Effect Of Non-Random Sampling Of SNPs On The Estimation Of Population Genetic Parameters

R. Negrini[*], R. Mazza*, Colli L.*, Pellecchia M.*, Bomba L.*, Stella A. [†], Williams JL.[†], Ajmone-Marsan P.*

Introduction

As a result of whole genome sequencing and HapMap projects, millions of SNPs have recently been discovered in several livestock species. From these, panels including tens of thousands of validated SNPs are already available to the scientific community (e.g in cattle, sheep and pig) or will likely be available in the near future (e.g. in chicken, goat, horse), permitting genome wide scans at a very low cost per data point (120-200 Euro for 50-60,000 markers).
The International Bovine HapMap Consortium has recently validated 35K SNPs on 497 animals belonging to 19 breeds (14 taurine, 3 indicine, 2 taurus x indicus crosses) one Anoa (*Bubalus quarlesi*) and one Water Buffalo (*Bubalus bubalis*) (The Bovine HapMap Consortium 2009). SNPs were discovered by re-sequencing random shotgun libraries from Holstein, Angus, Brahman, Limousin, Jersey, and Norwegian Red breeds and comparing sequence reads to release Btau20040927 of the bovine genome (a Hereford) (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp/Btau20040927/).
This, as well as larger SNP panels, will undoubtedly be very useful in association studies and in assisted breeding in industrialised breeds. However, their general utility in the investigation of within and between breed diversity worldwide is to be verified, since they risk to be subject to different degrees of ascertainment bias depending on the relationship between genotypes to be investigated and those used in SNP discovery. Biased SNP panels may produce distorted results when population genetic tools are applied. For example, Kreitman and Di Rienzo (2004) and Soldevila et al. (2005), analyzing the Human SNP HapMap panels, demonstrated that the apparent effect of balancing selection detected in the prion protein gene (PRPN) by Mead et al. (2003) were in fact an artefact caused by ascertainment bias. Moreover patterns of linkage disequilibrium (Nielsen and Signorovitch 2003), and level of population subdivision (Nielsen 2004) can be also severely affected.
Objective of this paper is to investigate the bias on a population genetics parameter caused by the use of SNP subsets discovered in different breeds.

Material and methods

Three SNP subset were extracted from the HapMap 35K full set (The bovine HapMap Consortium, 2009): a 14K panel of SNPs discovered in the taurine Holstein, a 6.5K discovered in the indicine Brahman and a 0.4K discovered in the taurine Limousine breeds.
Allele frequency estimates and observed heterozygosity were calculated on a total of 497

[*] Istituto di Zootecnica. Università Cattolica del Sacro Cuore, 29122 Piacenza, Italy
[†] Parco Tecnologico Padano, Polo Universitario, Via Einstein, 26900 Lodi, Italy

animals from 19 breeds (14 taurine, 3 indicine and 2 taurine x indicine crossbred) by direct counting.

