University of
Natural Resources and
Life Sciences, Vienna

Department of
Sustainable Agricultural Systems

Division of Livestock Sciences
WG Animal Breeding

Study of admixture and heterosis using molecular markers in a composite cattle breed

Negar Khayatzadeh

Doctoral thesis

Vienna, October 2017

Main supervisor

Univ.Prof. Dipl.-Ing. Dr. Johann Sölkner
University of Natural Resources and Life Sciences, Vienna (BOKU), Department of Sustainable Agricultural Systems, Division of Livestock Sciences (NUWI), Vienna, Austria

Co-supervisors

Prof. dr. sc. Ino Čurik
University of Zagreb, Faculty of Agriculture, Department of Animal Science, Zagreb, Croatia

Ass.Prof. Dr. Gábor Mészáros
University of Natural Resources and Life Sciences, Vienna (BOKU), Department of Sustainable Agricultural Systems, Division of Livestock Sciences (NUWI), Vienna, Austria

To my family

Acknowledgements

The successful completion of this dissertation has been possible with the support of people and I would like to express my sincere gratitude to all of them.

Foremost, I would like to express my deepest sense of **G**ratitude to my supervisor, **Professor Johann Sölkner**, for his valuable guidance, continuous support, patience and consistent encouragement I received throughout the last three years. It was of great significance to have free and unlimited access to his wealth of knowledge and experience. I thank him for his guidance and great effort in all the time of research and writing my manuscripts. I could not have imagined having a better advisor for my study.

Besides, I would like to express my sincere gratitude to my co-supervisor, **Dr. Gábor Mészáros**, for his insightful comments, encouragement and remarkable support to the thesis.

I would like to thank my co-supervisor, **Professor Ino Čurik**, for his valuable suggestions and comments on research papers of the thesis, and **Dr. Vlatka Čubrić Čurik**, for the help and support during my stay in Croatia.

I would like to give my sincere thanks to **Professor José Fernando Garcia** for arranging for me to spend a month in his lab in Araçatuba, Brazil.

I would like to thank **Yuri Tani Utsunomiya**, whom I greatly benefited from his keen scientific insight and his valuable suggestions for *R* program codes. I am grateful to **Dr. Maja Ferenčaković**, for her support and sharing her knowledge on the statistical data analysis for the last manuscript.

I would like to give my sincere thanks to **Dr. Ardeshir Nejati Javaremi**, for his encouragement and support.

I would like to acknowledge **Qualitas AG, Switzerland** and the group staff, **Dr. Urs Schnyder**, **Dr. Birgit Gredler** and **Dr. Franz Seefried**, for providing the genotype data during my study and lots of valuable interactions on the manuscripts.

Acknowledgment is extended to **Swissgenetics, Dr. Fritz Schmitz-Hsu** to provide the phenotype data for my study that was the basis for two exciting studies on estimation and mapping of “heterosis” effects.

Special thanks to all of the members at animal breeding working group at **BOKU** for their friendship, all the pleasant time and all the fun in the last three years and interesting discussions during coffee break.

Finally I take this opportunity to express the profound gratitude from my deep heart to my lovely family especially my father and my mother for their love and continuous support, and my all lovely friends.

Abstract

The identification of ancestry origin of chromosomal segments in crossbred population, termed local genetic ancestry, has been widely investigated in population genetics, for genetic disease mapping, admixture mapping and population history inference. A genome-wide perspective on a recently admixed population reveals that the ancestral contributions vary along the genome. Important sources of variations in the genome of the admixed population are natural selection as well as sampling errors and evolutionary fluctuations due to genetic drift and gene flow. Selection targets specific gene regions in contrast to random genetic drift influencing the entire genome. Extreme deviations of local genetic ancestries from the average genome-wide ancestry can be detected as signatures of selection having happened after admixture. Therefore, recent admixed populations provide an excellent opportunity to study post-admixture selection signatures. The aim of this thesis was to study ancestral contributions at local level in order to detect recent selection signature and to estimate heterosis components in admixed Swiss Fleckvieh cattle, a young composite of the two parental breeds Red Holstein Friesian (RHF) and Swiss Simmental (SI).

First, we estimated ancestry at both global and local level using Illumina® BovineSNP50k genotypes on 485 bulls, including admixed and two ancestral populations. The global RHF and SI proportions of ancestry were estimated 0.70:0.30. Local genetic ancestry estimations were used to detect selection signals. To identify the significant threshold for the detected signals, two approaches were employed based on permutations test and Bonferroni correction for extreme deviations from normal distribution. Both approaches resulted in similar thresholds. Two notable peaks, one on chromosome 13 (46.3-47.3 Mb) and another region on chromosome 18 (18.7-25.9) were identified as the recent selection signatures, according to both thresholds. Applying extended haplotype homozygosity (*EHH*) to explore pre- and post-admixture signals, revealed a signal on chromosome 18 (25.5-26.4 Mb) based on *iHS* statistics in RHF ancestral population and a wide region on chromosome 18 (6.6-24.6) based on *Rsb* statistics between admixed bulls and SI ancestry populations. Moreover, no considerable signal was detected by *Fst*. Wide admixture selection signals indicated that 1) the limited numbers of generations after admixture (~ 10-15) were not enough to sharpen signals; 2) comparison of pre- and post-admixture signals

was not very promising, and 3) vague candidates of genes under selection were found in the detected regions.

Second, local ancestries were estimated using two other different software tools (LAMP-LD and MULTIMIX), which require assumption of a parametric population genetic model, unlike LAMP used in the previous study, which trusts on clustering algorithms for local ancestry deconvolution. Different parameter settings such as phased data and window lengths were defined. The relatively high correlations were observed between LAMP-LD and MULTIMIX_MCMCgeno, where both used same phased reference panel and unphased genotypes of admixed animals with window lengths were 15 SNPs (0.81) and 23 SNPs (0.85). The highest correlations were observed between the results MULTIMIX_MCMC, using haplotypes on both reference panel and admixed animals and MULTIMIX_MCMCgeno with 15 SNPs (0.92) and 23 SNPs (0.85). Medium to low correlations between results of different software tools indicated that choosing the method of local ancestry inference and consequently inferred selection signals should be considered carefully and confirmation with alternative approaches is advised.

Local ancestry estimates were used to estimate the effects of dominance and epistatic loss (two definitions) as components of heterosis for sperm quality traits in admixed Swiss Fleckvieh bulls. Dominance component of heterosis was very significant and improved model accuracies of three out of four evaluated semen traits. Dominance components of heterosis were estimated 1.24 ml, 0.28 and 1.40 % for volume, transformed number of spermatozoa and percentage of live sperm. Although the epistatic effects have been reported in most of studies to have minor importance, we found significant levels for this effect for volume and transformed number of spermatozoa according to our new definition based on the extreme situation of losing breed specific epistatic combination.

Finally, genome-wide mapping of the dominance component of heterosis for percentage of live sperm was performed using an appropriate model with SNP effect, genomic breed percent and dominance effect. Some significant regions were found on chromosomes 3, 4, 5, 7, 13 and 14 hosted genes associated with spermatogenesis.

Keywords admixture, crossbreeding, dominance component of heterosis, epistatic loss, genome-wide mapping, global genetic ancestry, haplotype, heterosis, local genetic ancestry, permutation, phasing, selection signature, SNP

Table of contents

Chapter 1

General introduction

1-1 Introduction

Livestock products as the important agricultural commodity for global food security supply about 26 % of the proteins and 13 % of the calories consumed worldwide (FAO 2011). The global human population is expected to grow to almost 10 billion by 2050. Rapid population growth and transition in dietary patterns, due to urbanization and growth of income and social rank, increase global demand for higher quality food and protein sources (FAO 2017).

Effective management of farm animal genetic resources (FAnGR) requires understanding of population structure and genetic diversity within and among livestock populations as well as knowledge on geographical distribution and production environment. Insights to the genetic structure can contribute to improve the breeding programs (pure- and cross-breeding), understanding of environmental adaptation and conservation of the livestock breeds (Notter 1999; Groeneveld *et al.* 2010).

Crossbreeding is a promising strategy in modern livestock breeding programs to meet the food demand of the growing population. Considerable genetic improvement in economical traits (e.g., meat and milk) can be obtained by implementing of an optimized and efficient crossbreeding program. Applying crossbreeding rather than traditional straight-breeding programs is to establish new composite by combining the positive attributes of two or more different breeds to produce an end product that fits market requirements and takes advantage of heterosis (Gregory & Cundiff 1980; Bertram *et al.* 1993; Simm 1998).

1-2 Genetic structure of crossbred populations

The genome of crossbred (admixed) individual is a mosaic of ancestral haplotypes formed by recombination occurring at every generation (Sankararaman *et al.* 2008; Price *et al.* 2009a). In a recently admixed population, ancestral populations have been mixing for a relatively small number of generations, resulting in a new population with different proportions of the original populations. Due to recombination events, the genome of admixed individuals is fragmented into shorter genome regions of different ancestries (Figure 1-1).

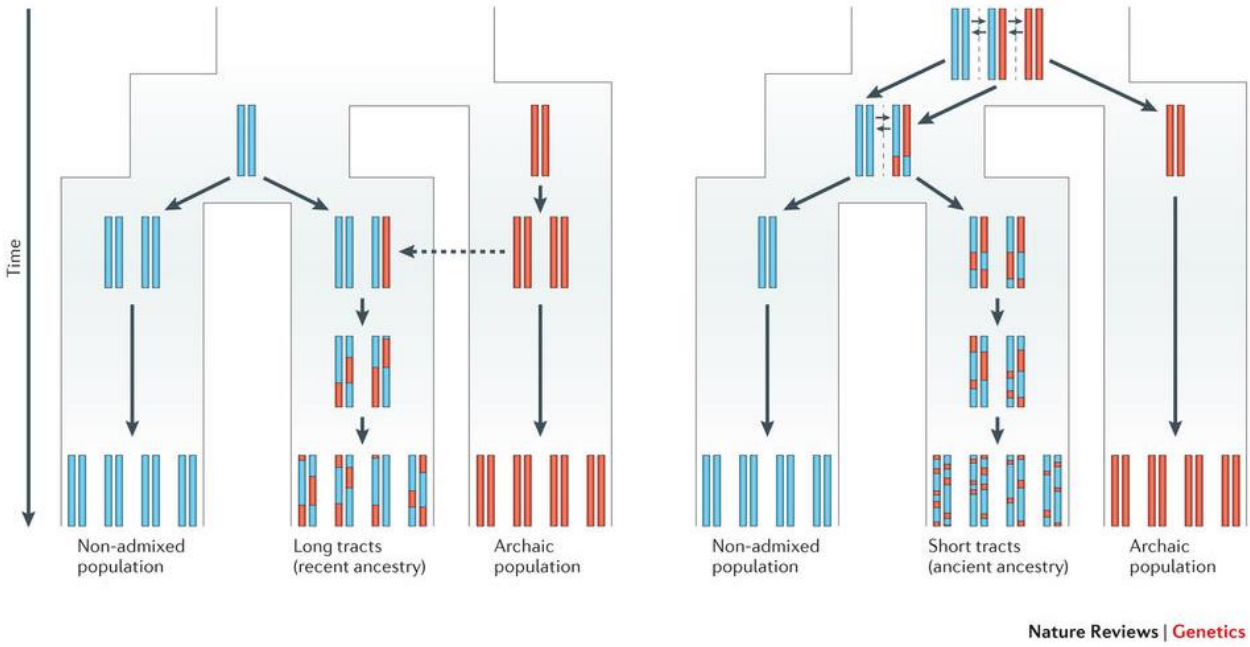


Figure 1-1 Expected lengths of ancestral segments in a recent admixed population (left panel) and in an ancient admixed population (right panel) (Racimo *et al.* 2015).

Estimation of the proportional contributions of ancestral populations in admixed (crossbred) individuals is important to clarify the population structure, historical background and pattern of admixture along the genome of admixed individuals. Recent advance in high-throughput genotype sequencing technology have provided unprecedented opportunities to learn about the evolutionary history of admixed populations at both global and local levels.

- **Global genetic ancestry** establishes ancestral proportions averaged across the genome of an individual.
- **Local genetic ancestry** is the identification of the ancestral origin of distinct chromosomal segments within an individual genome.

There is a growing concern in association studies about confounding effects, due to considerable discrepancy between the allele frequencies in the cases and the controls. An accurate inference of locus-specific ancestry in admixed populations has improved the genetic disease (Parkinson, Diabetic disease, Alzheimer and other diseases) association studies in human genetics (Sankararaman *et al.* 2008; Rosenberg *et al.* 2010; Seldin *et al.* 2011; Hu *et al.* 2013).

Moreover, admixed populations provide the special opportunity for studying recent selection signatures happened after admixture.

1-3 Selection signatures

In contrast to demographic processes of mutation, genetic drift and gene flow, which influence the entire genome, natural selection influences specific functionally important parts of genome (Bamshad & Wooding 2003; Oleksyk *et al.* 2010). Selection tends to cause specific changes in the patterns of variation among selected loci and in neutral linked loci as well, leaving its footprints in the adjacent chromosomal regions. There footprints are known as selection signatures (Kreitman 2000; Moradi *et al.* 2012; Gouveia *et al.* 2014).

The recent availability of high density single nucleotide polymorphism (SNP) markers and the development of analytical approaches offer the opportunity to screen the genome for evidence of selection. Analysis of F_{st} (Weir & Cockerham 1984), as the measure of genetic distance, was one of the first approaches to screen genome to detect the loci which exhibited high variation in allele frequency between populations. An alternative strategy is based on increased linkage disequilibrium (LD) and search for homozygous regions along the genome, where their frequency is more than expected (Voight *et al.* 2006; Tang *et al.* 2007b). This approach is called extended haplotype homozygosity (*EHH*) which is defined as the probability that two randomly chosen chromosomes carrying the core haplotype which are identical by decent (Sabeti *et al.* 2002).

In addition, the genome-wide distribution of ancestral segments in admixed individuals can be examined to detect selection signature happened after admixture (Tang *et al.* 2007a). Using admixed populations to study selection has a considerable history in human genetics, particularly analyzing African, Native American and Caucasian ancestries (Workman *et al.* 1963; Anderson & Reed 1969).

Under neutral evolution we expect each admixed individual's genome to represent an ensemble of ancestry blocks randomly sampled with a probability similar to genome wide average. However, ancestral contributions in the genome of recently admixed individuals vary at locus levels due to sampling error of the existing population, random evolutionary error of genetic drift and systematic biases of natural selection (Long 1991; Tang *et al.* 2007a; Oleksyk *et al.* 2010).

Extreme fluctuations in Δ ancestries, which are calculated by subtraction of the genome wide ancestry from locus-specific ancestry for each ancestry component, are unlikely to have occurred by chance and can exhibit a selection signature in admixed individuals (Tang *et al.* 2007a).

1-4 Methodologies for local ancestry estimates

The genome of an admixed individual comprises a mosaic of ancestral haplotypes formed by recombination occurring at every generation. The boundaries and origin of each ancestral segment can be reconstructed along each chromosome by statistical methods which can estimate ancestral allele and haplotype frequencies and their distribution in the admixed populations (Hu et al. 2013). Generally, the software tools for genetic ancestry estimates rely on multivariate statistical methods, using hidden Markov Models (HMM) and use the ancestral information as reference panel.

The approaches for local ancestry inference rely either on Li and Stephens (Li & Stephens 2003) framework, using an approximation to the coalescent with recombination, or on model-based clustering algorithms with no need to information on parametric population genetic model. Examples of algorithm using Li and Stephens (2003) include HAPMIX (Price *et al.* 2009b), LAMP-LD (Baran *et al.* 2012) and MULTIMIX (Churchhouse & Marchini 2013). Other method for local ancestry deconvolution is fundamentally based on the breaking genome into windows and then clustering relative to the reference panels; this is applied in LAMP (Sankararaman *et al.* 2008) and PCAdmix (Brisbin *et al.* 2012).

1-5 Heterosis; benefit of crossbreeding

Crossbred or admixed animals result from interbreeding, where sire and dam originate from different breeds or lines (Balding *et al.* 2007). An optimized crossbreeding program exploits the complementarity of the involved purebred parental populations based on breed additive genetic effects, termed “specific combining abilities” and makes use additional economic benefit of heterosis (Simm 1998; Gregory *et al.* 1999; Freyer *et al.* 2008).

Heterosis or hybrid vigor refers to the superiority in performance of the crossbreds compared to the average of their parents (Falconer & Mackay 1996). Heterosis and inbreeding depression are two manifestations of the same phenomenon, where the progenies of crossing of inbred lines show an increase of those characters suffering from inbreeding (Bourdon 1997; Lynch & Walsh 1998). Heterosis, like inbreeding depression, is most pronounced for fitness traits; fertility and longevity, all with relatively low heritability (Kristensen *et al.* 2005; Maki-Tanila 2007). Heterosis effects are intermediate for milk production, weight gain, feed efficiency, and body size; and lowest in carcass traits**.** Reproduction and maternal traits have low heritability and the traditional response to selection in breeding program will generally be slower compared to high heritability traits. However, significant improvement can be made through crossbreeding programs that maximize heterosis. For growth traits and milk traits with moderate heritability, genetic improvement can be achieved by applying both selection and crossbreeding programs. The amount of general heterosis for production traits in dairy cattle is reported 3 to 4 percent, while higher levels of heterosis are observed for functional and reproductive traits (VanRaden 2004; Freyer *et al.* 2008; Sorensen *et al.* 2008; Kargo *et al.* 2012).

1-6 Genetic basis of heterosis

The increased performance in crossbred animals is due to changes in non-additive genetic effects of dominance and epistasis components of heterosis. Dominance component of heterosis are caused by gene interaction within loci. The degree of heterosis for a specific breed combination, expressed in a crossbred animal, is equal to the chance that the animal, at a specific locus, has one gene from each of two breeds (Sorensen *et al.* 2008).

Epistatic effects are caused by gene interaction between loci and epistatic loss is considered as unfavorable gene effect in crossbred offspring due to breakdown of parental epistatic complex.

Under both natural and artificial selection, co-adapted positive gene complexes accumulate. However, favorable gene combinations established in the parental breeds may be lost by crossbreeding for traits that have been under selection. Different models for estimating effects of recombination caused by additive × additive (A×A) interaction have been proposed (Dickerson 1973; Hill 1982; Kinghorn 1983).

Kinghorn (1980, 1983) modeled dominance and two locus interaction, using epistatic term to describe effects of breakdown of parental combination. He considered heterozygosity is synonymous with dominance and epistatic loss is proportional to the probability that two non-allelic genes randomly chosen in diploid individual are of different breed origin.

The observed heterosis in first-generation crosses (F_1) is the sum of the dominance component of heterosis (normally positive) and the epistatic loss effects (normally negative). In a two breed crossbreeding program heterosis drops to 50% in the second-generation crosses (F_2) and continuing crosses between 2 breeds, 67% of the F1 heterosis will, on average, be expressed (Table 1-1). Including more parental breeds causes the more heterosis maintained after F_1 generations, while it causes the cross to be diluted for desired traits.

Amount of heterosis depends on degree of dominance and its direction, differences in allele frequencies of genetic variants contributed in heterosis between parents (genetic distance between parental population), number of involved parental populations and type of crossbreeding (Shull 1948; Falconer & Mackay 1996).

Table 1-1 Heterosis as a percentage of full heterosis in first generation (F_1) for different types of crosses

Type of cross	Heterosis %
F_1 (S×T)	100
F_2 (S×T) × (S×T)	50
Back cross S × (S×T) or T × (S×T)	50
Second generation of a rotational cross S × (T × (S×T))	75
Third generation of a rotational cross T × (S × (T × (S×T)))	62.5
Rotational cross after many generations	66.6

S and T denote on parental breeds, F_1 and F_2 denote first and second generations of crossbreeding

1-7 Crossbreeding and heterosis in dairy cattle

Crossbred offspring of more divergent and distant populations show more increased heterosis in comparison with offspring of crossing between more closely related populations. The most prominent example of crossing distant populations is *Bos Taurus* × *Bos Indicus*.

Results of designed studies in North America indicated that in crossbreeding programs involving the Holstein with Guernsey and Ayrshire, crossbred cows had better performance (e.g., milk fat and protein percentage, growth, survival, reproduction and lifetime yield) and exceeded pure Holsteins on the basis of income (Touchberry 1992; McAllister *et al.* 1994). Crossbreeding of Jersey cows in New Zealand and Australia, and Brown Swiss cow with Holsteins bulls are common, where breeders believe they can benefit from heterosis to increase protein and fat percentage with less calving difficulty and better morphology traits (leg and feet) in crossbred cows (VanRaden & Sanders 2003).

European dual purpose breeds are also good candidates for crossbreeding with more adapted local breeds in tropical climate (Mcdowell 1985). Crossing of European dual purpose breeds (Normande, Montbeliarde and Scandinavian Red as three dominant European breeds) with Holstein causes extra interest due to heterosis for milk components and to combine the better reproduction traits of European breeds. European breeds are typically good in survival rates in the first lactation and have better calving ease, conception rate and overall health as well (Heins *et al.* 2004; Heins *et al.* 2012).

Swiss Fleckvieh as a recent composite cattle breed is another example of crossbreeding between Holstein and European breeds. Swiss Fleckvieh, which is the case-study population in this thesis, is a dual-purpose cattle breed, with emphasis on milk production, as a composite of Red Holstein Friesian and Swiss Simmental. Its main breeding area is in Western and Northern Switzerland, suitable for grassland and grazing. The breeding program started with the purchase of Red Holstein Friesian bulls for artificial insemination of Simmental cows in the 1970s. The blood components of today's Swiss Fleckvieh population correspond to about 2/3 Red Holstein Friesian and 1/3 Simmental blood. Despite their similar origins with the Swiss Red Holstein population, Swiss Fleckvieh's focus is different. The basic objective is the combination of the economic advantages of the milk production of Red Holsteins Friesians as well as additional purposes such

as reproduction traits, beef value, fitness and longevity of Swiss Simmental origin (Swissherdbook Zollikofen, 2016).

1-8 Aim and objectives

The overall aim of this thesis was to analyze the genome of a composite cattle breed to detect selection signatures and estimate components of heterosis based on local genetic ancestries. Local genetic ancestries, using genomic markers and different software tools were estimated and compared. Local genetic ancestries were used to infer post-admixture selection signatures. Furthermore, local genetic ancestries were used to define coefficients of heterosis components (dominance and epistatic loss) to estimate these effects for semen traits in admixed Swiss Fleckvieh bulls. Finally, the dominance component of heterosis for percentage of live sperm was mapped to genomic regions contributing to heterosis.

Objectives

1) To perform a whole genome scan for selection signature in Swiss Fleckvieh bulls based on three different approaches:
   - Detection local excess or deficiency of ancestry from genome wide average ancestry
   - Examination of extended haplotype homozygosities (*EHH*)
   - Genetic distance between pure ancestral populations (F_{st})
2) To compare the results on local ancestry's estimates from different approaches to investigate how much the choice of different methods as well as parameter setting of the applied software can influence the estimations.
3) To define the heterosis components using local ancestries and estimate theses effects for semen traits in admixed bulls, using different models.
4) To use heterosis mapping to investigate whether specific genomic areas, contributing to heterosis, are associated with spermatogenesis in admixed bulls.

Thesis outline

This PhD thesis consists of four manuscripts (Chapters 2, 3, 4 and 5) with emphasis on study of genetic architecture of a composite cattle breed in order to: 1) detect selection signatures in admixed populations; 2) examine how much the choice of different algorithms can influence the local ancestry inferences; 3) estimate of non-additive genetic components of heterosis, and 4) perform heterosis mapping. The studied populations consisted of admixed Swiss Fleckvieh bulls and its two ancestral populations (e.g., Red Holstein Friesian and Simmental).

Chapter 1 provides general introduction, including introductory parts, aims and outline of thesis.

Chapter 2 focusses on the estimation of genetic ancestries at SNP level. Deviations of local genetic ancestries from genome-wide ancestry are calculated to detect extreme excess or deficiency along the genome, which were interpreted as selection signals in admixed population. To cope with defining the significant threshold for signals, permutation tests and extreme deviations from normal distribution for multiple hypotheses tests are performed. Extended haplotype homozygosity (*EHH*) was calculated to explore additional patterns of pre- and post-admixture selection signals. Genetic differentiation between two ancestral populations (F_{st}) is used as an alternative indicator of pre-admixture selection signal to find overlap with local ancestry.

In **Chapter 3** the results of how applying various algorithms can influence local ancestry estimations are presented. Genotypes of reference panel and admixed bulls are phased with two different approaches, implemented by *Shape-It* (Delaneau *et al.* 2012) and *AlphaPhase* (Hickey *et al.* 2011). Total 361 analyses are performed, using different settings and input files by three software tools (LAMP, LAMP-LD and MULTIMIX). The results of these programs are compared with the results of **Chapter 2** and used further to search for similar selection signatures in **Chapter 2** and any other signals.

Chapter 4 discusses about different models to estimate non-additive genetic components of heterosis. Genomic data based on local genetic ancestry, estimated by LAMP (**Chapter 2**) is used to define dominance and epistatic loss, with two definitions (Kinghorn, 1983 and our definition) for semen traits (volume, concentration, number of spermatozoa and percentage of

live sperm) in Swiss Fleckvieh bulls. Accuracy of different models of estimating heterosis components is compared.

In **Chapter 5** we use method suggested for inbreeding mapping (Ferencakovic *et al.* 2017) to study heterosis at genome-wide level for one of the semen traits (percentage of live sperm) in **Chapter 4**. Then we search for possible genes are associated with spermatogenesis in cattle.

Chapter 6 provides a critical reflection of results of previous chapters as well as suggestions for future work in the fields of local admixture, admixture selection signatures and heterosis of crossbred livestock.

References

Anderson E.N., Jr. & Reed T.E. (1969) Caucasian genes in American Negroes. *Science* **166**, 1353.

Balding D.J., Bishop M.J. & Cannings C. (2007) *Handbook of statistical genetics*. John Wiley, Chichester.

Bamshad M. & Wooding S.P. (2003) Signatures of natural selection in the human genome. *Nature Reviews Genetics* **4**, 99-111.

Baran Y., Pasaniuc B., Sankararaman S., Torgerson D.G., Gignoux C., Eng C., Rodriguez-Cintron W., Chapela R., Ford J.G., Avila P.C., Rodriguez-Santana J., Burchard E.G. & Halperin E. (2012) Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **28**, 1359-67.

Bertram J., Carrick M., Holroyd D., Lake M., Lehman W., Taylor K., Thompson R., Tierney M., Tyler R., Sullivan M., White R., Davis G. & Burrow H. (1993) Breeding for Profit. *The State of Queensland, Department of Primary Industries*.

Bourdon R.M. (1997) *Understanding animal breeding*. Prentice Hall, Upper Saddle River, N.J. ; London.

Brisbin A., Bryc K., Byrnes J., Zakharia F., Omberg L., Degenhardt J., Reynolds A., Ostrer H., Mezey J.G. & Bustamante C.D. (2012) PCAdmix: principal components-based

assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum Biol* **84**, 343-64.

Churchhouse C. & Marchini J. (2013) Multiway admixture deconvolution using phased or unphased ancestral panels. *Genet Epidemiol* **37**, 1-12.

Delaneau O., Marchini J. & Zagury J.F. (2012) A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179-81.

Dickerson G.E. (1973) Inbreeding and heterosis in animals. *Proc. Anim. Breed. and Genet. Symp. in Honor of Dr. J. L. Lush*, 24.

Falconer D.S. & Mackay T.F.C. (1996) *Introduction to quantitative genetics*. Longman, Harlow.

FAO (2011) World Livestock 2011 – Livestock in food security. *Rome, FAO*.

FAO (2017) The future of food and agriculture - Trends and challenges. *Rome*.

Ferencakovic M., Solkner J., Kaps M. & Curik I. (2017) Genome-wide mapping and estimation of inbreeding depression of semen quality traits in a cattle population. *Journal of Dairy Science* **100**, 4721-30.

Freyer G., Konig S., Fischer B., Bergfeld U. & Cassell B.G. (2008) Invited review: crossbreeding in dairy cattle from a German perspective of the past and today. *Journal of Dairy Science* **91**, 3725-43.

Gouveia J.J.D., da Silva M.V.G.B., Paiva S.R. & de Oliveira S.M.P. (2014) Identification of selection signatures in livestock species. *Genet Mol Biol* **37**, 330-42.

Gregory K.E. & Cundiff L.V. (1980) Crossbreeding in Beef-Cattle - Evaluation of Systems. *Journal of Animal Science* **51**, 1224-42.

Gregory K.E., Cundiff L.V., Koch R.M., United States. Agricultural Research Service. & University of Nebraska--Lincoln. Institute of Agriculture and Natural Resources. (1999) *Composite breeds to use heterosis and breed differences to improve efficiency of beef production*. U.S. Dept. of Agriculture, Available from National Technical Information Service, Washington, D.C.Springfield, VA.

Groeneveld L.F., Lenstra J.A., Eding H., Toro M.A., Scherf B., Pilling D., Negrini R., Finlay E.K., Jianlin H., Groeneveld E., Weigend S. & Consortium G. (2010) Genetic diversity in farm animals--a review. *Animal Genetics* **41 Suppl 1**, 6-31.

Heins B.J., Hansen L.B. & De Vries A. (2012) Survival, lifetime production, and profitability of Normande x Holstein, Montbeliarde x Holstein, and Scandinavian Red x Holstein crossbreds versus pure Holsteins. *Journal of Dairy Science* **95**, 1011-21.

Heins B.J., Hansen L.B. & Seykora A.J. (2004) Comparison of first-parity Holstein, Normande-Holstein crossbred, Montbeliarde-Holstein crossbred and Scandinavian-Holstein crossbred cows for dystocia and stillbirths. *Journal of Dairy Science* **87**, 282-.

Hickey J.M., Kinghorn B.P., Tier B., Wilson J.F., Dunstan N. & van der Werf J.H. (2011) A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet Sel Evol* **43**, 12.

Hill W.G. (1982) Dominance and epistasis as components of heterosis. *Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie-Journal of Animal Breeding and Genetics* **99**, 8.

Hu Y.N., Willer C., Zhan X.W., Kang H.M. & Abecasis G.R. (2013) Accurate Local-Ancestry Inference in Exome-Sequenced Admixed Individuals via Off-Target Sequence Reads. *American Journal of Human Genetics* **93**, 891-9.

Kargo M., Madsen P. & Norberg E. (2012) Short communication: Is crossbreeding only beneficial in herds with low management level? *Journal of Dairy Science* **95**, 925-8.

Kinghorn B. (1980) The Expression of Recombination Loss in Quantitative Traits. *Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie-Journal of Animal Breeding and Genetics* **97**, 138-43.

Kinghorn B. (1983) Genetic-Effects in Crossbreeding .3. Epistatic Loss in Crossbred Mice. *Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie-Journal of Animal Breeding and Genetics* **100**, 209-22.

Kreitman M. (2000) Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics* **1**, 539-59.

Kristensen T.N., Sorensen A.C., Sorensen D., Pedersen K.S., Sorensen J.G. & Loeschcke V. (2005) A test of quantitative genetic theory using Drosophila- effects of inbreeding and rate of inbreeding on heritabilities and variance components. *J Evol Biol* **18**, 763-70.

Li N. & Stephens M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213-33.

Long J.C. (1991) The genetic structure of admixed populations. *Genetics* **127**, 417-28.

Lynch M. & Walsh B. (1998) *Genetics and analysis of quantitative traits*. Sinauer, Sunderland, Ma.

Maki-Tanila A. (2007) More precise utilization of dominance variation. *Journal of Animal Breeding and Genetics* **124**, 175.

McAllister A.J., Lee A.J., Batra T.R., Lin C.Y., Roy G.L., Vesely J.A., Wauthy J.M. & Winter K.A. (1994) The influence of additive and nonadditive gene action on lifetime yields and profitability of dairy cattle. *Journal of Dairy Science* **77**, 2400-14.

Mcdowell R.E. (1985) Crossbreeding in Tropical Areas with Emphasis on Milk, Health, and Fitness. *Journal of Dairy Science* **68**, 2418-35.

Moradi M.H., Nejati-Javaremi A., Moradi-Shahrbabak M., Dodds K.G. & McEwan J.C. (2012) Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. *Bmc Genetics* **13**.

Notter D.R. (1999) The importance of genetic diversity in livestock populations of the future. *Journal of Animal Science* **77**, 61-9.

