Short communication: Population structure of the South African Bonsmara beef breed using high density SNP genotypes

L. Bosman, E. van Marle-Köster, R.R. van der Westhuizen, C. Visser and D.P. Berry

Department of Animal and Wildlife Sciences, University of Pretoria, Pretoria, South Africa

South African Stud Book, South Africa

Teagasc, Moorepark, Ireland

Highlights

- Establishment of a training population for SA Bonsmara beef cattle.
- High density SNP genotypes of 583 cattle used in analysis.
- Eight putative sub-populations identified via ADMIXTURE, but no concrete clusters.
- Results indicate high levels of diversity and a heterogeneous population.

Abstract

The composite Bonsmara beef cattle breed in South Africa is in the process of assembling a reference population for genomic selection. For genomic selection to be a success within the breed, which already has large numbers of genotypes and phenotypes available, the marker effects must be accurately determined by estimating the genetic variance in the breed. To this end the population structure of the Bonsmara reference population was studied using 583 genotypes with 19119 SNPs in linkage disequilibrium ($r^2 < 0.2$). It was found that the reference population is largely non-homogenous, and while 121 breeders contributed toward the reference population, a strong herd/breeder effect could also be observed. Results indicate that a larger reference population may be required for genomic selection in the Bonsmara to be practiced accurately.

Keywords: genetic diversity, training population, Single Nucleotide Polymorphism,

[#] Corresponding author: Este van Marle-Köster: evm.koster@up.ac.za

Introduction

The Bonsmara composite breed was developed using a targeted approach between 1937 and 1963 in South Africa, with the aim of breeding a beef animal with improved growth adpated to sub tropical conditions (Bonsma, 1980). This breed - the most numerous in the country - has gained popularity beyond the borders of South Africa and genetic material has been exported to Namibia, Argentina, Australia and Brazil (Bignardi et al., 2014). All registered Bonsmara cattle are subjected to mandatory performance recording. The relatively large datasets available on several production traits (fertility, growth and carcass traits) have made routine genetic evaluations possible within the breed and contributed to a number of quantitative studies based on their performance data (Maiwashe et al., 2002; Steyn et al., 2014).

The discovery of single nucleotide polymorphism (SNP) markers has led to development of commercial SNP chips for application in genomic selection (GS) (Wickham et al., 2012). Besides the usefulness of GS for selection of sex-limited, lowly heritable and difficult to measure traits (Fan et al., 2010; Eggen, 2012), GS has the advantage of increasing the accuracy of selection at an early age and consequently the rate of genetic progress in the population (Eggen, 2012).

To engage in GS, a reference population within the specific breed is a prerequisite, with large numbers of genotypes and corresponding phenotypes. It is important to account for marker effects in the population (Blasco and Toro, 2014), and thus they should be estimated with high accuracy in order to estimate the level of genetic variance explained (Pszczola et al., 2012a; Boddhireddy et al., 2014). The degree of genetic variance explained by the markers is influenced by effective population size (N_e), with larger N_e resulting in lower prediction accuracy (Pszczola et al., 2012a; Boddhireddy et al., 2014). Sampling across a wide range of genotypes and phenotypes for the reference population may also yield more reliable predictions (Calus, 2010). Studies have shown that including cows in the genomic reference population increases the ability to select for traits such as female fertility (Clark et al., 2012; Calus et al., 2013). In this study the aim was to asses the population structure and genetic background of the Bonsmara breed for establishment of a training population for application in GS.

High impact sires in the Bonsmara with Estimated Breeding Values (EBV) accuracies of at least 65% for maternal weaning weight were identified for this study by the South African Stud Book

Association. Hair samples of 583 Bonsmara, including 388 bulls and 195 cows older than 6 years that have weaned at least 3 calves with recorded weights, were collected by the breeders. The samples were genotyped with the GeneSeek® Genomic Profiler Bovine HD™ Chip (GGP-HD) 80K chip by GeneSeek (Neogen, Lincoln, NE, USA). Ethical approval for the use of the genotypic data was granted by the University of Pretoria and the Bonsmara Cattle Breeders' Society (EC 150304-004 & 005).

