Proceedings of the World Congress on Genetics Applied to Livestock Production, 11.585

Genetic composition, divergence, admixture and use of ancestry informative markers in indigenous breeds of cattle in India

S.G. Gajjar[1], B. Guldbrandtsen[2], N.G. Nayee[1], A. Sudhakar[1], K.R. Trivedi[1], M.S. Lund[2] & G. Sahana[2]

[1]National Dairy Development Board, Anand-388001, Gujarat, India
[2]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark
goutam.sahana@mbg.au.dk (Corresponding Author)

Summary

A total of 14 *Bos indicus* cattle breeds of India (indigenous cattle) were studied to investigate their genetic composition, divergence and admixture. Descriptive statistics on genotype data revealed that Amritmahal, Tharparkar and Kangayam had relatively low expected heterozygosity than other indigenous breeds. Geography and F_{ST} were tightly connected. Principal component analysis did not reveal any major substructures within the indigenous breeds; however, more principal components than total number of breeds (30 components were highly significant) were required to explain considerable variability amongst breeds studied. All indigenous breeds except Hariana, Rathi, Khillar and Hallikar had distinct origins in our Admixture analysis. Stable proportions of four populations each in Khillar and Hallikar suggested a distant origin by admixture for these breeds, while Hariana and Rathi showed signs of recent admixture. Patterns of divergence of the selected indigenous breeds of cattle of India have been presented in the paper and compared with other *Bos indicus* and *Bos taurus* breeds in other parts of the world by TreeMix analysis. No evidence for migration to or from indigenous breeds of India to other cattle breeds was found. A set of 500 ancestry informative markers for Gir, Sahiwal, Kankrej, Red Sindhi, Holstein Friesian and Jersey breeds were identified and found suitable for tracing ancestries of these purebred as well as their crossbreds. Implications of these studies have been described with emphasis on efficient implementation of breed improvement programs.

Keywords: indicus cattle*, population structure, admixture, ancestry informative marker*

Introduction

India has rich cattle genetic diversity consisting of 41 recognized cattle breeds (NBAGR, 2017) as well as large proportion of non-descript cattle. Besides dairy breeds, there are draft and dual purpose breeds with distinct roles in rural livelihoods. Additionally, *Bos taurus* cattle, mostly Holstein Friesian (HF) and Jersey and their crosses with indigenous cattle (primarily Sahiwal, Gir, Red Sindhi and Kankrej) also contribute significantly to total milk production in India. Several breed improvement programs are in operation for indigenous as well as crossbred cattle, mostly limited to selection of bulls based on dam's and/or daughters' phenotypic information. In the field, some indigenous breeds are often admixed due to rearing patterns of rural livestock owners. The crossbreds are expected to comprise varying proportions of taurine inheritance, due to scarce pedigree records since inception of crossbreeding programs. The current study was undertaken to understand the genetic composition of 14 indigenous cattle breeds of India, their ancestry patterns and admixture. The study also aimed at identifying ancestry informative

markers for 4 major indigenous dairy cattle breeds, along with HF and Jersey, to ease identification of breed purity in purebreds and ancestry proportions in their crossbreds.

Materials and Methods

Blood/semen samples from 829 animals from 14 indigenous breeds were collected, avoiding any known close relatives. Bulls with semen samples at semen stations were preferentially included. In addition, 777K HD chip (Illumina Inc., San Diego, CA) genotype data on 20 each from Danish Jersey and Nordic Holstein breeds provided by Aarhus University, Denmark, and 777K HD chip genotype data on 2 *Bos gaurus* samples and 50K chip (Illumin Inc., San Diego, CA) genotype data on 29 cattle breeds - both maintained by WIDDE (http://widde.toulouse.inra.fr/widde/) were included for the study at various stages of analysis.