Oleksyk T.K., Smith M.W. & O'Brien S.J. (2010) Genome-wide scans for footprints of natural selection. *Philos Trans R Soc Lond B Biol Sci* **365**, 185-205.

Price A.L., Tandon A., Patterson N., Barnes K.C., Rafaels N., Ruczinski I., Beaty T.H., Mathias R., Reich D. & Myers S. (2009a) Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *Plos Genetics* **5**.

Price A.L., Tandon A., Patterson N., Barnes K.C., Rafaels N., Ruczinski I., Beaty T.H., Mathias R., Reich D. & Myers S. (2009b) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* **5**, e1000519.

Racimo F., Sankararaman S., Nielsen R. & Huerta-Sánchez E. (2015) Evidence for archaic adaptive introgressionin humans. *Nature Reviews Genetics* **16**, 359-371.

Rosenberg N.A., Huang L., Jewett E.M., Szpiech Z.A., Jankovic I. & Boehnke M. (2010) Genome-wide association studies in diverse populations. *Nature Reviews Genetics* **11**, 356-66.

Sabeti P.C., Reich D.E., Higgins J.M., Levine H.Z., Richter D.J., Schaffner S.F., Gabriel S.B., Platko J.V., Patterson N.J., McDonald G.J., Ackerman H.C., Campbell S.J., Altshuler D., Cooper R., Kwiatkowski D., Ward R. & Lander E.S. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-7.

Sankararaman S., Sridhar S., Kimmel G. & Halperin E. (2008) Estimating local ancestry in admixed populations. *American Journal of Human Genetics* **82**, 290-303.

Seldin M.F., Pasaniuc B. & Price A.L. (2011) New approaches to disease mapping in admixed populations. *Nature Reviews Genetics* **12**, 523-8.

Shull G.H. (1948) What Is Heterosis. *Genetics* **33**, 439-46.

Simm G. (1998) *Genetic improvement of cattle and sheep*. Farming, Ipswich.

Sorensen M.K., Norberg E., Pedersen J. & Christensen L.G. (2008) Invited review: crossbreeding in dairy cattle: a Danish perspective. *Journal of Dairy Science* **91**, 4116-28.

Swissherdbook Zollikofen (2016). retrieved from https://www.swissherdbook.ch/unsere-rassen/swiss-fleckvieh/

Tang H., Choudhry S., Mei R., Morgan M., Rodriguez-Cintron W., Burchard E.G. & Risch N.J. (2007a) Recent genetic selection in the ancestral admixture of Puerto Ricans. *American Journal of Human Genetics* **81**, 626-33.

Tang K., Thornton K.R. & Stoneking M. (2007b) A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol* **5**, e171.

Touchberry R.W. (1992) Crossbreeding effects in dairy cattle: the Illinois Experiment, 1949 to 1969. *Journal of Dairy Science* **75**, 640-67.

VanRaden P.M. (2004) Invited review: selection on net merit to improve lifetime profit. *Journal of Dairy Science* **87**, 3125-31.

VanRaden P.M. & Sanders A.H. (2003) Economic merit of crossbred and purebred US dairy cattle. *Journal of Dairy Science* **86**, 1036-44.

Voight B.F., Kudaravalli S., Wen X. & Pritchard J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72.

Weir B.S. & Cockerham C.C. (1984) Estimating F-Statistics for the Analysis of Population Structure. Evolution 38, 1358-70.

Workman P.L., Blumberg B.S. & Cooper A.J. (1963) Selection, Gene Migration and Polymorphic Stability in a U. S. White and Negro Population. American Journal of Human Genetics 15, 429-37.

Chapter 2

Locus-specific ancestry to detect recent response to selection in admixed Swiss Fleckvieh cattle

N. Khayatzadeh[*], G. Mészáros[*], Y. T. Utsunomiya[†], J. F. Garcia[†‡], U. Schnyder[§], B. Gredler[§], I. Curik[¶] and J. Sölkner[*]

[*]Division of Livestock Science, Department of Sustainable Agricultural Systems, University of Natural Resources and Life Sciences, Vienna, Gregor-Mendel-Straße 33, A-1180, Vienna, Austria.

[†]Departamento de Medicina Veterinária Preventiva e Reprodução Animal, Faculdade de Ciências Agrárias e Veterinárias, UNESP – Univ Estadual Paulista, Jaboticabal, São Paulo, Brazil.

[‡]Departamento de Apoio, Saúde e Produção Animal, Faculdade de Medicina Veterinária de Araçatuba, UNESP – Univ Estadual Paulista, Araçatuba, São Paulo, Brazil.

[§]Qualitas AG, Chamerstrasse 56, CH-6300, Zug, Switzerland.

[¶]Department of Animal Science, Faculty of Agriculture, University of Zagreb, Svetošimunska cesta 25, 10000 Zagreb, Croatia

Animal Genetics (2016) 47(6):637-646

Summary

Identification of selection signatures is one of the current endeavors of evolutionary genetics. Admixed populations may be used to infer post-admixture selection. We calculated local ancestry for Swiss Fleckvieh, a composite of Simmental (SI) and Red Holstein Friesian (RHF), to infer such signals. Illumina BovineSNP50 BeadChip data for 300 admixed, 88 SI and 97 RHF bulls were used. The average RHF ancestry across the whole genome was 0.70. To identify regions with high deviation from average, we considered two significance thresholds, based on permutation test and extreme deviation from normal distribution. Regions on chromosomes 13 (46.3-47.3 Mb) and 18 (18.7-25.9 Mb) passed both thresholds in direction of increased SI. Extended Haplotype Homozygosity within (*iHS*) and between (*Rsb*) populations was calculated to explore additional patterns of pre- and post-admixture selection signals. The *Rsb* score of admixed and SI was significant in a wide region of chromosome 18 (6.6-24.6 Mb) overlapped with one area of strong local ancestry deviation. *FTO*, with pleiotropic effect on milk and fertility, *NOD2* on dairy, *NKD1* and *SALL1* on fertility traits are located there.

Genetic differentiation of RHF and SI (F_{st}), an alternative indicator of pre-admixture selection in pure populations, was calculated. No considerable overlap of peaks of local ancestry deviations and F_{st} was observed.

We found two regions with significant signatures of post-admixture selection in this very young composite, applying comparatively stringent significance thresholds. The signals cover relatively large genomic areas and do not allow pinpointing the gene(s) responsible for the apparent shift in ancestry proportions.

Keywords admixture, extended haplotype homozygosity, F_{st}, *iHS*, local ancestry deviation, permutation, *Rsb*, selection signature, SNP, Swiss Fleckvieh

2-1 Introduction

Genetic exchange between two or more previously separated populations causes admixture of genetic material (Decker *et al*. 2009; Racimo *et al*. 2015). With limited number of recombinations taking place each generation, a mosaic of ancestral segments is formed in the genome of admixed individuals (Sankararaman *et al*. 2008; Hu *et al*. 2013; Zhang & Stram 2014).

Estimates of ancestry proportions at population and individual levels are widely used to study the population structure in many species and breed composition in livestock. Various methods have been developed to estimate global (genome-wide) ancestry. Principal component analysis (PCA) is frequently used to infer genetic structure in domesticated cattle breeds (Bovine Genome *et al*. 2009; Gautier *et al*. 2010). Frkonja *et al*. (2012) estimated the global admixture proportions in Swiss Fleckvieh cattle with model-based clustering, partial least squares and Bayesian regression. McTavish *et al*. (2013) studied the population structure of some US cattle breeds using PCA and model-based clustering. Decker *et al*. (2014) investigated the population structure of domesticated cattle and calculated Asian indicine (*Bos indicus*), Eurasian taurine and African taurine (both *Bos taurus*) ancestry proportions using similar procedures.

In global ancestry estimation, similar ancestry proportions contributed by each pure population on each locus are implicitly assumed (Long 1991). However, admixture proportions vary among loci and local ancestry of admixed individuals deviates from global admixture because of a very limited number of potential ancestral configurations (0, 0.5 and 1 for any ancestry) when considering the two alleles of a single locus. From an evolutionary perspective, the most important sources of variation in admixture estimates are genetic drift, gene flow and selection (Long 1991; Tang *et al*. 2007a; Jin *et al*. 2012; Jones & Wang 2012). In contrast to the demographic process of genetic drift and gene flow influencing the whole genome, selection targets only functional elements in specific gene regions (Oleksyk *et al*. 2010).

When selection acts in an admixed population, selected alleles are expected to have higher frequencies after some generations of admixture, causing local ancestry to deviate from genome-wide average (Bhatia *et al*. 2014). These deviations in the genome of admixed individuals (excess or deficiency) can be used to detect signals of recent selection response. The effect on

those regions is cumulative over several generations and therefore may be interpreted as signals of selection after admixture (Long 1991; Tang *et al*. 2007a). Genetic drift as a random source of variation after admixture may also produce large deviations in local ancestry (here considered as noise) and thus should be accounted for as a factor influencing the local ancestry (Tang *et al*. 2007a; Oleksyk *et al*. 2010; Gautier & Naves 2011; Bhatia *et al*. 2014).

Since the advent of high throughput single nucleotide polymorphism (SNP) genotyping, inferring selection signatures from differences in local admixture levels has received considerable attention in human genetics (Tang *et al*. 2007a; Jin *et al*. 2012; Bhatia *et al*. 2014). Similar studies in livestock investigated local ancestry levels of New World Creole cattle (Gautier & Naves 2011; Flori *et al*. 2014) and selection signatures in dairy cattle in East Africa, resulting from admixture of European breeds (Kim & Rothschild 2014), and in East African short horn Zebu (Bahbahani *et al*. 2015). The ancestry proportions of indicine Zebu, considered trypano-susceptible, and taurine Baoule, trypano-tolerant, in trypanosoma tolerance candidate regions versus the background genome were calculated for admixed cattle in Burkina Faso (Smetko *et al*. 2015).

Crossbreeding is one of the key concepts of modern livestock breeding. Systematic breeding for distinct characteristics and classification of livestock species into breeds commenced only around 250-300 years ago (Felius *et al*. 2011). Systems of terminal crossbreeding, with first or second generation crossbreds producing the bulk of marketable livestock products, are very common, particularly in pig and poultry. Alternatively, breeders make use of heterosis and breed complementarity by forming composite populations by initially crossing parental breeds and then performing mating of the crosses (Felius *et al*. 2015).

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 traits, beef value, fitness and longevity of the Simmental breed.

In this study, we searched for post-admixture signals of selection by applying local ancestry deviation suggested by Tang *et al*. (2007a), Gautier and Naves (2011) and Bhatia *et al*. (2014). In addition, we estimated extended haplotype homozygosities (*EHH*) within each ancestral and

admixed populations, as well as between populations (RHF vs. SI, RHF and SI vs. admixed populations) to detect pre- and post-admixture signals of selection (Sabeti *et al*. 2002; Voight *et al*. 2006; Tang *et al*. 2007b). We also investigated whether regions with strong signals of post-admixture selection coincided with pre-admixture signatures of selection, based on population differentiation (F_{st}) of the parental breeds.

2-2 Materials and methods

The genotype data from the Illumina Bovine SNP50 BeadChip were available for 101 pure RHF, 91 pure SI and 308 admixed bulls, provided by Swissherdbook cooperative Zollikofen. Doses of sperm routinely collected for artificial insemination provided the tissue used for genotyping; therefore no ethics statement was required for collecting genetic material. Formally, animals are categorized Swiss Fleckvieh when their pedigree admixture level is 0.125-0.875 RHF, animals with < 0.125 RHF are part of the Simmental section of the herd book and those > 0.875 are in the Holstein Friesian section of the herd book. For the purpose of our study, we did not respect this definition and considered all admixed animals along the range of pedigree composition of 0.02-0.99 RHF. Quality control of the data was performed with PLINK 1.07 (Purcell *et al*. 2007). The dataset was controlled to exclude monomorphic SNPs, those with call rate of < 95% and those that deviated from Hardy Weinberg Equilibrium with Fisher's exact *P*-value < 10^{-6}. Animals with more than 5% missing genotypes, SNPs on sex chromosomes and no location information were also removed. After applying the quality control criteria 39 525 SNPs and 485 (97 RHF, 300 admixed, 88 SI) animals were left for the analysis.

Unsupervised global ancestry estimation was performed with the full SNP set using the ADMIXTURE software (Alexander *et al*. 2009) with the number of ancestral populations fixed at two. To estimate local ancestry for admixed animals, the LAMP 2.5 program (Sankararaman *et al*. 2008) was used in LAMPANC mode. In LAMP configuration, we defined a constant recombination rate of 1e-8 based on the assumption that 0.01 recombination occur per Mb (equivalent to 1 cM), given that no accurate genetic map is currently available for cattle. As the SNP density is not very high in the 50k chip, we did not include Linkage Disequilibrium (LD) in this analysis. The locus-specific ancestry was estimated for each admixed individual with respect to pure breeds, representing the proportion of each involved ancestry (0, 0.5, and 1) for each

SNP. For 300 admixed animals, we computed the average locus-specific ancestry level across each chromosome separately.

Furthermore, we searched for the excess and deficiency of local ancestry with respect to RHF breed using the approach proposed by Tang *et al*. (2007a), using LAMP similar to Gautier and Naves (2011).

The 'Δ ancestry' was calculated by subtracting the genome wide ancestry as a baseline from the average locus-specific ancestry for each of the two ancestry components. The Δ ancestry for ancestral population *k* at each SNP *m* is defined as:

$$\delta_k^m = \frac{1}{I}\sum_{i=1}^{I}\left(q_k^{i,m} - \bar{q}_k^i\right) = \tilde{q}_k^m - \bar{q}_k$$

where $q_k^{i,m}$ is the locus-specific ancestry of animal *i* at SNP *m*, estimated by LAMP, $\bar{q}_k^i$ is mean of locus-specific ancestry for individual *I*, $\tilde{q}_k^m$ is the mean of ancestry at SNP *m* averaged over all admixed animals; and $\bar{q}_k$ is the mean of locus-specific ancestry across the whole genome for admixed population *k*.

We scaled δ_k^m values by their standard deviation (SD, 0.040). On the basis of the extent of admixture LD in admixed populations, we determined genome-wide threshold of signals of selection by correction for multiple hypothesis testing (based on Bonferroni correction) assuming 5000 and 1000 independent segments along the whole genome. Local ancestry deviations > 4.42 SDs (*P*-value < 1×10^{-5}) corresponding to 5000 hypotheses and > 4.06 SDs (*P*-value < 5×10^{-5}) corresponding to 1000 hypotheses tested were considered significant, following the study of human admixture by Bhatia *et al*. (2014). As LD is higher and therefore the number of independent segments of the genome is smaller in bovine populations compared to human populations, we consider these thresholds conservative (Hayes *et al*. 2003).

Moreover, we performed permutation tests (Doerge & Churchill 1996) to evaluate the significance level for the excess or deficiency of the SNPs over the whole genome of admixed animals for each pure ancestry proposed by Tang *et al*. (2007a), and implemented in cattle by Gautier and Naves (2011) and poultry by Qanbari *et al*. (2012). For each animal, we concatenated the local ancestry estimations of all 29 autosomes and then permuted the circularized genome by cutting at a random location and rearranging the two resulting pieces of

the genome for each individual independently. This type of permutation preserves the extent of LD, assuming that it is homogeneously distributed over the whole genome. We implemented 20 000 permutations and further added a percentage quantile transformation step. In each permutation test, the SD of the distribution of the permuted statistics (trimming 0.05 end of each test) was multiplied by a scale factor to match with corresponding observed distribution. We further computed the minimum and maximum values of each permutation. The distributions of maximum and minimum over all permutations were then used to define 1% and 5% threshold levels that indicated significant deviation of the observed local ancestries from the genome-wide average ancestry (Tang *et al*. 2007a; Gautier & Naves 2011).

For the next approach of finding the selection signals in admixed animals, we calculated *iHS* and *Rsb* statistics suggested by Sabeti *et al*. (2002); (Voight *et al*. 2006; Tang *et al*. 2007b). At first we phased our dataset with SHAPEIT *v2.r790* (Delaneau *et al*. 2012) and then used REHH package in R developed by Gautier and Vitalis (2012) with some minor adaptations (Utsunomiya *et al*. 2013). The two scores measure the segments through the genome that show unexpected high levels of haplotype homozygosity within and between populations respectively. We standardized *iHS* to have mean 0 and variance 1. Moreover, in *Rsb*, the ratio of corresponding populations (RHF/SI, admixed/RHF and admixed/SI cluster at each SNP site was calculated and transformed to logarithmic form, then standardized with mean 0 and variance 1. Because we are interested in both tails of this distribution, two-sided *P*-values were calculated based on Gaussian cumulative density function. Following Bhatia *et al.* (2014) and extension of LD in the genome of admixed individuals, we decided to use two thresholds, using Bonferroni thresholds based on 5000 and 1000 hypothesis tests as significance levels.

Allele frequency differentiation between two original populations can provide information on selection before admixture. F_{st} statistics have been calculated to detect selection signatures in cattle (Mancini *et al*. 2014; Bahbahani *et al*. 2015; Zhao *et al*. 2015) and sheep populations (Moradi *et al*. 2012). We calculated locus-specific F_{st} (Weir & Cockerham 1984) using the DIVERSITY R package (Keenan *et al*. 2013). F_{st} values were averaged for 500-kb windows on each chromosome to identify candidate regions of high Δ ancestry after admixture and to check whether these regions were also regions of high differentiation in the ancestral breeds.

2-3 Results

The individual admixture proportions using the full set of 39 525 SNPs were estimated for all pure and admixed animals. Individual admixture levels based on SNP chip data calculated by ADMIXTURE are presented in Fig. 2-1 with animals ordered from the highest to lowest RHF ancestry proportions according to the pedigree information. The average ancestry proportions were estimated at 0.68 RHF and 0.32 SI (SD = 0.19).

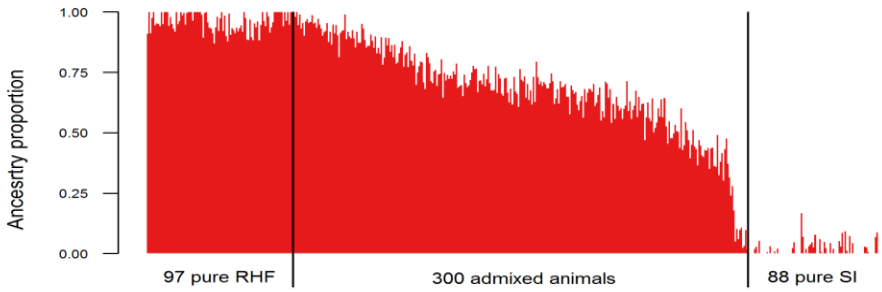


Figure 2-1 Ancestry proportions for all animals with the full set of 39 525 single nucleotide polymorphisms (SNPs).

The average ancestry estimation for every single SNP was performed across 29 autosomes separately with LAMP. In the body of this paper we show the results for 6 autosomal chromosomes (2, 3, 7, 13, 18 and 29) giving the most extreme patterns of Δ ancestry; information for all chromosomes is provided in Fig. S2-1. The average ancestry across chromosomes 2, 3, 7, 13, 18 and 29 for all 300 admixed animals are shown in Fig. 2-2, keeping the same order based on the pedigree RHF ancestry as in Fig. 2-1.

The average RHF ancestry along the whole genome was calculated to be 0.70 (SD = 0.04), taking the average of all SNPs across the 29 autosomes. The RHF ancestry was also calculated across each chromosome by averaging the ancestry proportions of all SNPs across each chromosome.

The average RHF ancestry proportion across chromosome 2 was estimated at 0.73, which is larger than average RHF ancestry across the whole genome (0.70). Average RHF ancestry estimates along chromosomes 3 and 7 were 0.71 and 0.70 respectively, close to the average genome-wide ancestry. On the other hand, the average RHF ancestry levels on chromosomes 13, 18 and 29 were estimated 0.65, 0.63 and 0.67 respectively.

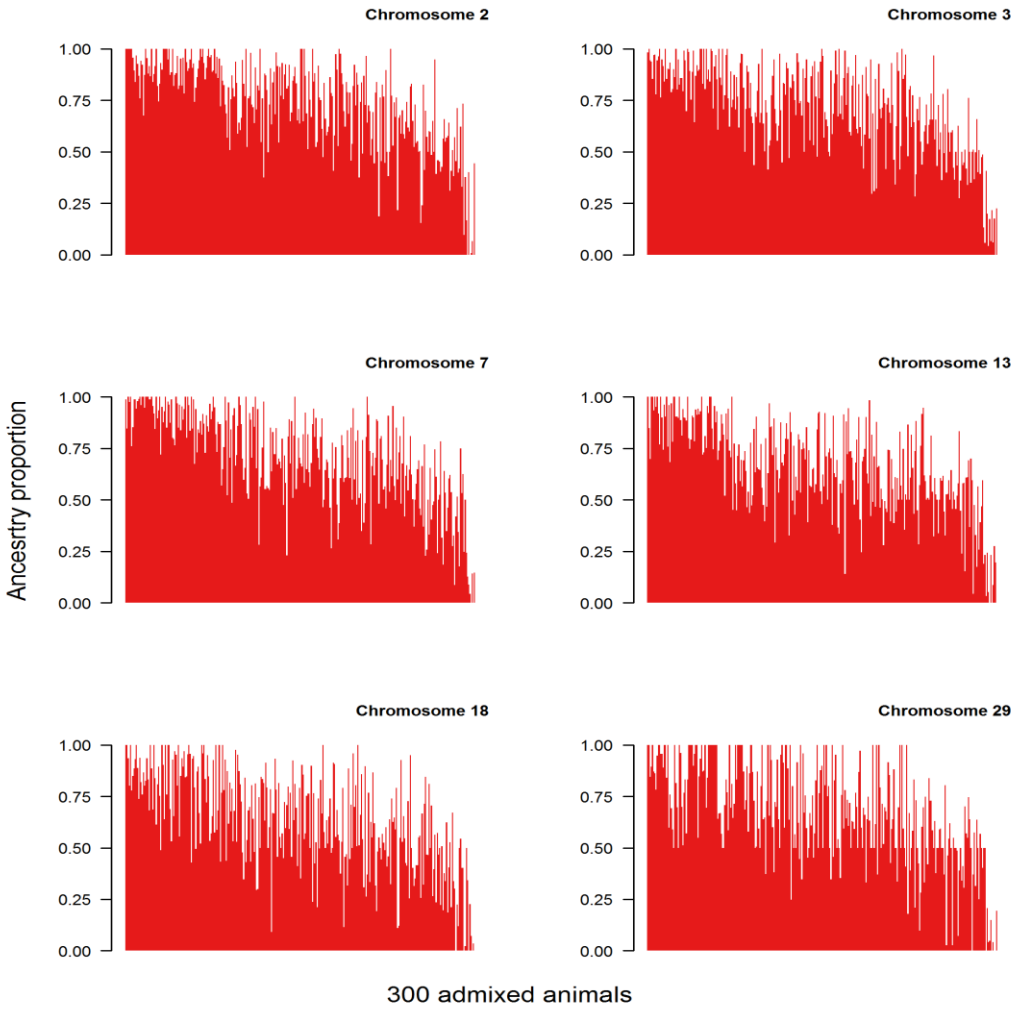


Figure 2-2 Average ancestry proportions across chromosomes 2, 3, 7, 13, 18 and 29 for all 300 admixed animals as determined by LAMP.

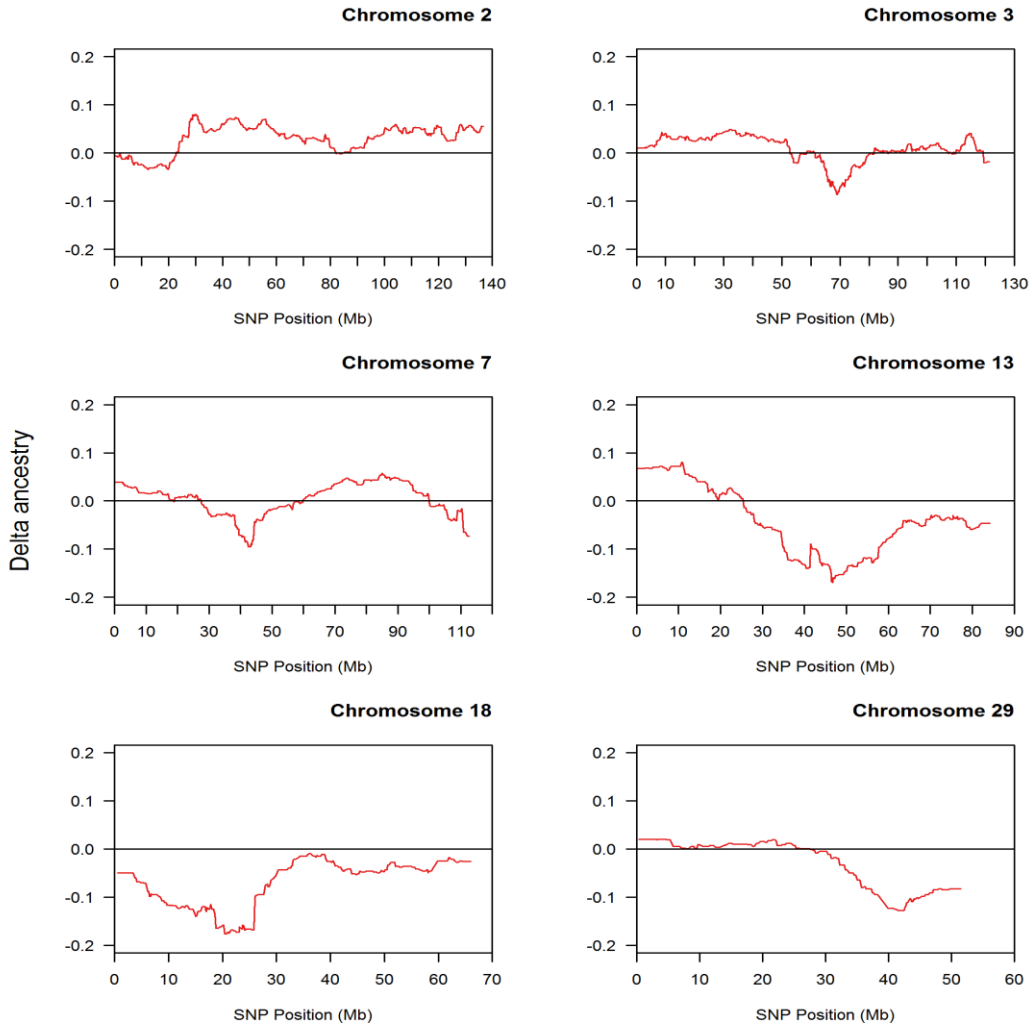


Figure 2-3 Local ancestry deviations from average of genome-wide ancestry at the corresponding single nucleotide polymorphism (SNP) positions on chromosomes 2, 3, 7, 13, 18 and 29, averaged over all 300 admixed animals.

The average excess or deficiency of RHF local ancestry from the average locus-specific ancestry across the whole genome on chromosomes 2, 3, 7, 13, 18 and 29 for all admixed animals is shown in Fig. 2-3. On chromosome 2 some wide peaks of Δ ancestry in favor of RHF (0.07-0.08) are observable between 28.2-30.8 Mb, 41.7-45.9 Mb and 55.6-56.1 Mb. On chromosome 3, excess in favor of SI ancestry (0.09) around 68.8-69.1 Mb was detected. The excess level in favor of SI on chromosome 7 reached 0.10 at 42.4-43.3 Mb.

In comparison, most notable peaks were seen on chromosomes 13 and 18. On chromosome 13, excess of RHF ancestry (0.07-0.08) was detected in a wide region along the first part of the chromosome (1-11 Mb) and the excess of SI (i.e., deficiency in RHF) ancestry with a very wide peak (0.10-0.17) around 35.2-57.6 Mb was observed. The most extreme values (0.16-0.17) were located at 46.3-47.3 Mb.

Likewise, a wide peak (0.10-0.18) on chromosome 18 between 8.0 and 26.1 Mb was detected. Extreme excess in favor of SI reached 0.17-0.18 on chromosome 18 at 20.4-23.1 Mb. Another relatively wide peak in excess of SI (0.10-0.13) was on chromosome 29 (39.9-45.1 Mb).

Genome-wide graphs of Δ ancestry (a) at the original scale and (b) scaled by SDs are provided in Fig. 2-4. The 5% and 1% genome-wide significance thresholds according to the permutation test are given in Fig. 2-4a. Considering the 1% genome-wide threshold (-0.174, 0.169), we found a significant region on chromosome 18. Based on a 5% genome-wide threshold (-0.157, 0.153) another region on chromosome 13 was also significant. Applying the alternative hypothesis test exploring extreme deviations from the normal distribution of local admixture deviations, based on multiple tests with 5000 and 1000 hypotheses (Bhatia *et al*. 2014), threshold lines are given in Fig. 2-4b. We found these two above-mentioned regions on chromosomes 13 and 18 to surpass the 1000 hypotheses significance line.

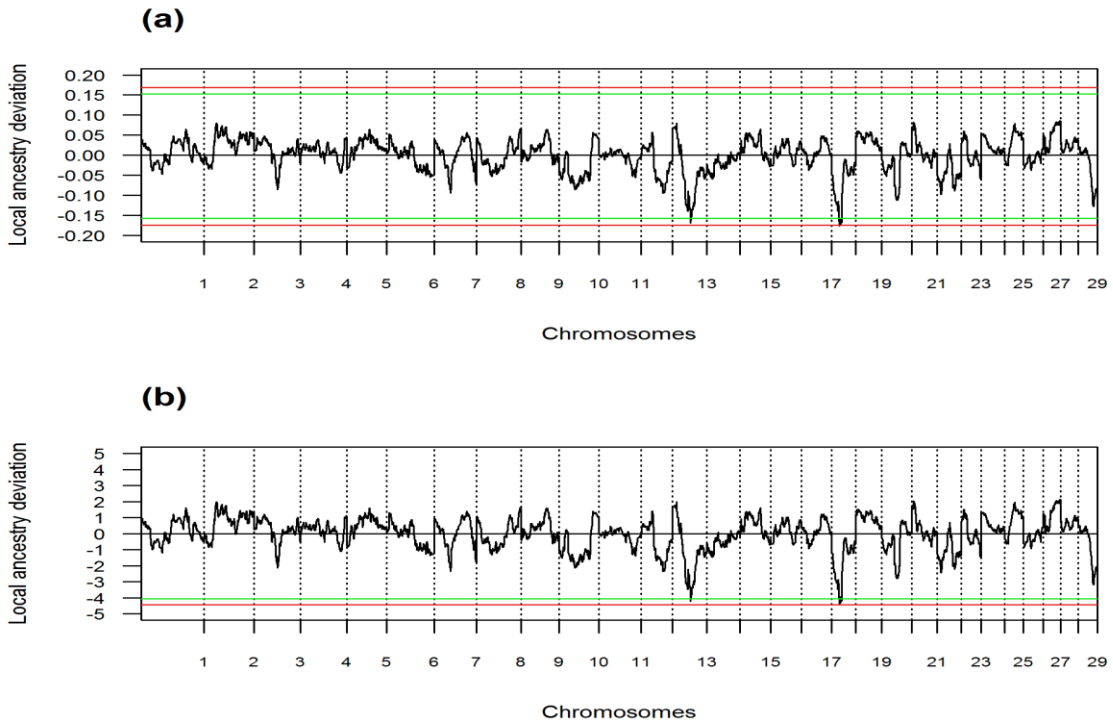


Figure 2-4 (a) Local ancestry deviations based on the permutation threshold. Green and Red lines signify 5% and 1% genome-wide thresholds respectively. (b) Standardized local ancestry deviations based on normal distribution hypotheses tests. Green and red lines are threshold lines based on *P*-value < 5×10^{-5} (4.06 SDs) and *P* -value < 1×10^{-5} (4.42 SDs) respectively.

Manhattan Plots of *iHS* scores for admixed, RHF and SI and *Rsb* of RHF/SI, admixed/RHF and admixed/SI for 29 autosomes are illustrated in Fig. 2-5. We again used thresholds of 4.42 SDs and 4.06 SDs, considering normal distribution with 5000 and 1000 independent segments of the genome (Bhatia *et al*. 2014). Considering *iHS* graphs and based on normal distribution for 5000 and 1000 independent tests, several regions displayed significant scores in RHF and SI ancestral populations. Regards to RHF with 5000 hypotheses, two SNPs on chromosome 18 (25.5 and 26.4 Mb) and one SNP on chromosome 8 (61.9 Mb) passed the threshold. Based on 1000 hypotheses, another SNP on chromosome 18 (23.5 Mb) passed the threshold.

