A NOVEL NUMERICAL METHOD TO QUANTIFY THE CONTRIBUTION OF GENES TO THE POPULATION STRUCTURE

P. Kasarapu[1], L. R. Porto-Neto[1], M. R. S. Fortes[2], S. A. Lehnert[1], M. A. Mudadu[3], L. Coutinho[4], L. Regitano[5], A. George[6] and A. Reverter[1]

[1] CSIRO Agriculture and Food, Queensland Bioscience Precinct, Brisbane, Queensland, Australia
[2] School of Chemistry and Molecular Biosciences, The University of Queensland, Australia
[3] Embrapa Agricultural Informatics, Av. Andre Tosello, 209 Campinas, Sao Paulo, Brazil
[4] University of Sao Paulo, Brazil
[5] Embrapa Southeast Livestock, Rodovia Washington Luiz, Km 234, Sao Carlos, Brazil
[6] CSIRO, DATA61, Ecosciences Precinct Brisbane, Queensland, Australia

SUMMARY

Principal component analysis (PCA) using genome-wide single nucleotide polymorphism (SNP) genotype data is traditionally used to determine distinct groups in a population. We present a novel numerical approach to quantify the importance of each gene to the emerging clusters as informed by PCA. Our method is based on modelling the coefficients (SNP weights) of the first principal component using mixtures of Normal distributions. We applied our approach to three distinct datasets of cattle, chicken, and sheep. We were able to identify subsets of genes in the cattle and chicken genomes that are likely to be important determinants for understanding the phenotypic differences among various disparate livestock populations.

INTRODUCTION

The utility of PCA in discriminating individuals according to breed differences has been well documented in the literature. The Bovine HapMap Consortium used PCA as a central method in the elucidation of genetic structure across biologically diverse breeds (Gibbs *et al.* 2009). Other studies relied on PCA to measure the genetic divergence between indicine and taurine cattle (Bertolini *et al.* 2015). PCA also informs machine learning based classification methods to predict the individual ancestry of cattle (Bertolini *et al.* 2015).

In our study, we used PCA as a starting point to identify a set of genes that have the discriminatory power to identify the lineage of a particular population. We build a model based on the contributions of the SNP to the first principal component (PC1). These are the coefficients of the PC1 which are produced as part of the PCA. The empirical distribution of the SNP reveals distinct modes. We used the output from PCA as a first step to project the data on to the maximum variable direction and used statistical machine learning based mixture modelling to quantify the contribution of genes to the respective lineages.

MATERIALS AND METHODS

Animals and genotypes. We tested our method on three datasets - cattle, chicken, and sheep.

Cattle: We used data from 18,363 animals and 19 breeds belonging to the Beef CRC (http://www.beefcrc.com) and Nelore data from Mudadu *et al.* 2016. The cattle belong to a spectrum of lineages ranging from pure *Bos indicus* (BI; N=5,536 cattle) to pure *Bos taurus* (BT; N=7,589). Additionally, we have 5,238 cattle that are crossbred or tropically adapted composites which are classified as *Bos taurus* – *Bos indicus* (BTI) breeds. The original data had genotypes for 729,068 SNP. We considered SNP located in autosomal chromosomes and mapped within 1Kb of a known gene to capture SNP associated with protein-coding regions. We further pre-processed the data so that we retained those genes that have at least the median number of 6 SNP to ensure that the genes are minimally represented. The final dataset contained 246,864 SNP in 8,631 genes.

Chicken: The data were from 988 chickens from 4 commercial lines of broilers (Hudson *et al.* 2015), denoted as Lines A (N = 204), B (N = 244), C (N = 254), and D (N = 286). Lines A and B have been generated to select females, whereas lines C and D are to select males. The data had genotypes for 51,713 SNP. After removing monomorphic SNP and retaining those within 20 Kb of a gene (Reyer *et al.* 2015), we considered 36,395 SNP located in 12,642 genes.

Sheep: We used data from the Sheep Hapmap project (http://www.sheephapmap.org/) including 1,222 animals distributed across 9 regions and genotypes for 49,034 SNP. We considered SNP that were not monomorphic and those within 30 Kb of a known protein coding gene (Miller *et al.* 2011), which resulted in 26,077 SNP spanning 12,737 genes.

Principal Component Analysis and Gene contribution to lineage. We used PLINK (Chang *et al.* 2015) to perform the principal component analysis (PCA) and considered only the PC1 as it explains the maximum variability in the data and extracted the weights of each SNP to that component. We used mixture modelling to quantify the contribution of the genes to the lineages.

