by filtering subjects with more than 5% missing SNPs and SNPs with more than 5% missing typings, the final dataset included 2266 individuals and 45087 SNPs. Among these 43771 were located on the 29 autosomes and BTAX and 1316 remained not anchored to the Btau 4.0 version of the bovine sequence assembly. These latter were excluded from further analyses.

2.3 F_{st}

F_{st} index was calculated as $F_{st}=1–H_s/H_t$, where H_s is the Hardy-Weinberg equilibrium (HWE) heterozygosity within subdivisions, averaged across subpopulations and H_t is the HWE heterozygosity for the total population, assuming no genetic differentiation among subpopulations. F_{st} values were averaged along sliding windows of nine consecutive SNPs, irrespectively on the relative distance between adjacent markers. Each chromosome contained a number of sliding windows equal to $SW_i=N_i-(N_{sw}-1)$, where SW_i is the number of sliding windows on chromosome i, N_i is the number of markers on chromosome i and N_{sw} is the number of markers included in the sliding window. In total, 43531 sliding windows were assembled.

2.4 Permutations

To estimate the 5% genome-wide significance thresholds of F_{st} values, markers were first randomly shuffled across the genome. Then, the distribution of average F_{st} value of groups of 9 randomly selected markers was computed. Finally the F_{st} values separating 5% of the distribution were recorded. The highest values among permutation runs were used as F_{st} thresholds to evaluate the significance of F_{st} calculated on markers ordered along chromosomes.

2.5 Genomic prediction

Genomic predictions of breeding values (DGVs) were obtained using a BayesA model (Meuwissen *et al.*, 2001). A total of 20.000 runs of iteration were performed on each analysis. First 10.000 iterations were discarded as burn-in and no thinning interval was considered. The model included a polygenic term for taking the population structure into account. Accuracies were obtained as Pearson correlations between DGVs and breeding values obtained from progeny testing (EBVs).

3. Results

3.1. Selection signatures

Average F_{st} values of individual markers varied between 0.034±0.049 in BRW vs PIM to 0.057±0.080 in BRW vs FRI. F_{st} values of sliding windows had same average and smaller values of SD, spanning from 0.023 in BRW vs PIM to 0.035 in BRW vs FRI. Sliding windows spanned on average 483±263 Kb, with a maximum of 1382 Kb and a minimum of 4 Kb. In all comparisons involving Italian Brown 8944 sliding windows had F_{st} significantly higher than the 5% threshold established by the permutation approach. These sliding windows are not equally distributed across breed comparisons (Table 1). Surprisingly the comparison with Italian Friesian was the one in which the highest number of signatures was detected. A remarkable number of signatures was found consistent across all comparisons or across comparisons between BRW and the beef and dual purpose breeds. Selection signatures were not equally distributed across chromosomes as well. Numbers ranged between 1076 on BTA6 and 4 on BTAX.

Table 1. Selection signatures detected in the comparison between Italian Brown cattle and all four, three beef/dual purpose breeds and each single breed investigated.

Comparison involving Italian Brown	N. sliding windows with signature (P≤5%)
Piedmontese	1770
Marchigiana	1795
Italian Simmental	1728
Italian Friesian	1999
Three beef/dual purpose breeds	869
All four breeds	463

3.2. Genomic predictions

Table 3 summarises the correlations between EBVs and DGVs calculated using subsets of markers carrying signatures of selection. Triplicate random sets of markers having the same size of subsets investigated were also used as control. Using 8944 markers, correlations were slightly higher with markers under selection compared to the average of three runs with random subsets, but always lower than with the random subset giving the highest correlations. With the 869 Italian Brown specific subset correlations were always lower than with random subsets.

Table 2. Correlation between EBVs and DGVs estimated by different marker subsets in Italian Brown cattle.