Data were filtered (average call rate < 90% for individual and for locus, MAF < 0.01) using PLINK (Purcell *et al*., 2007). Pairwise within-breed relationships were estimated using GCTA (Yang *et al*., 2011). Principal component analysis (PCA) was done using SMARTPCA, part of EIGENSOFT 7.2.1 (Patterson *et al*., 2006). Tracy-Widom statistics (Tracy & Widom, 2015) were calculated using TWSTATS, part of EIGENSOFT 7.2.1, to identify the number of principal components significantly explaining variability amongst the breeds included in PCA. Admixture analysis was done after processing the final data used for PCA using ADMIXTURE 1.3 (Alexander *et al*., 2009) and admixture graphs were plotted using "Distruct" (Rosenberg, 2014). We also studied the population split and gene flow between various cattle breeds using TreeMix (Pickrell & Pritchard, 2012) with two data sets: 14 indigenous breeds and 2 *Bos gaurus* samples (TreeMix-I), and combined with global populations in WIDDE (TreeMix-II). The criteria for deciding number of migration edges was based on proportion of variance in relatedness between breeds explained by the model (Pickrell & Pritchard, 2012). Data generated as described above for PCA were used to identify ancestry informative markers (AIMs). Six major breeds contributing to crossbreding in India were retained for analysis, namely HF, Jersey, Gir, Sahiwal, Red Sindhi and Kankrej. Further details in each step are mentioned in Appendix 1.

Results and Discussion

Amongst indigenous breeds, all major dairy breeds *viz*. Gir, Sahiwal, Red Sindhi, Kankrej, Rathi and Hariana had relatively high heterozygosity (Figure 1). Two major draft breeds, Hallikar and Khillar were also highly heterozygous. Interestingly, Siri had the highest heterozygosity amongst all indigenous breeds. Lowest heterozygosity was noted for Kangayam. A relatively large number of SNPs (> 30%) were monomorphic for Kangayam, Amritmahal and Tharparkar. Amongst the major dairy breeds (excluding Tharparkar), Sahiwal had highest proportion of monomorphic SNPs (21.5%).

No major substructures were detected within indigenous breeds (Figure 2). Each breed was found to be well-clustered, based on visual inspection of graphs for principal components (Figures 2, 3, 4). Accounting for the fact that crossbreds were also included in PCA, Tracy-widom statistics indicated first 30 principal components to be significant ($P < 0.0001$). A clear relevance of geography of breed origin could be noted in lower F_{ST} values for the breeds (details not presented).

Figure 5 depicts results of Admixture analysis with $K = 12$, which was found to be optimal value to separate the breeds as distinct populations. Hallikar, Hariana, Khillar and Rathi did not appear as distinct populations. Amritmahal, Kangayam and Sahiwal showed least evidence of admixture. A few Gir, Tharparkar and Red Sindhi showed evidence of recent admixture with Kankrej. However, many samples of Kankrej showed signs of admixture as well, more consistently with Gir. Deoni revealed evidence of recent admixture from Gir. Hallikar and Khillar showed uniform admixture from populations related to Amritmahal, Ongole, Deoni and Kankrej indicating distant admixture events and further

divergence at a later stage. Crossbreds could well be identified, in accordance with known patterns of crossbreeding. As many indigenous breeds show evidence of admixture, screening and selection for admixture in candidate bulls for breed purity can improve breeding programs.

For determining ancestry informative markers, initial PCA and Tracy-widom statistics revealed that only 5 principal components were significant and clustered the 6 breeds. A total of 500 informative markers *viz*. 50 with maximum weightings and 50 with minimum weightings on each principal component were optimal to identify ancestry. Figure 6 describes admixture analysis on validation set along with training set animals, using the identified 500 AIMs. All 60 pure breed animals under validation set were assigned to the correct populations. Though, they deviated from assignment as per admixture analysis done on 777K HD data, no sample deviated by more than 30%. Validation for crossbreds also followed a similar pattern. Identified ancestry informative markers, if incorporated in to the custom chip, can be used as a potential tool to routinely screen breed purity in purebreds and ancestry proportions in crossbreds; more importantly - at low input cost.