The SNPs passing significance level based on 5000 hypotheses regarding to SI were on chromosomes 5 (61.32 and 61.36 Mb) and 14 (12.5 Mb). Based on 1000 hypothesis, sporadic SNPs on chromosomes 3 (50.8 Mb), 5 (55.5 Mb) and 11 (90.3 Mb) passed the threshold.

The results for *iHS* in admixed animals showed that no regions across the genome passed the threshold. Yet some regions on chromosome 18 (22.5 to 23.5 Mb) were near the threshold line.

Regarding *Rsb* between the two pure populations, a relatively wide region (9.06-23.2 Mb) passed the first threshold and another wide region along chromosome 18 (8.4-24.1 Mb) and two SNPs on chromosome 5 (55.5 and 65.9 Mb) passed thresholds based on 1000 hypotheses. *Rsb* scores between admixed and each ancestral population indicated that one SNP on chromosome 10 (13.8 Mb) surpassed the first threshold line and others on chromosomes 1 (62.8 and 63.06 Mb) and 10 (13.9 Mb) passed the second threshold line, related to RHF.

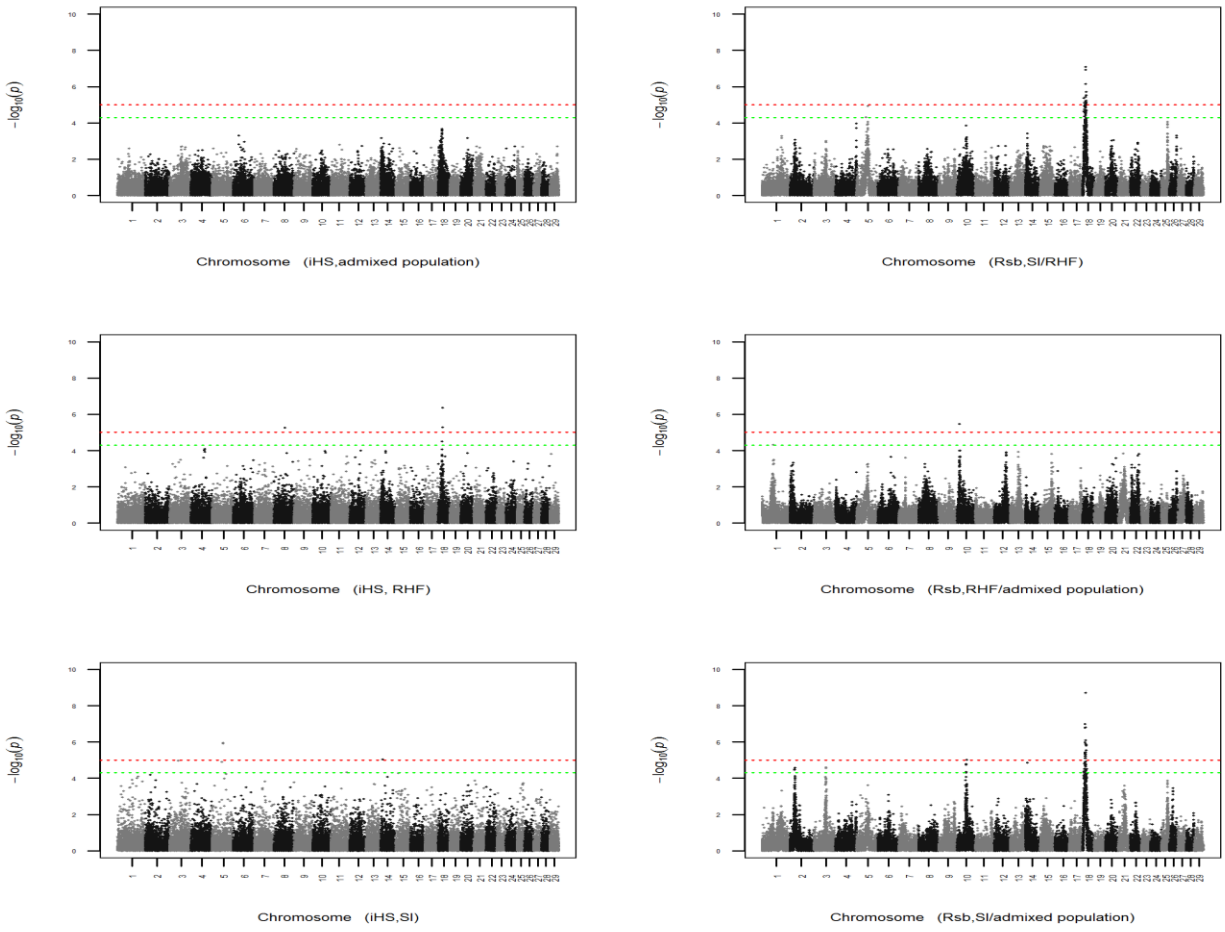


Figure 2-5 Manhattan plots of 29 autosomes for *iHS* on admixed, RHF and SI populations. *Rsb* between RHF/SI, between admixed and RHF/SI respectively. Green and red lines represent two thresholds based on *P*-value < 5×10^{-5} (4.06 SDs) and *P*-value < 1×10^{-5} (4.42 SDs) respectively.

Based on *Rsb* scores calculated for SI and admixed, a wide region on chromosome 18 (6.6-22.8 Mb) and one SNP on chromosome 10 (54.08 Mb) passed the threshold line regarding to 5000 hypotheses. Another wide region on chromosome 18 (9.06-24.6 Mb) and some SNPs on chromosomes 2 (24.5, 26.2 and 29.2 Mb), 3 (66.9, 67.9 and 68.7 Mb) and 10 (52.1, 52.2 and 52.7 Mb) passed the threshold related to 1000 hypotheses.

The average Δ ancestry, *iHS* and *Rsb* related to candidate genes over the two candidate regions is given in Table 2-1.

Table 2-1 Description of the regions harboring signals of selection based on RHF ancestry proportion and extended haplotype homozygosities (*EHH*) values (*iHS* and *Rsb*).

Chr	Δ ancestry location (Mb)	Δancestry (RHF)	*iHS* admixed	*iHS* RHF	*iHS* SI	*Rsb* RHF/SI	*Rsb* admixed/RHF	*Rsb* admixed/SI	Genes of interest
13	46.3-47.3	**-0.16** **-0.17**	-1.21 2.19	-1.06 1.49	-1.94 2.47	-3.22 -0.13	1.17 3.95 (46.5 Mb)	-1.56 -0.79	*IDI1* *GTBP4* *ZMYND11*
18	18.7-25.9	**-0.16** **-0.18**	-3.71 2.36	**-4.56*** **(25.5 Mb)** 1.94	-2.27 2.60	**-4.76*** **(22.3 Mb)** -1.39	-0.09 2.54	**-4.84*** **(18.9Mb)** -0.73	*FTO* *NOD2* *NKD1* *SALL1*
18	6.6-18.7	-0.09 -0.14	-3.45 2.17	-3.29 1.47	-3.35 3.15	**5.36*** **(16.4 Mb)** -0.65	-1.23 2.40	**-6.01*** **(16.4 Mb)** -1.51	*MC1R*

Significant Δ ancestries based on both permutation and hypotheses test are bold.
*Significant *EHH* values based on *P*-value < 1×10^{-5} (4.06 SDs)

Population differentiation between two pure populations along each SNP (F_{st}) was calculated (see Fig. S2-2), averaged over 500-kb windows, between pure RHF and SI, and absolute values of deviations of local ancestry were calculated for the six chromosomes inspected in detail. There was no strong overlap between results on F_{st} and local ancestry deviations. Pearson's correlation of Δ ancestry and F_{st} was low (0.08).

The most notable exception was a region on chromosome 7, where the highest admixture deviation (0.09) was at 42.2-43.7 Mb. Maximum F_{st} was 0.35 and is located at 44.4-44.8 Mb and another peak of F_{st} (0.27) was also detected at 43.2-43.7 Mb, which is near to the peak of local ancestry deviation.

2-4 Discussion

In the current study, we performed locus-specific ancestry estimation across the chromosomes and calculated the excess and deficiency of RHF ancestry at SNP level, compared to the average RHF ancestry across the autosomal genome. As has been the case with many other studies using high throughput genomic data (Sham & Purcell 2014), determination of which regions of the genome deviate significantly from the average global ancestry levels was not straightforward. It was hard to determine how many independent tests should be considered, and there also were forces other than selection, such as drift, causing deviation of local ancestry from global ancestry levels. We applied two approaches of significance testing that have been suggested independently and that we considered appropriate.

A permutation test of circularizing the genome by concatenating the SNPs of all autosomes in a single string, cutting this string once and rearranging the two resulting segments, as proposed by Tang *et al*. (2007a), was used. Distinguishing between the effect of natural selection and demographic events on the genome is difficult. Because the permutation approach destroys not only effects of selection, but also local effects of genetic drift, the threshold is considered to be non-conservative. Nevertheless, based on simulations (Tang *et al*. 2007a) outliers are unlikely to be due to genetic drift. Therefore, this procedure is considered robust to correct for multiple testing to find significant signals for selection.

Bhatia *et al*. (2014) proposed a simple method, looking for extreme deviations from the assumed normal distribution of Δ ancestry values, scaled by SDs. They considered the number of effective independent hypotheses in their analysis of data of African American humans to be somewhere in the range of 1000-5000, resulting in significance thresholds of 4.06 and 4.42 SDs. They reported that simulations with their data suggested a number of independent hypotheses in the range of 1000-1500. We applied the same thresholds in our study, considering this approach conservative, given the much smaller effective population sizes of cattle compared to human populations (Hayes *et al*., 2003).

As visible in Fig. 2-4, both approaches of determining significance provided very similar thresholds. They indicated almost identical regions on chromosome 13 (46.3-47.3 Mb, based on permutation test, 5% genome wide significance level; and 46.3-46.8 Mb, based on deviation

from normal distribution test, 1000 hypotheses) and chromosome 18 (18.7-25.9 Mb, based on both tests, with significance levels as above) to be candidates for signals of selection after admixture. Both signals were in the direction of increased SI ancestry.

We also identified selection signals using the *EHH* method, which relies on unexpected homozygous haplotypes within and between populations (*iHS* and *Rsb* statistics). Based on the results of *iHS*, in RHF, we found a region on chromosome 18 (23.5-26.4 Mb) that showed unexpected long haplotype homozygosity. Based on *iHS* value of admixed animals, no considerable signal was detected, which is consistent with the recent admixture of this population, not giving enough time to establish population specific homozygous haplotypes. On the other hand, we calculated *Rsb* scores, which is expected to be more powerful to identify selection based on variants that are close to fixation in one population (Gautier & Naves 2011).

Related to *Rsb* scores between RHF and SI, a wide region on chromosome 18 (8.4-24.1 Mb) indicated a difference between the ancestral populations. This wide peak (6.6-24.6 Mb) was also observed based on the *Rsb* between admixed and SI. A significant deficiency of RHF was detected on chromosome 18 at 18.7-25.9 Mb based on the results from Δ ancestry. We searched for genes that are located in the region of overlap. *FTO* (*fat mass and obesity associated*) is a gene that is responsible for homeostasis and expenditure and reported mostly for obesity, with negative correlation with fertility and semen quality in human (Landfors *et al*. 2016). In German Holstein cattle, this gene was found to be responsible for milk composition, milk fat and protein yield, which represents a high amount of energy secreted during lactation. This is a gene with pleiotropic effects for milk yield, milk composition and fertility. In that region, there are also some other genes: *NKD1*, *NOD2* related to fertility and *SALL1*, related to dairy traits (Rothammer *et al*. 2013).

In the region of chromosome 13 significant for Δ ancestry, *IDI1* and *ZMYND11* have been found to be related to fertility in bovine. *IDI1* is responsible for nutrient transfer to milk secretion in mammary gland (Connor *et al*. 2008) and has regulatory role in follicle development (Liu *et al*. 2009). *ZMYND11* is also responsible for fetus and placenta development (Smith *et al*. 2009). Conserving a high proportion of SI segments in this region may have resulted in better fertility for the respective admixed animals. *GTPBP4* on chromosome13 is a gene responsible for

morphology traits (Ramayo-Caldas *et al*. 2014). Patterns of *Rsb* and Δ ancestry for the two regions on chromosomes 13 and 18 are visualized in Fig. S2-3.

For recent composites, like this population, admixture selection signals are necessarily wide. The limited numbers of recombination events in 10-15 generations of crossbreeding are not enough to sharpen the signals in such a way to point to any one particular gene responsible for the signal. This is also reflected by the chromosome-wide Δ ancestry graphs for all chromosomes in Fig. S2-1. The genes reported here are therefore all comparatively vague candidates for being drivers of selection in this particular composite population.

Analyzing F_{st} between ancestral populations is one way of investigating selection signals in these populations before admixture. We used average F_{st} and average Δ ancestry within 500-kb windows to explore overlap of signals (see Fig. S2-2). The correlation of the two metrics across 4997 windows was 0.079, indicating a very weak positive association. There was no obvious overlap in the two regions on chromosomes 13 and 18, significant for Δ ancestry. Given the great width of Δ ancestry signals, a comparison of selection signatures before and after admixture in recently admixed populations is not very promising (Jin *et al*. 2012).

2-5 Conclusions

In this study we explored the variability of local ancestry for detection of admixture selection signatures along the genome of a recent composite of two taurine cattle breeds. Based on two types of thresholds - a 5% genome-wide threshold according to a permutation test, and a hypothesis test exploring extreme deviations from the normal distribution of Δ ancestry, based on multiple testing with 1000 hypotheses - two regions on chromosome 13 and 18 were found to be significant and are regarded as recent selection signatures. The signals found were wide, which is consistent with the small number of generations since the start of crossbreeding and not enough generations having passed for narrowing the signatures of selection.

Acknowledgements

We would like to thank the Swissherdbook cooperative Zollikofen for providing genotypes for the analysis. Requests for access to the genotype to facilitate replication of the published results can be addressed to Swissherdbook.

References

Alexander D.H., Novembre J. & Lange K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655-64.

Bahbahani H., Clifford H., Wragg D., Mbole-Kariuki M.N., Van Tassell C., Sonstegard T., Woolhouse M. & Hanotte O. (2015) Signatures of positive selection in East African Shorthorn Zebu: A genome-wide single nucleotide polymorphism analysis. *Sci Rep* **5**, 11729.

Bhatia G., Tandon A., Patterson N. *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**, 437-44.

Bovine Genome Sequencing and Analysis Consortium, Elsik C.G., Tellam R.L. *et al*. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522-8.

Connor E.E., Siferd S., Elsasser T.H., Evock-Clover C.M., Van Tassell C.P., Sonstegard T.S., Fernandes V.M. & Capuco A.V. (2008) Effects of increased milking frequency on gene expression in the bovine mammary gland. *Bmc Genomics* **9**.

Decker J.E., McKay S.D., Rolf M.M., Kim J., Alcala A.M., Sonstegard T.S., Hanotte O., Gotherstrom A., Seabury C.M., Praharani L., Babar M.E., Regitano L.C.D., Yildiz M.A., Heaton M.P., Liu W.S., Lei C.Z., Reecy J.M., Saif-Ur-Rehman M., Schnabel R.D. & Taylor J.F. (2014) Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. *Plos Genetics* **10**.

Decker J.E., Pires J.C., Conant G.C., McKay S.D., Heaton M.P., Chen K., Cooper A., Vilkki J., Seabury C.M., Caetano A.R., Johnson G.S., Brenneman R.A., Hanotte O., Eggert L.S., Wiener P., Kim J.J., Kim K.S., Sonstegard T.S., Van Tassell C.P., Neibergs H.L.,

McEwan J.C., Brauning R., Coutinho L.L., Babar M.E., Wilson G.A., McClure M.C., Rolf M.M., Kim J., Schnabel R.D. & Taylor J.F. (2009) Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proc Natl Acad Sci U S A* **106**, 18644-9.

Delaneau O., Marchini J. & Zagury J.F. (2012) A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179-81.

Doerge R.W. & Churchill G.A. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285-94.

Felius M., Koolmees P.A., Theunissen B., Consortium E.C.G.D. & Lenstra J.A. (2011) On the Breeds of Cattle—Historic and Current Classifications. *Diversity* **3**, 660.

Felius M., Theunissen B. & Lenstra J.A. (2015) Conservation of cattle genetic resources: the role of breeds. *Journal of Agricultural Science* **153**, 152-62.

Flori L., Thevenon S., Dayo G.K., Senou M., Sylla S., Berthier D., Moazami-Goudarzi K. & Gautier M. (2014) Adaptive admixture in the West African bovine hybrid zone: insight from the Borgou population. *Molecular Ecology* **23**, 3241-57.

Frkonja A., Gredler B., Schnyder U., Curik I. & Solkner J. (2012) Prediction of breed composition in an admixed cattle population. *Animal Genetics* **43**, 696-703.

Gautier M., Laloe D. & Moazami-Goudarzi K. (2010) Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. *Plos One* **5**.

Gautier M. & Naves M. (2011) Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Molecular Ecology* **20**, 3128-43.

Gautier M. & Vitalis R. (2012) rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**, 1176-7.

Hayes B. J., Visscher P. M., McPartlan H. C. and Goddard M.E. (2003) Novel multi locus measure of linkage disequilibrium to estimate past effective population size. *Genome Research* **13**, 635-43.

Hu Y.N., Willer C., Zhan X.W., Kang H.M. & Abecasis G.R. (2013) Accurate Local-Ancestry Inference in Exome-Sequenced Admixed Individuals via Off-Target Sequence Reads. *American Journal of Human Genetics* **93**, 891-9.

Jin W.F., Xu S.H., Wang H.F., Yu Y.G., Shen Y.P., Wu B.L. & Jin L. (2012) Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Research* **22**, 519-27.

Jones O.R. & Wang J.L. (2012) A comparison of four methods for detecting weak genetic structure from marker data. *Ecology and Evolution* **2**, 1048-55.

Keenan K., McGinnity P., Cross T.F., Crozier W.W. & Prodohl P.A. (2013) diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods in Ecology and Evolution* **4**, 782-8.

Kim E.S. & Rothschild M.F. (2014) Genomic adaptation of admixed dairy cattle in East Africa. *Front Genet* **5**, 443.

Landfors M., Nakken S., Fusser M., Dahl J.A., Klungland A. & Fedorcsak P. (2016) Sequencing of FTO and ALKBH5 in men undergoing infertility work-up identifies infertility-associated variant and two missense mutations. *Fertility and Sterility*.

Liu Z.L., Youngquist R.S., Garverick H.A. & Antoniou E. (2009) Molecular Mechanisms Regulating Bovine Ovarian Follicular Selection. *Molecular Reproduction and Development* **76**, 351-66.

Long J.C. (1991) The Genetic-Structure of Admixed Populations. *Genetics* **127**, 417-28.

Mancini G., Gargani M., Chillemi G., Nicolazzi E.L., Marsan P.A., Valentini A. & Pariset L. (2014) Signatures of selection in five Italian cattle breeds detected by a 54K SNP panel. *Molecular Biology Reports* **41**, 957-65.

McTavish E.J., Decker J.E., Schnabel R.D., Taylor J.F. & Hillis D.M. (2013) New World cattle show ancestry from multiple independent domestication events. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E1398-E406.

Moradi M.H., Nejati-Javaremi A., Moradi-Shahrbabak M., Dodds K.G. & McEwan J.C. (2012) Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. *Bmc Genetics* **13**.

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**, 185-205.

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 tool set for whole-

genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559-75.

Qanbari S., Strom T.M., Haberer G., Weigend S., Gheyas A.A., Turner F., Burt D.W., Preisinger R., Gianola D. & Simianer H. (2012) A High Resolution Genome-Wide Scan for Significant Selective Sweeps: An Application to Pooled Sequence Data in Laying Chickens. *Plos One* **7**.

Racimo F., Sankararaman S., Nielsen R. & Huerta-Sanchez E. (2015) Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics* **16**, 359-71.

Ramayo-Caldas Y., Fortes M.R.S., Hudson N.J., Porto-Neto L.R., Bolormaa S., Barendse W., Kelly M., Moore S.S., Goddard M.E., Lehnert S.A. & Reverter A. (2014) A marker-derived gene network reveals the regulatory role of PPARGC1A, HNF4G, and FOXP3 in intramuscular fat deposition of beef cattle. *Journal of Animal Science* **92**, 2832-45.

Rothammer S., Seichter D., Forster M. & Medugorac I. (2013) A genome-wide scan for signatures of differential artificial selection in ten cattle breeds. *Bmc Genomics* **14**.

Sabeti P.C., Reich D.E., Higgins J.M., Levine H.Z., Richter D.J., Schaffner S.F., Gabriel S.B., Platko J.V., Patterson N.J., McDonald G.J., Ackerman H.C., Campbell S.J., Altshuler D., Cooper R., Kwiatkowski D., Ward R. & Lander E.S. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-7.

Sankararaman S., Sridhar S., Kimmel G. & Halperin E. (2008) Estimating local ancestry in admixed populations. *American Journal of Human Genetics* **82**, 290-303.

Sham P.C. & Purcell S.M. (2014) STUDY DESIGNS Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics* **15**, 335-46.

Smetko A., Soudre A., Silbermayr K., Muller S., Brem G., Hanotte O., Boettcher P.J., Stella A., Meszaros G., Wurzinger M., Curik I., Muller M., Burgstaller J. & Solkner J. (2015) Trypanosomosis: potential driver of selection in African cattle. *Front Genet* **6**, 137.

Smith S.L., Everts R.E., Sung L.Y., Du F.L., Page R.L., Henderson B., Rodriguez-Zas S.L., Nedambale T.L., Renard J.P., Lewin H.A., Yang X.Z. & Tian X.C. (2009) Gene Expression Profiling of Single Bovine Embryos Uncovers Significant Effects of In Vitro Maturation, Fertilization and Culture. *Molecular Reproduction and Development* **76**, 38-47.

Stella A., Ajmone-Marsan P., Lazzari B. & Boettcher P. (2010) Identification of Selection Signatures in Cattle Breeds Selected for Dairy Production. *Genetics* **185**, 1451-U498.

Tang H., Choudhry S., Mei R., Morgan M., Rodriguez-Cintron W., Burchard E.G. & Risch N.J. (2007a) Recent genetic selection in the ancestral admixture of Puerto Ricans. *American Journal of Human Genetics* **81**, 626-33.

Tang K., Thornton K.R. & Stoneking M. (2007b) A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol* **5**, e171.

Utsunomiya Y.T., O'Brien A.M.P., Sonstegard T.S., Van Tassell C.P., do Carmo A.S., Meszaros G., Solkner J. & Garcia J.F. (2013) Detecting Loci under Recent Positive Selection in Dairy and Beef Cattle by Combining Different Genome-Wide Scan Methods. *Plos One* **8**.

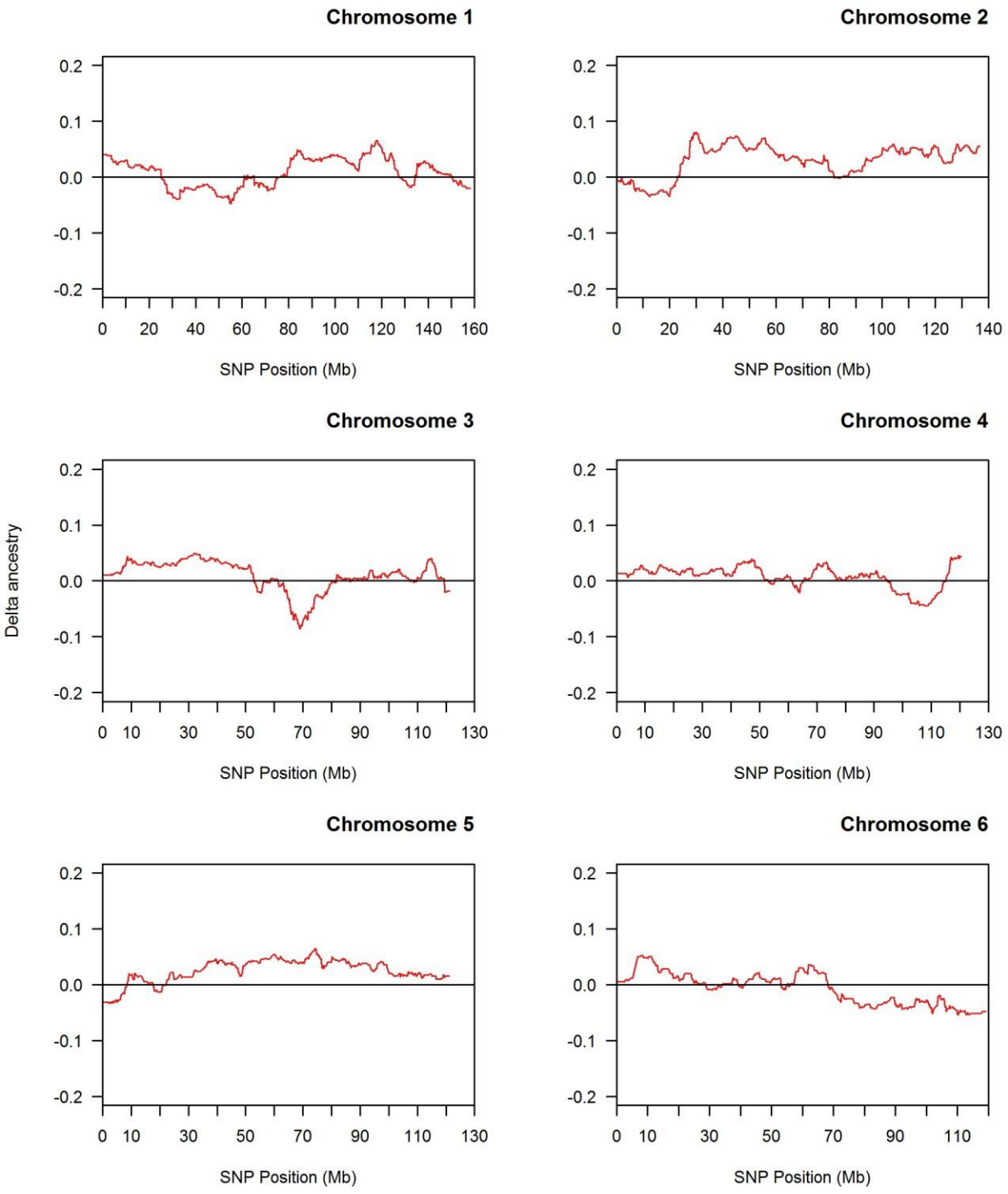
Voight B.F., Kudaravalli S., Wen X. & Pritchard J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72.

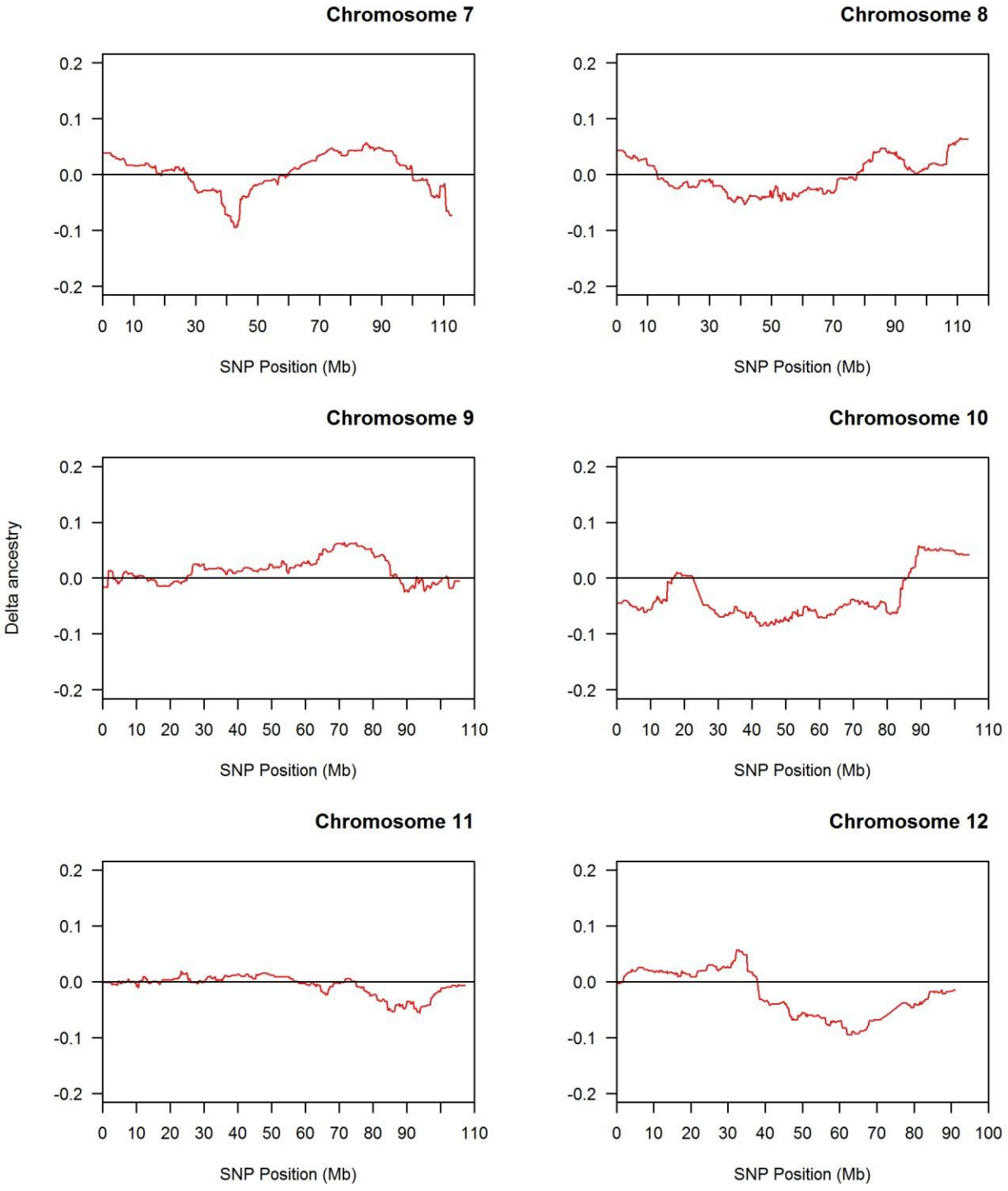
Weir B.S. & Cockerham C.C. (1984) Estimating F-Statistics for the Analysis of Population-Structure. *Evolution* **38**, 1358-70.

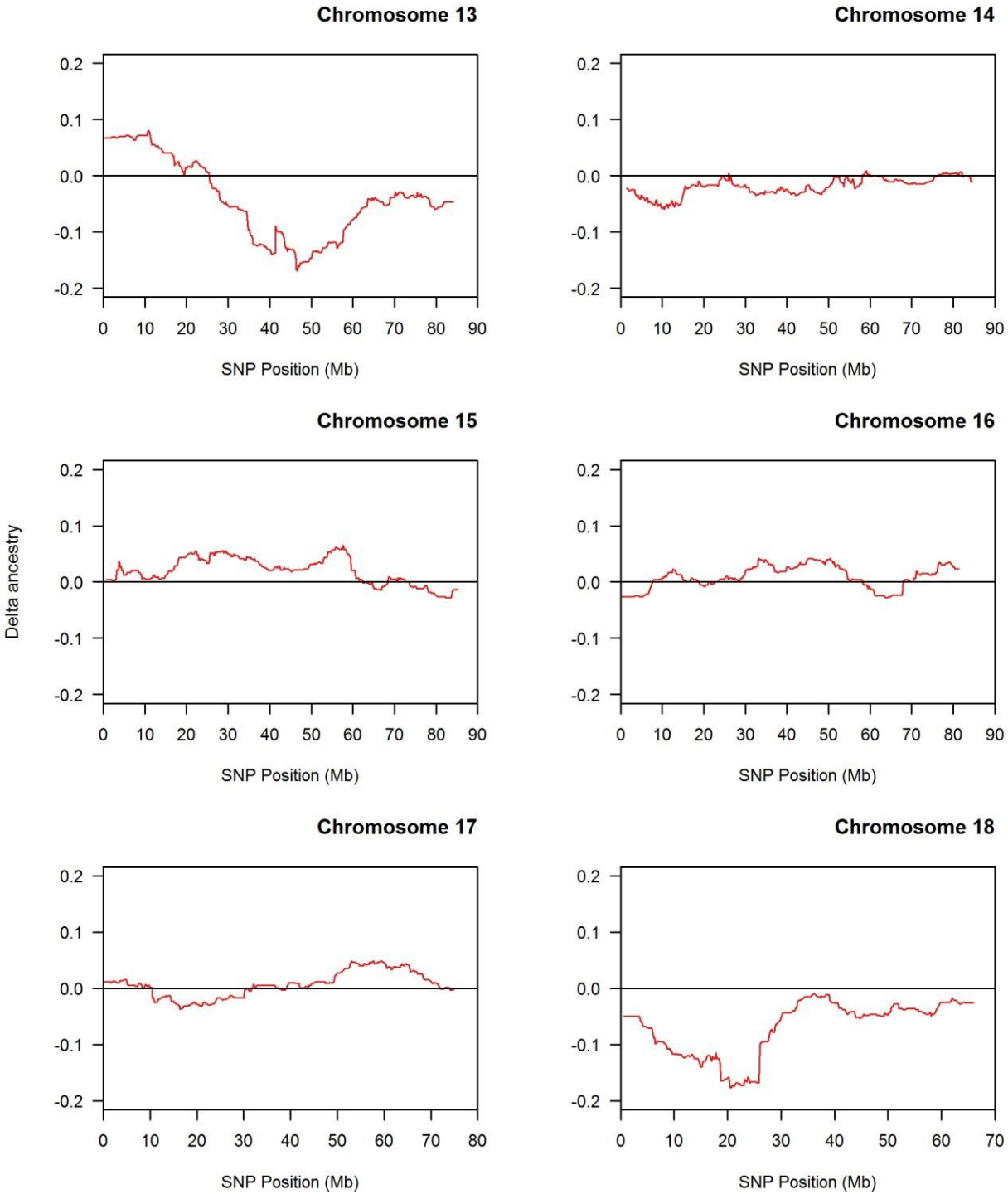
Zhang J.Q. & Stram D.O. (2014) The Role of Local Ancestry Adjustment in Association Studies Using Admixed Populations. *Genetic Epidemiology* **38**, 502-15.