Quality control was performed using PLINK (Purcell et al., 2007; Chang et al., 2015), pruning the dataset on minor allele frequency (>0.01), SNP missingness (>0.05) and individual missingness (>0.1). The dataset was further pruned for SNPs in linkage disequilibrium using PLINK, excluding SNPs with a pairwise genotypic $r^2 > 0.2$, (Novembre et al., 2008). A 50 SNP sliding window was used, with a 5 SNP increment between windows. A principal component analysis was performed using GCTA (Yang et al., 2011) based on autosomal SNPs. The relationships and population structure of the three breeds were analyzed using ADMIXTURE (Alexander et al., 2009).

The 583 samples were contributed by 121 breeders, of which 43 submitted only a single sample, while four breeders submitted more than 40 samples each. Of the 71129 possible autosomal SNPs on the 80K chip, 56248 SNPs remained after quality control resulting in a call rate across the dataset of 99.7%. Despite the fact that the Bonsmara was not used in the validation of the bovine SNP chip from its inception (Illumina, 2016), the call rate achieved with the 80K chip compared favorably with those published studies (Cooper et al., 2013; McClure et al., 2013), which allowed further downstream applications.

The average MAF of 0.280 in the Bonsmara was slightly higher compared to the average published during the validation of the chip (Illumina, 2016), and higher than the 0.230 reported by Qwabe et al. (2013). This could be due to the larger number of animals present in the current study, as more minor alleles could be identified. No significant deviation between the observed (0.361) and expected (0.365) heterozygosity values were observed for the Bonsmara, and were similar to Makina et al. (2014). After pruning the dataset for SNPs in linkage disequilibrium (LD), 19119 SNPs were retained in the dataset, averaging 659 SNPs per autosomal chromosome. The mean SNP density in the Bonsmara of 1 SNP/ ~43kb decreased to 1 SNP/ ~90kb after LD pruning, and compared favorably to

that observed in the Blonde d'Aquitaine (1 SNP/ ~ 61kb) (Beghain et al., 2013). The number of SNPs retained after LD pruning was comparable to the study by Makina et al. (2014), where a 50K chip and only 44 animals were used. This suggests that the application of a higher density commercial chip did not affect the number of SNPs LD in this study.

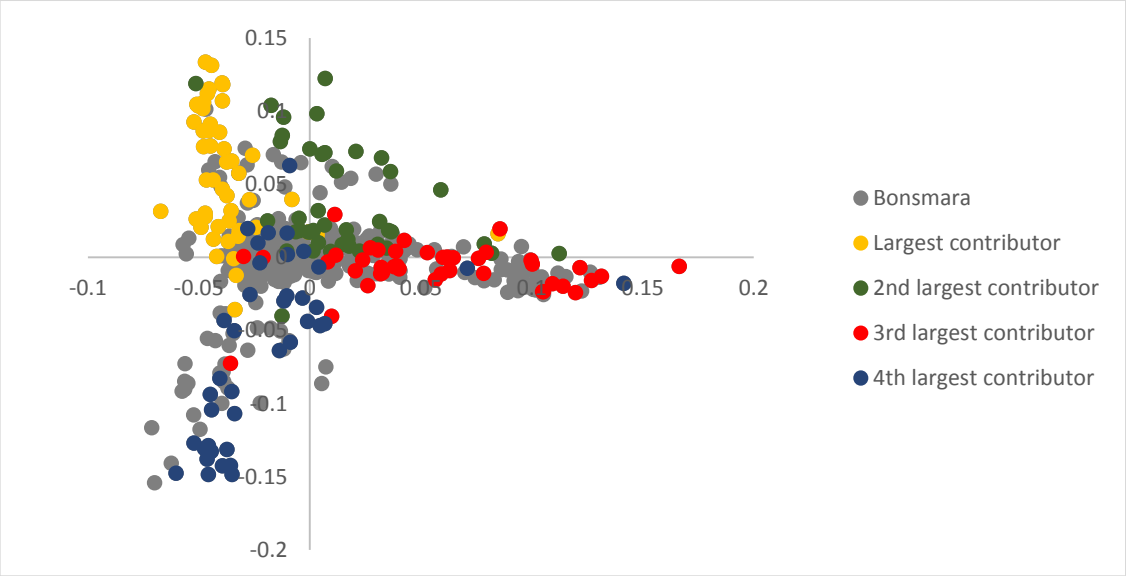

Figure 1 Principal component analysis (PCA) of the Bonsmara reference population (PC1vsPC2)