Marker subset (N markers)	Milk yield	Prot. %	Udder score	ITE
Sel. signature all (8944)	0.131	0.423	0.270	0.511
Random (mean) (8944)	0.127	0.407	0.256	0.579
Sel. signature Italian Brown specific (869)	0.010	0.105	0.017	0.245
Random (mean) (869)	0.160	0.219	0.291	0.294

4. Discussion

In this paper we have scanned the genome of the Italian Brown dairy cattle searching for signatures of selection. Among possible indexes, we used F_{st} to study selection because it is robust, easy to calculate and widely used for this purpose (e.g. Barendse *et al.,* 2009). Single marker F_{st} values varied substantially even among SNPs very close to each other and had standard deviations even higher than the means. Therefore, we adopted a sliding windows approach to avoid excessive noisiness (Weir *et al.,* 2005). We decided to include in sliding windows an homogeneous number of markers rather than using a predetermined genome size. This to avoid having windows including only one or a few markers. The use of 9 markers was a subjective choice but also facilitates the comparison wih published data using the same or similar sliding window size (e.g. Stella *et al.,* 2010). On average the 43771 windows spanned genomic regions of 500Kb and among these 1195 regions larger than 1Mb and 118 larger than 2Mb, providing a rather detailed survey of the cattle genome.

F_{st} values were calculated in pairwise comparisons in which the dairy Italian Brown was contrasted with the dairy Italian Friesian, the dual purpose Italian Simmental and the beef Piedmontese and Marchigiana cattle breeds. The highest average F_{st} across markers was found in the BRW vs FRI. This is likely the result of the combined effect of different origin, reduced gene flow and small effective population size of Italian Friesian compared to the beef and dual purpose breeds investigated. However, ascertainment bias is possibly contributing to this divergence, given that a relevant number of SNPs included in the array have been developed to be highly informative in the Holstein population. A permutation approach permitted the identification of significant selection signatures in each pairwise comparison. In total 8944 were under selection in at least one of the comparisons involving Italian Brown (Table 1). Contrary to expectation, the highest number of signatures were found in the contrast between the two dairy breeds, rather than between Italian Brown and the beef breeds. In pairwise comparisons signatures are due to selection in either breed or to divergent selection in both breeds. Markers having consistent outlier behaviour in multiple comparisons involving the same breed are likely to be under selection in that same breed. Using this rationale, we have isolated 869 markers specific to selection in BRW. Markers under selection might include those associated to traits included in selection indexes and hence be informative for genomic prediction of genetic merit. However DGV of three production traits estimated in young BRW bulls based on markers under selection were no better and often worse than those calculated from an equal number of random markers (Table 2). With the current approach only the historical and strongest effects of selection could be detected, probably on genes close to fixation and having either a qualitative or a major effect on traits that have been selected since Italian Brown breed formation. Therefore, most selection signatures likely correspond to genomic regions subjected to historical selection that do not match with those in which genes controlling economic traits are segregating in modern populations. The selection signature approach is useful in the reconstruction of the interesting process of breed formation, but seems to have little application in the choice of marker subsets that can be profitably used in genomic selection.

5. References

Aulchenko, Y.S., Ripke, S., Isaacs, A. & van Duijn, C.M. 2007. GenABEL: an R package for genome-wide association analysis. *Bioinformatics 23,* 1294-1296.

Barendse, W., Harrison, B.E., Bunch, R.J., Thomas, M.B. & Turner, L.B. 2009. Genome wide signatures of positive selection: the comparison of independent samples and the identification of regions associated to traits. *BMC Genomics 10:*178.

Bovine HapMap Consortium *et al.* 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science 324,* 528-532.

Ennis, S. 2007. Linkage disequilibrium as a tool for detecting signatures of natural selection. *Methods in Molecular Biology 376,* 59-70.

Gulcher, J. & Stefansson, K. 1998. Population genomics: laying the groundwork for genetic disease modeling and targeting. *Clin. Chem. Lab. Med. 36,* 523-527.

Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics 157,* 1819-1829.

Stella, A., Ajmone-Marsan, P., Lazzari, B. & Boettcher, P. 2010. Identification of Selection Signatures in Cattle Breeds Selected for Dairy Production. *Genetics* (in press).

Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. 2006. A map of recent positive selection in the human genome. *PLoS Biology* 4:e72.

Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M. & Hill, W.G. 2005. Measures of human population structure show heterogeneity among genomic regions. *Genome Research 15,* 1468-1476.