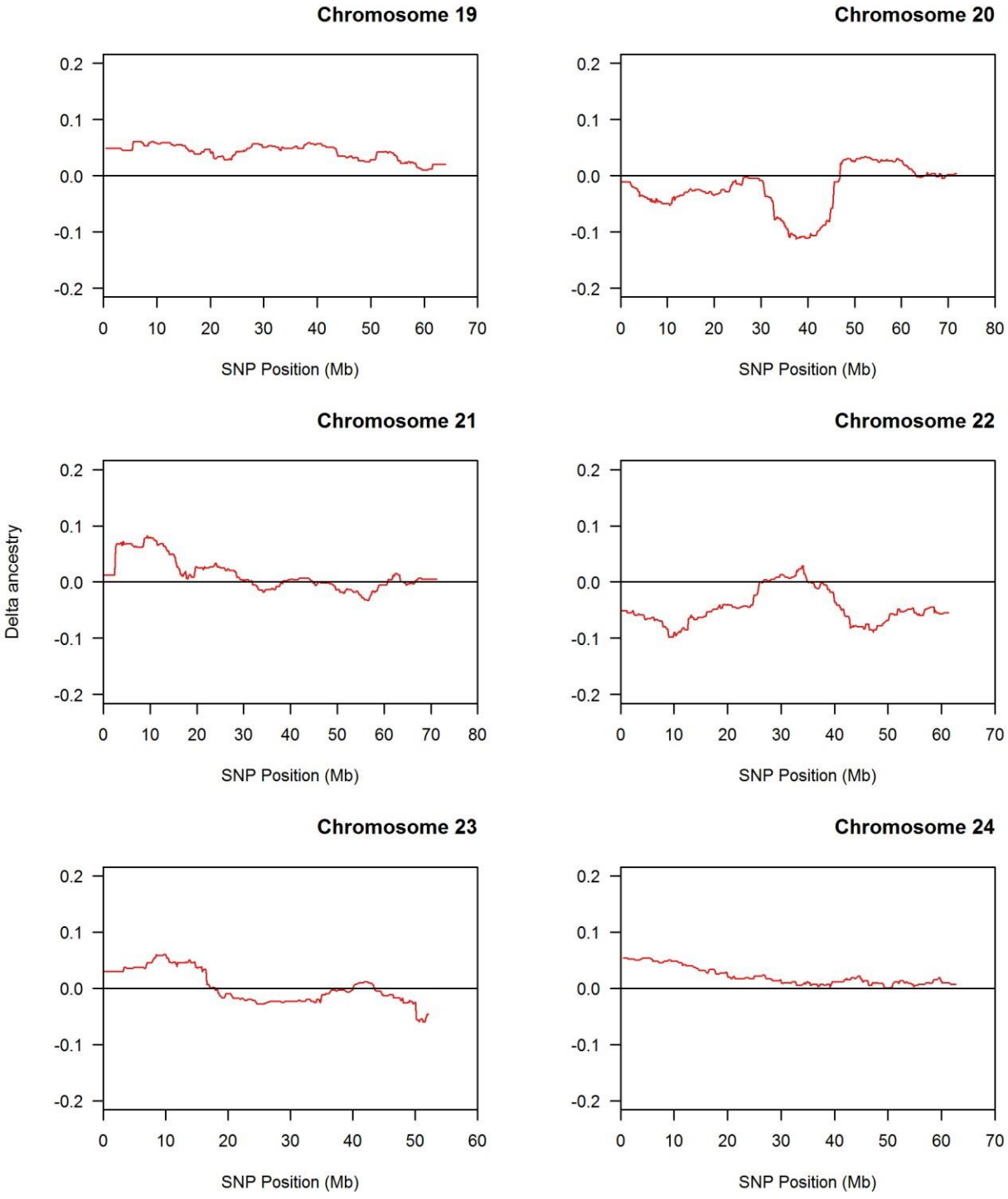
Zhao F.P., McParland S., Kearney F., Du L.X. & Berry D.P. (2015) Detection of selection signatures in dairy and beef cattle using high-density genomic information. *Genetics Selection Evolution* **47**.

Figure S2-1 Local ancestry deviations from genome-wide ancestry at the corresponding SNP positions on all 29 autosomes averaged over all 300 admixed cattle.




Chromosome 7
Chromosome 8
Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12
Delta ancestry
SNP Position (Mb)


Chromosome 13
Chromosome 14
Chromosome 15
Chromosome 16
Chromosome 17
Chromosome 18
Delta ancestry
SNP Position (Mb)
0.2
0.1
0.0
-0.1
-0.2
0 10 20 30 40 50 60 70 80 90
0 10 20 30 40 50 60 70 80
0 10 20 30 40 50 60 70


Chromosome 19
Chromosome 20
Chromosome 21
Chromosome 22
Chromosome 23
Chromosome 24
Delta ancestry
SNP Position (Mb)

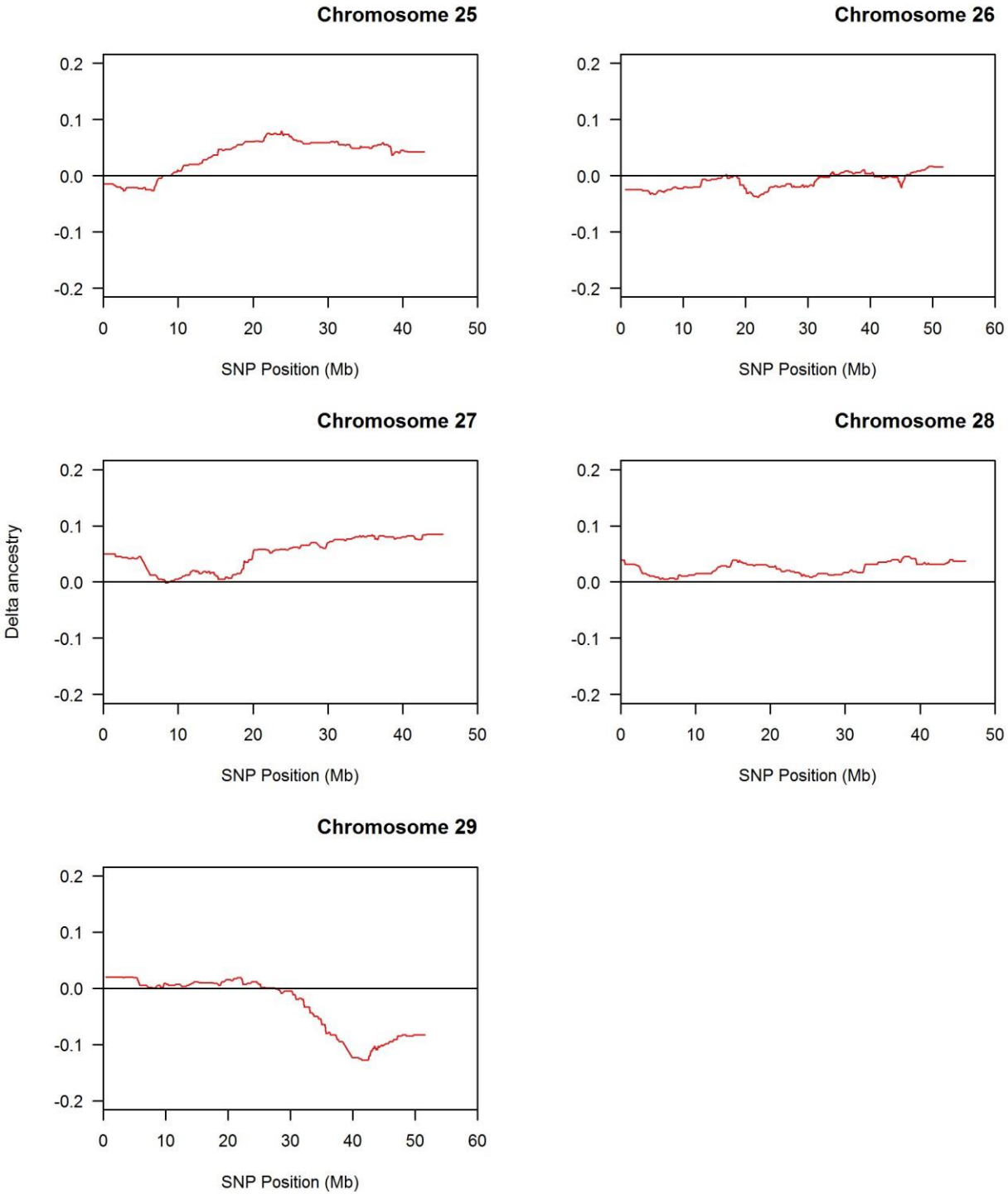
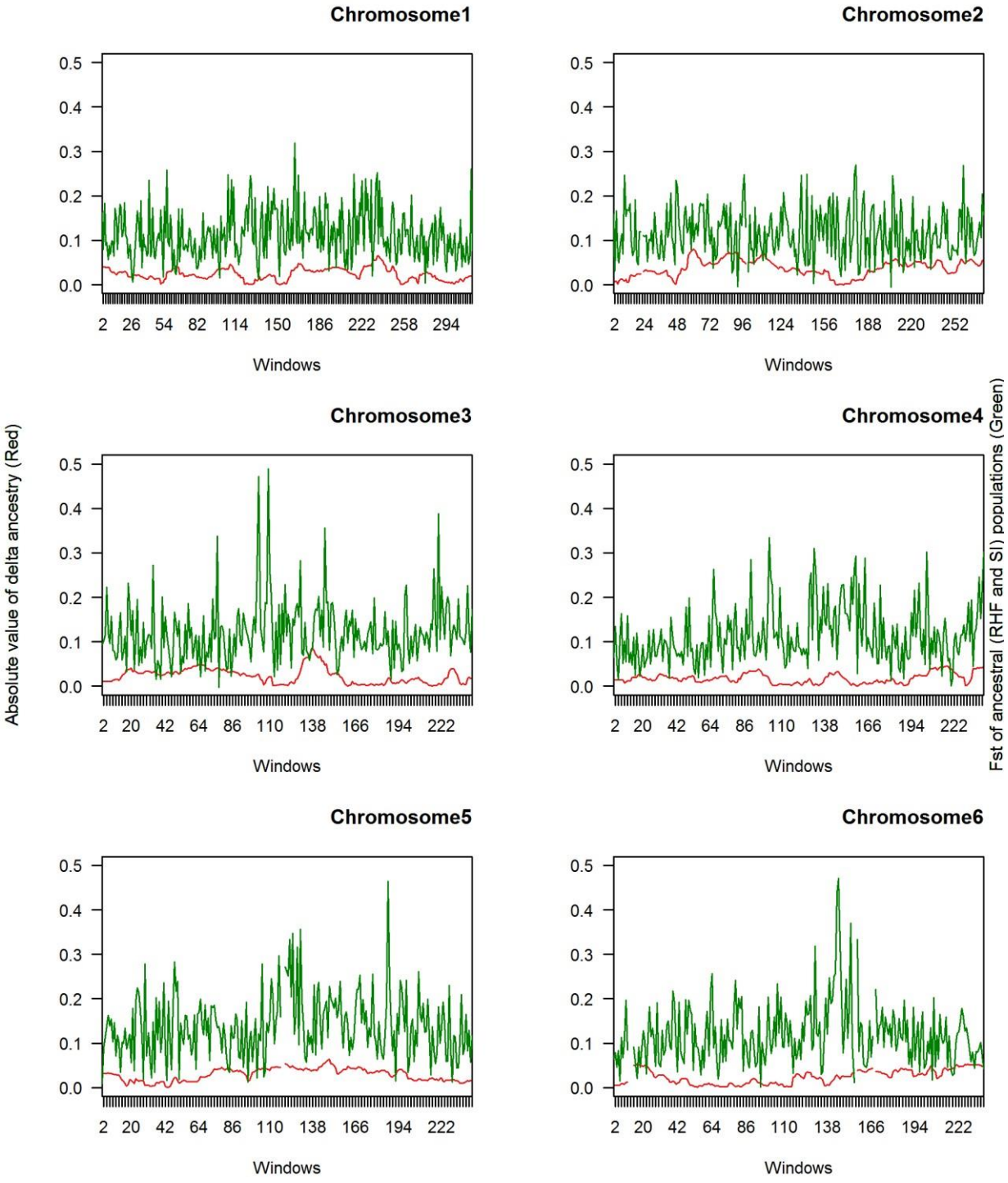
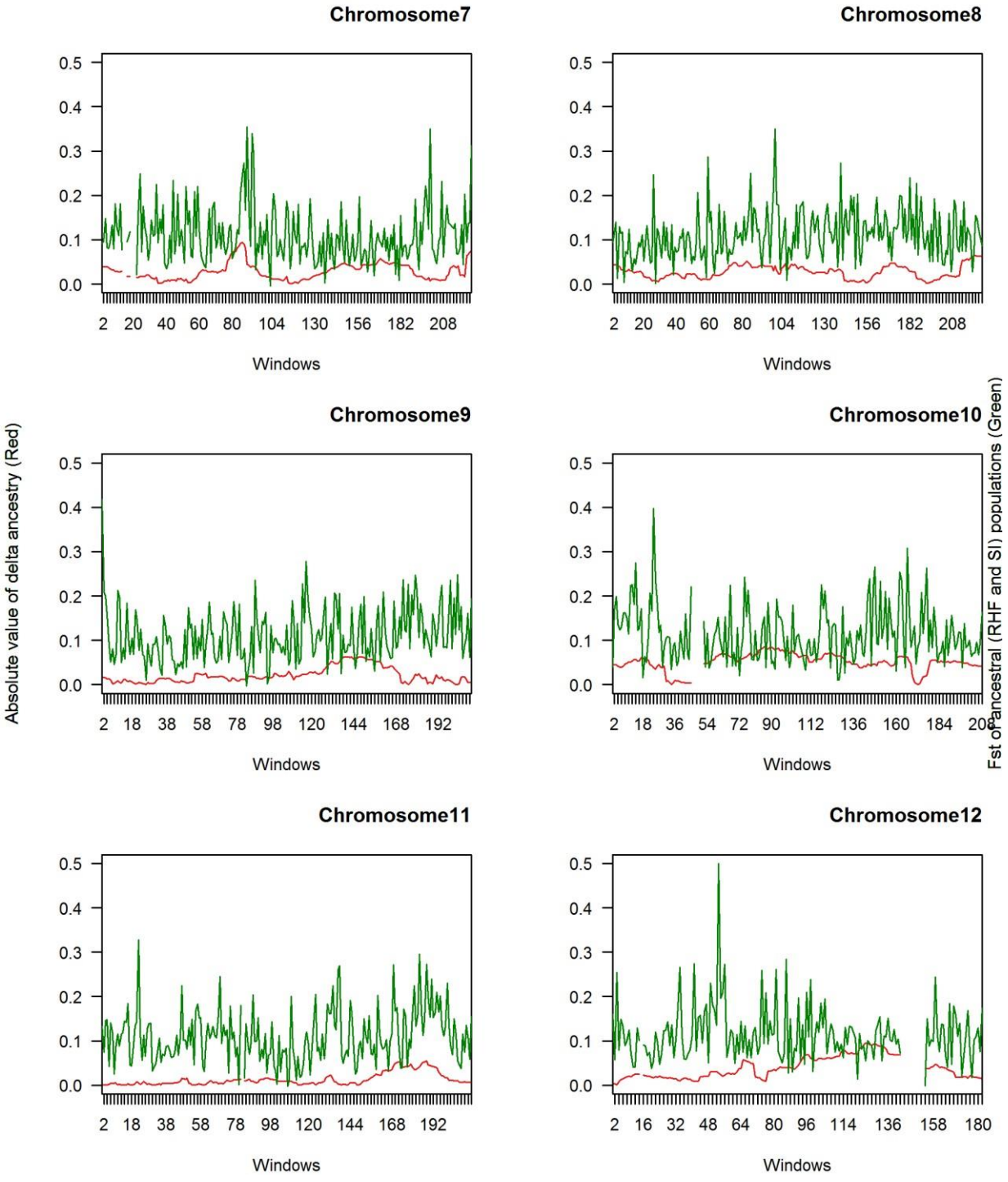
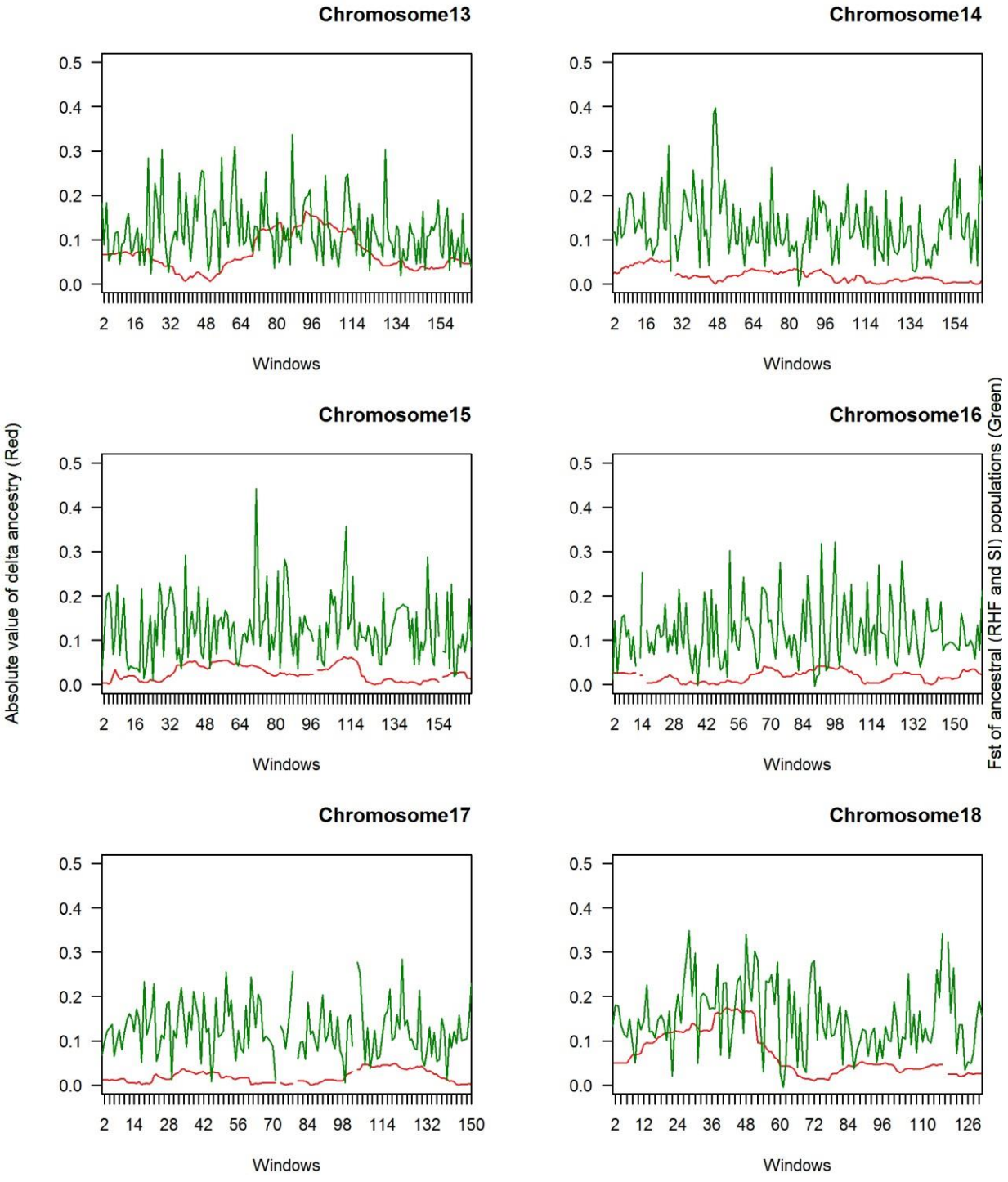
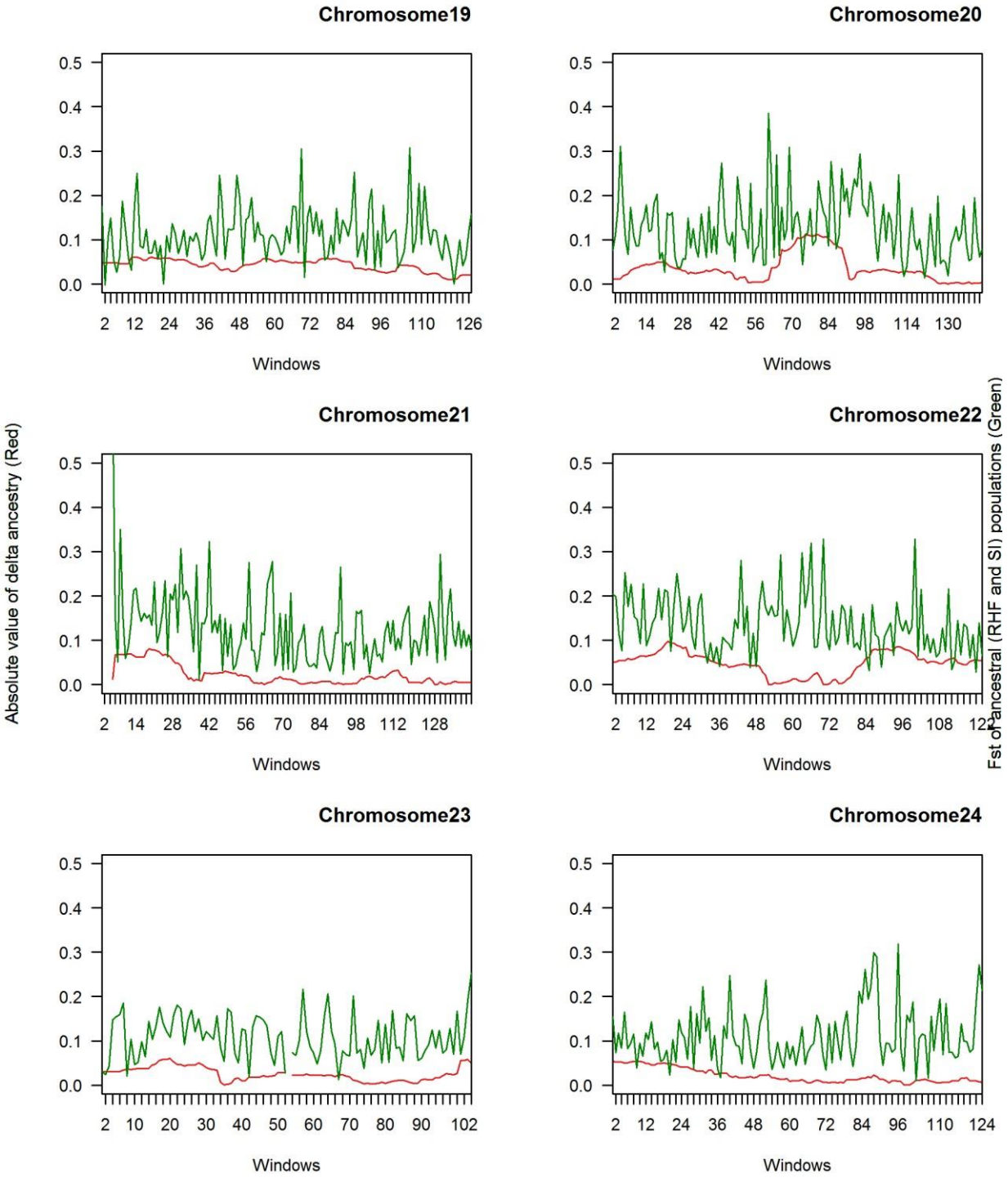

Chromosome 25
Chromosome 26
Chromosome 27
Chromosome 28
Chromosome 29
Delta ancestry
SNP Position (Mb)
0.2
0.1
0.0
-0.1
-0.2
0
10
20
30
40
50
60

Figure S2-2 Genomic distribution for F_{st} of ancestral (RHF and SI) populations (Green) and absolute Δ ancestry (Red) averaged across 500 kb window on 29 autosomes.




Chromosome7
Chromosome8
Chromosome9
Chromosome10
Chromosome11
Chromosome12
Absolute value of delta ancestry (Red)
Fst of ancestral (RHF and SI) populations (Green)
Windows
0.0
0.1
0.2
0.3
0.4
0.5
2 20 40 60 80 104 130 156 182 208
2 18 38 58 78 98 120 144 168 192
2 18 36 54 72 90 112 136 160 184 208
2 16 32 48 64 80 96 114 136 158 180


Chromosome13
Chromosome14
Chromosome15
Chromosome16
Chromosome17
Chromosome18
Absolute value of delta ancestry (Red)
Fst of ancestral (RHF and SI) populations (Green)
Windows
0.0
0.1
0.2
0.3
0.4
0.5
2 16 32 48 64 80 96 114 134 154
2 14 28 42 56 70 84 98 114 132 150
2 12 24 36 48 60 72 84 96 110 126


Chromosome19
Chromosome20
Chromosome21
Chromosome22
Chromosome23
Chromosome24
Absolute value of delta ancestry (Red)
Fst of ancestral (RHF and SI) populations (Green)
Windows
0.0
0.1
0.2
0.3
0.4
0.5
2 12 24 36 48 60 72 84 96 110 126
2 14 28 42 56 70 84 98 114 130
2 14 28 42 56 70 84 98 112 128
2 12 24 36 48 60 72 84 96 108 122
2 10 20 30 40 50 60 70 80 90 102
2 12 24 36 48 60 72 84 96 110 124

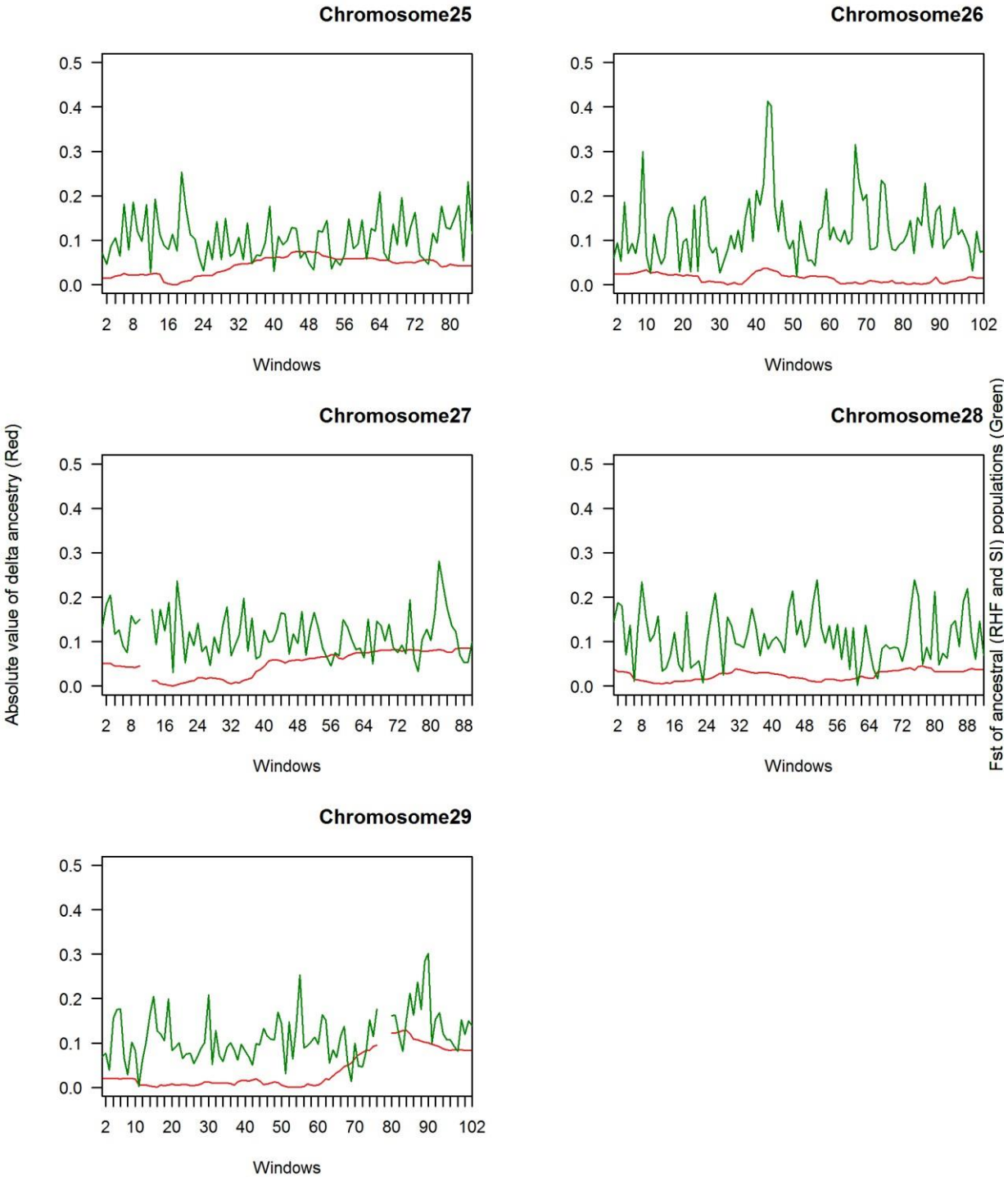
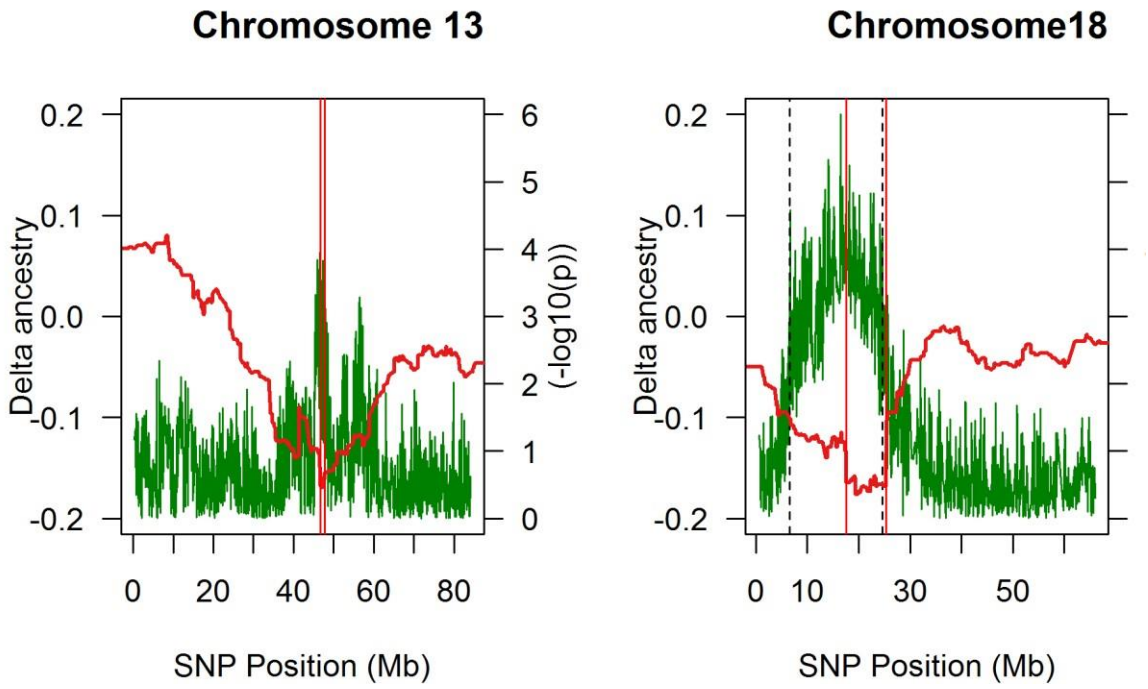

Chromosome25
Chromosome26
Chromosome27
Chromosome28
Chromosome29
Absolute value of delta ancestry (Red)
Fst of ancestral (RHF and SI) populations (Green)
Windows
0.0
0.1
0.2
0.3
0.4
0.5
2 8 16 24 32 40 48 56 64 72 80
2 10 20 30 40 50 60 70 80 90 102
2 8 16 24 32 40 48 56 64 72 80 88
2 10 20 30 40 50 60 70 80 90 102

Figure S2-3 Comparison of significance chromosome regions (13 and 18) as selection signals by Δ ancestry deviations (red) and absolute value of *Rsb* (green). The significance region based on Δ ancestry deviations are shown by vertical red lines and based on *Rsb* by dash black vertical lines.