The principal component analysis is the measure of the genetic variation seen in a population (Aschard et al., 2014), and was performed using the LD pruned dataset. In Figure 1, it is shown that most animals clustered together. Several outliers were identified however, but not at a distance from the other clusters that suggested the presence of a separate population. The four breeders that contributed the largest number of samples towards the reference population have been marked separately, and some of the outliers can be attributed to these individual breeders. These outliers could possibly be the result of differences in breeding goals practiced by different breeders. The outliers may not necessarily have an adverse effect on the reference population, as the aim of compiling the reference population is to increase the relationship strength between the reference population and the rest of the breed (Clark et al., 2012; Calus et al., 2013). Most of these outliers can be attributed to the four breeders however, whom contributed 33.2% of the animals in the reference population, and therefore possible bias may be present.

Artificial insemination is not commonly used in extensively produced beef cattle, and therefore family sizes are smaller. This is important for the eventual calculation of genomic enhanced breeding values (GEBV), as the accuracy of such values will be influenced if an animal has weak a genetic linkage with the reference population. Breeders that contributed more animals to the reference population will therefore have a greater relationship to it, and will have more reliable GEBVs compared to those with limited linkage (Pszczola et al., 2012a). As more genotypes are added to the reference population, relationships with the rest of the breed will improve.

Historically five sup-populations of Bonsmara were recognized (Strydom et al., 2001), namely the Edelheer, T-49, Wesselsvlei, Roodebos and Belmont Red strains. These were distinguished mainly due to the relationships of the animals to respective ancestral strains within the breed. Results from the ADMIXTURE analysis, plotted with Genesis (Buchmann, 2014) is shown in Figure 2, displaying the population structure when plotted with eight populations. The number of possible populations (K) in the dataset were modelled between K=2 to K=15, and cross validation of the results were done 5-fold. Analysis of the cross validation errors indicated that eight populations (clusters) within the breed was the best modelling choice (Alexander et al., 2009).

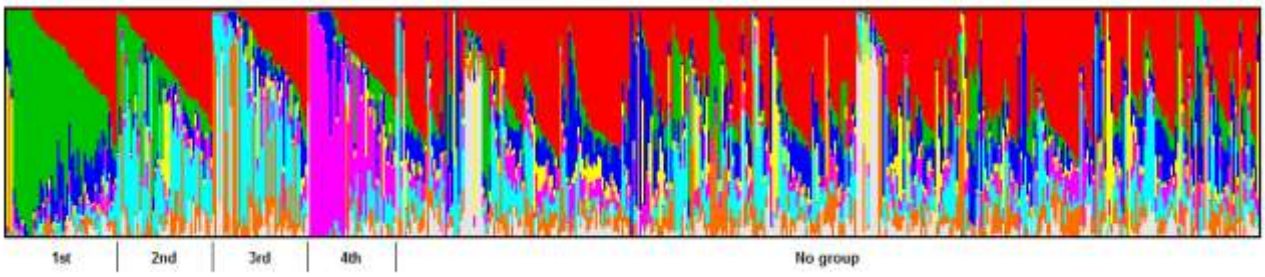

Figure 2 ADMIXTURE analysis of Bonsmara reference population for K=8, plotted by herd, and showing the top four contributing breeders

The Bonsmara reference population is largely non-homogenous, although strong herd effects could be observed. The four top contributing breeders can be seen to form clusters, of which the first and the fourth largest contributors formed clusters not shared widely with the rest of the population. The average proportion of the population found within the eight clusters identified was 30% (red), 13% (light blue), 12% (green), 12% (blue), 10% (grey), 9% (pink), 7% (yellow) and 7% (orange).

