COBISS: 1.08
Agris category code: L10

ESTIMATION OF LOCAL GENETIC ANCESTRY IN AN ADMIXED CATTLE POPULATION APPLYING DIFFERENT METHODS

Negar KHAYATZADEH [1], Gábor MÉSZÁROS [2], Birgit GREDLER [3], Urs SCHNYDER [4], Ino CURIK [5], Johann SÖLKNER [6]

ABSTRACT

In current study, we estimated local genetic ancestry for Swiss-Fleckvieh dairy cattle population using two alternative approaches implemented in LAMP and MULTIMIX software. Swiss-Fleckvieh is a composite descending from Simmental and Red Holstein Friesian. Illumina BovineSNP50 Beadchip data for 485 animals were used. Genetic ancestries averaged across admixed animals indicated a high correlation of individual admixture at global levels (0.99), yet the correlation between genetic ancestries averaged across 7914 windows (5 SNPs each) was 0.55. The highest correlations of window-wise estimates were along chromosomes 2, 5, 12, 15, 21, 27 and 29 (≥0.80) and the lowest was −0.71 on chromosome 24. Based on LAMP results, a region on chromosome 13 (46.3–46.7 Mb) and three adjacent regions on chromosome 18 (18.8–19.2, 20.1–25.7 and 23.9–25.7 Mb) passed the threshold based on deviation from normal distribution of local ancestries assuming 1000 independent segments. Regarding MULTIMIX, another region on chromosome 22 (9.3–10.3 Mb) was significant. Although estimates from both programs are comparable at the global levels, there was considerable difference at the local levels.

Key words: cattle, breeds, Swiss-Fleckvieh, genetics, SNP, global genetic ancestry, local genetic ancestry, LAMP, MULTIMIX, selection signature

1 INTRODUCTION

The study of population structure based on genetic ancestry is an important component of many genetic studies. Genetic ancestry is a broad concept that is concerned with 1) defining the number of ancestral populations in admixed populations, 2) assigning ancestral population proportions to admixed individuals and 3) identifying the genetic ancestry of distinct chromosomal segments within an individual (Liu *et al.*, 2013). Progressive advances in high-throughput single nucleotide polymorphism (SNP) genotyping in the recent decade provide a proper opportunity to calculate the ancestral origins of distinct chromosomal segments (termed local ancestry) as well as ancestry proportions averaged across the genome of an individual (termed global ancestry).

Local genetic ancestry estimates vary among loci along the genome of admixed individuals and therefore deviate from genome wide average ancestries. The most important sources of variation in local genetic ancestries are genetic drift and selection (Long, 1991). Genetic drift is a demographic process that influences on the local ancestry proportions along the whole genome, but selection targets only specific gene regions (Oleksyk *et al.*, 2010; Tang *et al.*, 2007). Recently admixed populations represent attractive biological models to study the genome response to adaptive selection. Indeed, during the admixture, the frequencies of the favorable variants are

[1] BOKU, Department of Sustainable Agricultural Systems, Division of Livestock Sciences (NUWI), Gregor-Mendel-Strasse, 1180 Vienna, Austria, e-mail: negar.khayatzadeh@students.boku.ac.at
[2] Same address as 1, e-mail: gabor.meszaros@boku.ac.at
[3] Qualitas AG, Chamerstrasse 56, Ch-6300, Zug, Switzerland, e-mail: birgit.Gredler@qualitasag.ch
[4] Same address as 3, e-mail: urs.schnyder@qualitasag.ch
[5] Department of Animal Science, Faculty of Agriculture, University of Zagreb, Svetošimunska cesta 25, 10000 Zagreb, Croatia, e-mail: icurik@agr.hr
[6] Same address as 1, e-mail: johann.soelkner@boku.ac.at

expected to increase and the levels of the admixture in these regions deviate from the expected ancestry proportions (Bhatia *et al.*, 2014; Gautier and Naves, 2011).

Moreover, local genetic ancestry estimates tend to be considered in mapping of genes associated with diseases to remove the spurious associations due to admixture (Chen *et al.*, 2014; Kang *et al.*, 2009; Seldin, 2007).

