Tracing The Origin Of Cattle Breeds With PCA-based Ancestry Informative SNPs

J. Lewis[*], *Z. Abas*[†], C. Dadousis[†], D. Lykidis[§], P. Paschou[§] and P. Drineas[*]

Introduction

The recent release of data by the Bovine HapMap Consortium (2009) represents the most detailed survey of bovine genetic diversity to date. Our goal was to first attempt to accurately classify individual cattle from the Bovine HapMap dataset by ancestral population using all available genotype data (30,000 SNPs), and then to further examine our ability to perform such classification using only small panels of Ancestry Informative Markers (AIMs) that can reproduce the same results. We used Principal Components Analysis (PCA), which has emerged as a powerful tool for the characterization and analysis of the structure of genome-wide datasets. In prior work, we described an algorithm that can be used to select small subsets of genetic markers (SNPs) that correlate well with population structure, as captured by PCA (PCA Informative Markers – PCAIMs) (Paschou et al. (2007); Paschou et al. (2008)). Our method can be used to detect SNPs that differentiate individuals from different populations, without any prior hypotheses.

Here, we chose to split the main task of classifying individuals by ancestry into hierarchical levels, splitting the entire cattle population into nested groups, organized into a decision tree. Groups were chosen by visually determining which populations are most easily differentiated along the higher principal components and recursively looking at the principal components of each subgroup in the same way. Applying our SNP selection algorithms (Paschou et al. 2007; Paschou et al. 2008), we chose small subsets of SNPs that almost perfectly reproduce population structure as identified by PCA and can be used to accurately assign individuals to one of 19 breeds.

Material and methods

Dataset. We analyzed the Bovine HapMap dataset (Bovine HapMap Consortium (2009)). Of the 501 individual cattle and 34,884 genotyped SNPs, 497 cattle and 30,501 SNPs were used in our analysis (13 taurine, three indicine and three admixed breeds). We removed from our analysis the cattle populations Anoa and Water Buffalo (comprising 4 cattle in total) as well as all SNPs with more than 10% missing entries (approximately 4000 SNPs).

Selecting PCA-Informative Markers (PCAIMs). We viewed individual genotypes as a matrix of *m* individuals genotyped for *n* SNPs in order to run PCA. We then computed PCA

[*] Dept. of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[†] Dept. of Agricultural Development, Democritus University of Thrace, Orestiada 68200, Greece
[§] Dept. of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupoli 68100, Greece

scores for each SNP using the algorithm of Paschou et al. (2007), and we selected the SNPs with the highest scores (PCAIMs).

Removing redundancy. The computation of the above scores *does not* take into account the high LD between SNPs, thus resulting in the selection of redundant SNPs. In order to remove redundant SNPs, we employ a simple, clustering-based algorithm that first clusters the selected PCAIMs and then keeps only one representative PCAIM from each cluster. This algorithm is similar to the one described in Paschou et al. (2008) (they both solve the Column Subset Selection Problem, which corresponds to a theoretical formulation of the redundancy removal problem), but seems to perform better in this dataset.

Five Nearest Neighbors (5-NN) classification algorithm. In order to assign an individual to a population, we used a simple 5-NN algorithm. The number of significant principal components used in the neighbor distance calculations was estimated experimentally, running the pre-validation experiment on multiple values and choosing the one that achieved the best results with the fewest principal components. Given a target individual, we identify its five nearest neighbors. If at least three of the five nearest neighbors (a majority) belong to the same population, we assign the target individual to that population.

Results and discussion

At the highest level in the decision tree (figure 1), individual cattle are broadly classified into one of three groups: bos taurus, bos indicus, or admixed breeds. Moving down the decision tree, individuals are more specifically classified into sub-groups, proceeding until they are finally assigned to the deepest membership that can be inferred using the given dataset. The number of classification levels depends on the complexity of the population structure within the initial group, and successive subgroups.

We ran a complete leave-one-out experiment using all 30,000 available markers in order to infer ancestry using all SNPs. Classification was performed by looking at the distances to neighbors of each individual in the space spanned by the significant top principal components of the genotype data. As demonstrated in figure 2, for most nodes in the decision tree the classification accuracy was over 98% using the whole 30k dataset.

In our next experiment we attempted to evaluate whether there exist small panels of AIMs that could accurately reproduce the excellent results of ancestry inference using all 30k available markers. Within each level, we selected the top 2,000 PCA-Informative markers, using the number of significant PCs of figure 1 and repeated the full leave-one out validation test at each level. Indeed, the light blue bars in figure 2 indicate that these panels are roughly as accurate as the full 30k panels. Our next step was the removal of redundant markers via the clustering technique described in Methods. We experimented with numerous panel sizes and we report here our complete leave-one-out cross-validation results on three different panels of 10, 25, and 50 SNPs (P1, P2, and P3) for each node in our decision tree. Even our smallest panels of AIMs (P1) achieve very high accuracy at most nodes of our decision tree.