Mixture modelling is a statistical method to construct a probability distribution by combining the effects due to several component probability distributions. We considered the Normal component distributions to model the probability of the SNP weights in PC1, which were best modelled using two component distributions. Formally, a two-component mixture is defined as

$$\Pr(x) = \underbrace{w\ \mathcal{N}(x;\mu_1,\sigma_1)}_{p_1} + \underbrace{(1-w)\ \mathcal{N}(x;\mu_2,\sigma_2)}_{p_2}$$

where x corresponds to the data (SNP weights in PC1), $\Pr(x)$ is the probability distribution of the mixture, w is the weight of the first component in the mixture, μ_1,μ_2 and σ_1,σ_2 denote the means and the standard deviations of the two Normal ($\mathcal{N}$) components, respectively. As part of statistical inference, the mixture parameters, that is, $w,\mu_1,\mu_2,\sigma_1,\sigma_2$ were estimated using the EMMIX software (McLachlan *et al.* 1999). After estimating the mixture parameters, the contribution of each SNP to each components is given by its *posterior probability*, that is,

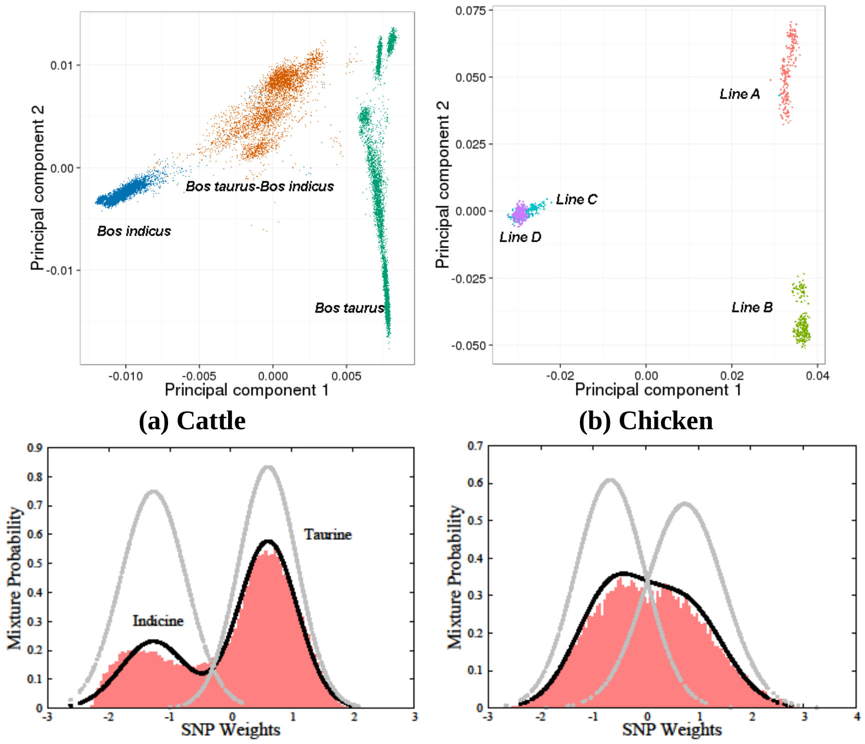
$$m_1 = \frac{p_1}{p_1+p_2} \text{ and } m_2 = \frac{p_2}{p_1+p_2}$$

where m_1 and m_2 are the posterior probabilities of the given SNP to belong to the first and second component, respectively. The values p_1 and p_2 constitute the two parts of Pr(x). Note that $m_1 + m_2 = 1$ which implies that for a given gene, m_1 and m_2 correspond to the contributions (memberships) of that gene to the two components of the mixture. As an example, to estimate a gene's contribution to the indicine content in bovine genome, we average the posterior probabilities (m_1 values) of the corresponding SNP in its coding region. A gene contributes to both the indicine and taurine components of the bovine genome. The value m_1 denotes the amount of contribution (as a percentage) to the indicine lineage. We infer that the left mode corresponds to *Bos indicus* because the animals with negative SNP weights are Nelore/Brahman cattle.

RESULTS AND DISCUSSION

The PCA of the cattle and chicken datasets reveals distinct clusters based on their respective lineages; the *Bos indicus, Bos taurus* and *Bos taurus – Bos indicus* breeds are separately clustered (Figure 1a). The PC1 and PC2 explain 21.8% and 2.3% variation in the data respectively. Similarly, for the chicken dataset, we observe Lines A and B distinctly clustered whereas Lines C and D are overlapping (Figure 1b). Mixture modelling of the SNP weights along PC1, (Figure 1c,d) resulted in distinct modes corresponding to the indicine and taurine components of the bovine genome. The estimate of the mixing proportion is $w = 0.31$ establishing an effective membership of 31% *Bos indicus* and 69% *Bos taurus* genes for this particular population. For the chicken data, PC1 and PC2 explain 22% and 3.6% variation in the data, respectively. The value $w = 0.49$ implies an almost equal number of genes contributing to male and female lines.

Further, 64 and 718 genes have a contribution of at least 95% to the indicine and taurine components, respectively. In the chicken genome, there are 1,072 and 1,386 genes with least 95% contribution. The study of these candidate genes can aid our understanding of ancestry-related differences in gene expression and susceptibility of a given lineage to exhibit a certain phenotype.