There are various software tools that have been developed to calculate the genetic ancestry in both global and local levels. In general, they rely either on multivariate statistical methods to like principal component analysis (PCA) and clustering algorithms, or explicit genetic models to inference genetic ancestry. LAMP (Local Ancestry in adMixed Populations) is a program that infers locus-specific ancestry in admixed populations using sliding windows of contiguous SNPs (Sankararaman *et al.*, 2008). It does not require the genotype of ancestral populations as input. The prior knowledge is the allele frequencies in each ancestral population that can constitute an ancestry informative markers panel (AIMs) to calculate genetic ancestry. AIMs are the markers that represent a high difference in their frequencies. The idea of LAMP is to select a suitable window length and then a clustering algorithm is used to estimate the maximum likelihood to infer the ancestry within this window.

Likewise, MULTIMIX is another program to infer local genetic ancestry (Churchhouse and Marchini, 2013). It needs phased data as input to calculate haplotypes. Differences in haplotype frequencies between populations are used to infer the origin of observed haplotypes in admixed individuals. It calculates the genetic ancestry for a set of SNPs (window-wise). Levels of linkage disequilibrium (LD) between a subset of SNPs within windows also need to be taken into account. This program uses a Markov process to model switches between ancestries along the genome based on information on LD between SNPs.

In this study, we used LAMP and MULTIMIX to estimate local ancestry with different approaches in Swiss Fleckvieh cattle population. We further searched for selections signatures happened after admixture by applying the thresholds suggested by Bhatia *et al.* (2014).

2 MATERIALS AND METHODS

Swiss Fleckvieh is a composite breed of Simmental (SI) and Red Holstein Friesian (RHF) that was established over the last forty years in Switzerland with the emphasis on high milk production derived from the Holstein Friesian as well as additional purposes such as reproduction, beef value and longevity of the Simmental breed. The genotype data from the Illumina Bovine SNP50 beadChip on 101 pure RHF, 91 pure SI and 308 admixed bulls, provided by Swissherdbook cooperative Zollikofen, were used for this study. According to the formal definition, animals with pedigree admixture level of 0.125 to 0.875 RHF are categorized as Swiss Fleckvieh. In the current study, we considered all admixed animals along the range of pedigree admixture level of 0.02 to 0.99 RHF. Quality control of the data was performed with PLINK 1.9 (Purcell *et al.*, 2007). SNPs with call rate less than 95 %

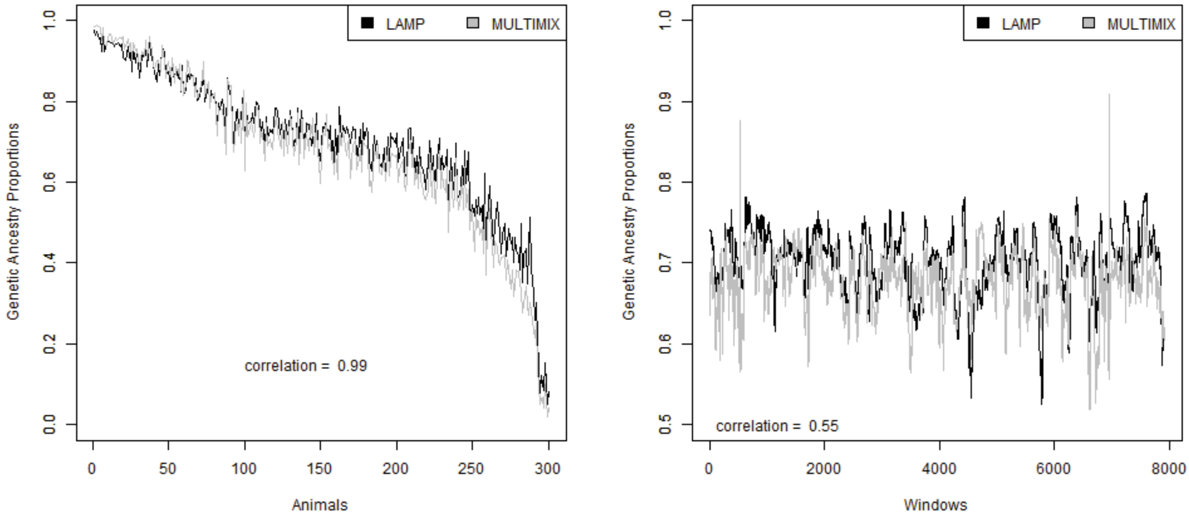

Figure 1: *Genome wide ancestry across the whole genome for 300 admixed animals (left). Average locus specific ancestries calculated across 7,914 windows (right).*

that were monomorphic and also deviated from Hardy-Weinberg equilibrium (*p*-value $< 10^{-6}$) were excluded from the dataset. We also excluded animals with more than 5 % missing genotypes, SNPs with no information on their positions and those that were located on the X chromosome. After quality control 485 animals (97 RHF, 88 SI and 300 admixed bulls) and 39,525 SNPs were left.