Figures 7 & 8 present the two TreeMix analysis results. One migration edge for TreeMix-I (14 *B. indicus* breeds with *B. gaurus* as an out group) and no migration edges for TreeMix-II (*B. taurus* and *B. indicus* breeds) were required. TreeMix-I reveals distinct population splitting and drift parameters amongst the indigenous breeds. Siri was the most distinct breed. Rathi and Sahiwal share common history, distinct from that of Red Sindhi. Gir and Kankrej also shared common evolutionary history. Tharparkar had drifted to considerable extent. Ongole and Deoni shared no common branch with any indigenous breeds. Further, the draft breeds *viz*. Hallikar and Khillar shared common history distinguishing them from Amritmahal, followed by Kangayam, which has drifted the most from other indigenous breeds. Strong evidence of migration was noted from the branch including Gir to Deoni. In TreeMix-II, no hybrid cattle were included and no migration edges were required to be fitted, indicating absence of gene migration between the *B. taurus* and *B. indicus* breeds in this study. Description of TreeMix-II results for *B. taurus* breeds has been described earlier by Decker *et al*. (2014).

Conclusions

The study provides useful understanding of the population structure of *B. indicus* breeds of India. Most indigenous breeds were fairly homogenous and no major substructures for theses breed were identified. Indigenous breeds could be traced to distinct populations, except Hariana, Rathi, Khillar and Hallikar. An evidence of migration from Gir to Deoni was noted, however, no evidence of migration was noted to or from *B. indicus* breeds of India to other global breeds included here. A set of 500 ancestry informative markers for major dairy cattle breeds in India identified which can trace ancestry of major dairy purebred as well as crossbred animals.

Acknowledgements

This study is part of the Indo-Danish collaboration to explore and implement genomic selection in Indian cattle populations between the Department of Molecular Biology and Genetics, Aarhus University, Denmark and the National Dairy Development Board, Anand, Gujarat.

List of References

Alexander, D.H., J. Novembre & K. Lange, 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19: 1655-1664.

Chang, C.C., C.C. Chow, L.C.A.M. Tellier, S. Vattikuti, S.M. Purcell & J.J. Lee, 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:7: doi: 10.1186/s13742-015-0047-8. eCollection 2015.

NBAGR (2017). National Bureau of Animal Genetic Resources, Karnal, Haryana, India. Registered breeds of cattle. URL: www.nbagr.res.in/regcat.html (accession date: 19 September 2017).

Purcell, S. & C. Chang, 2017. Package: PLINK 1.9; URL: www.cog-genomics.org/plink/1.9/.

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick & D. Reich, 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: 904-909.

Patterson, N., A.L. Price & D. Reich, 2006. Population structure and eigenanalysis. PLoS Genetics 2(12): e190. doi:10.1371/journal.pgen.0020190.

Pickrell, J.K. & J.K. Pritchard, 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics 8(11): e1002967. doi:10.1371/journal.pgen.1002967.

Rosenberg, N.A., 2004. DISTRUCT: a program for the graphical display of population structure Molecular Ecology Notes 4:137–138, doi: 10.1046/j.1471-8286.2003.00566.x.

Sempéré, G., K. Moazami-Goudarzi,, A. Eggen, D. Laloë, M. Gautier, & L. Flori, 2015. WIDDE: a Web-Interfaced next generation Database for genetic Diversity Exploration, with a first application in cattle. BMC Genomics 16: 940.

Tracy, C.A. & H. Widom, 2002. Distribution functions for largest eigenvalues and their applications. Proc. International Congress of Mathematicians (Beijing, 2002), 1, Beijing: Higher Ed. Press, pp. 587–596, MR 1989209.

Yang, J., S.H. Lee, M.E. Goddard & P.M. Visscher, 2011. GCTA: a tool for Genome-wide Complex Trait Analysis. Am. J. Hum. Genet. 88(1): 76-82.

Decker, J.E., S.D. McKay, M.M. Rolf, J. Kim & A. Molina Alcalá, *et al*., 2014. Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. PLoS Genet. 10(3): e1004254. doi:10.1371/journal.pgen.1004254.

Matukumalli, L.K., C.T. Lawley, R.D. Schnabel, J.F. Taylor, M.F. Allan, *et al*., 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS ONE 4(4): e5350. doi:10.1371/journal.pone.0005350.

Gautier, M., D. Laloë, K. Moazami-Goudarzi, 2010. Insights into the Genetic History of French Cattle from Dense SNP Data on 47 Worldwide Breeds. PloS ONE 5(9): e13038. doi:10.1371/journal.pone.0013038.

