Application of SNP reduction approaches and random forest for the identification of population informative markers in cosmopolitan and local cattle breeds

Journal:	*Italian Journal of Animal Science*
Manuscript ID	TJAS-2017-0288
Manuscript Type:	Abstract Submission
Date Submitted by the Author:	27-Feb-2017
Complete List of Authors:	Bertolini, Francesca; University of Bologna, Department of Agricultural and Food Sciences Galimberti, Giuliano; University of Bologna MASTRANGELO, Salvatore; Universita degli Studi di Palermo, Scienze Agrarie e Forestali Di Gerlando, Rosalia; Universita degli Studi di Palermo, Scienze Agrarie e Forestali Bagnato, Alessandro; Universita degli Studi di Milano, Strillacci, Maria; University of Milano Portolano, Baldassare; University of Palermo Fontanesi, Luca; University of Bologna, Dept. of Agricultural and Food Sciences - Division of Animal Sciences
Abstract:	In livestock, single nucleotide polymorphism genotyping arrays have been used to differentiate breeds and populations for several downstream applications, including breed allocation of individuals, breeds of origin of crossbred animals, authentication of mono breed products, comparative analyses of selection signatures among several other uses. We already tested a combination of principal component analysis (PCA), used as pre-selection method, and random forest (RF) used as classification method to

assign cosmopolitan Italian breeds with no or very low error rate. In this work, we increased the number of breeds and approaches, to have a more comprehensive view of the strategies available and the applicability to local Italian breeds. The most common cosmopolitan dairy or dual purpose breeds (Holstein, Brown, Simmental) and 3 local breeds subjected to limited or no breeding programs (Reggiana, Modicana and Cinisara) were analyzed comparing several methods of SNPs pre-selection (Delta, Fst and PCA) in addition to RF classifications. From these classifications, two panels of 96 and 48 SNPs that contained the most discriminant SNPs were created for each pre-selection method. The results showed that the 96-SNP panels were generally more able to discriminate all breeds, while for the 48- SNP panels the error rate increased mainly for autochthonous breeds, particularly for Cinisara. This was probably a consequence of limited selection pressure, admixed origin, and ascertain bias on the construction of the SNP chip that was not designed considering these breeds. Several selected SNPs are located nearby genes affecting breed-specific traits (e.g. coat color and stature) or associated to production traits. The 96-SNP panel obtained after a preselection chromosome by chromosome, and used in the previous work with cosmopolitan breeds only, could identify informative SNPs that were particularly useful for the assignment of minor breeds. This panel reached the lowest value of out of bag (OOB) error in the RF test 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. Our results showed the usefulness and power of the combination of PCA pre-selection and RF also for the discrimination of local cattle breeds.

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

Application of SNP reduction approaches and random forest for the identification of population informative markers in cosmopolitan and local cattle breeds

Francesca Bertolini[1,2], Giuliano Galimberti[3], Salvatore Mastrangelo[4], Rosalia Di Gerlando[4], Maria Giuseppina Strillacci[5], Alessandro Bagnato[5], Baldassare Portolano[4], Luca Fontanesi[1]

[1]*Dipartimento di Scienze e Tecnologie Agroalimentari, Università di Bologna, Bologna, Italy*
[2]*Department of Animal Science, Iowa State University, Ames, Iowa, USA*
[3]*Dipartimento di Scienze Statistiche "Paolo Fortunati", Università di Bologna, Italy*
[4]*Dipartimento Scienze Agrarie e Forestali, Università di Palermo, Palermo, Italy*
[5]*Dipartimento di Medicina Veterinaria, Università degli Studi di Milano, Milano, Italy*
Corresponding author: luca.fontanesi@unibo.it

In livestock, single nucleotide polymorphism genotyping arrays have been used to differentiate breeds and populations for several downstream applications, including breed allocation of individuals, breeds of origin of crossbred animals, authentication of mono breed products, comparative analyses of selection signatures among several other uses. We already tested a combination of principal component analysis (PCA), used as pre-selection method, and random forest (RF) used as classification method to assign cosmopolitan Italian breeds with no or very low error rate. In this work, we increased the number of breeds and approaches, to have a more comprehensive view of the strategies available and the applicability to local Italian breeds. The most common cosmopolitan dairy or dual purpose breeds (Holstein, Brown, Simmental) and 3 local breeds subjected to limited or no breeding programs (Reggiana, Modicana and Cinisara) were analyzed comparing several methods of SNPs pre-selection (Delta, F_{st} and PCA) in addition to RF classifications. From these classifications, two panels of 96 and 48 SNPs that contained the most discriminant SNPs were created for each pre-selection method. The results showed that the 96-SNP panels were generally more able to discriminate all breeds, while for the 48- SNP panels the error rate increased mainly for autochthonous breeds, particularly for Cinisara. This was probably a consequence of limited selection pressure, admixed origin, and ascertain bias on the construction of the SNP chip that was not designed considering these breeds. Several selected SNPs are located nearby genes affecting breed-specific traits (e.g. coat color and stature) or associated to production traits. The 96-SNP panel obtained after a preselection chromosome by chromosome, and used in the previous work with cosmopolitan breeds only, could identify informative SNPs that were particularly useful for the assignment of minor breeds. This panel reached the lowest value of out of bag (OOB) error in the RF test 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. Our results showed the usefulness and power of the combination of PCA pre-selection and RF also for the discrimination of local cattle breeds.

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