We used LAMP 2.5 to calculate locus specific ancestry. The constant recombination rate was set to 10^{-8}. We assumed that 0.01 recombination events occur per Mb. We also did not consider linkage disequilibrium between our SNPs. The prior ancestry proportions were set to 0.68 and 0.32 obtained with ADMIXTURE software (Alexander *et al.*, 2009). The locus specific ancestries were estimated for each admixed animal with respect to pure breeds, representing the proportion of each involved ancestry for each SNP. The fraction of alleles from RHF population is respectively 1, 0.5 and 0, if both alleles, one allele and none of them are from the RHF ancestry population.

To calculate the local ancestry with MULTIMIX, we phased the dataset on ancestral and admixed population using SHAPEIT *v2.r790* (Delaneau *et al.*, 2012). Then we used MULTIMIX with chunk length 5 SNPs per window

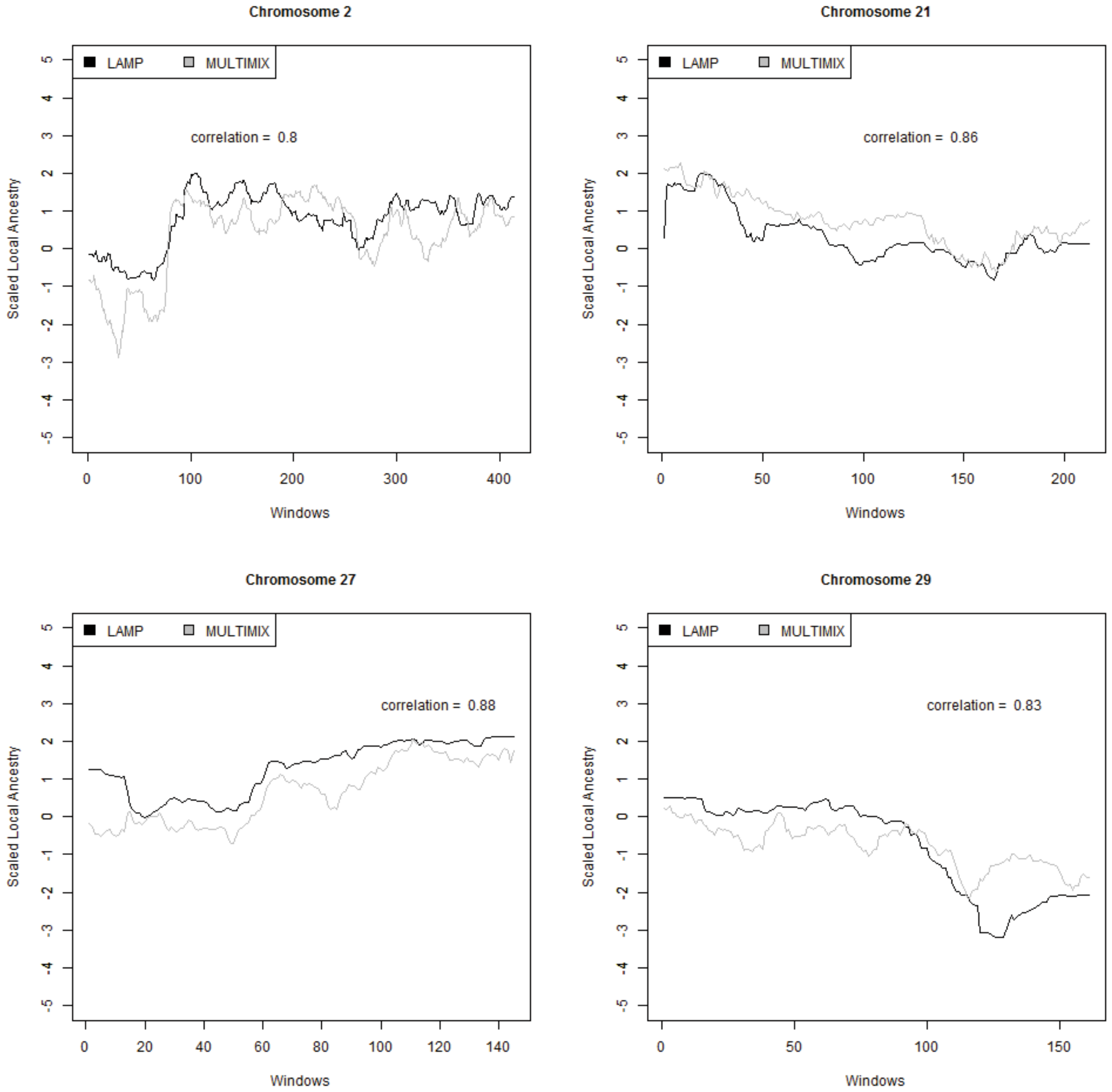

Figure 2: *Standardized local ancestries based on mean 0.70 and SD 0.040 estimated by LAMP and mean 0.68 and SD 0.038 estimated by MULTIMIX on chromosomes 2, 21, 27 and 29.*

and misfit probabilities 0.68 and 0.32 and calculate the local ancestries for each chromosome separately. The genetic ancestries by MULTIMIX were calculated percentagewise. We changed them to the proportions (1, 0.5 and 0) similar to LAMP. So, if both haplotypes showed RHF probability greater than 0.68, we put RHF genetic ancestry to 1. We defined 0.5 if one of the haplotypes had the probability of RHF greater than 0.68 and the other one had the probability of RHF less than 0.32. If both haplotypes showed the probability of RHF ancestry less than 0.32, the RHF genetic ancestry was set to 0. The other values in between were regarded as missing values.

To detect the selection signals following Bhatia *et al.* (2014), we standardized the local ancestries calculated by LAMP (by 0.70 mean and 0.040 SD) and MULTIMIX (0.68 mean and 0.038 SD).

On the basis of the extended linkage disequilibrium (LD) in the genome of admixed individuals, we defined a genome wide significant selection signals by correction for multiple hypothesis testing considering 5000 and 1000 independent segments along the whole genome (Bhatia *et al.*, 2014). Therefore, standardized local ancestries greater than 4.42 (p-value < 1×10^{-5}) correspond to 5000 hypotheses and 4.06 (p-value < 5×10^{-5}) correspond