Almost all cattle in the reference population had some relationship to the largest cluster (ranging from 0.001% to 67.853%). The distribution of the reference population among the eight clusters identified is given in **Table 1**. In comparing the results represented in **Figure 2** and **Table 1**, it was noted that while clusters 2, 3 and 6 shared a similar proportion of the reference population, clusters 3 and 6 were more evenly distributed throughout. A similar observation could be made regarding clusters 4, 5, 7 and 8, where 4, 7 and 8 were evenly distributed throughout. Clusters 2 and 5 could almost be solely attributable to the first and the fourth largest contributors. In all clusters, apart from cluster 1, individuals could be observed that fell almost exclusively within each cluster, with a maximum relationship between 84.93% and 99.993%.

Table 1 Proportionate membership of the reference population to the eight clusters

	Color coding	Average proportion	Minimum relationship	Maximum relationship
Cluster 1	Red	30%	0.001%	67.853%
Cluster 2	Green	12%	0.001%	99.807%
Cluster 3	Blue	12%	0.001%	99.993
Cluster 4	Yellow	7%	0.001%	99.993
Cluster 5	Pink	9%	0.001%	99.993%
Cluster 6	Light blue	13%	0.001%	84.93%
Cluster 7	Orange	7%	0.001%	91.003%
Cluster 8	Grey	10%	0.001%	88.14%

Based on the analyses of the animals currently included in the reference population there is limited evidence for the presence of the historical five strains. The high diversity seen in the reference population indicates that there are a large number of independent genomic segments (Pszczola et al., 2012a). The aim of sampling widely for the reference population have been met, but the variation seen on a genomic level may have a detrimental effect on accurate GEBV estimation, due to possible weak genetic relationships in the reference population (Pszczola et al., 2012b). The variation on a

molecular level indicates that a larger reference population may be needed for accurate GEBV estimation (Pszczola et al., 2012a).

Conclusion

For successful implementation of GS, large reference populations are required (Blasco and Toro, 2014) and the Bonsmara breed is the most likely candidate to establish a reference population within a reasonable time frame due to its available phenotypes and biological samples. Results indicate that further sampling should consider increasing the relationships among animals within the reference population. Some bias may be present in the current reference population due to 3% of the contributing breeders having contributed 33.2% of the genotypes, and this should be addressed to ensure that the rest of the breed genomic diversity is represented in the reference population, to facilitate accurate genomic predictions..

Acknowledgements

Appreciation is expressed to SA Bonsmara Breeders' Society and SA Stud Book for providing the genotypic data and pedigree records. RMRDT (Genomic selection for SA beef industry669197469) for funding. NRF-THRIP (Project number: TP13073024535 for research support

Author contributions

LB performed the statistical analyses and prepared the draft. EVMK and CV conceptualized and refined the manuscript. RVDW and DB provided insights leading to improved interpretation of results. All authors contributed scientific content and approved the final manuscript.

References

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

Aschard, H., Vilhjalmsson, B.J., Greliche, N., Morange, P.E., Tregouet, D.A., Kraft, P., 2014. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. Am. J. Hum. Genet. 94, 662-676.

Beghain, J., Boitard, S., Weiss, B., Boussaha, M., Gut, I., Rocha, D., 2013. Genome-wide linkage disequilibrium in the Blonde d'Aquitaine cattle breed. J. Anim. Breed. Genet. 130, 294-302.

Bignardi, A.B., Santana, M.L., Eler, J.P., Ferraz, J.B.S., 2014. Models for genetic evaluation of growth of Brazilian Bonsmara cattle. Livest. Sci. 162, 50-58.

Blasco, A., Toro, M.A., 2014. A short critical history of the application of genomics to animal breeding. Livest. Sci. 166, 4-9.

Boddhireddy, P., Kelly, M.J., Northcutt, S., Prayaga, K.C., Rumph, J., DeNise, S., 2014. Genomic predictions in Angus cattle: Comparisons of sample size, response variables, and clustering methods for cross-validation. J. Anim. Sci. 92, 485-497.