(a) Cattle **(b) Chicken**

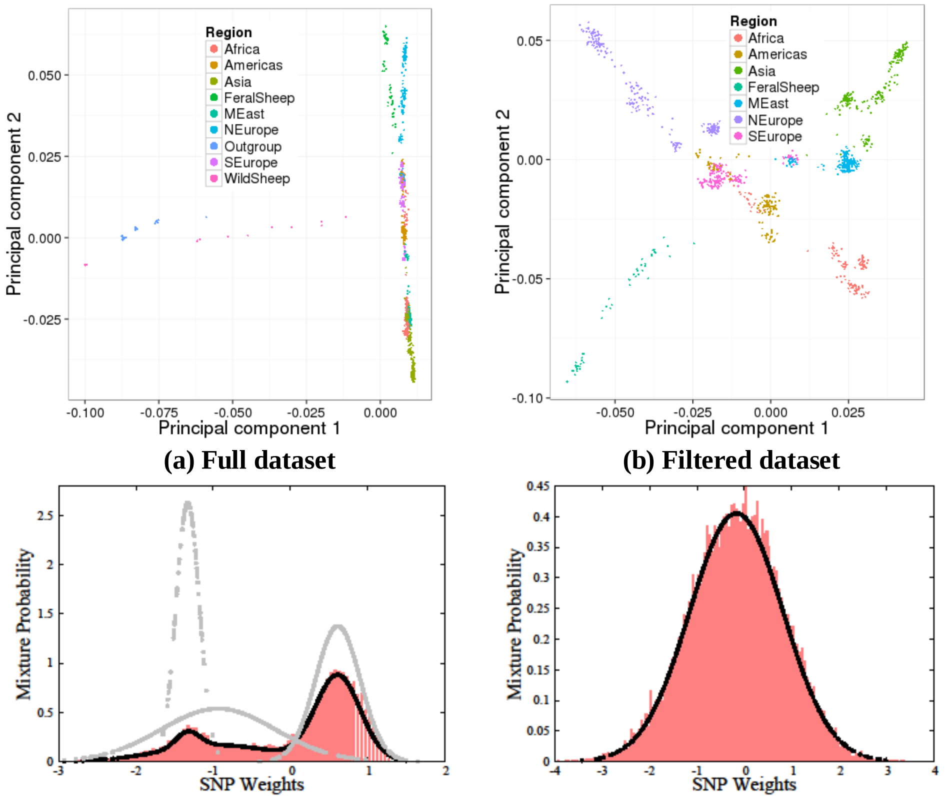
(c) Cattle **(d) Chicken**

Figure 1. (a)-(b) PCA of SNP genotypes resulting in distinct clusters of animals based on their lineages. (c)-(d) Mixture modelling of SNP weights along PC1. Red indicates the actual distribution of SNP weights, grey curves are the individual Normal distributions, and black curve is the mixture model obtained by combining the two Normal distributions based on the mixing proportions.

The PCA of the sheep data revealed a cluster with sheep (Outgroup, Wildsheep) widely scattered and having negative PC1 values (Figure 2a). On removing these outliers, we note a star-shaped cluster (Figure 2b) with PC1 and PC2 accounting for 4.5% and 2.2% variation, respectively. Mixture modelling shows three distinct modes for the full data (Figure 4c), whereas there is a clear unimodal distribution for the filtered data (Figure 4d). This finding highlights the importance of pre-processing the data prior to our analysis. Kijas *et al.* (2012) suggests the absence of distinct lineages and strong historic mixing, in agreement with our observation of a unimodal distribution.

CONCLUSIONS

Our method based on the mixture modelling of SNP weights captures the genes responsible for the underlying population structure and potentially serves to establish a relationship between the evolutionary structure and phenotypic variation in livestock populations.



(a) Full dataset **(b) Filtered dataset**

(c) Full dataset **(d) Filtered dataset**

Figure 3. PCA and mixture modelling of SNP weights for the Sheep Hapmap data. The full and filtered datasets consist of 1,222 and 1,105 animals respectively.

ACKNOWLEDGEMENT

We acknowledge the support of Cobb-Vantress Inc. in allowing us to use their chicken data.

REFERENCES

Bertolini F., Galimberti G., Calo D.G., Schiavo G., Matassino D., *et al*. (2015) *Journal of Animal Breeding and Genetics*, **132**: 346.

Chang C.C., Chow C.C., Tellier L.C., Vattikuti S., Purcell S.M., *et al*. (2015) *GigaScience*, **4**.

Gibbs R.A., Taylor J.F., Tassell C.P.V., Barendse W., Eversoie K.A., *et al.* (2009) *Science*, **324**: 528.

Hudson N. J., Hawken R., Sapp R., Reverter A. (2015) *Proc. AAABG*, **21**: 153.

Kijas J.W., Lenstra J.A., Hayes B., other members of International Sheep Genomics Consortium (2012) *PLOS Biology*, **10**: e1001258.

McLachlan G.J., Peel D., Basford K.E., Adams P. (1999) *Journal of Statistical Software*, **4**: 1.

Miller J.M., Poissant J., Kijas J.W., Coltman D.W., International Sheep Genomics Consortium (2011) *Molecular Ecology Resources*, **11**: 314.

Mudadu M.A., Porto-Neto L.R., Mokry F.B, Tizioto P.C., Oliveira P.S.N, *et al.* (2016) *BMC Genomics*, **17**: 235.

Reyer H., Hawken R., Murani E., Ponsuksili S., Wimmers K. (2015) *Scientific Reports*, **5**: 16387.