Chapter 3

Inference of local ancestry by different algorithms using phased and unphased genotypes in admixed Swiss Fleckvieh cattle

Negar Khayatzadeh[1], Gábor Mészáros[1*], Yuri Tani Utsunomiya[2], Urs Schnyder[3], Birgit Gredler[3], José Fernando Garcia[2, 4], Ino Curik[5] and Johann Sölkner[1]

[1] Division of Livestock Science, Department of Sustainable Agricultural Systems, University of Natural Resources and Life Sciences, Vienna, Gregor Mendel Straße 33, 1180 Vienna, Austria

[2] Departmento de Medicina Veterinária Preventiva e Reprodução Animal, Faculdade de Ciências Agrárias Veterinárias, UNESP – Univ Estadual Paulista, Jaboticabal, São Paulo, Brazil

[3] Qualitas AG, Chamerstrasse 56, CH-6300, Zug, Switzerland

[4] Departamento de Apoio, Saúde e Produção Animal, Faculdade de Medicina Veterinária de Araçatuba, UNESP – Univ Estadual Paulista, Araçatuba, São Paulo, Brazil

[5] Department of Animal Science, Faculty of Agriculture, University of Zagreb, Svetošimunska cesta 25, 10000 Zagreb, Croatia

To be submitted

Abstract

Background The genome of a recently admixed individual resembles a mosaic of haplotypes from the ancestral populations. Variations in local ancestry proportions from genome-wide ancestry along chromosomal segments arise from demographic process of genetic drift and selection. Local ancestry estimations can be used to detect post-admixture selection signature in recent admixed cattle populations. Development of various software tools for estimation of local ancestries from genomic data provides the possibility of comparing different statistical methods. The main aim of this study was to apply different algorithms using both unphased and phased genotypes to examine how various methods influence local ancestry estimations and selection signature detections in Swiss Fleckvieh cattle.

Results We performed 361 analyses, using three different methods implemented in LAMP, LAMP-LD and MULTIMIX for local ancestry estimations. Ancestral haplotypes phased by S*hapeIt* and *AlphaPhase* were used as the input files for LAMP-LD and MULTIMIX_MCMCgeno. MULTIMIX_MCMC applied phased admixed samples beside ancestral haplotypes. Correlations between estimations using haplotypes phased with different window sizes in *ShapeIt* were close to unity. Estimations from phased haplotypes using *AlphaPhase* with cores length 100 to 150 SNPs showed correlations > 0.95 which had also high correlations with *ShapeIt* results (0.99). Regards to local ancestry's estimates with different parameters of each program, high correlations (0.98) were observed between local ancestry estimations of LAMP-LD using window lengths 15 to 30 SNP. The highest correlations between results of MULTIMIX software were observed between windows lengths 15 and 23 SNP, 23 and 30 SNP for MULTIMIX_MCMCgeno (> 0.95) and MULTIMIX_MCMC (> 0.89). Comparison between analyses set by different methods showed highest correlation between MULTIMIX_MCMCgeno and MULTIMIX_MCMC with 15 (0.92) and 23SNP (0.85), LAMP-LD and MULTIMIX_MCMCgeno with 23 SNP (0.85). Autosome-wise comparison ranged from negative correlations between estimations of LAMP and the other methods (chromosomes 3, 14, 24 and 26) to correlations higher than 0.90 between LAMP-LD and MULTIMIX_MCMCgeno and MULTIMIX_MCMCgeno and MULTIMIX_MCMC (3, 4, 6, 8, 13, 14, 23 and 29). Two selection signals were detected by LAMP on chromosomes 13 and 18 based on multiple hypothesis tests, yet no similar signals were detected by the other approaches.

Conclusions Choosing the method for estimation of local genetic ancestries should be considered carefully.

Keywords admixture, cattle, haplotype, local ancestry, phasing, selection signature, SNP, Swiss Fleckvieh

3-1 Background

The genome of an admixed individual comprises a mosaic of ancestral haplotypes formed by recombination occurring at every generation [1-5]. Study of the patterns of DNA sequence variation shaped by admixture can be considered at both "global" (genome-wide) and "local" (locus-specific) levels. Global admixture is a relative proportion of ancestral haplotypes averaged across the entire genome of an individual, while local admixture deals with identification of the ancestral origin of distinct chromosomal segments within an individual [4, 6-8]. Local ancestry (LANC) proportions diverge from genome wide ancestries as a result of demographic processes of selection, small population size (random genetic drift) and sampling variability. The extreme deviations of local ancestries from genome wide ancestry are interpreted as selection signatures happened after admixture [9, 10]. Local ancestries as a measure of post-admixture selection signature have been estimated in recently admixed human [9, 11, 12] and livestock populations [13-18].

The assessment of fine-scale ancestry has been recently facilitated by the development of statistical methods implemented in different software tools to estimate admixture from genomic data [7, 19]. The approaches for local ancestry inference rely either on Li and Stephens [20] framework using an approximation to the coalescent and ancestral recombination graphs or on model-based clustering algorithms.

The methods applied by LAMP-LD [21], HAPMIX [2] and MULTIMIX [22] model an admixed individual genome as a noisy mixture of the ancestral haplotypes using hidden Markov models (HMM). These algorithms can be applied to both unphased and phased data, provided that phased reference ancestral haplotypes are available. These phased haplotypes can be obtained by various specific software tools, such as PHASE [23, 24], fastPHASE [25], Beagle [26] or *ShapeIt* [27]. These programs use haplotype frequencies and identical by descent (IBD) segment probabilities to model linkage disequilibrium (LD) using HMM, which are computationally intensive and time-consuming. Among these software tools, *ShapeIt* has faster implementation in terms of run-time [28]. An alternative to HMM-based phasing is the use of deterministic algorithm based on long range haplotypes [29], which uses the concept of surrogate parents to determine chromosomal phased haplotypes. This method was expanded and combined by adding

haplotype library imputation, using pedigree information for livestock populations in *AlphaPhase* software [30].

In contrast to programs using genetic parameters model, an approximation of ancestral recombination graph, for LANC inference, there are other software tools such as LAMP (Local Ancestry in adMixed populations) which infer local ancestries by breaking the genome into sliding windows and clustering SNPs within each window based on allele frequencies of ancestry informative markers (AIMs). Optimal window size is selected internally based on the number of generations since admixture, recombination rate and SNP density [1, 7, 19].
Local ancestries as a measure of selection signature have been estimated using LAMP program in New World Creole cattle [13-15], East African short horn Zebu [17] and Swiss Fleckvieh cattle [18]. Kim and Rothschild [16] used LAMP-LD to estimate local ancestries in East African dairy cattle.
In the present study, we took advantage of the availability of different phasing algorithms and LANC inference software tools to investigate how much the choice of different methods as well as parameter settings of the applied software tools can affect the estimations in Swiss Fleckvieh, a cattle population with ~ 10 generations of admixture history. Search for post-admixture selection signatures was performed by multiple hypothesis tests. LANC estimations from different approaches were compared to see the consistency of the results of different methods.

3-2 Methods

3-2-1 Animals

Illumina® BovineSNP50k v2 (50k) genotypes of 91 Simmental, 101 Red Holstein Friesian and 308 admixed bulls from Swissherdbook cooperative Zollikofen were used in this study. These genotypes build on previously published data [18, 31]. Swiss Fleckvieh is a composite breed of Simmental (SI) and Red Holstein Friesian (RHF) that was established in 1970 in Switzerland with the aim of combining the high milk production of the Red Holstein Friesian with the high fertility, beef value and longevity of the Simmental breed. According to the formal definition, animals with pedigree admixture levels of 0.125-0.875 RHF are categorized as Swiss Fleckvieh.

We considered all admixed animals along the range of pedigree admixture level of 0.02 to 0.99 RHF in the current study.

3-2-2 Quality control

Quality control of the data was performed with PLINK 1.90 [32, 33]. Markers that were monomorphic, unmapped, non-autosomal, those presented call rate below 95% or deviated from Hardy Weinberg Equilibrium (Fisher's exact *P*-value less than 10^{-6}) were excluded. Animals with more than 5% missing genotypes were also removed. After applying these quality control criteria, 39525 SNPs and 485 (97 RHF, 300 admixed and 88 SI) animals were retained for analysis.

3-2-3 Phasing

We used two different algorithms, namely *ShapeIt v2.r837* and *AlphaPhase 1.2*, to define haplotypes of admixed and ancestral samples.

ShapeIt uses a hidden Markov model (HMM) and builds an imperfect mosaic of haplotypes (**H**) underlying genotypes (**G**). In the first step, all of *K* haplotypes in **H** are collapsed into a graph structure $\mathbf{H_g}$ by splitting haplotypes into *J* disjoint segments, which are considered as states. Each marker is labeled either with allele 1 or allele 0 (node) and at edges is weighed by the number of haplotypes in **H** that traverse it. All available haplotypes are kept in a compact HMM. In the second step, pairs of compatible haplotypes for **G,** with linear complexity are sampled from $\mathbf{H_g}$. Genotypes in **G** are partitioned into disjoint segments, and the then all haplotypes which are compatible with **G** enumerated as the compatible haplotypes in each disjoints segment, putting into a graph structure $\mathbf{S_g}$ as a compact representation of the possible haplotypes. Then transition probabilities between segments are computed using a forward-backward algorithm, and pairs of compatible haplotypes can be sampled [27, 34].

Model parameters of *ShapeIt* are conditioning states and indicate the number of disjoint segments (*J*), window size (*W*), genetic map and effective population size. *ShapeIt* uses 100 states by default, but in this study we have increased the number of states to 200 to achieve greater accuracy. Different lengths (0.5, 1, 1.5 and 2 Mb) of window sizes were used in these calculations, replacing the default value of 2 Mb. The effective population size (N_e) for admixed

animals and ancestral populations was computed using *SNeP* [35], estimated through LD. The N_e of RHF and SI ancestral populations were estimated at 137 and 188, respectively. Based on N_e for each ancestral population and considering the guidelines in the *ShapeIt* documentation, the N_e of the mixed sample was set to 153 depending on the proportion of each population in the data set. For genetic map we considered 1 Mb ≈ 1 cM and constant recombination rate of 1 cM /Mb. We increased default values for number of burn-in, pruning and main iterations from 7 to 10, 8 to 10, and 20 to 50 iterations to increase accuracy.

AlphaPhase employs the long range phasing algorithm of Kong et al. [29] to phase a string of consecutive SNPs, termed "core" in the program's terminology, by identifying surrogate parents of each individual, termed "proband". Surrogate parents share a haplotype with a proband with no opposing homozygote genotypes. These parents could be one degree (Erdös 1) or more than one degree (Erdös 2 or more) removed from the proband on the basis of haplotype identity. Layers older than Erdös 1 for surrogates do not have shared haplotypes with proband, but do have shared haplotypes with more recent layers. In a follow up step, the parental surrogates are partitioned to paternal and maternal gametes based on pedigree information. Inference of the phase for the proband is attempted by stepping through the surrogates until one is found that is homozygous at that locus. Adjacent tails to either end of each core are defined to provide additional information about surrogacy especially near the end of the core. For compensation of the lack of surrogate parents due to recombination, a haplotype library is built to impute phase for unphased individuals [30].

The authors recommended in *AlphaPhase* documentation, a core length of 100 SNP and a core and tail length of 300 to 500 SNP for 60K SNP density. However, as the *AlphaPhase* algorithm is robust to small variations in terms of core and tail length, we used longer cores (150, 200 and 300 SNP) as we expected longer shared haplotypes in the recent admixed population to find optimum phasing length. A total of 7 different phasing analyses in terms of core and tail lengths were performed. Core and tail length settings were 1: (100, 100), 2: (100, 150), 3: (100, 200), 4: (300, 100), 5: (200, 200), 6: (150, 200) and 7: (250,200).The phasing analyses were run twice considering offset and not-offset between successive cores (50% overlap), which gave rise to 14 phasing analyses in total. The threshold for disagreement between homozygous genotypes to

identify surrogate parents, due to genotyping error was set to 1%. The threshold for disagreement between surrogate parents haplotypes and proband genotypes was also set to 1%.

3-2-4 Local genetic ancestry estimation

LAMP 2.5 applies a clustering algorithm to infer locus-specific ancestries at the chromosome level in admixed individuals. It works based on allele frequencies of a reference panel with no need of individual genotypes from the ancestral populations. The idea is to select a suitable window length that is long enough to enable ancestry estimation but short enough such that the window on average does not contain recombinants of purebred haplotypes. We used LAMP in LAMPANC mode. The following configuration parameters were set: admixture proportions (alpha) = 0.68 RHF and 0.32 SI based on the global ancestry estimation using ADMIXTURE [36], number of generations since admixture (g) = 7, recombination rate (r) = 10^{-8}, fraction of overlap between adjacent windows (offset) = 0.2. LAMP relies on a predefined set of AIMs that are in low LD ($r^2 < 0.1$) for each pair of selected SNPs. Because of the sparsity of the 50k Beadchip, we did not exclude SNPs based on LD in this analysis. The locus-specific ancestry was estimated for each admixed individual with respect to pure breeds, representing the proportion of each involved ancestry (0, 0.5, and 1) for each SNP. For the 300 admixed animals, we computed the average locus-specific ancestry level across each chromosome separately.

LAMP-LD models an admixed chromosome as a set of haplotypes from K ancestral populations after g generations, considering crossing over that occurs in each generation. The recombination with average rate ρ thorough g generations breaks the ancestries and inserts ancestry breakpoints. Each segment between break points is modeled as an independent draw from ancestral populations with probabilities given by the admixture fraction α. For the sake of simplicity, LAMP-LD assumes constant recombination rate and physical positions are scaled based on 1 Mb $\approx$ 1 cM. The model consists of HMMs that emit genotypes in non-overlapping windows and models LD structure of ancestral segments. The hidden states are local ancestries on each chromosome within each window. The model structure is described by S in terms of number of states and constant window length L ($S \times L$ states in total), with emission and transition probabilities estimated from reference haplotypes. Intuitively larger S induces better modelling of haplotype with increase in runtime, but fixing state into moderately small numbers usually results in improvements in run time with very modest reduction in accuracy. Window lengths (L)

of 200-400 Kb, together with 10-15 numbers of states (*S*) are sufficient to estimate local ancestries with high accuracy. We defined different lengths 5 (~300 Kb), 8 (~500 Kb), 15(~1 Mb), 23 (~1.5 Mb) and 30 (~2 Mb) in terms of number of SNP for windows, similar to windows sizes for phasing haplotypes by *ShapeIt*, and number of states 10 and 15 to analyze the influence of different lengths and states on local genetic ancestries.

MULTIMIX similarly to LAMP-LD uses **H** source haplotypes from *k* ancestral populations to estimate local ancestries as hidden states. The chromosomes of the individuals in a recent admixed population is considered as a series of segments consisting of alleles with shared ancestry, as neighboring loci tend to be inherited together during meiosis. Therefore, it may be possible to carry out inference at a coarse scale (longer haplotypes) without noticeable loss of accuracy in the recent admixed population. Each chromosome was split into W = L/n contiguous windows (chunks) with the assumption that ancestry was constant within each window, considering the age of admixture. Within a given window due to difference in haplotype frequencies between populations an observed haplotype tends to be more likely from one source population than others. MULTIMIX models observed haplotype frequencies and allele frequencies in case of unphased genotypes, as a discrete multivariate distribution, using a coalescent model given fitted haplotype ancestries. Within the *j*th window, covariance matrix of SNPs from the *k*th ancestral population is estimated. Constant small value λ is added to the variances to ensure that covariance matrix is positive definite and invertible. MULTIMIX models breakpoint ancestries along a chromosome using Markov process. The other model parameter is misfitting probabilities to make a distinction between true and fitted ancestral populations to avoid spurious switches in ancestries.

We applied MULTIMIX to both phased and unphased samples with phased references. We used MULTIMIX_MCMC algorithm for phased samples and MULTIMIX_MCMCgeno for unphased samples with misfitting probabilities equal to the estimation from MULTIMIX_CEM. The initial values for global ancestry estimations were set to 0.68 and 0.32 for RHF and SI, calculated by ADMIXTURE [36].

3-2-5 Statistical analysis

We searched for excess and deficiency of LANC with respect to RHF ancestry using the approach proposed by Tang et al. [9]. The 'Δ ancestry' was calculated by subtracting the average local ancestries at SNP *m* from the genome-wide ancestry for each of the two ancestry components, which is defined as:

$$\delta_k^m = \frac{1}{I}\sum_{i=1}^{I}\left(q_k^{i,m} - \bar{q}_k^i\right) = \tilde{q}_k^m - \bar{q}_k$$

where $q_k^{i,m}$ is the local ancestries of animal *i* at SNP *m*, $\bar{q}_k^i$ is mean of local ancestries for individual *I*, $\tilde{q}_k^m$ is the mean of ancestry at SNP *m* averaged over all admixed animals; and $\bar{q}_k$ is the mean of local ancestries across the entire whole genome for population. We scaled δ_k^m values by their standard deviation. Following Bhatia et al. [12] in their analysis of recent admixed human populations, we determined genome-wide threshold of signals of selection as LANC deviation greater than 4.42 SDs (*P*-value < 1×10^{-5}), applying Bonferroni correction for 5000 hypotheses and 4.06 SDs (*P*-value < 5×10^{-5}) corresponding to 1000 hypotheses. As LD is higher and therefore the number of independent segments of the genome is smaller in bovine populations compared to human populations, we consider these thresholds conservative [37].

3-3 RESULTS

3-3-1 Effect of phasing algorithms

In order to assess the influence of different phasing algorithms on the estimates of LANC along the genome, we applied the *ShapeIt* and *AlphaPhase* with different set of parameters summarized in Table 3-1. Chromosome 2 was chosen as an example, since it consists of considerable number of SNPs in comparison with other chromosomes. We phased ancestral haplotypes using *ShapeIt* with different window sizes (0.5, 1, 1.5 and 2 Mb) and then calculated local ancestries using LAMP-LD (window length of 5 SNP (~300 Kb) with 10 states).

Table 3-1 Overview of parameters used for different phasing algorithms and methods to estimate local genetic ancestries.

Phase.Alg[a]	Win.S (Mb)[b]	LANC.Alg[c]	States	Win.L (SNP)[e]	Offse	Analysis Numbers
-	-	LAMP				A0
ShapeIt	0.5Mb	LAMP-LD	10-15	5, 8, 15, 23, 30		A1-A10
		MULTIMIXgeno	-	5, 8, 15, 23, 30		A11-A15
		MULTIMIX	-	5, 8, 15, 23, 30		A16-A20
	1Mb	LAMP-LD	10-15	5, 8, 15, 23, 30		A21-A30
		MULTIMIXgeno	-	5, 8, 15, 23, 30		A31-A35
		MULTIMIX	-	5, 8, 15, 23, 30		A36-A40
	1.5Mb	LAMP-LD	10-15	5, 8, 15, 23, 30		A41-A50
		MULTIMIXgeno	-	5, 8, 15, 23, 30		A51-A55
		MULTIMIX	-	5, 8, 15, 23, 30		A56-A60
	2Mb	LAMP-LD	10-15	5, 8, 15, 23, 30		A61-A70
		MULTIMIXgeno	-	5, 8, 15, 23, 30		A71-A75
		MULTIMIX	-	5, 8, 15, 23, 30		A76-A80
	CTL[g]					
AlphaPhase	100,100	LAMP-LD	10-15	5, 8, 15, 23, 30	Both	A81-A100
		MULTIMIXgeno	-	5, 8, 15, 23, 30	Both	A101-A110
		MULTIMIX	-	5, 8, 15, 23, 30	Both	A111-A120
	100,150	LAMP-LD	10-15	5, 8, 15, 23, 30	Both	A121-A140
		MULTIMIXgeno	-	5, 8, 15, 23, 30	Both	A141-A150
		MULTIMIX	-	5, 8, 15, 23, 30	Both	A151-A160
	100,200	LAMP-LD	10-15	5, 8, 15, 23, 30	Both	A161-A180
		MULTIMIXgeno	-	5, 8, 15, 23, 30	Both	A181-A190
		MULTIMIX	-	5, 8, 15, 23, 30	Both	A191-A200
	300,100	LAMP-LD	10-15	5, 8, 15, 23, 30	Both	A201-A220
		MULTIMIXgeno	-	5, 8, 15, 23, 30	Both	A221-A230
		MULTIMIX	-	5, 8, 15, 23, 30	Both	A231-A240
	200,200	LAMP-LD	10-15	5, 8, 15, 23, 30	Both	A241-A260
		MULTIMIXgeno	-	5, 8, 15, 23, 30	Both	A261-A270
		MULTIMIX	-	5, 8, 15, 23, 30	Both	A271-A280
	150,200	LAMP-LD	10-15	5, 8, 15, 23, 30	Both	A281-A300
		MULTIMIXgeno	-	5, 8, 15, 23, 30	Both	A301-A310
		MULTIMIX	-	5, 8, 15, 23, 30	Both	A311-A320
	250,200	LAMP-LD	10-15	5, 8, 15, 23, 30	Both	A321-A340
		MULTIMIXgeno	-	5, 8, 15, 23, 30	Both	A341-A350
		MULTIMIX	-	5, 8, 15, 23, 30	Both	A351-A360

Note: Phase.Alg[a] is phasing algorithms used for phasing haplotypes. Win.S (Mb)[b] window size (Mb) used for phasing data with *ShapeIt*. LANC.Alg[c] is algorithm used to estimate local genetic ancestries. States[d] refers to number of states in LAMP-LD. Win.L (SNP)[e] refers to window length for local genetic ancestry's estimations. Offset[f] refers to considering offset between consecutive cores for phasing data with *AlphaPhase*. CTL[g] is general length of core together with its adjacent tails for phasing data with *AlphaPhase*.

The results indicated that applying different window sizes to phase haplotypes by *ShapeIt* did not substantially influence the estimation of local ancestries. Pearson's correlations between local ancestries, using different window sizes for phasing ancestral haplotypes were close to unity (> 0.99).

Pearson's correlations of LANC proportions, using the phasing algorithm implemented by *AlphaPhase* were in the range of 0.91 to 0.99. The method of phasing ancestral haplotypes with or without offset did not have any notable influence on the results (Pearson's correlations of 0.98 to 0.99 between analyses with and without offset). Correlations of less than 0.95 were observed in dataset with core length 300 and general core and tail length 500 SNPs, considering offset between cores. Figure 3-1a and 3-1b shows LANC proportions estimated by LAMP-LD at SNP level, using *ShapeIt* and *AlphaPhase*.

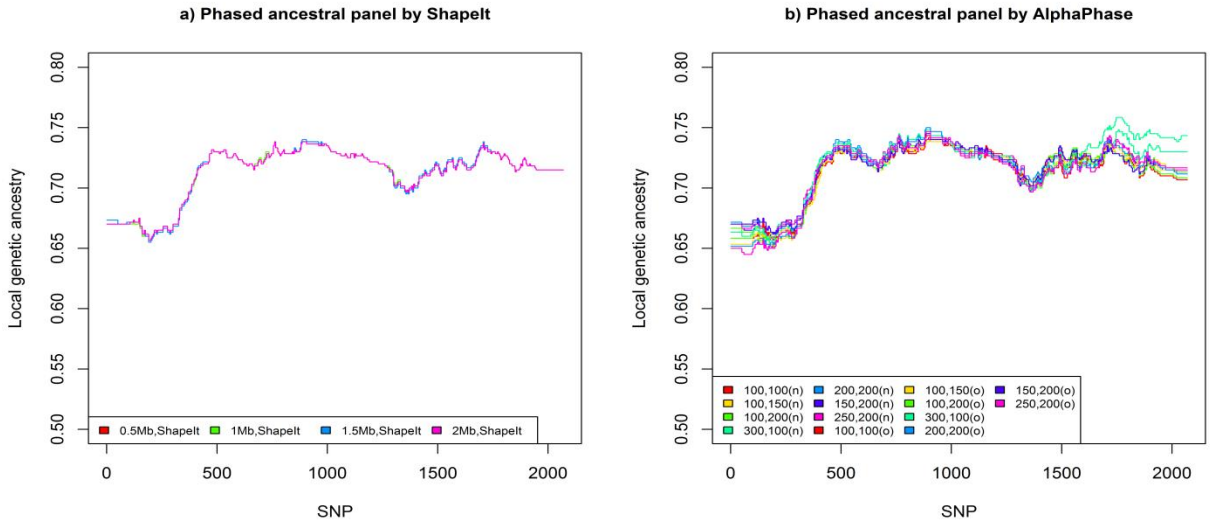


Fig. 3-1 LANC proportions by LAMP-LD with window length 5 SNP (300 kb in average) and number of states 10, using ancestral panels phased **a**. by *ShapeIt* considering different window size (0.5, 1, 1.5 and 2 Mb) **b**. by *AlphaPhase* considering different combinations of general core and tail lengths and not-offset and offset between cores.

3-3-2 Effect of settings within algorithms applied for LANC prediction

LAMP-LD

In the second step, two different methodologies applied by LAMP-LD and MULTIMIX were used for estimation of local genetic ancestries. The different settings of LAMP-LD were

investigated, in terms of window length, number of states, and their influence on change of LANC estimations along the chromosome.

Based on the setting in Table 3-1, the correlations between LANC estimations, using phased ancestral haplotypes with different window sizes in *ShapeIt* (0.5, 1, 1.5 and 2 Mb) and within each window length defined for LAMP-LD (5, 8, 15, 23 and 30 SNPs) were greater than 0.99. Moreover, increasing number of states to 15 did not change the results considerably and high pairwise correlations were observed between all comparisons ($r \sim 0.99$) (Additional file 3-1: Figure S3-1). The results showed that using different sets of phased ancestral haplotypes by *ShapeIt* does not have a remarkable impact on LANC estimations along different window lengths in LAMP-LD. We then compared the averaged LANC proportions through different window in LAMP-LD (Fig. 3-2.a). With regards to the length, highest correlations (> 0.98) were observed between windows with 15, 23 and 30 SNPs (Table 3-2). Since Swiss Fleckvieh is a recent composite with small number of generations after admixture, the expected length of ancestral haplotypes are relatively wide and choosing the segments with 5 to 8 SNP length can cause of noises in estimations. Therefore, we took the average of windows 15, 23 and 30 in order to have one set of local genetic ancestries estimated by LAMP-LD, using ancestral haplotypes phased by *ShapeIt* (Fig. 3-2.c).

We estimated LANC proportions with window lengths and number of states similar to previous stage with LAMP-LD, using ancestral haplotypes phased with different general core and tail lengths by *AlphaPhase* (Table 3-1). Estimated local ancestries using different core and tail length have correlations in the range of 0.91 to 0.99, 0.94 to 0.99, 0.96 to 0.99, 0.97 to 0.99, and 0.96 to 0.99 within window lengths 5, 8, 15, 23 and 30 SNPs respectively. Considering no offset and offset between pairwise comparisons, we found correlation greater than 0.95 between all comparisons. Correlations between estimations were decreased (< 0.95) by using cores with length longer than 100 SNP. (Additional file 3-2: Figure S3-2).

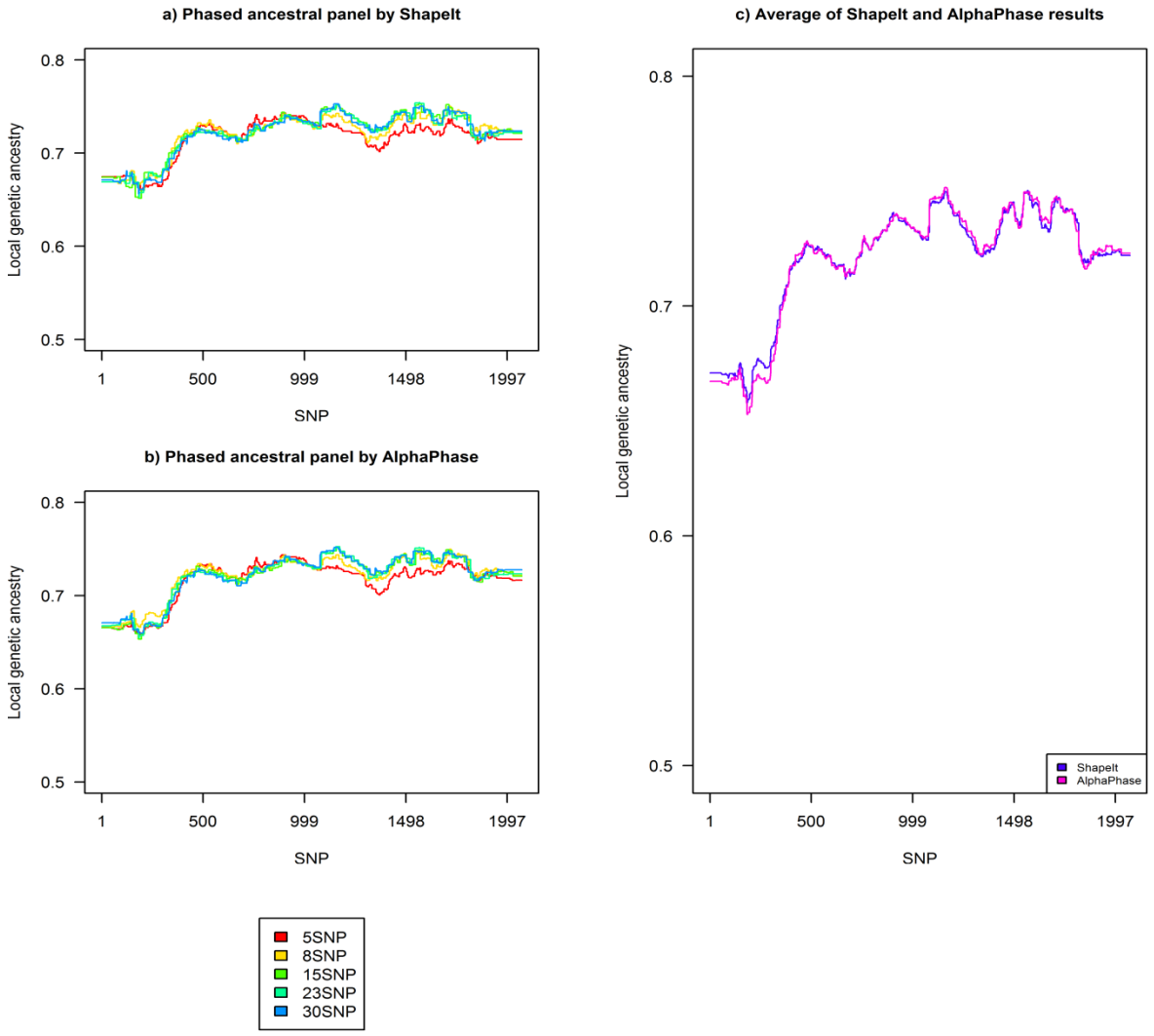


Fig. 3-2 LANC estimated by LAMP-LD averaged within each window .Ancestral haplotypes were phased by a) *ShapeIt*, b) *AlphaPhase*. c) Averaged of LANC estimations between windows with 15, 23 and 30 SNP with different phase algorithm.