Buchmann, R.W., 2014. Genesis - PCA and Admixture Plot Viewer. University of the Witwatersrand, Johannesburg, South Africa. http://www.bioinf.wits.ac.za/software/genesis.

Calus, M.P., 2010. Genomic breeding value prediction: methods and procedures. Animal 4, 157-164.

Calus, M.P.L., de Haas, Y., Veerkamp, R.F., 2013. Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies. J. Dairy Sci. 96, 6703–6715.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7.

Clark, S.A., Hickey, J.M., Daetwyler, H.D., van der Werf, J.H., 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44, 1-9.

Cooper, T.A., Wiggans, G.R., VanRaden, P.M., 2013. Short communication: relationship of call rate and accuracy of single nucleotide polymorphism genotypes in dairy cattle. J. Dairy Sci. 96, 3336-3339.

Eggen, A., 2012. The development and application of genomic selection as a new breeding paradigm. Anim. Front. 2, 10-15.

Fan, B., Du, Z.-Q., Gorbach, D.M., Rothschild, M.F., 2010. Development and application of high-density SNP arrays in genomic studies of domestic animals. Asian-Aust. J. Anim. Sci. 23, 833-847.

Maiwashe, A.N., Bradfield, M.J., Theron, H.E., van Wyk, J.B., 2002. Genetic parameter estimates for body measurements and growth traits in South African Bonsmara cattle. Livest. Prod. Sci. 75, 293-300.

Makina, S.O., Muchadeyi, F.C., van Marle-Köster, E., MacNeil, M.D., Maiwashe, A., 2014. Genetic diversity and population structure among six cattle breeds in South Africa using a whole genome SNP panel. Front. Genet. 5, 333.

McClure, M.C., Sonstegard, T.S., Wiggans, G.R., Van Eenennaam, A.L., Weber, K.L., Penedo, C.T., Berry, D.P., Flynn, J., Garcia, J.F., Carmo, A.S., Regitano, L.C., Albuquerque, M., Silva, M.V., Machado, M.A., Coffey, M., Moore, K., Boscher, M.Y., Genestout, L., Mazza, R., Taylor, J.F., Schnabel, R.D., Simpson, B., Marques, E., McEwan, J.C., Cromie, A., Coutinho, L.L., Kuehn, L.A., Keele, J.W., Piper, E.K., Cook, J., Williams, R., Bovine HapMap, C., Van Tassell, C.P., 2013. Imputation of microsatellite alleles from dense SNP genotypes for parentage verification across multiple Bos taurus and Bos indicus breeds. Front Genet 4, 176.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M., Bustamante, C.D., 2008. Genes mirror geography within Europe. Nature 456, 98-101.

Pszczola, M., Strabel, T., Mulder, H.A., Calus, M.P.L., 2012a. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95, 389–400.

Pszczola, M., Strabel, T., van Arendonk, J.A.M., Calus, M.P.L., 2012b. The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection. J. Dairy Sci. 95, 5412–5421.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am. J. of Hum. Genet. 81, 559-575.

Qwabe, S.O., Van Marle-Köster, E., Maiwashe, A., Muchadeyi, F.C., 2013. Evaluation of the BovineSNP50 genotyping array in four South African cattle populations. S. Afr. J. Anim. Sci. 43, 64-67.

Steyn, Y., van Marle-Köster, E., Theron, H.E., 2014. Residual feed intake as selection tool in South African Bonsmara cattle. Livest. Sci. 164, 35-38.

Strydom, P.E., Naude, R.T., Smith, M.F., Kotze, A., Scholtz, M.M., Van Wyk, J.B., 2001. Relationships between production and product traits in subpopulations of Bonsmara and Nguni cattle. S. Afr. J. Anim. Sci. 31, 181-194.

Wickham, B.W., Amer, P.R., Berry, D.P., Burke, M., Coughlan, S., Cromie, A., Kearney, J.F., Mc Hugh, N., Mc Parland, S., O'Connell, K., 2012. Industrial perspective: capturing the benefits of genomics to Irish cattle breeding. Anim. Prod. Sci. 52, 172-179.