Gautier, M., L. Flori, A. Riebler, F. Jaffrézic, D. Laloé, I. Gut, K. Moazami-Goudarzi and J. Foulley, 2009. A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. BMC Genomics 10:550. doi: 10.1186/1471-2164-10-550.

Appendix 1. Details on Materials and Methods

Sample selection and genotyping

DNA was extracted using commercially available kits (Qiagen) following the manufacturer's protocols at NDDB R&D laboratory, Hyderabad, India. Samples having at least 50 ng/µl DNA concentration, 260/280 OD ratios of 1.8-2.0, and good quality gel picture were selected for genotyping with the BovineHD Beadchip (777K HD; Illumina, San Diego, CA) at M/s Sandor Life Sciences Ltd., Hyderabad, India.

Principal component analysis

Variants with more than 10% missing genotype per variant and per sample and minor allele frequency below 1% were removed. One variant each from 55 duplicate variants with identical genomic positions were removed. Only autosomal variants were retained. One sample from each pair of individuals related by more than 30% was removed. To avoid over-representation of breeds with larger sample sizes, randomly chosen individuals from breeds with larger sample size were excluded (to ensure that no breed consisted more than double the number of samples of breed with minimum sample size). A total of 385 individuals and 699,068 SNP variants were retained for PCA analysis using SMARTPCA, part of EIGENSOFT 7.2.1 (Patterson *et al*., 2006). F_{ST} statistics were used to quantify genetic variability between the breeds.

Admixture analysis

LD based variant pruning was done in PLINK prior to analysis with window size of 100 variants, step size of 10 variants and r^2 threshold of 0.9. A total of 462,104 variants were retained from 385 samples for further analysis. Analysis was done with values of *K* between 1 and 20.

TreeMix analyses

Animals shown to be of crossbred origin in admixture analyses were removed from further analysis. For TreeMix-I, data used for Admixture analysis was merged with *Bos gaurus* samples, thus only retaining SNPs shared between datasets. For TreeMix-II, removal of genetically related samples and LD based pruning, as described in earlier analyses, were repeated. Datasets A, B and C were merged, thus only retaining shared SNPs. Merging of data was done using PLINK. A total of 294 samples were used for TreeMix-I, while 836 samples were used for TreeMix-II. *Bos gaurus* was used as outgroup for TreeMix-I, while Bali cattle was used as outgroup for TreeMix-II. Initially 0 to 5 migration events were allowed accounting for preliminary linkage disequilibrium using groups of 1000 SNPs in both TreeMix analyses. With required migration edges, a further 9 analyses each for TreeMix-I and TreeMix-II, with different -k values *viz*. 1, 5, 10, 50, 100, 500, 1000, 5000 and 10000, as required in TreeMix, were conducted to decide on number of SNPs to group to account for LD. Finally, to check for consistency of results, both TreeMix analyses were repeated for ten times with optimized parameters.

Ancestry informative markers

All pure animals as revealed by preliminary admixture analyses were used for six major breeds. Samples retained were randomly divided into training and validation sets. The training set was used to derive ancestry informative markers (AIMs) and validation set was used to test the performance of the AIMs.

PCA under this section was done using SMARTPCA. SNP weightings of each principal component were derived initially using data on training set. Tracy-Widom statistics were further calculated using TWSTAT, to decide the number of significant principal components. SNPs were ranked based on their maximum and minimum weightings on each of the selected principal components. Different numbers of top and bottom ranked SNPs pertaining to each principal component were selected as different SNP sets. Admixture analysis using ADMIXTURE 1.3 was done on validation set using each of the SNP sets, to decide optimum number of SNPs to use as AIMs, to identify correct ancestry (compared to ancestry of respective animals as depicted by earlier admixture analysis using 777K HD data).