Results of phasing haplotypes with core lengths 100 to 150 SNP represented highest correlations with each other. Increasing core length to 200 or 300 SNPs made deviations of estimations from results of haplotypes phased with core length 100 to 150 SNPs. Therefore, we took the average of these within each window, and we then compared the averaged local ancestries between different windows (Fig. 3-2.b). Comparisons between different window lengths in LAMP-LD showed that the highest correlations (0.99) were observed between estimations of window lengths bigger than 15 SNPs, where ancestral haplotypes phased by AlphaPhase (Table 3-2). As

the increasing number of states was not relevant (with 0.99 correlations), we selected only estimations with number of states 15 for comparing the results of LANC estimates by LAMP-LD, applying different phasing algorithms.

To compare LAMP-LD results when data were phased with different algorithms, we took the average between different windows, except windows lengths of 5 and 8 SNP. LANC estimates of differently phased sets by *ShapeIt* were in the range of 0.65 to 0.76 along the chromosome. Local ancestries averaged across different sets ranged from 0.65 to 0.75. The range of LANC estimations of different phased sets from *AlphaPhase* was 0.65-0.76 and their averages ranged from 0.66 to 0.75. Results showed there was high correlation between LANC estimations, when genotypes were phased with *ShapeIt* or *AlphaPhase* (Fig. 3-2.c).

MULTIMIX_MCMCgeno

Similar to LAMP-LD, we used the same set of haplotypes which were phased by different window sizes with *ShapeIt* to estimate LANC proportions. We used the MULTIMIX_MCMCgeno algorithm with unphased admixed samples and phased reference panel (RHF and SI) with misfitting probabilities equal to the estimation from MULTIMIX_CEM. We set length of the chunks similar to LAMP-LD in previous section to 5, 8, 15, 23 and 30 SNPs per window. Local genetic ancestries which were estimated using different set of ancestral haplotypes by *ShapeIt* represented notably high correlation (0.99) comparing different chunk lengths (Additional file 3-3: Figure S3-3). Due to high correlations between estimations from different ancestral haplotype setting within each window in MULTIMIXgeno, we took the average of estimations within each window. Comparisons among different chunk lengths are given in Table 3-2. Correlations were 0.95 between chunk length 15 and 23 SNP, 23 and 30 SNP. Misfitting probabilities matrix which were estimated using MULTIMIX_CEM were in the range of $\begin{pmatrix} 0.91 & 0.09 \\ 0.12 & 0.88 \end{pmatrix}$ for chunk length 5, 8 and 15 SNPs and $\begin{pmatrix} 0.96 & 0.04 \\ 0.16 & 0.84 \end{pmatrix}$ for chunks longer than 23 SNP. Lambda (λ) was 0.05 for all analysis.

We used the same general cores and tail lengths, which were used by LAMP-LD, considering not-offset and offset between core ant tails to phase haplotypes. Correlations between local

ancestries calculated by MULTIMIX based on different sets for phasing ancestral haplotype were in the range of 0.90 to 0.99, 0.87 to 0.99, 0.86 to 0.99, 0.83 to 0.99, and 0.78 to 0.99 for chunk length 5, 8, 25, 23 and 30 SNPs respectively (Additional file 3-4: Figure S3-4). The correlations were lower when we used long general core and tail length (core length 300 SNP and general core and tail length greater than 500 SNP). Moreover, with increase in chunk length to 30, the cores longer than 150 SNPs showed deviation from shorter cores. In addition, only small differences were observed between offset and not-offset (0.99) among cores with 100 to 150 SNP; with lower correlations belonging to long segments. Due to the more similarity between results of core lengths 100, we took the average of not-offsets for the first three phased lengths within each chunk. Correlations among different chunk lengths in MULTIMIX_MCMCgeno ranged between 0.67 and 0.95. Chunk length 5 and 8 had lower correlation with the other chunks and highest correlation (> 0.95) were observed between chunk length 15 and 30 SNP, 23 and 30 SNP (Table 3-2). Misfitting probabilities matrices, which were used for haplotypes phased by *AlphaPhase*, were $\begin{pmatrix} 0.90 & 0.10 \\ 0.10 & 0.90 \end{pmatrix}$ for chunks with length 5 to 15 SNP and $\begin{pmatrix} 0.97 & 0.03 \\ 0.13 & 0.87 \end{pmatrix}$ for longer chunks. Value of λ increased to 0.1 for phased haplotypes with core length 200 and 300 SNP.

The results from MULTIMIXgeno, with the ancestral haplotypes phased by *ShapeIt* and *AlphaPhase*, showed high correlations among the same window lengths (0.99, 0.99, and 0.97 between chunk lengths 15, 23 and 30 SNPs, respectively). These high correlations indicated that different phasing algorithms do not influence the LANC estimation by MULTIMIXgeno (Fig. 3-3).

MULTIMIX_MCMC

The same sets of haplotypes phased with *ShapeIt* were used to calculate LANC proportions with MULTIMIX. We used the MULTIMIX_MCMC algorithm to phased samples with phased reference panel with misfitting probabilities equal to the estimation from MULTIMIX_CEM. Local ancestries which were estimated using different set of ancestral haplotypes by *ShapeIt* represented notably high correlation (> 0.99) within different chunk lengths (5, 8, 15, 23 and 30

SNPs) (Additional file 3-5: Figure S3-5). Due to high correlations between estimations from different ancestral haplotype setting within each window in MULTIMIX, we took the average of estimations within each window. Comparisons between different chunk lengths are given in Table 3-2. The highest correlations which were observed were between chunk length 15 and 23 SNP, 23 and 30 SNP (0.91). We used the same general core and tail lengths, which were used by LAMP-LD and MULTIMIX_MCMCgeno as well, considering not-offset and offset between core and tails to phase haplotypes. Correlations between local genetic ancestries calculated by

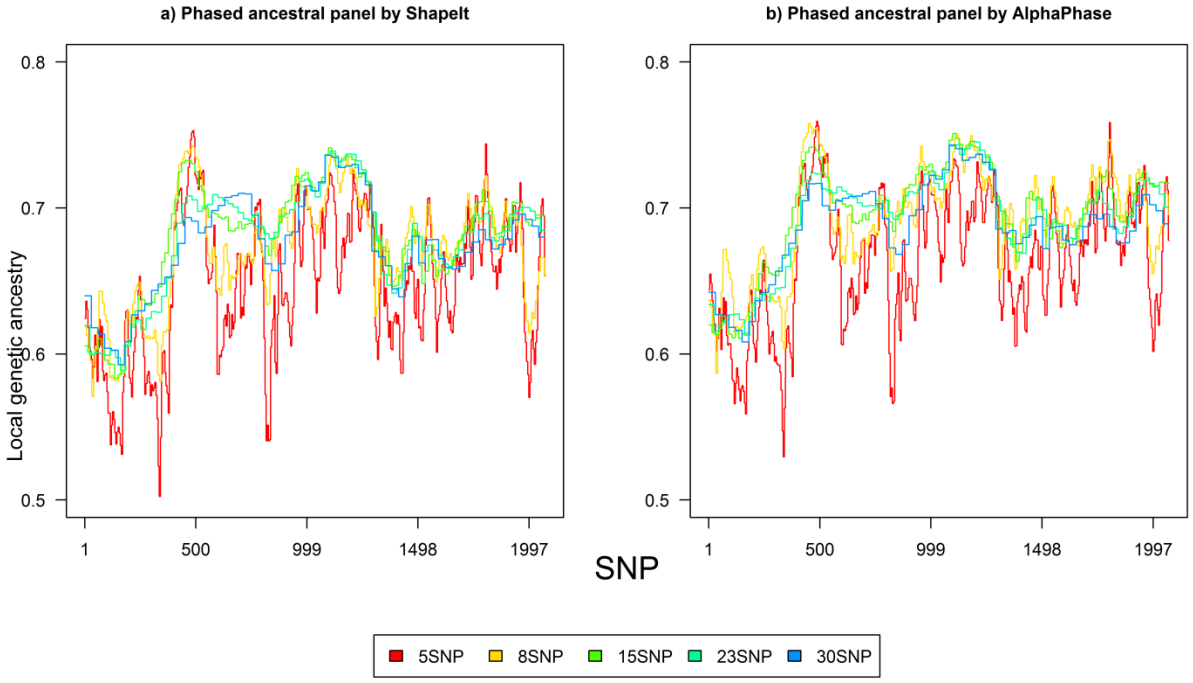


Fig. 3-3 LANC estimated by MULTIMIXgeno averaged within each window. Data were phased by a) *ShapeIt* and b) *AlphaPhase.*

MULTIMIX based on different sets for phasing ancestral haplotype were in the range of 0.84 to 0.99, 0.68 to 0.99, 0.60 to 0.99, 0.53 to 0.97, and 0.31 to 0.95 for chunk length 5, 8, 25, 23 and 30 SNPs respectively. The less correlated results emerged when we used cores longer than 300 SNP. Moreover, with increase in chunk length to 30, the other long length (core length 250 and general core and tail length 650) showed the more deviations (Additional file 3-6: Figure S3-6). Correlations between not-offset and offset decreased, when the chunk lengths increased to 23

and 30 SNPs for the data set phased by longer general core and tail length. For the sake of convenience, we took the average of not-offsets for first 3 phased lengths (with core length 100 SNP) along each chunk. Correlations between different chunk length were ranged from 0.30 to 0.89 (Table 3-2).

Comparison the results of different chunk lengths on MULTIMIX, which haplotypes were phased by *ShapeIt* and *AlphaPhase*, represented 0.95, 0.92, 0.92, and 0.90 correlations between chunk lengths 8, 15, 23 and 30 SNPs, respectively (Fig. 3-4).

Table 3-2 Correlations among LANC estimated by different window lengths with LAMP-LD and MULTIMIX, phased haplotypes from *ShapeIt* and *AlphaPhase*.

Correlations	LAMP-LD ShapeIt	LAMP-LD *AlphaPhase*	MULTIMIXgeno ShapeIt	MULTIMIXgeno *AlphaPhase*	MULTIMIX ShapeIt	MULTIMIX *AlphaPhase*
5SNP & 8SNP	0.94	0.94	0.87	0.87	0.69	0.67
5SNP & 15SNP	0.90	0.93	0.74	0.73	0.41	0.42
5SNP & 23SNP	0.90	0.92	0.67	0.68	0.32	0.30
5SNP & 30SNP	0.91	0.92	0.64	0.67	0.25	0.28
8SNP & 15SNP	0.96	0.98	0.85	0.84	0.74	0.76
8SNP & 23SNP	0.97	0.98	0.78	0.80	0.70	0.66
8SNP & 30SNP	0.97	0.96	0.70	0.75	0.59	0.60
15SNP & 23SNP	0.99*	0.99*	0.95	0.95	0.91	0.89
15SNP & 30SNP	0.98*	0.99*	0.90	0.93	0.86	0.82
23SNP & 30SNP	0.99*	0.99*	0.95	0.95	0.91	0.89

* indicated correlations equal and greater than 0.99

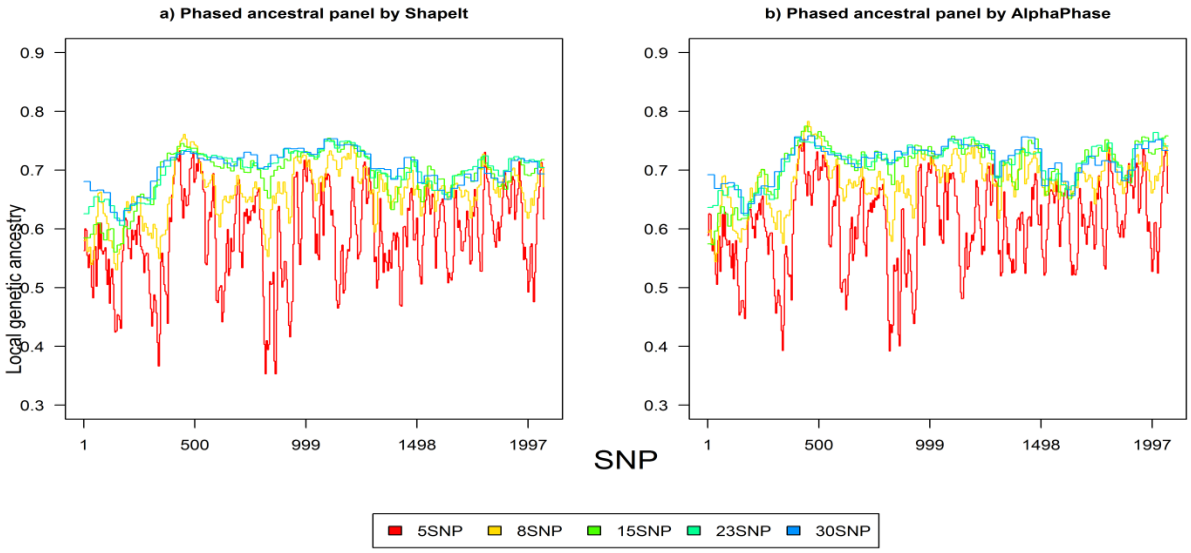


Fig. 3-4 LANC estimated by MULTIMIX_MCMC averaged within each window. Data were phased by a) *ShapeIt* and b) *AlphaPhas*.

3-3-3 Comparisons among LAMP, LAMP-LD and MULTIMIX

The results of these comparisons are given in Table 3-3. The highest correlations were observed between MULTIMIX_MCMCgeno and MULTIMIX_MCMC with window lengths 15 SNP (0.92) and 23 SNP (0.85). Correlation between LAMP-LD and MULTIMIX_MCMCgeno with chunk length 23 SNP was 0.85. Correlations between LAMP and the LAMP-LD (0.76), LAMP and MULTIMIXgeno with 23 SNP (0.72) were moderate, but it was lower with the other models of MULTIMIX_MCMC. A second statistic applied for comparison was the maximum absolute difference between LANC estimations (abs_diff) derived from different methods, which were minimum between MULTIMIX_MCMC, with 15 and 23SNP chunk lengths (0.054 and 0.066), and LAMP and LAMP-LD (0.059), LAMP-LD and MULTIMIX_MCMCgeno with 23 SNP (0.068). Comparisons of maximum absolute values with mean of absolute difference were also at the minimum level for the mentioned analyses. The highest difference between absolute difference and mean of absolute difference was between LAMP and MULTIMIX_MCMC with 15 and 30 SNP, LAMP-LD and MULTIMIX_MCMC with 15 and 30 SNP. Means of absolute difference were at minimum between LAMP-LD and MULTIMIX with 23 and 30 SNP, MULTIMIX_MCMC and MULTIMIX_MCMCgeno with 15 SNP, LAMP and LAMP-LD which showed 0.019 mean of absolute difference as a measure for least absolute error between

estimations of two compared methods. Comparison between the results of LAMP, LAMP-LD, MULTIMIX_MCMCgeno and MULTIMIX_MCMC with 23 SNPs are shown in Fig. 3-5.

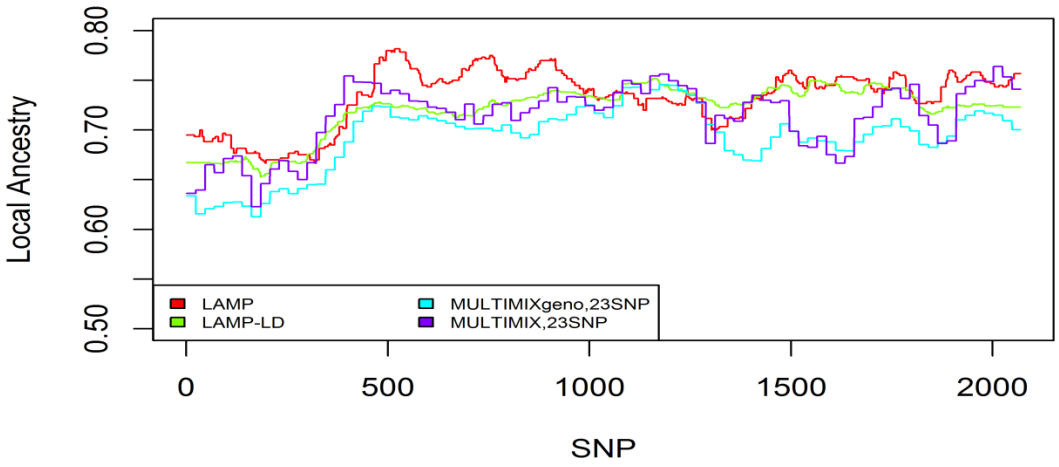


Fig. 3-5 LANC estimated by LAMP, LAMP-LD, MULTIMIXgeno and MULTIMIX_MCMC with 23 SNP in terms of window size. Data were phased by *AlphaPhase*.

3-3-4 Genome wide scale of local genetic ancestries

Here we calculated local genetic ancestries along the autosomes using LAMP, LAMP-LD, and MULTIMIX. Haplotypes were phased by *AlphaPhase* with core length 100 SNP and general core and length 300 SNP. We chose for window length 23 SNPs and number of states 15 to calculate local genetic ancestries using LAMP-LD. In addition, we ran MULTIMIX setting chunk length 23 SNPs (Fig. 3-6). In Additional file 3-7: Figure S3-7, Δ ancestries without standardized with standard deviations are shown. Hypothesis tests for deviation of Δ ancestry from normality with Bonferroni correction tests were employed to find selection signatures happened after admixture. For this, we calculated genome wide ancestries (local genetic ancestries averaged along autosomes) and then we standardized local genetic ancestries by the respective means and standard deviations derived from each program separately. Mean genome wide RHF ancestry estimates were 0.70, 0.69, 0.66 and 0.67, and standard deviations were 0.042, 0.027, 0.035 and 0.038 by LAMP, LAMP-LD, MULTIMIX_MCMCgeno and MULTIMIX_MCMC respectively (Fig.3-6).

The consistency of LANC estimation from different programs was not high, and different correlations were observed for different chromosomes. Comparisons between LAMP and LAMP-LD showed correlations > 0.80 for chromosomes 1, 17, 27 and 29, correlations of 0.70 to 0.80 were for chromosomes 2, 6, 11, 13, 16, 21, 23 and 25. Correlations for the rest of chromosomes were < 0.60 and for chromosomes 3, 14, 24 and 26 negative correlations were identified (Table 3-4).

Comparison of LAMP and MULTIMIX_MCMCgeno showed correlations > 0.80 for chromosomes 12, 20 and 29 and ranges from 0.7 to 0.8 for chromosomes 1, 2, 6, 16, 17, and 27. Correlations between Lamp and MULTIMIX_MCMC were at highest point for chromosomes 27 (0.86), 17 (0.85) and 29 (0.78). Results on LAMP-LD and MULTIMIX_MCMCgeno in most cases were higher than 0.7, except for chromosomes 20, 21 and 25. Similarly high correlations were observed between LAMP-LD and MULTIMIX_MCMC except for chromosomes 1, 9, 10 11, 16 19 to 26 and 28, but not as high as between LAMP-LD and MULTIMIX_MCMCgeno. Correlations between results on MULTIMIX_MCMCgeno and MULTIMIX_MCMC were high along most of the chromosomes except chromosomes 19 (0.50), 24 (0.66) and 28 (0.67) (Table 3-4). In case of using MULTIMIX_MCMCgeno and MULTIMX_MCMC misfitting probabilities were in most cases similar with averaged values of $\begin{pmatrix} 0.96 & 0.04 \\ 0.20 & 0.80 \end{pmatrix}$. Lambda also was in the range of 0.005 to 0.016.

Applying the hypothesis tests for deviation from normality with 5000 and 1000 hypotheses, two significant signals were detected on chromosomes 13 and 18 based on the results from LAMP program.

The regions on chromosome 13 (46.3-47.3 Mb) and chromosome 18 (18.7-25.9 Mb) surpass the threshold lines based on 1000 and 5000 hypotheses tests respectively and were regarded as signals of selection [18]. No other significant signals were detected from the result of LAMP-LD and MULTIMIX.

Table 3-3 Comparison among different algorithm to estimate local genetic ancestries on chromosome 2.

	Correlation[a]	Max abs_diff[b]	Mean abs_diff[c]
LAMP,LAMPLD	0.76	0.059	0.019
LAMP,MULTIMIX_MCMCgeno,15SNP	0.65	0.087	0.039
LAMP,MULTIMIX_MCMCgeno,23SNP	0.72	0.084	0.042
LAMP,MULTIMIX_MCMCgeno,30SNP	0.64	0.095	0.046
LAMP,MULTIMIX_MCMC,15SNP	0.57	0.122	0.033
LAMP,MULTIMIX_MCMC,23SNP	0.61	0.088	0.027
LAMP,MULTIMIX_MCMC,30SNP	0.54	0.098	0.027
LAMPLD,MULTIMIX_MCMCgeno,15SNP	0.81	0.075	0.025
LAMPLD,MULTIMIX_MCMCgeno,23SNP	0.85	0.068	0.028
LAMPLD,MULTIMIX_MCMCgeno,30SNP	0.79	0.069	0.033
LAMPLD,MULTIMIX_MCMC,15SNP	0.67	0.094	0.024
LAMPLD,MULTIMIX_MCMC,23SNP	0.72	0.081	0.017
LAMPLD,MULTIMIX_MCMC,30SNP	0.65	0.086	0.017
MULTIMIX_MCMCgeno,MULTIMIX_MCMC,15SNP	0.92	0.054	0.017
MULTIMIX_MCMCgeno,MULTIMIX_MCMC,23SNP	0.85	0.066	0.023
MULTIMIX_MCMCgeno,MULTIMIX_MCMC,30SNP	0.79	0.077	0.029

Correlation[a] is Pearson correlation. Max abs_diff[b] is absolute difference between maximum values estimated from pairs comparing methods. Mean abs_diff[c is] mean of absolute difference between values estimated from pairs comparing methods

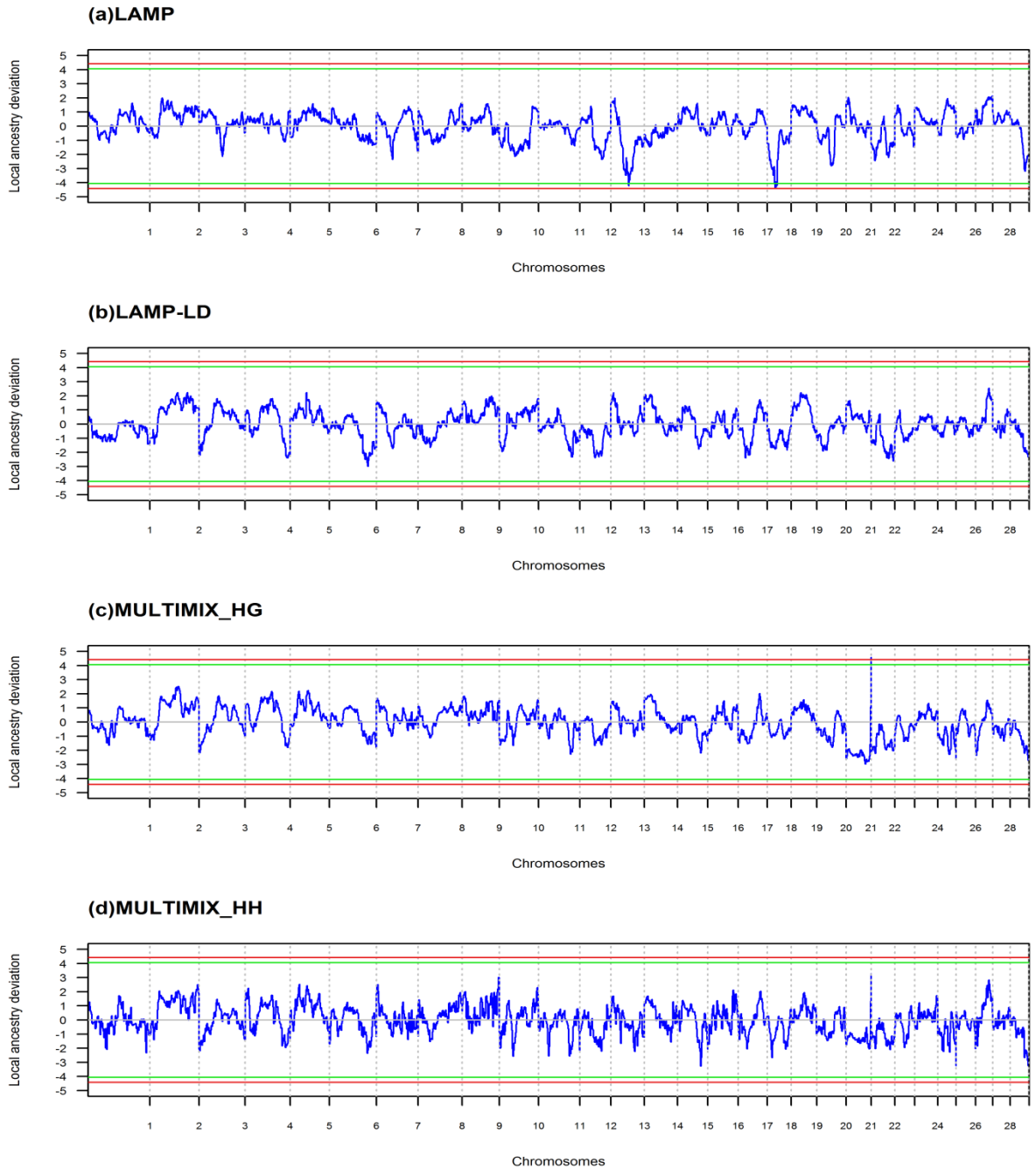


Fig. 3-6 LANC standardized by its mean and standard deviations along whole 29 autosomes, estimated by a) LAMP, b) LAMP-LD, c) MULTIMIXgeno and d) MULTIMIX_MCMC with 23 SNP in terms of window size. Data were phased by *AlphaPhase* (100, 100, not-offset). Green and red lines are thresholds based on p-value < 5×10^{-5} (4.06 SDs) and p-value < 1×10^{-5} (4.42 SDs) respectively based on hypothesis tests.

Table 3-4 Comparison among whole 29 autosomes (should become smaller and more informative)

Chromosomes	C1	C2	C3	C4	C5	C6
1	0.84	0.74	0.55	0.78	0.62	0.79
2	0.74	0.71	0.63	0.84	0.76	0.85
3	-0.45	-0.53	-0.40	0.93	0.93	0.95
4	0.26	0.32	0.46	0.97	0.85	0.91
5	0.23	0.56	0.43	0.78	0.84	0.89
6	0.77	0.72	0.55	0.93	0.81	0.87
7	0.67	0.48	0.37	0.89	0.74	0.85
8	0.62	0.56	0.69	0.90	0.76	0.87
9	0.14	0.16	-0.08	0.74	0.60	0.73
10	0.36	0.32	0.54	0.87	0.68	0.80
11	0.72	0.61	0.37	0.75	0.49	0.85
12	0.66	0.89	0.64	0.79	0.78	0.76
13	0.71	0.65	0.55	0.88	0.76	0.91
14	-0.70	-0.72	-0.56	0.94	0.82	0.82
15	0.54	0.67	0.54	0.87	0.86	0.93
16	0.76	0.79	0.59	0.88	0.69	0.83
17	0.86	0.78	0.85	0.89	0.85	0.89
18	0.63	0.47	0.38	0.80	0.84	0.77
19	0.68	0.53	0.02	0.86	0.32	0.50
20	0.24	0.82	0.60	0.50	0.53	0.85
21	0.79	0.18	0.07	0.30	0.00	0.76
22	0.13	0.52	0.55	0.72	0.77	0.84
23	0.71	0.67	0.42	0.90	0.63	0.83
24	-0.35	-0.61	-0.39	0.71	0.25	0.66
25	0.79	-0.04	0.12	0.24	0.48	0.78
26	-0.42	-0.32	-0.50	0.77	0.56	0.72
27	0.87	0.70	0.86	0.80	0.87	0.87
28	0.68	0.31	0.47	0.77	0.69	0.67
29	0.92	0.86	0.78	0.94	0.89	0.87

Columns 1 to 6 are related to Pearson's correlations between LAMP and LAMP_LD, LAMP and MULTIMIXgeno, LAMP and MULTIMIX_MCMC, LAMP_LD and MULTIMIXgeno, LAMP_LD and MULTIMIX_MCMC, MULTIMIXgeno and MULTIMIX_MCMC.

3-4 Discussion

In the present study, we used three different software tools which apply different methods for LANC inference using whole-genome SNP genotypes in order to investigate the extent of overlap between estimations. Extreme deviations of LANC from average genome wide ancestry are interpreted as post-admixture selection signatures in recently admixed populations [9-11].

We applied developed analyses panel containing 361 different analyses. We ran LAMP with one set of parameters regarding recombination rate and number of generations since admixture. Window size is optimized internally with this software. LAMP-LD and MULTIMIX_MCMCgeno were applied to phased ancestral haplotypes and unphased samples genotypes whereas MULTIMIX_MCMC was run on both phased ancestral and admixed samples. Results from different methods were very comparable at the global level (> 0.99). However, there was considerable difference at local estimates [38, 39].

Correlations near to unity were observed between estimations when genotypes were phased using different length of windows by *ShapeIt*. Results of phasing with *AlphaPhase* had high correlations when data were phased by core lengths 100 to 150 SNPs and total core tail lengths of 300 to 500 SNPs. They were also highly correlated with *ShapeIt* results.

Using MULTIMIX for LANC inference with haplotypes phased by longer cores gave deviations of phasing results and caused increase in λ, a quantity required for making covariance matrix invertible. Increasing λ reduces fitting the model and smooths errors [22]. Therefore the results were less correlated and we decided not to consider the longer cores further.

Results of LAMP-LD showed consistency when window lengths were 15, 23 and 30 SNPs. Defining small windows with 5 and 8 SNPs produced noise in estimations. Since Swiss Fleckvieh is a crossbred population with small number of generations, very short haplotypes cannot be expected.

Results of both MULTIMIX_MCMCgeno and MULTIMIX_MCMC showed high difference between chunk lengths. MULTIMIX estimates LANC at window level based on the fact that recently admixed individuals have long stretches of loci with same shared ancestry. Therefore

the results are different and very susceptible to changes in chunk length. Defining short chunk length (5 and 8 SNP) can cause noise as observed for LAMP-LD.

Estimates of local ancestries using LAMP-LD and MULTIMIX were not highly correlated with LAMP estimations (Table 3-3). The most comparable analyses were MULTIMIX_MCMCgeno and MULTIMIX_MCMC with 15 SNP (0.92) and 23 SNP (0.85) as well as between LAMP-LD and MULTIMIX_MCMCgeno with 15 SNP (0.85). In general the correlations of MULTIMIX_MCMCgeno with MULTIMIX_MCMC and LAMP-LD with MULTIMIX_MCMCgeno results, they were not high, implying big deviations in LANC patterns. Unfortunately, computation under the coalescent model with recombination is difficult to find number of ancestral recombination graphs compatible with admixed samples. The other key limitation of these methods is that they assume uniformly exponential distribution of ancestral haplotypes along the genome of admixed individuals, which is a reasonable assumption for low rates of admixture. However, the admixture tracts are stochastically larger than expected under exponential distribution in the recent admixed populations [19, 40-41]. The advantage of LAMP and PCAdmix is that they do not require assumption of parametric genetic model (ancestral recombination graphs). Moreover the benefit of LAMP is that it chooses for the optimal window sizes internally based on assumed admixture parameters (admixture intensity and age of admixture) [19].

Finally we used multiple hypothesis tests for normality of local ancestry deviations, assuming 1000 and 5000 independent segments to tag the entire genome. In our previous study [18], significant signals were detected for local genetic ancestries estimated with LAMP on chromosome 13 (46.3-47.3 Mb) and on chromosome 18 (18.7-25.9 Mb), where a similar signal but wider were found on chromosome 18 (9.06-24.6 Mb) based on extend haplotype homozygosity approach [18]. These signals were not confirmed by local ancestry deviation with LAMP-LD and MULTIMIX and no other significant signals were found with those two approaches in the current study.