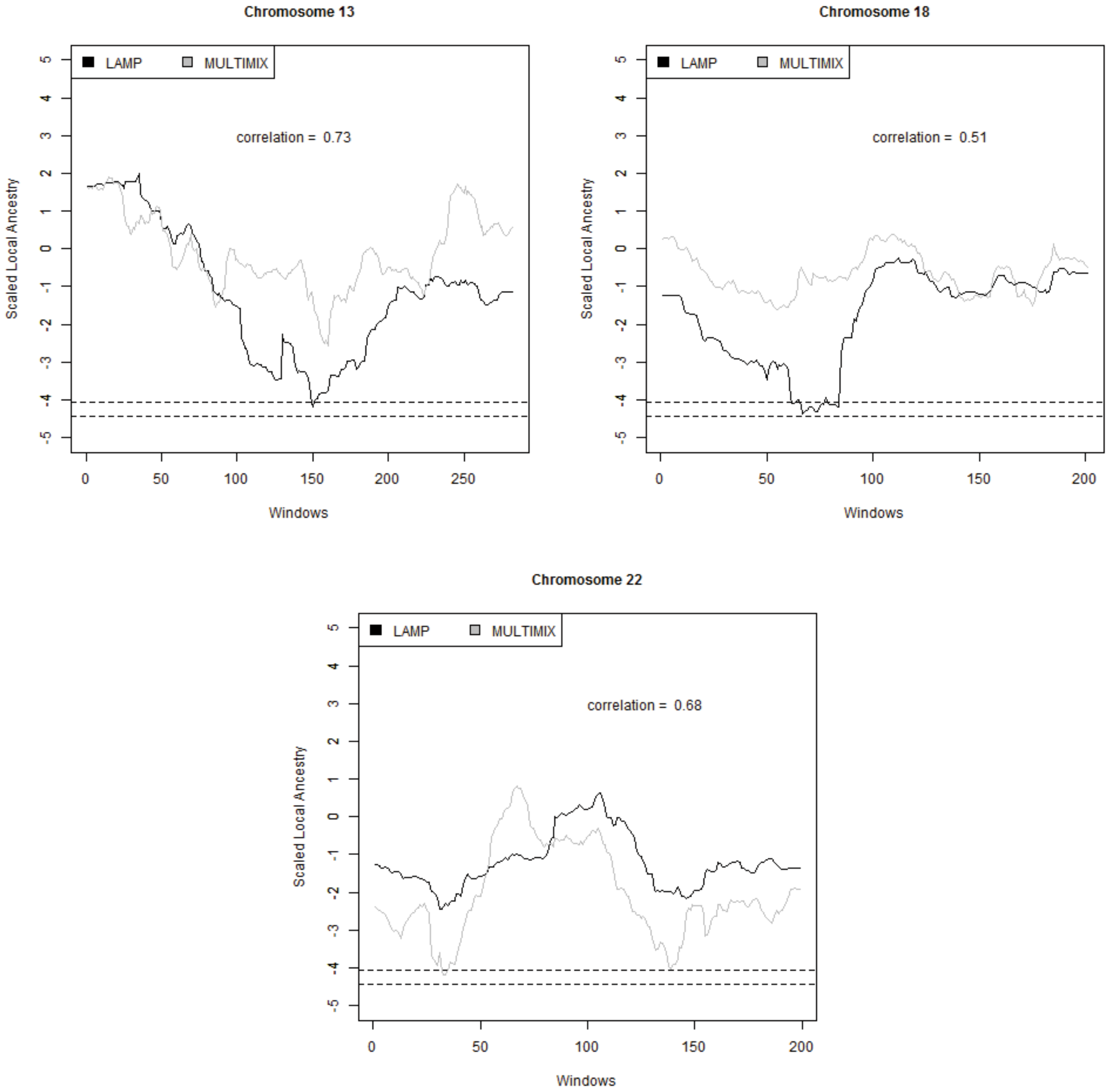

Figure 3: *Standardized local ancestries based on normal distribution hypotheses tests. Two dashed lines are threshold lines based on 1000 (−4.06) and 5000 (−4.42) independent tests.*

to 1000 hypotheses were considered as significant thresholds. As LD is higher and therefore the number of independent segments of the genome is smaller in bovine populations compared to human populations analyzed by Bhatia *et al.* (2014), we consider these thresholds conservative (Hayes *et al.*, 2003).

3 RESULTS AND DISCUSSION

Local genetic ancestries were calculated across contiguous windows with the length of 5 SNPs. In total the genome was divided into 7914 windows. The genome wide ancestry across the whole genome for each admixed animal (global admixture) was calculated by LAMP and MULTIMIX programs. Pearson's correlation for genetic ancestries across individuals from both programs was considerably high (0.99). The average locus specific ancestries were calculated across the admixed animals for 7914 windows. The results from both programs showed moderate correlation (0.55) (Fig. 1).

We searched through chromosomes, to see the pattern of the results derived from the two programs and also how the results from both programs correlated across each chromosome. The correlation between the local genetic ancestry estimates along each chromosome from both programs ranged from −0.71 on chromosome 24 with 977 SNPs to 0.88 on chromosome 27 with 723 SNPs. Correlations greater than 0.80 were observed on chromosomes 2, 5, 12, 15, 21, 27 and 29. The results on chromosomes 2, 15, 21 and 27 were shown in Figure 2.

Figure 3 displays the average local ancestry at each SNP along chromosomes 13, 18 and 22 derived from LAMP and MULTIMIX. Regards to the effective number of independent segments in the genome of admixed animals, we decided for two thresholds (values greater than 4.06 and greater than 4.42). Based on LAMP results, two regions on chromosomes 13 and 18 passed the first threshold (1000 independent segments or hypotheses). Based on MULTIMIX results, this region does not overlap with LAMP results. However, another region on chromosome 22 passed the threshold.

A relatively wide significant signal based on 1000 independent hypotheses was detected on 46.3–46.7 Mb (windows 149 and 150) by LAMP program. The locus specific ancestries from MULTIMIX in this region were in the same direction with signals detected by LAMP, but they did not pass the threshold.