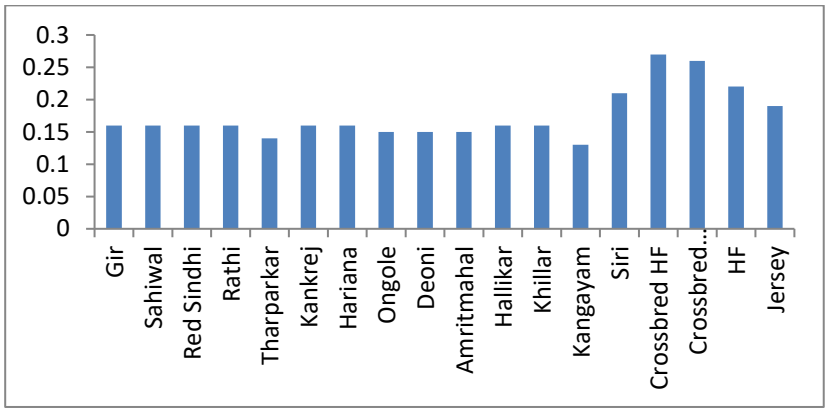

Figure 1. Histogram of breed-wise average minor allele frequencies

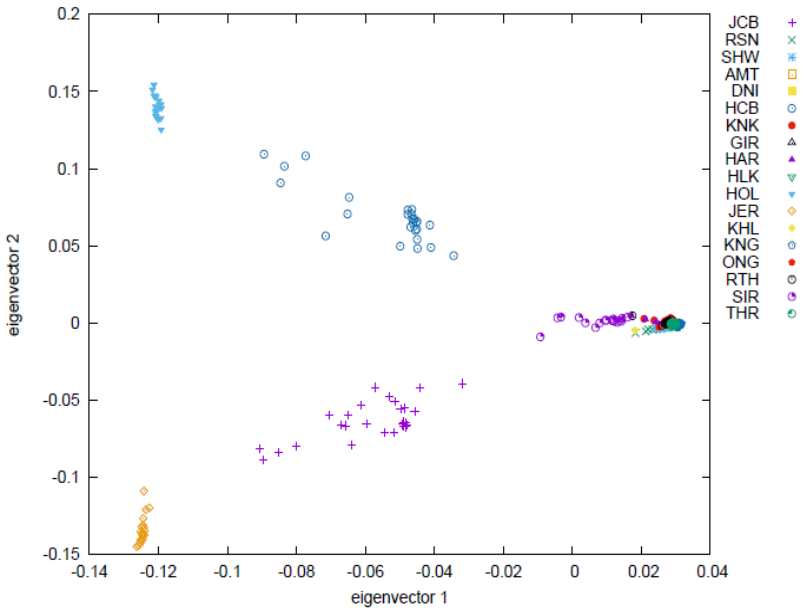

Figure 2. PCA for principal components 1 and 2. JCB:Crossbred Jersey, RSN: Red Sindhi, SHW: Sahiwal, AMT: Amritmahal, DNI: Deoni, HCB: Crossbred HF, KNK: Kankrej, GIR: Gir, HAR: Hariana, HLK: Hallikar, HOL: HF, JER: Jersey, KHL: Khillar, KNG: Kangayam, ONG: Ongole, RTH: Rathi, SIR: Siri