3-5 Conclusions

This study considered different methodologies to estimate local genetic ancestries as indicator of recent selection signature in Swiss Fleckvieh composite breed. The results of a recent study on the same data set [18] from clustering method applied by LAMP showed signals on chromosomes 13 and 18 which were some similar signals by *EHH* methodology. Using alternative methodology with phased haplotypes using coalescent model in LAMP-LD and MULTIMIX could not capture these or any other signals across the autosome. The results suggest that care should be taken when interpreting selection signature based on local ancestry detected by a single method and confirmation with alternative approaches is advisable.

References

1. Sankararaman S, Sridhar S, Kimmel G, Halperin E: **Estimating local ancestry in admixed populations**. *Am J Hum Genet* 2008, **82**(2):290-303.

2. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S: **Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations**. *Plos Genet* 2009, **5**(6).

3. Jin WF, Li R, Zhou Y, Xu SH: **Distribution of ancestral chromosomal segments in admixed genomes and its implications for inferring population history and admixture mapping**. *Eur J Hum Genet* 2014, **22**(7):930-937.

4. Padhukasahasram B: **Inferring ancestry from population genomic data and its applications**. *Frontiers in genetics* 2014, **5**.

5. Zhang JQ, Stram DO: **The Role of Local Ancestry Adjustment in Association Studies Using Admixed Populations**. *Genet Epidemiol* 2014, **38**(6):502-515.

6. Hu YN, Willer C, Zhan XW, Kang HM, Abecasis GR: **Accurate Local-Ancestry Inference in Exome-Sequenced Admixed Individuals via Off-Target Sequence Reads**. *Am J Hum Genet* 2013, **93**(5):891-899.

7. Liu YS, Nyunoya T, Leng SG, Belinsky SA, Tesfaigzi Y, Bruse S: **Softwares and methods for estimating genetic ancestry in human populations**. *Hum Genomics* 2013, **7**.

8. Thornton TA, Bermejo JL: **Local and Global Ancestry Inference and Applications to Genetic Association Analysis for Admixed Populations**. *Genet Epidemiol* 2014, **38**:S5-S12.

9. Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ: **Recent genetic selection in the ancestral admixture of Puerto Ricans**. *Am J Hum Genet* 2007, **81**(3):626-633.

10. Oleksyk TK, Smith MW, O'Brien SJ: **Genome-wide scans for footprints of natural selection**. *Philos T R Soc B* 2010, **365**(1537):185-205.

11. Jin WF, Xu SH, Wang HF, Yu YG, Shen YP, Wu BL, Jin L: **Genome-wide detection of natural selection in African Americans pre- and post-admixture**. *Genome Res* 2012, **22**(3):519-527.

12. Bhatia G, Tandon A, Patterson N, Aldrich MC, Ambrosone CB, Amos C, Bandera EV, Berndt SI, Bernstein L, Blot WJ *et al*: **Genome-wide Scan of 29,141 African Americans Finds No Evidence of Directional Selection since Admixture**. *Am J Hum Genet* 2014, **95**(4):437-444.

13. Gautier M, Naves M: **Footprints of selection in the ancestral admixture of a New World Creole cattle breed**. *Mol Ecol* 2011, **20**(15):3128-3143.

14. Qanbari S, Strom TM, Haberer G, Weigend S, Gheyas AA, Turner F, Burt DW, Preisinger R, Gianola D, Simianer H: **A High Resolution Genome-Wide Scan for Significant Selective Sweeps: An Application to Pooled Sequence Data in Laying Chickens**. *Plos One* 2012, **7**(11).

15. Flori L, Thevenon S, Dayo GK, Senou M, Sylla S, Berthier D, Moazami-Goudarzi K, Gautier M: **Adaptive admixture in the West African bovine hybrid zone: insight from the Borgou population**. *Mol Ecol* 2014, **23**(13):3241-3257.

16. Kim ES, Rothschild MF: **Genomic adaptation of admixed dairy cattle in East Africa**. *Frontiers in genetics* 2014, **5**.

17. Bahbahani H, Clifford H, Wragg D, Mbole-Kariuki MN, Van Tassell C, Sonstegard T, Woolhouse M, Hanotte O: **Signatures of positive selection in East African Shorthorn Zebu: A genome-wide single nucleotide polymorphism analysis**. *Sci Rep-Uk* 2015, **5**.

18. Khayatzadeh N, Meszaros G, Utsunomiya YT, Garcia JF, Schnyder U, Gredler B, Curik I, Solkner J: **Locus-specific ancestry to detect recent response to selection in admixed Swiss Fleckvieh cattle**. *Animal genetics* 2016, **47**(6):637-646.

19. Schraiber JG, Akey JM: **Methods and models for unravelling human evolutionary history**. *Nat Rev Genet* 2015, **16**(12):727-740.

20. Li N, Stephens M: **Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data**. *Genetics* 2003, **165**(4):2213-2233.

21. Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, Rodriguez-Cintron W, Chapela R, Ford JG, Avila PC *et al*: **Fast and accurate inference of local ancestry in Latino populations**. *Bioinformatics* 2012, **28**(10):1359-1367.

22. Churchhouse C, Marchini J: **Multiway admixture deconvolution using phased or unphased ancestral panels**. *Genet Epidemiol* 2013, **37**(1):1-12.

23. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data**. *Am J Hum Genet* 2001, **68**(4):978-989.

24. Stephens M, Scheet P: **Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation**. *Am J Hum Genet* 2005, **76**(3):449-462.

25. Scheet P, Stephens M: **A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase**. *Am J Hum Genet* 2006, **78**(4):629-644.

26. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering**. *Am J Hum Genet* 2007, **81**(5):1084-1097.

27. Delaneau O, Marchini J, Zagury JF: **A linear complexity phasing method for thousands of genomes**. *Nat Methods* 2012, **9**(2):179-181.

28. Browning SR, Browning BL: **Haplotype phasing: existing methods and new developments**. *Nat Rev Genet* 2011, **12**(10):703-714.

29. Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, Olason PI, Ingason A, Steinberg S, Rafnar T *et al*: **Detection of sharing by descent, long-range phasing and haplotype imputation**. *Nature genetics* 2008, **40**(9):1068-1075.

30. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, van der Werf JH: **A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes**. *Genet Sel Evol* 2011, **43**:12.

31. Frkonja A, Gredler B, Schnyder U, Curik I, Solkner J: **Prediction of breed composition in an admixed cattle population**. *Animal genetics* 2012, **43**(6):696-703.

32. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ *et al*: **PLINK: A tool set for whole-genome association and population-based linkage analyses**. *Am J Hum Genet* 2007, **81**(3):559-575.

33. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets**. *Gigascience* 2015, **4**.

34. Delaneau O, Coulonges C, Zagury JF: **Shape-IT: new rapid and accurate algorithm for haplotype inference**. *BMC bioinformatics* 2008, **9**:540.

35. Barbato M, Orozco-terWengel P, Tapio M, Bruford MW: **SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data**. *Frontiers in genetics* 2015, **6**:109.

36. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals**. *Genome Res* 2009, **19**(9):1655-1664.

37. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME: **Novel multilocus measure of linkage disequilibrium to estimate past effective population size**. *Genome Res* 2003, **13**(4):635-643.

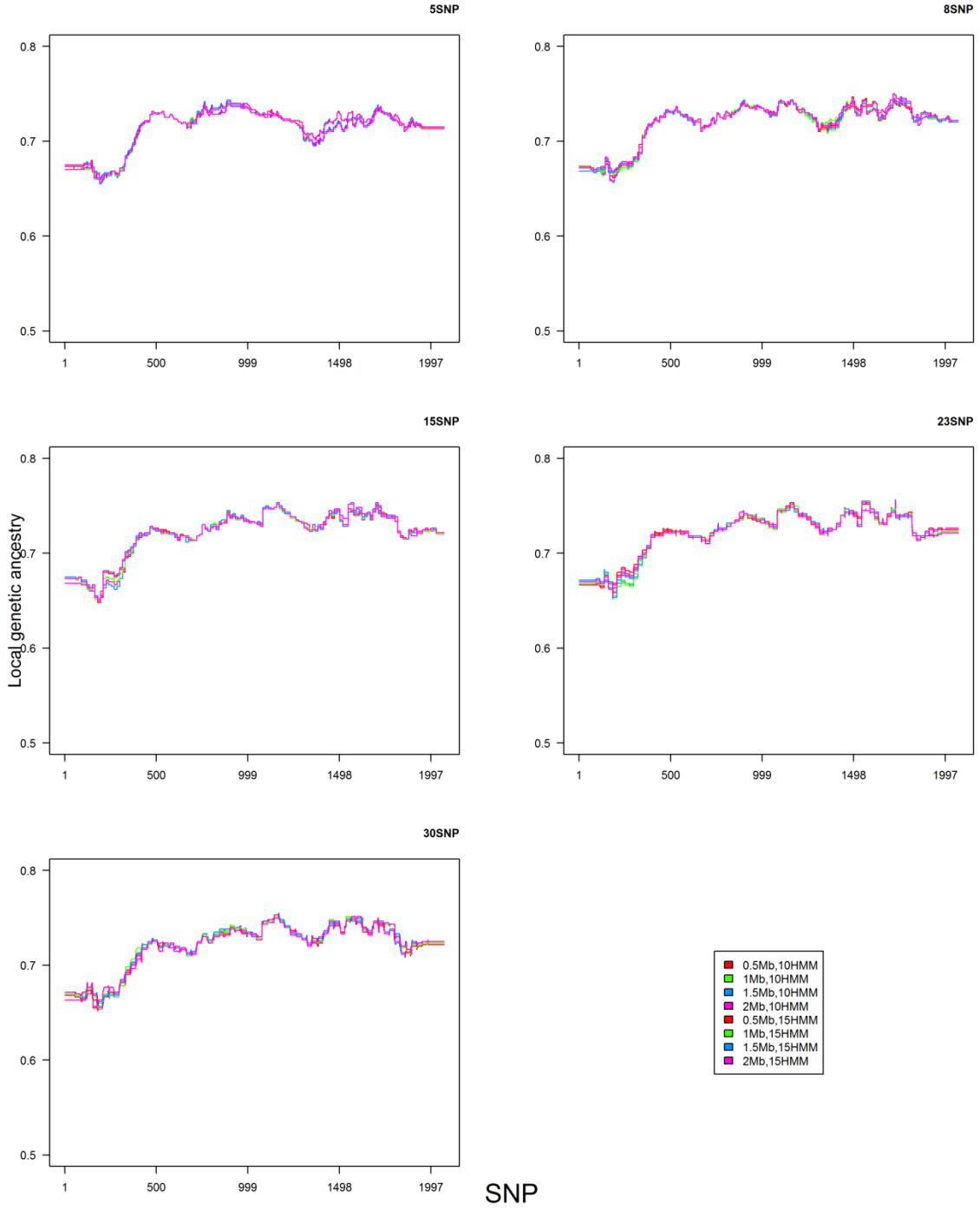
38. Chen M, Yang C, Li C, Hou L, Chen X, Zhao H: **Admixture mapping analysis in the context of GWAS with GAW18 data**. *BMC proceedings* 2014, **8**(Suppl 1):S3.

39. Khayatzadeh N, Meszaros, G., Gredler B., Schnyder U., Curik I., Soelkner, J.: **Estimation of local genetic ancestry in an admixed cattle population applying different methods**. *Acta agriculturae Slovenica* 2016, **5**:6.

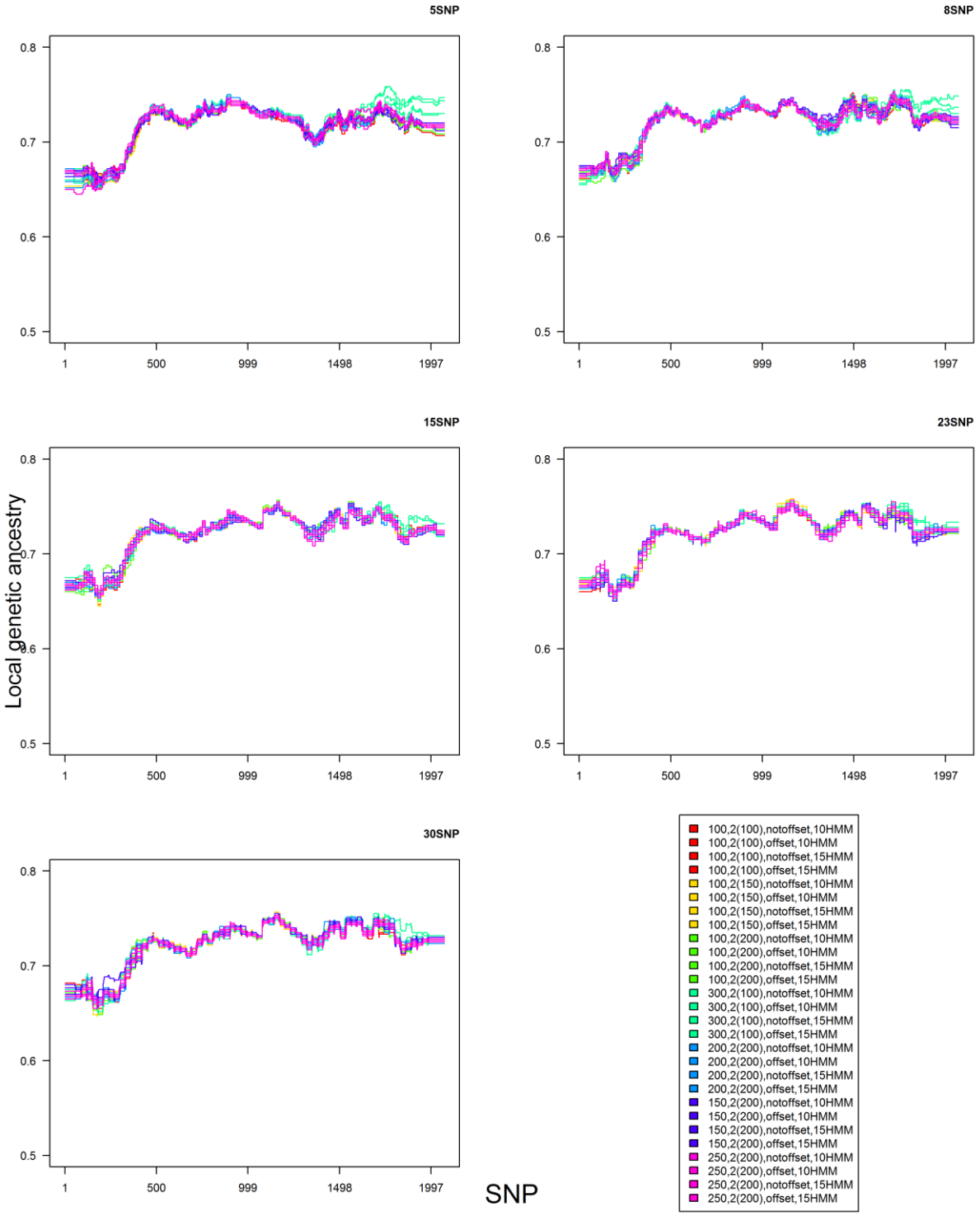
40. Pool J, Nielsen R.: **Inference of historical changes in migration rate from the lengths of migrant tracts**. *Genetics* 2009, **181**:711-719.

41. Liang M, Nielsen R.: **The lengths of admixture tracts**. *Genetics* 2014, **197**:953-967

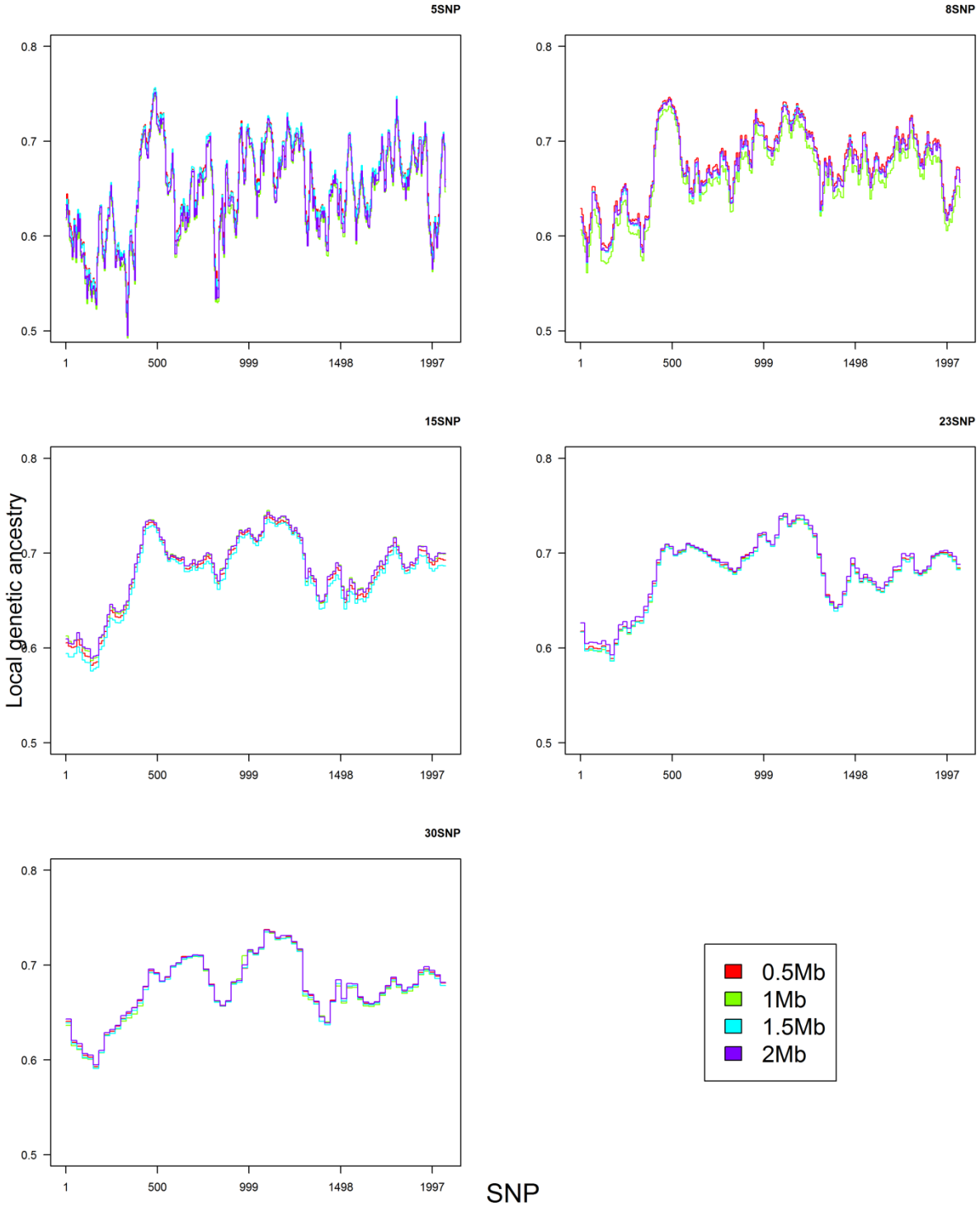
Additional file 3-1: Figure S3-1.
LANC proportions were estimated by LAMP-LD using different window size (5, 8, 15, 23 and 30 SNP) and 10 and 15 number of states. Ancestral haplotypes were phased by *ShapeIt* with different window lengths (0.5, 1, 1.5 and 2 Mb).



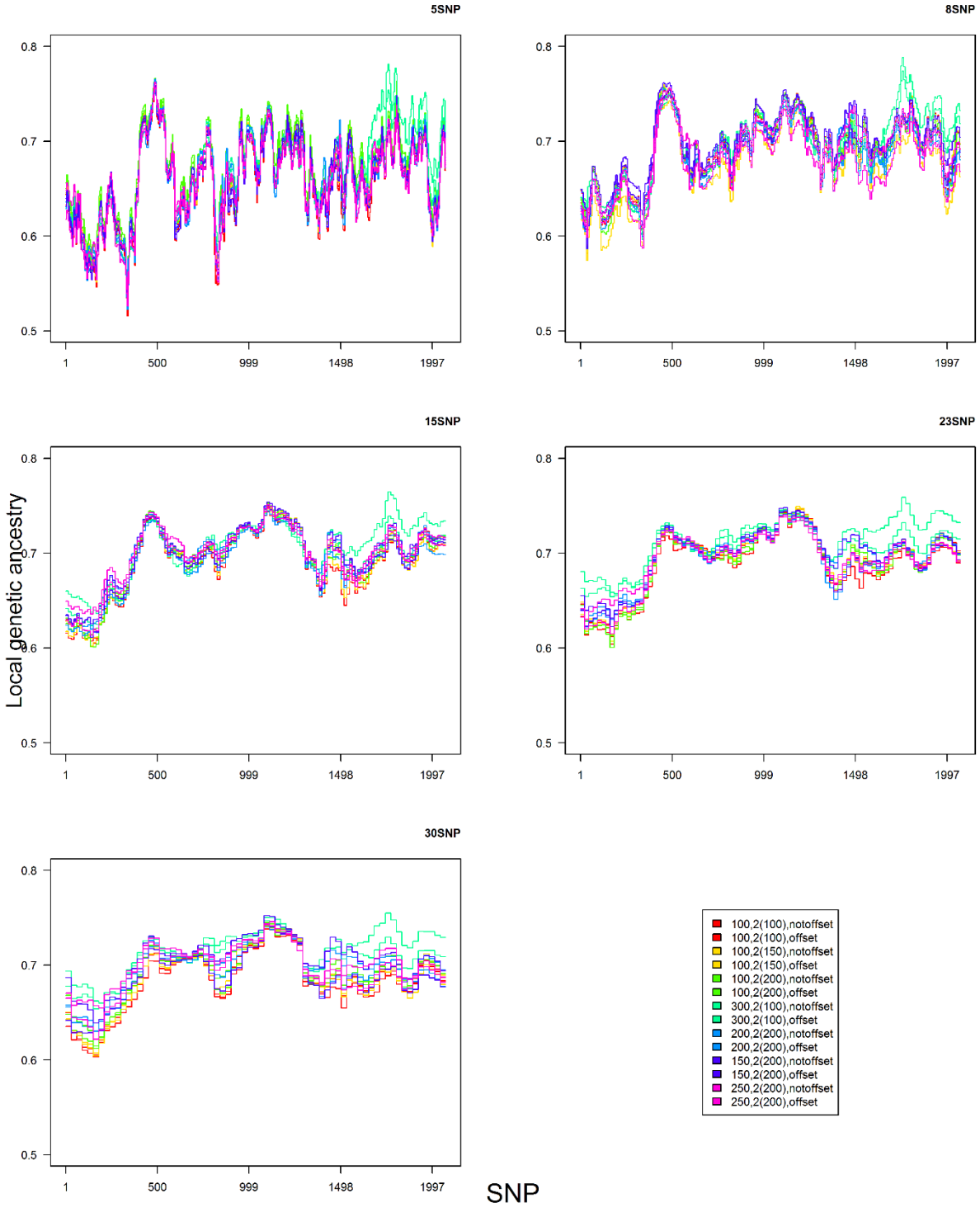
Additional file 3-2: Figure S3-2.
LANC proportions were estimated by LAMP-LD using different window size (5, 8, 15, 23 and 30 SNP) and 10 and 15 number of states. Ancestral haplotypes were phased by *AlphaPhase* with different general core and tail lengths.



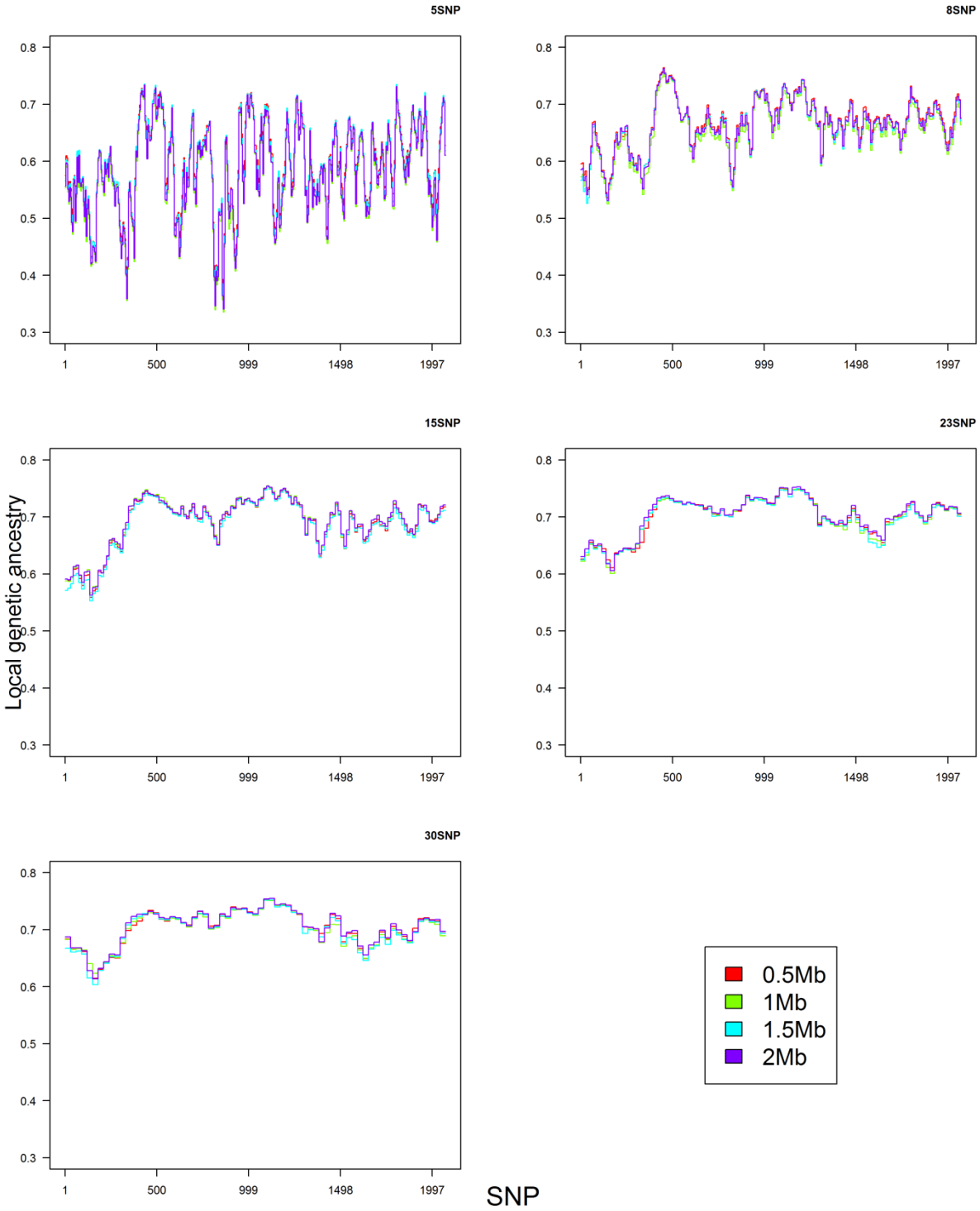
Additional file 3-3: Figure S3-3.
LANC proportions were estimated by MULTIMIXgeno using different window size (5, 8, 15, 23 and 30 SNP). Ancestral haplotypes were phased by *ShapeIt* with different window lengths (0.5, 1, 1.5 and 2 Mb).



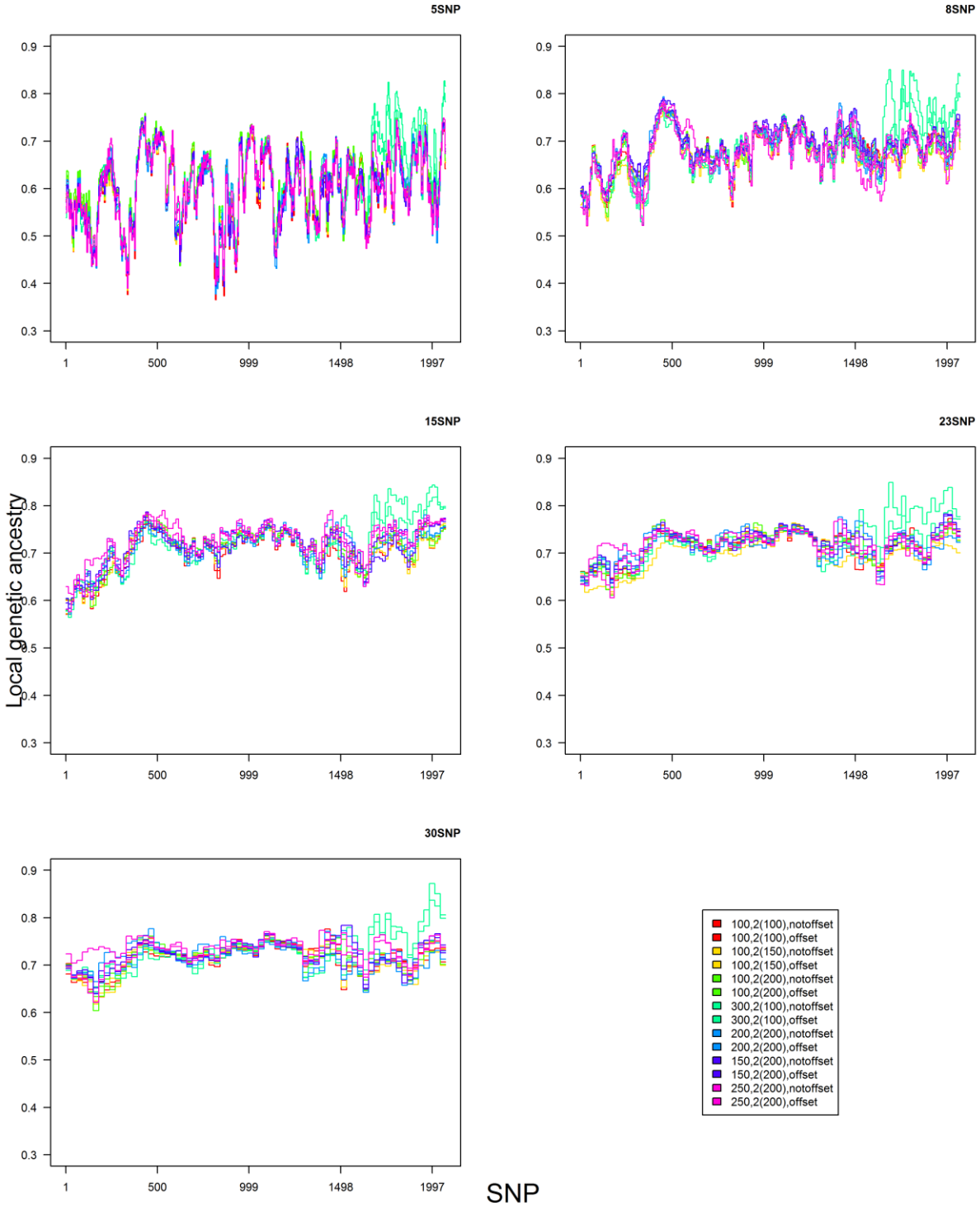
Additional file 3-4: Figure S3-4.
LANC proportions were estimated by MULTIMIXgeno using different window size (5, 8, 15, 23 and 30 SNP). Ancestral haplotypes were phased by *AlphaPhase* with different general core and tail lengths.