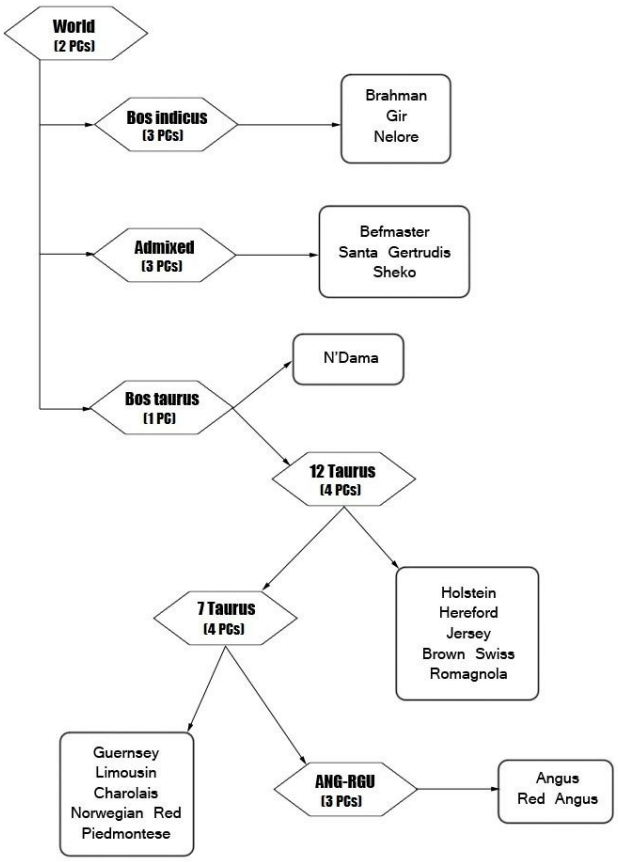

Figure 1: The decision tree for individual assignment to a particular population (or population group) of origin using the Bovine HapMap data. We also report the number of principal components that we chose in order to separate populations at each level. For example, in order to classify an individual as RGU, we first determine whether the individual is part of the Bos Taurus group. We then decide whether the individual belongs to the NDA population, or to the 12 Taurus group. We then proceed to differentiate between the HOL, HFD, JER, BSW and RMG populations and the 7 Taurus group. The 7 Taurus level of the hierarchy allows us to differentiate further between the GNS, LMS, CHL, NRC, and PMT populations, and the RGU-ANG group. Finally we distinguish between RGU and ANG.

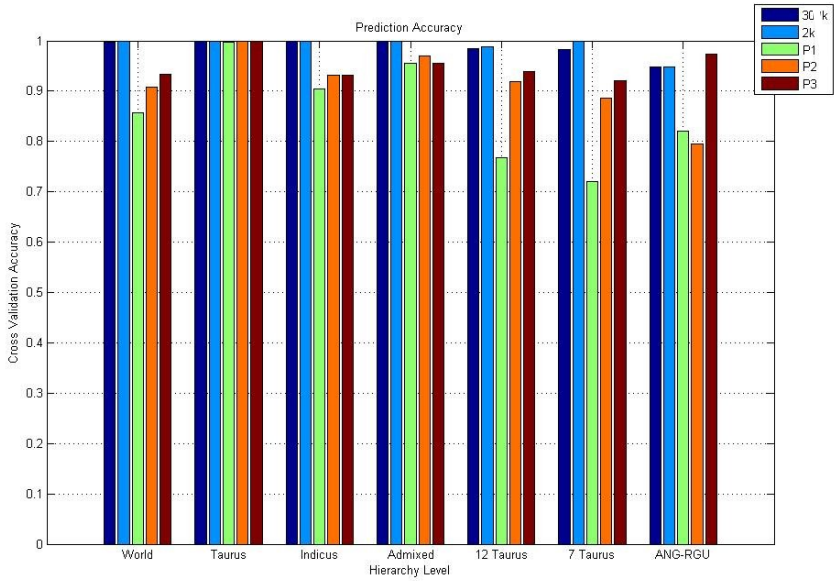

Figure 2: Classification accuracy of our complete leave-one-out validation experiment at all nodes of our decision tree using all available SNPs as well as PCA Informative Markers (PCAIMs). Five different panel sizes are evaluated, with 30k corresponding to the whole Bovine HapMap dataset, 2k corresponding to the top 2,000 AIMs, and P1, P2, and P3 corresponding to 10, 25, and 50 SNPs. These smaller panels emerged by removing redundant markers from the top 2,000 AIMs.

Conclusion

Our results clearly demonstrate that it is indeed feasible to accurately assign individual cattle to breed of origin, using in most cases less than 100 carefully selected SNPs. The method that we introduce requires no modeling or prior assumptions about the data and has the potential to become an important tool for the traceability of cattle products.

References

Paschou, P., Drineas P., Lewis J. *et al.* (2008). *PLoS Genet.*, 4: e1000114.

Paschou, P., Ziv, E., Burchard E.G. *et al.* (2007). *PLoS Genet.*, 3: 1672-1686.

The Bovine HapMap Consortium (2009). *Science*, 324:528–532.