Three wide regions on chromosome 18 passed the threshold based on 1000 hypotheses by LAMP. The first region was located on 18.8–19.2 Mb. The second region was detected on 20.1–23.7 Mb and the third significant signal was detected on 23.9–25.7 Mb. No considerable signals were detected in these regions by MULTIMIX.

On chromosome 22, a significant signal was detected on 9.3–10.3 Mb based on 1000 independent tests based on MULTIMIX. Regards to LAMP results, the locus specific ancestries were in the same direction as MULTIMIX, but they were not significant.

Although estimates from MULTIMIX and LAMP were comparable at the global level, there was considerable difference at the local level (see Figures 2 and 3). LAMP does not use LD background of data to calculate the local ancestries and it works well when the data density is not very high. On the other hand, MULTIMIX takes into account linkage disequilibrium (LD) and is capable of multiway admixture deconvolution. Based on a study by Chen *et al.* (2014), the results from MULTIMIX are more accurate when the size of windows are not high (Chen *et al.*, 2014).

4 CONCLUSIONS

This study considered two methods implemented by LAMP and MULTIMIX programs to calculate the local admixture along the genome of Swiss Fleckvieh cattle, a composite of Simmental and Red Holstein Friesian. Both programs showed high correlation in estimation at global level, yet at local level the correlation was not high. Therefore, one should be careful when interpreting signals of selection based on results produced by a single software.

5 ACKNOWLEDGEMENTS

We would like to thank the Swissherdbook cooperative Zollikofen for providing genotypes for analysis.

6 REFERENCES

Alexander, D. H., Novembre, J., Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research, 19*(9), 1655–1664.

Bhatia, G., Tandon, A., Patterson, N., Aldrich, M. C., Ambrosone, C. B., Amos, C., Bandera, E. V., *et al.* (2014). Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. *American Journal of Human Genetics, 95*(4), 437–444.

Chen, M., Yang, C., Li, C., Hou, L., Chen, X., Zhao, H. (2014). Admixture mapping analysis in the context of GWAS with GAW18 data. *BMC Proceedings, 8*(Suppl 1), S3. Retrieved from http://bmcproc.biomedcentral.com/articles/10.1186/1753-6561-8-S1-S3.

Churchhouse, C., Marchini, J. (2013). Multiway admixture deconvolution using phased or unphased ancestral panels. *Genetic Epidemiology, 37*(1), 1–12.

Delaneau, O., Marchini, J., Zagury, J. F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods, 9*(2), 179–181.

Gautier, M., Naves, M. (2011). Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Molecular Ecology, 20*(15), 3128–3143.

Hayes, B. J., Visscher, P. M., McPartlan, H. C., Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research, 13*(4), 635–643.

Kang, S. J., Larkin, E. K., Song, Y., Barnholtz-Sloan, J., Baechle, D., Feng, T., Zhu, X. (2009). Assessing the impact of global versus local ancestry in association studies. *BMC Proceedings, 3*(Suppl 7), S107. Retrieved from http://bmcproc.biomedcentral.com/articles/10.1186/1753-6561-3-S7-S107.

Liu, Y. S., Nyunoya, T., Leng, S. G., Belinsky, S. A., Tesfaigzi, Y., Bruse, S. (2013). Softwares and methods for estimating genetic ancestry in human populations. *Human Genomics, 7*(1). Retrieved from http://humgenomics.biomedcentral.com/articles/10.1186/1479-7364-7-1.

Long, J. C. (1991). The genetic structure of admixed populations. *Genetics, 127*(2), 417–428.

Oleksyk, T. K., Smith, M. W., & O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B-Biological Sciences, 365*(1537), 185–205.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maler, J., *et al.* (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics, 81*(3), 559–575.

Sankararaman, S., Sridhar, S., Kimmel, G., Halperin, E. (2008). Estimating local ancestry in admixed populations. [Comparative Study]. *American Journal of Human Genetics, 82*(2), 290–303.

Seldin, M. F. (2007). Admixture mapping as a tool in gene discovery. *Current Opinion in Genetics & Development, 17*(3), 177–181.

Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E. G., Risch, N. J. (2007). Recent genetic selection in the ancestral admixture of Puerto Ricans. *American Journal of Human Genetics, 81*(3), 626–633.