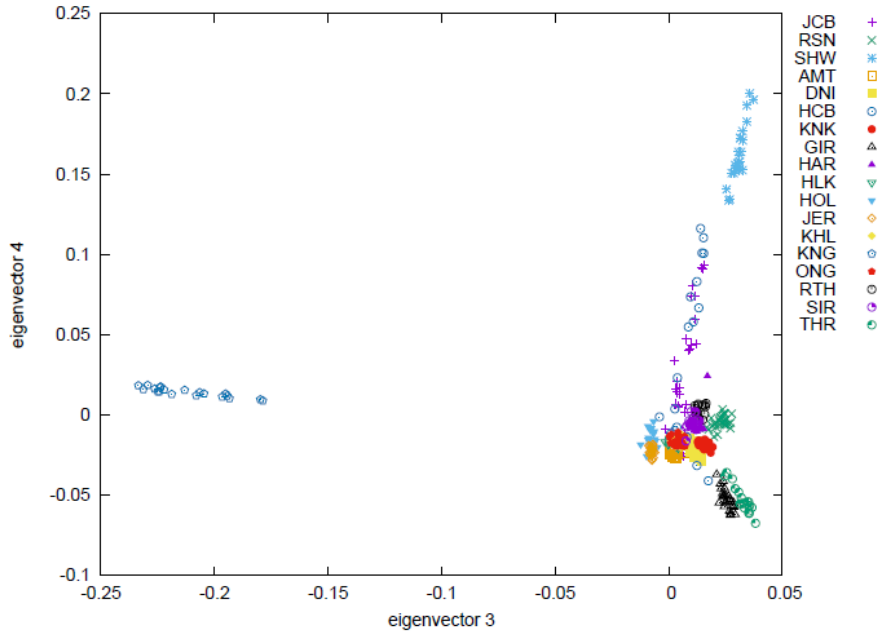

Figure 3. PCA for principal components 3 and 4. Breed names are as Figure 2.

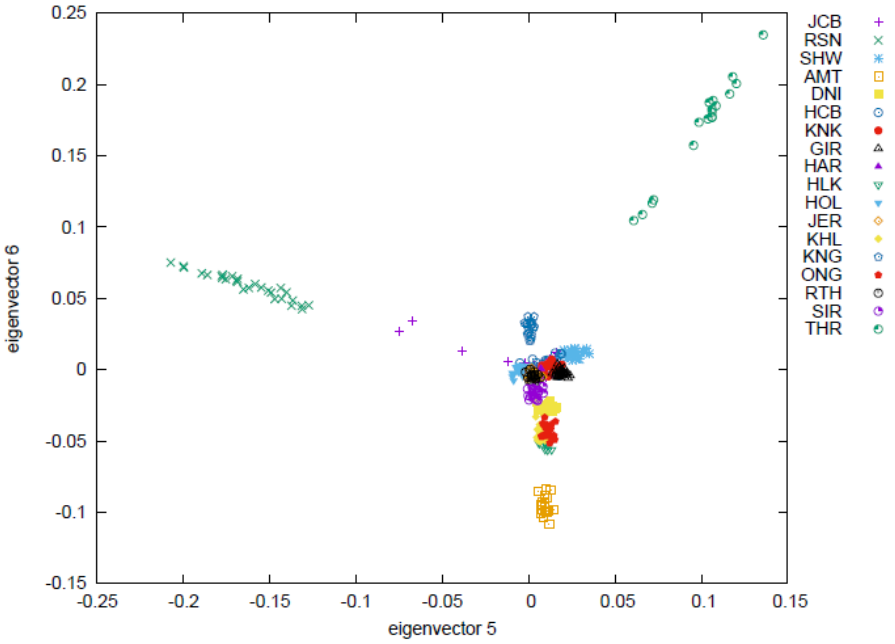

Figure 4. PCA for principal components 5 and 6. Breed names are as Figure 2.