Results and discussion

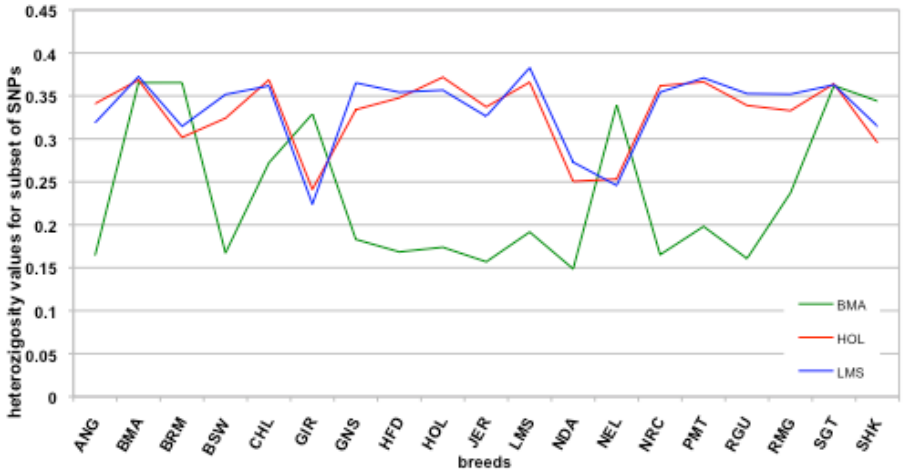

Figure 1 Observed heterozygosity calculated in 19 breeds using three different subsets of SNPs. ANG=Angus (taurus); BMA=Beefmaster (hybrid); BRM=Brahman (indicus); BSW=Brown Swiss (taurus); CHL=Charolais (taurus); GIR=Gir (indicus); GNS=Guernsey (taurus); HFD=Hereford (taurus); HOL=Holstein (taurus); JER=Jersey (taurus); LMS=Limousin (taurus); NDA=N'Dama (taurus); NEL=Nelore (indicus); NRC=Norwegian Red (taurus); PMT=Piedmontese (taurus); RGU=Red Angus (taurus); RMG=Romagnola (taurus); SGT=Santa Gertrudis (hybrid); SHK=Sheko (taurus)

In figure 1 is reported the plot of heterozygosity observed in the HapMap 19 breed set using three subsets of SNP discovered in Holstein, Brahman and Limousine. The 14K Holstein markers indicate Holstein as the most heterozygous of the 19 breeds and indicates the three indicine and two crossbred breeds as the less variable. A similar pattern is observed using Limousine markers, in this case Limousine appearing the most heterozygous breed. Holstein and Limousine panels also detect different levels of heterozygosities in other breeds, as Angus and Romagnola, however the significance of this difference remains to be verified. When markers discovered in Brahman (*Bos indicus*) are used, Brahman is the most heterozygous breeds and all taurine breeds appear to have a substantially lower diversity compared to indicine or mixed breeds. The observed pattern allows to make inferences on the ascertainment bias scheme: because rare SNPs are missing due to the usual low number of individuals in the discovery panel, the average heterozygosity of the sites that are polymorphic, and of the breed in which they have been discovered is inflated. Conversely,

the average heterozygosity across all sites and breeds not comprised in the discovery panel is underestimated because either some loci are monomorphic and private SNPs of those breed are lost. Such ascertainment bias scheme can be minimized at the beginning of the discovery process by including a larger number of breeds or at least a good representation of the most variable breeds in the SNP discovery panel. In some cases the bias can also be corrected afterwards by applying proper mathematical algorithms (mainly based on Maximum Likelihood) that, for example, adjust the frequency spectrum in the whole dataset using the allele frequencies in the ascertainment samples.

Conclusion

Our results indicate that existing SNP panels, developed in breeds unrelated or poorly related to those used in extensive agriculture, are biased tools to investigate population structure. As a result, they may not be fully informative to detect selective sweeps along the genome in many breeds. In this sense whole genome sequencing will represent a significant progress in comparison to existing marker sets, since it is not affected by subjective and biased sample pre-selection choice, overcoming the limitations of microsatellite and SNP markers and providing information on neutral and selected polymorphisms: useful attributes for decision making in FAnGR conservation.
SNP ascertainment bias can be minimised by including a large number of breeds or at least a good representation of the most variable breeds in the SNP discovery panel, and should typically include those subjected to low artificial selection pressure reared in different agro-climatic areas and nearby domestication centres. Hence, additional sequencing effort is needed to include a proper representation of diversity in SNP panels.

References

Kreitman, M., and Di Rienzo, A. (2004). *Trends Genet*. 20: 300-4.

Soldevila, M., Calafell, F., Helgason, A., Stefansson, K., and Bertranpetit, J. (2005) *Trends Genet.* 21: 389-391

Mead, S., Stumpf, M. P., Whitfield, J., Beck, J. A., Poulter, M. et al. (2003). *Science* 300: 640-643.

Nielsen, R., Signorovitch, J. (2003). *Theor. Popul. Biol*. 63: 245-55.

Nielsen, R., (2004) *Hum. Genomics* 1: 218-224.