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

Supplementary material

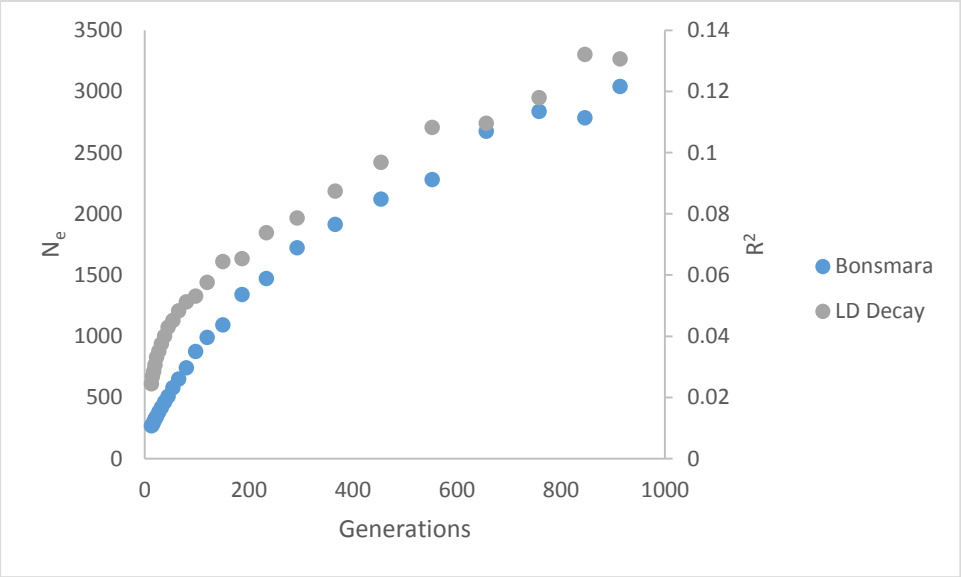

Figure S1 Effective population (N_e) size of the Bonsmara reference population over generations, and the corresponding decay in marker linkage disequilibrium (LD) according to the R^2 values

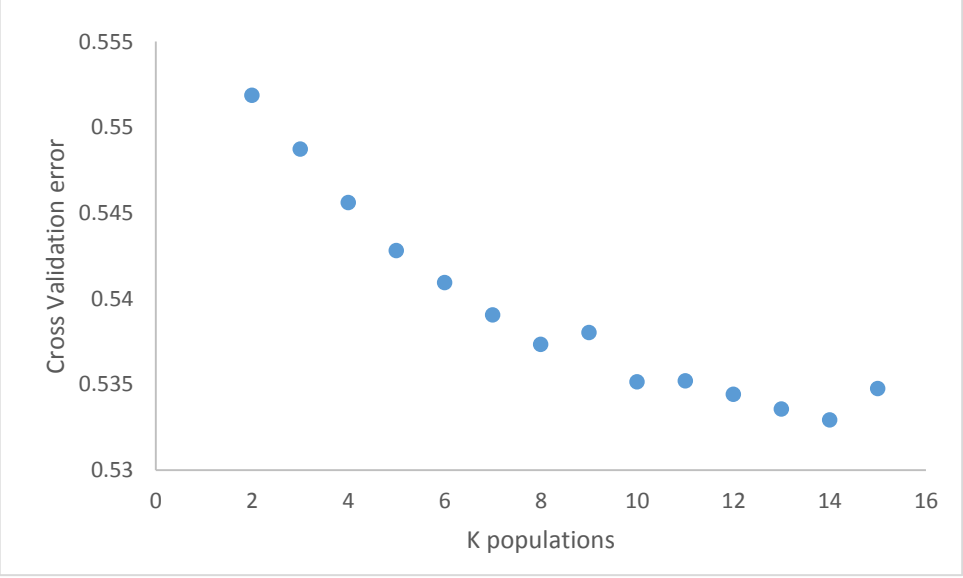

Figure S2 The most likely K for the Bonsmara reference population is determined by plotting the mean cross-validation (CV) errors for each ADMIXTURE run, and finding the inflection point at the lowest CV error (K=8)