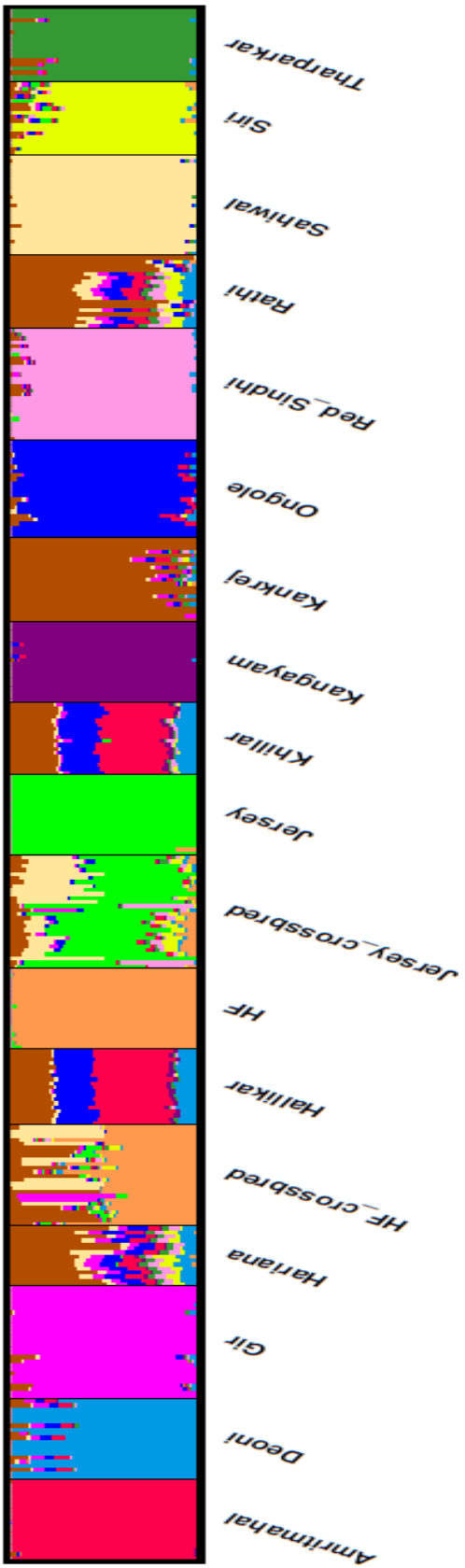

Figure 5. Graph on Admixture analysis with K = 12

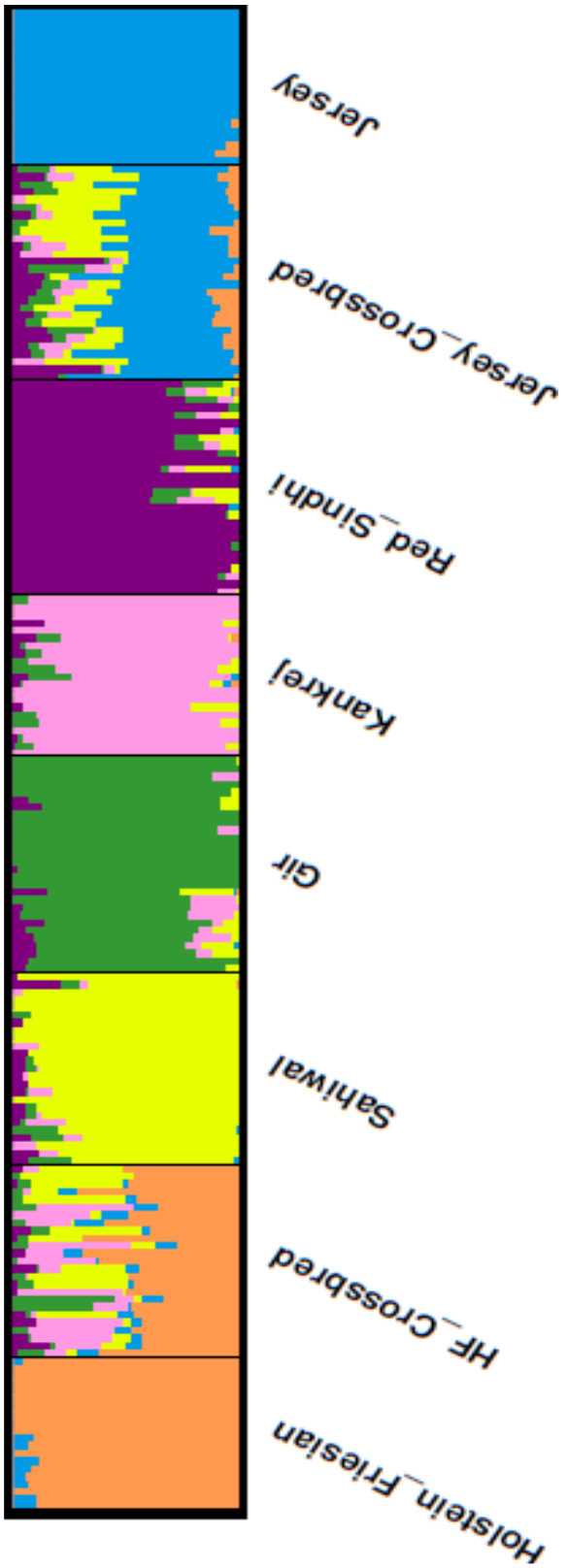

Figure 6. Results of Admixture analysis using ancestry informative markers

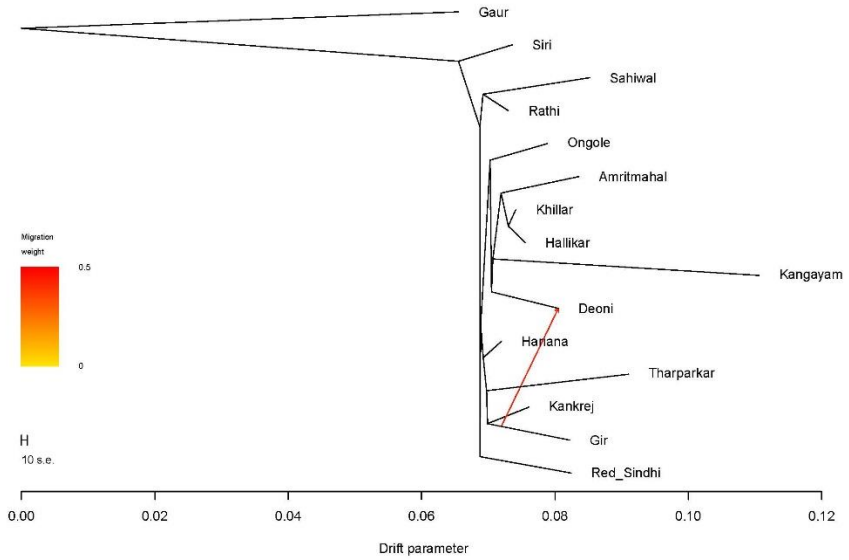

Figure 7. TreeMix for Bos indicus breeds from India

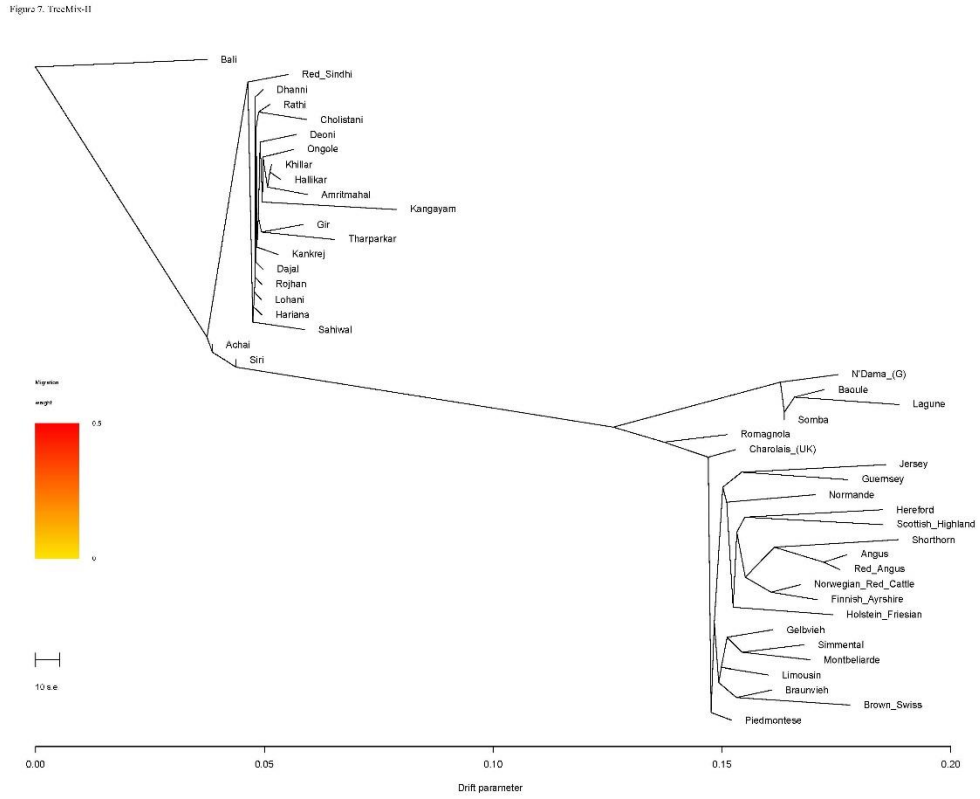

Figure 8. TreeMix for Bos indicus and Bos taurus breeds