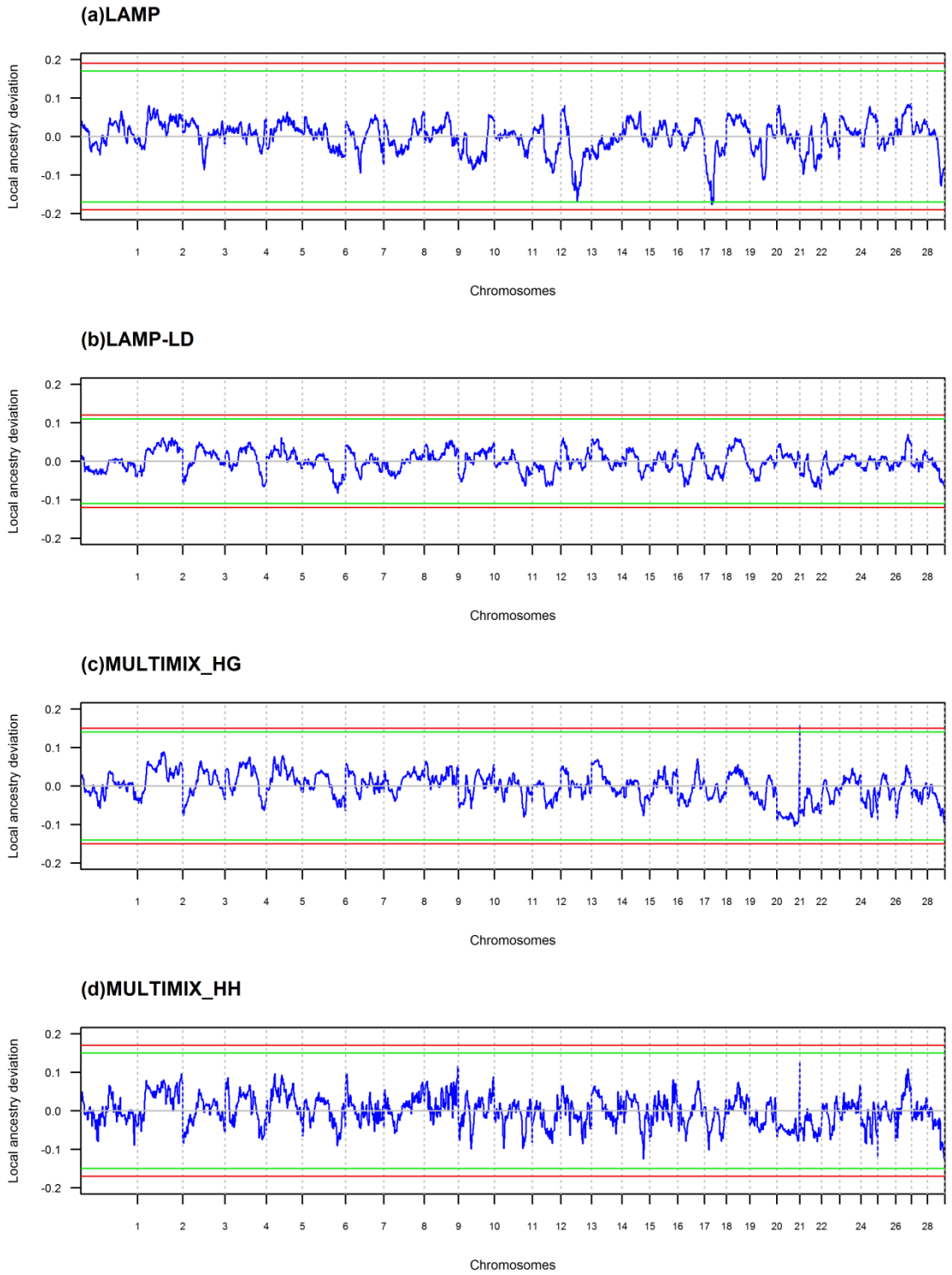
Additional file 3-5: Figure S3-5.
LANC proportions were estimated by MULTIMIX_MCMC using different window size (5, 8, 15, 23 and 30 SNP). Ancestral haplotypes were phased by *ShapeIt* with different window lengths (0.5, 1, 1.5 and 2 Mb).



Additional file 3-6: Figure S3-6.
LANC proportions were estimated by MULTIMIX_MCMC using different window size (5, 8, 15, 23 and 30 SNP). Ancestral haplotypes were phased by *AlphaPhase* with different general core and tail lengths.



Additional File 3-7: Figure S3-7.
Δ ancestries along whole 29 autosomes, estimated by a) LAMP, b) LAMP-LD, c) MULTIMIXgeno and d) MULTIMIX_MCMC with 23 SNP in terms of window size. Data were phased by *AlphaPhase* (100, 100, not-offset). Green and red lines are thresholds based on p-value < 5×10^{-5} and p-value < 1×10^{-5} respectively based on hypothesis tests.



Declarations

Abbreviations
LANC: Local Ancestry; RHF: Red Holstein Frisian; SI: Simmental

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

Availability of data and materials
Requests for access to the genotypes to facilitate replication of the published results can be addressed to Swissherdbook.

Authors' contributions
JS conceived the original idea of the study and together with NKH, GM and YTU further developed the idea and decided on the set of analysis. NKH did the statistical analysis and wrote the text. US and BG collected the data for the analysis and together with IC and JFG critically reviewed the text. All authors approved the final version of the manuscript.

Competing interests
The authors declare they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate

Doses of sperm routinely collected for artificial insemination provided the tissue used for genotyping. Therefore, no ethics statement was required for collecting genetic materials.

Chapter 4

Effects of breed proportion and components of heterosis for semen traits in a composite cattle breed

N. Khayatzadeh [(1)] | G. Mészáros [(1)] | Y. T. Utsunomiya [(2)] | F. Schmitz-Hsu [(3)] | F. Seefried [(4)] | U. Schnyder [(4)] | M. Ferenčaković [(5)] | J.F. Garcia [(2, 6)] | I. Curik [(5)] | J. Sölkner [(1)]

(1) BOKU, Department of Sustainable Agricultural Systems, Division of Livestock Sciences (NUWI), Gregor-Mendel-Strasse, 1180 Vienna, Austria
(2) Departamento de Medicina Veterinária Preventiva e Reprodução Animal, Faculdade de Cinêcias Agrárias Veterinárias, UNESP – Univ Estadual Paulista, Araçatuba, São Paulo, Brazil
(3) Swissgenetics, Meielenfeldweg 12, Postfach 466,3052 Zollikofen, Switzerland
(4) Qualitas AG, Chamerstrasse 56, Ch-6300, Zug, Switzerland
(5) Department of Animal Science, Faculty of Agriculture, University of Zagreb, Svetošimunska cesta 25, 10000 Zagreb, Croatia
(6) Departamento de Apoio, Saúde e Produção Animal, Faculdade de Medicina Veterinária de Araçatuba, UNESP – Univ Estadual Paulista, Araçatuba, São Paulo, Brazil

Journal of Animal Breeding and Genetics
Accepted

Summary

The aim of this study was to estimate the non-additive genetic effects of the dominance component of heterosis as well as epistatic loss on semen traits in admixed Swiss Fleckvieh, a composite of Simmental and Red Holstein Friesian cattle. Heterosis is the additional gain in productivity or fitness of crossbred progeny over the mid purebred parental populations. Intra-locus gene interaction usually has a positive effect while epistatic loss generally reduces productivity or fitness due to lack of evolutionarily established interactions of genes from different breeds. Genotypic data on 38,205 SNP of 818 admixed, as well as 148 Red Holstein Friesian and 213 Simmental bulls as the parental breeds were used to predict breed origin of alleles. The genome wide locus-specific breed ancestries of individuals were used to calculate effects of breed difference as well as the dominance component of heterosis while proxies for two definitions of epistatic loss were derived from 100,000 random pairs of loci. The average Holstein Friesian ancestry in admixed bulls was estimated 0.82. Results of fitting different linear mixed models showed including the dominance component of heterosis considerably improved the model adequacy for three of the four traits. Inclusion of epistatic loss increased the accuracy of the models only for our new definition of the epistatic effect for two traits while the other definition was so highly correlated with the dominance component that statistical separation was impossible.

Keywords breed composition, breed heterosis, epistatic loss, semen traits, Swiss Fleckvieh

4-1 | INTRODUCTION

Crossbreeding is widely used in animal and plant breeding to optimize average genetic merit of performance traits by introducing the favorable parental genes, decreasing inbreeding depression and maintaining the gene interactions that cause heterosis (VanRaden & Sanders, 2003). Systematic crossbreeding programs in dairy cattle have improved functional traits, which deteriorate with high selection pressure in pure breeding, inbreeding depression and antagonistic genetic correlations between functional and production traits (Ferencakovic et al., 2017; Freyer et al., 2008; Rauw et al., 1998; Sorensen et al., 2008).

Heterosis is the opposite of inbreeding depression and results from an increase in heterozygosity. It measures the degree that offspring exceed the average of the performance of their parental populations (Falconer & Mackay, 1996; Shull, 1948). The extent of heterosis depends on the genetic distance between the parental populations, the number of involved parental populations and the type of crossbreeding program. The genetic mechanisms underlying heterosis are favorable non-additive gene effect due to dominance and over dominance which is attributed to advantageous combinations of alleles at heterozygous loci and epistasis as a result of interaction among loci (Amuzu-Aweh et al., 2013; Lynch & Walsh, 1998).

The generally positive intra-locus component of heterosis is based on the relation $d \times y^2$, d being the dominance effect and y the allele frequency difference between parental populations (Falconer & Mackay, 1996). Unfavorable gene effect due to the breakdown of the beneficiary associated parental gene complex, which accumulated within breeds through long-term selection in purebreds, is called “epistatic loss” (Koch et al., 1985). Kinghorn (1983) established the term epistatic loss (e_x) as the probability that two non-allelic genes in diploid individuals (derived from either one or both parents) chosen at random originate from different breeds.

The amount of general heterosis for production traits in dairy cattle was reported in the range of 3 to 4 percent, while higher levels of heterosis were observed for functional and reproductive traits (Freyer et al., 2008; Kargo et al., 2012; Sorensen et al., 2008; VanRaden & Sanders, 2003).

The objective of the present study was to develop animal mixed models with breed ancestry proportion, the dominance component of heterosis and epistatic loss effects as fixed covariates based on genome wide data on semen traits of bulls in a data set including admixed Swiss

Fleckvieh as well as the two pure parental breeds Simmental and Red Holstein Friesian. We compared different genetic models to find the most appropriate mixed model for semen traits in admixed Swiss Fleckvieh.

4-2 | MATERIALS AND METHODS

4-2-1 | Phenotypic data, herd and management description

Swiss Fleckvieh is a composite cattle breed, which was introduced over the last 40 years in Switzerland with the aim of combining the strengths of both originating breeds, Simmental (SI) and Red Holstein Frisian (RHF). Swiss Fleckvieh breeding program relies on high milk yield derived from RHF ancestry as well as dual-purpose characters, functional and fitness traits of the SI breed.

Animals with a pedigree-based Red Holstein blood share of 0.125-0.875 are registered as Swiss Fleckvieh. Animals with < 0.125 RHF are considered Simmental and those with > 0.875 RHF are categorized as Red Holstein Frisian in Swiss herd book. In the current study, we did not follow the formal definition of Swiss Fleckvieh and considered all of the bulls with a range of 0.01-0.99 RHF blood proportion as admixed animals.

Phenotypic records on semen production and semen quality traits were made available by Swissgenetics from Mülligen artificial insemination (AI) station in Switzerland. Bulls are generally kept in tie-stalls. Semen is collected twice a week, using a teaser bull to prepare bulls and then semen is collected with a dummy and artificial vagina. In this AI station, ejaculates are collected 1 or 2 times per day for the same bull. Records from 2000 to 2015 were considered. The routinely recorded traits for each ejaculate were volume (ml), concentration (10^9/ml) and percentage of live sperm. Total number of spermatozoa for each ejaculate was calculated by multiplying volume with concentration. Percentage of live sperm is determined by visual assessment. In total 68,475 records were available from 1298 bulls (171 RHF, 226 SI and 901 admixed Swiss Fleckvieh bulls) born between 1990 and 2014. Bulls with at least 10 records were kept for analysis. Ejaculates volumes in the range of 1-25 ml and concentrations in the range of 0.1-3×10^9/ml were kept for analysis. Ejaculates with interval less than three days since recent

ejaculation were removed. Since number of spermatozoa were not normally distributed, the following transformation was performed (Box & Cox, 1964).

$$transformed\ total\ number\ of\ spermatozoa = (total\ number\ of\ spermatozoa^{0.3} - 1)/0.3$$

Observations on transformed number of spermatozoa and percentage of live sperm beyond the range mean ± 3 standard deviations were discarded. Number of records, means and standard deviations for semen traits are presented in Table 4-1.

TABLE 4-1 Overall phenotypic mean (± standard deviation), number of bulls and number of records on semen traits.

Semen traits	Mean (±standard deviation)	No. of bulls	No. of records
Ejaculate volume (ml)	5.56±2.50	1177	42442
Ejaculate concentration (10^9/ml)	1.39±0.46	1176	42355
Number of spermatozoa(transformed)	2.62±0.94	1176	42270
Live spermatozoa (%)	86.28±3.38	1169	41749

4-2-2 | Genotype data

Bulls were genotyped using Illumina® Bovine SNP 50k, 150k and 777k BeadChip (1420 bulls). The genotypic data consisted of a subset of 44,999 SNP for 1411 bulls.

The standard quality control for genotypic data was performed with PLINK 1.90 (Chang et al., 2015; Purcell et al., 2007) to remove monomorphic SNP with call rate < 0.95 and those SNP deviated from Hardy Weinberg equilibrium with *P*-value < 10^{-5} from data set; 38,205 SNP for 1179 (148 RHF, 213 SI and 818 admixed) bulls were used for the analyses. To estimate the locus-specific ancestry at each SNP position, we used LAMP 2.5 (Sankararaman et al., 2008) in LAMP ancestry mode with similar configurations applied in our previous study (Khayatzadeh et al., 2016).

4-2-3 | Additive genetic (breed), heterosis effect and epistatic loss effects

Breed composition for pure RHF bulls was coded as 1 and for pure SI bulls was coded as 0. Breed composition for admixed bulls was computed by taking the average RHF proportions for all SNPs across the 29 autosomes based on the LAMP results. For a single individual and a

single locus, values could be 0, 0.5 and 1. The intra-locus ("dominance") component of heterosis as a fixed covariate was also calculated based on LAMP results, and was set to 1 where both alleles at each single SNP derived from different ancestral populations and 0 where both alleles came from the same breed origin. Values were averaged across the autosomes for admixed bulls while this quantity was set to 0 for purebred bulls.

Epistatic loss was modelled following the definition of Kinghorn (1983) where the epistatic effect is proportional to the probability that two non-allelic genes randomly chosen in the diploid individual are of different breed origin. We randomly chose 100,000 times one allele each from two different SNP along the autosome, derived from either one or both parents, for each admixed bull. Epistatic loss was set to 0 when the two non-allelic genes had the same ancestral origin and were set to 1 when they came from different ancestral populations.

In a second definition, to our knowledge not previously applied, we randomly sampled 100,000 times two different SNP for every admixed bull. Epistatic loss was set to 1 only when both alleles of one SNP derived from one breed and both alleles of the second SNP derived from the other breed following the local ancestry algorithm. This setting reflects the extreme situation of losing breed specific epistatic combinations. In the case of Kinghorn's definition described above, at least one favorable combination of non-allelic genes derived from the same population is possible.

4-2-4 | Statistical analyses

The fixed effects considered were age of bull (< 16 months, 16-72 months and > 72 months), assistant (semen collector), contemporary group (year-season of collection) and interval days between two consecutive ejaculates. The elapse between two consecutive ejaculations was also categorized into three different levels (3-6 days, 7-9 days and > 9 days interval). Season effect was defined as categorical variable (February to May, June to September and October to January). The four models considering different combinations of genetic effects were:

(1) $y_{ijklmn} = \mu + \alpha_i + age_j + contempgroup_k + elapse_l + assitant_m + \varepsilon_{ijklmn}$

(2) $y_{ijklmno} = \mu + \alpha_i + age_j + contempgroup_k + elapse_l + assitant_m + breedpercent_n + \varepsilon_{ijklmno}$

(3) $y_{ijklmnop} = \mu + \alpha_i + age_j + contempgroup_k + elapse_l + assitant_m + breedpercent_n + domhet_o + \varepsilon_{ijklmnop}$

(4a) $y_{ijklmnopq} = \mu + \alpha_i + age_j + contempgroup_k + elapse_l + assitant_m + breedpercent_n + domhet_o + epstloss1_p + \varepsilon_{ijklmnopq}$

(4b) $y_{ijklmnopq} = \mu + \alpha_i + age_j + contempgroup_k + elapse_l + assitant_m + breedpercent_n + domhet_o + epstloss2_p + \varepsilon_{ijklmnopq}$

where $y_{ijklmn(opq)}$ denotes observations for each bull in a population containing both purebred ancestral and admixed bulls, μ is the overall mean, α_i is the random permanent environmental effect of each bull, age_j, $contempgroup_k$, $elapse_l$, $assitant_m$ are the fixed effects related to age of bulls, year season (contemporary groups), ejaculate intervals and assistant. $breedpercent_n$, $domhet_o$, $epstloss1_p$ and $epstloss2_p$ are the regression coefficients for breed ancestry proportion (proportion of RHF), dominance component of heterosis, epistatic loss (Kinghorn, 1983), epistatic loss (our definition) and $\varepsilon_{ijklmn(opq)}$ is the random error associated with each observation in animal models. The analyses were performed using proc MIXED and maximum likelihood method in SAS (Garoia et al., 2007), following methodology applied by (Ferencakovic et al., 2017).

The goodness of fit of each of the genetic models (including ancestral breed proportion, breed heterosis and different definitions of epistatic loss) was tested by likelihood ratio test (-2logliklihood) which is asymptotically Chi-square. The degree of freedom is associated with ratio test is equal to the difference between number of parameters in two compared models (Wilks, 1938). Moreover, a second-order bias correction of Akaike information criterion is reported, which is corrected based on samples size (Anderson, 2008):

$$AICc = -2\log likelihood + 2k(k+1)/(n-k-1)$$

where *k* is number of estimated parameters and *n* is the number of observations. The model with the smallest AICc value is usually the preferred model (Akaike, 1974).The difference of AICc of different models was calculated using (Burnham & Anderson, 2002) model selection. ΔAIC smaller than 2 indicates no substantial differences between models. There is less support for

models with ΔAIC between 3 and 7, while ΔAIC greater than 7 indicate substantially less fitting models.

4-3 | RESULTS

4-3-1 | Genomic breed proportions and general statistics

In this study we applied genomic data to predict the ancestral breed proportions at both global and local (SNP) level in admixed Swiss Fleckvieh bulls. To specify the global ancestry proportions contributed from each ancestral population in admixed Swiss Fleckvieh bulls, unsupervised clustering analysis was performed by ADMIXTURE (Alexander et al., 2009). The average ancestry proportions were estimated 0.82 RHF and 0.18 SI (0.16 SD) which were highly correlated (0.97) with pedigree ancestry proportions with average 0.85 RHF and 0.15 SI, respectively.

Descriptive statistics including phenotypic means, standard deviation, number of bulls and number of observations for the studied traits are summarized in Table 4-1.

4-3-2 | Breed difference

Differences between both RHF and SI ancestral populations were estimated for semen traits. Comparison between purebred bulls using univariate animal models revealed that there were significant differences for volume and transformed number of spermatozoa ($p < 0.01$) in favor of SI ancestry, and for percentage of live sperm ($p < 0.05$) in favor of RHF. No significant difference between purebreds was detected for concentration. Estimates of mean difference between parental purebreds are presented in Table 4-2.

TABLE 4-2 *F*-value and means estimates for breed ancestry proportions between RHF and SI purebred bulls for semen traits.

Semen traits	*F*-value	Estimates	LS means ± standard errors	
			RHF	SI
Volume (ml)	9.63 **	-0.63 (±0.20)	5.62±0.18	6.23±0.14
Concentration (10^9/ml)	3.55 n.s	-0.07 (±0.04)	1.38±0.03	1.45 ±0.03
No. of spermatozoa (transformed)	8.39 **	-0.20 (±0.07)	2.70±0.07	2.90±0.05
Live spermatozoa (%)	4.90 *	0.68 (±0.31)	85.79±0.27	85.11±0.22

LS means denotes the least square means

4-3-3 | Fixed effects

The analysis of variance of model 1 (base model) showed significant effects of bull age, contemporary group, time elapsed between ejaculations and semen collector on all evaluated traits (Table 4-4).

4-3-4 | Models comparisons

Breed ancestry proportions, dominance component of heterosis, epistatic loss1 (Kinghorn, 1980, 1983) and epistatic loss2 were added as fixed genetic covariates in animal models for all evaluated semen traits in order to find the importance of the inclusion of these effects in the model. Table 4-3 shows likelihood ratio test results, AICc and ΔAIC between models with different fitted covariates.

The base model (model 1) included fixed environmental effects and the random effect of bull. Breed ancestry proportion effect was included in model 2 by calculating the average genome-wide RHF ancestry proportions estimated by LAMP along autosomes for each admixed bull. In model 3, the effect of the dominance component of heterosis was considered. In the last group of models, epistatic loss1 (model 4a) or epistatic loss2 (model 4b) were included as well.

Likelihood ratio results indicated considerably high significant differences between models 2 and 1 for percentage of live sperm ($p < 0.0001$). Comparatively less improvement in model accuracy was observed for volume ($p < 0.01$) and transformed number of spermatozoa ($p < 0.05$) by involving breed proportion effect in the model. There was no improvement for concentration.

Comparison of models 3 and 2 for volume, transformed number of spermatozoa and percentage of live sperm represented significant improvement in model accuracy ($p < 0.0001$), while concentration was not significantly affected by the dominance component of heterosis.

Inclusion of the effect of epistatic loss 1, model 4a (Kinghorn, 1983) yielded no substantial increase in the accuracy for any of the traits considered.

For model 4b, we used our definition for epistatic loss. Including epistatic loss 2 improved considerably the model accuracy for volume ($p < 0.0001$). Significant improvement was also

observed for transformed number of spermatozoa ($p < 0.05$), while no improvement was detected for percentage of live sperm and concentration.

With AICc model comparison, model 4b including epistatic loss 2 clearly outperformed all other models for volume. The lowest ΔAIC was 11.4, for model 3. For concentration, models including genetic effects related to crossbreeding did not improve the model fit, compared to the base model 1. Model 4b was identified as the best fitting for transformed number of spermatozoa. The model including breed proportion and heterosis was still reasonable (ΔAIC = 3.8), better than model 4a (ΔAIC = 5.5). Models with breed proportion only (model 2) or the base model (model 1) fitted substantially worse (ΔAIC >13). For percentage of live sperm, models with the dominance component of heterosis (model 3) and epistatic effects (models 4a and 4b) showed all similar fit, substantially improved over the base model and the model including breed proportion only (ΔAIC > 13).

TABLE 4-3 Model comparisons based on likelihood ratio tests and AICc for by including different covariates in the model.

Traits	Models	Log likelihood	L.Ratio	AICc	ΔAIC
Volume	Model 1	-88279.31		176692.80	38.20
(ml)	Model 2	-88275.35	7.92^{**}	176686.90	32.30
	Model 3	-88263.96	22.78^{****}	176666.20	11.40
	Model 4a	-88263.96	$0.0028^{n.s}$	176668.20	13.60
	Model 4b	-88257.19	13.55^{****}	176654.60	0
Concentration	Model 1	-17096.00		34325.90	0.20
(10^9/ml)	Model 2	-17095.00	$2.22^{n.s}$	34325.70	0
	Model 3	-17094.74	$0.01^{n.s}$	34327.70	2.00
	Model 4a	-17094.53	$0.41^{n.s}$	34329.30	3.60
	Model 4b	-17094.61	$0.26^{n.s}$	34329.40	3.70
No. of spermatozoa	Model 1	-48862.53		97859.30	18.10
(transformed)	Model 2	-48859.38	6.30^{*}	97855.00	13.80
	Model 3	-48853.38	12.00^{****}	97845.00	3.80
	Model 4a	-48853.25	$0.26^{n.s}$	97846.70	5.50
	Model 4b	-48850.48	5.81^{*}	97841.20	0
Live sperm	Model 1	-98750.00		197634.90	30.60
(%)	Model 2	-98741.93	16.85^{****}	197620.10	15.80
	Model 3	-98733.05	17.76^{****}	197604.30	0
	Model 4a	-98732.88	$0.33^{n.s}$	197606.00	1.70
	Model 4b	-98732.96	$0.18^{n.s}$	197606.20	1.90

- L.Ratio is likelihood ratio values
- AICc is the Akaike information criteria (second order bias correction)and ΔAIC is the difference of the model with minimum AICc with the other models
- $^{****}p < 0.0001$, $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{n.s}p > 0.1$

4-3-5 | Estimates of genetic effects

Mean estimates (± standard error) for all models including different fixed genetic effects are represented in Table 4-4.

Based on model 2, breed ancestry proportion were significant for volume, transformed number of spermatozoa and percentage of live sperm and regression coefficients for breed ancestry proportions were estimated -0.40 (0.14), -0.11 (0.04) and 0.76 (0.18), similar to breed difference estimates using purebred animals only (see Table 4-2).

Including the dominance component of heterosis in the model was highly significant for volume, number of spermatozoa and percentage of live sperm. It increased the volume 1.24 (±0.26) ml, transformed number of spermatozoa 0.28 (±0.08) and live spermatozoa 1.40 (±0.33) %.

Considering epistatic loss 1 in the model did not significantly improve the model and the estimates for these effects were not significant for any of the traits. This effect in the model was confounded with dominance component of heterosis. Epistatic loss 2 in the model was estimated significant for volume and transformed number of spermatozoa. Regression coefficients of epistatic loss 2 for these two traits were -7.6 (±2.05) and -1.57 (0.66), respectively which indicated 7.6 ml less volume and 1.58 less transformed number of spermatozoa are expected in admixed population, due to epistatic loss.

TABLE 4-4 Analysis of variance and mean estimates for breed ancestry proportion, dominance component of heterosis, and epistatic loss for (a) volume, (b) concentration, (c) no. of spermatozoa and (d) percentage of live spermatozoa

(a) Volume (ml)

Models	Fixed effects	Df	*F*-value	Estimates	Standard error
Model 1	Age	2	1406.81****		
	Contemporary group	47	25.06****		
	Elapse	2	95.19****		
	Assistant	13	4.41****		
Model 2	Breed proportion	1	7.97**	-0.40	0.14
Model 3	Breed proportion	1	14.37****	-0.55	0.14
	Dominance	1	22.95****	1.24	0.26
Model 4a	Breed proportion	1	13.45***	-0.55	0.15
	Dominance	1	1.48$^{n.s}$	1.30	1.06
	Epistatic loss (1)	1	0.00$^{n.s}$	-0.07	1.28
Model 4b	Breed proportion	1	15.73****	-0.57	0.14
	Dominance	1	36.44****	2.03	0.34
	Epistatic loss (2)	1	13.55***	-7.56	2.05

(b) Concentration (10^9/ml)

Models	Fixed effects	df	*F*-value	Estimates	Standard error
Model 1	Age	2	204.11****		
	Contemporary group	47	7.88****		
	Elapse	2	14.52****		
	Assistant	13	1.83*		
Model 2	Breed proportion	1	2.22$^{n.s}$	-0.04	0.03
Model 3	Breed proportion	1	2.18$^{n.s}$	-0.04	0.03
	Dominance	1	0.01$^{n.s}$	0.00	0.05
Model 4a	Breed proportion	1	2.51$^{n.s}$	-0.04	0.03
	Dominance	1	0.36$^{n.s}$	-0.11	0.19
	Epistatic loss (1)	1	0.41$^{n.s}$	0.15	0.23
Model 4b	Breed proportion	1	2.10$^{n.s}$	-0.04	0.03
	Dominance	1	0.07$^{n.s}$	-0.02	0.06
	Epistatic loss (2)	1	0.26$^{n.s}$	0.19	0.37

(c) Transformed no. of spermatozoa

Models	Fixed effects	Df	*F*-value	Estimates	Standard error
Model 1	Age	2	1802.51****		
	Contemporary group	47	14.96****		
	Elapse	2	20.11****		
	Assistant	13	3.92****		
Model 2	Breed proportion	1	6.31*	-0.11	0.04
Model 3	Breed proportion	1	10.21**	-0.15	0.05
	Dominance	1	12.01***	0.28	0.08
Model 4a	Breed proportion	1	10.39**	-0.15	0.05
	Dominance	1	0.12$^{n.s}$	0.12	0.34
	Epistatic loss (1)	1	0.26$^{n.s}$	0.21	0.41
Model 4b	Breed proportion	1	10.95***	-0.15	0.04
	Dominance	1	17.68****	0.45	0.11
	Epistatic loss (2)	1	5.81*	-1.57	0.65

(d) Percentage of live sperm

Models	Fixed effects	df	*F*-value	Estimates	Standard error
Model 1	Age	2	626.10****		
	Contemporary group	47	14.13****		
	Elapse	2	136.70****		
	Assistant	13	1.88*		
Model 2	Breed proportion	1	16.96****	0.76	0.18
Model 3	Breed proportion	1	10.31**	0.60	0.19
	Dominance	1	17.91****	1.40	0.33
Model 4a	Breed proportion	1	10.61**	0.62	0.19
	Dominance	1	2.53$^{n.s}$	2.16	1.36
	Epistatic loss (1)	1	0.33$^{n.s}$	-0.95	1.64
Model 4a	Breed proportion	1	10.16**	0.59	0.19
	Dominance	1	12.43***	1.52	0.43
	Epistatic loss (2)	1	0.18$^{n.s}$	-1.12	2.63

- $^{****}p < 0.0001$, $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{n.s}p > 0.1$
- df is degree of freedom
- Dominance represents the dominance component of heterosis

4-3-6 Relationships of levels of heterosis and different definitions of epistatic loss

To evaluate the potential confounding of genetic effects related to crossbred performance, we calculated the correlations between the dominance components of heterosis and the two definitions of epistatic loss. The correlations between breed heterosis effects and epistatic loss 1 and epistatic loss 2 were estimated 0.97 and 0.64, respectively. The relationship is graphically represented in Figure 4-1).

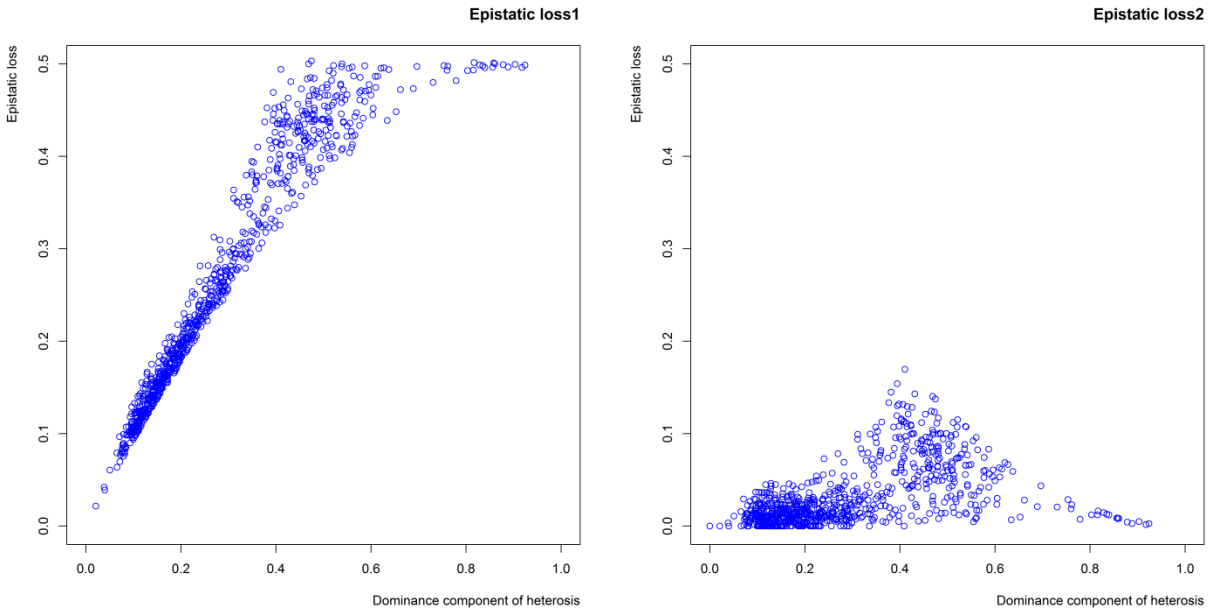


Figure 4-1 Scatter plot and Pearson's correlation of epistatic loss 1 and 2 with breed heterosis

4-4 | DISCUSSION

In the present study, contributions of non-additive genetic components of heterosis were tested for semen traits in admixed Swiss Fleckvieh bulls with different mixed models. We used genomic information to define these effects in the model. For this, we worked exclusively with breed ancestry information of SNPs, not the actual genotypes. This means that a SNP may be homozygous in state but "breed-heterozygous" in our definition because the two alleles were derived from different breeds, and vice versa. Including this definition of the dominance component of heterosis improved significantly the model accuracy for volume, number of spermatozoa and percentage of live sperm with very clear positive effects. The two definitions of epistatic effects yielded considerably different results. Applying the definition of Kinghorn

(1983) yielded non-significant effects. Considering the proportion of pairs of loci with one locus having both alleles derived from one breed and the other with both alleles with ancestry origin from the opposite parental population, significant negative effects were found for volume and number of spermatozoa.

Please note that we did not include Dickerson's (1973) classical definition of epistatic effects, which he called recombination loss, in this study. Deriving the corresponding values requires knowledge of the parental phase at each locus and across chromosomes, which we could not derive with accuracy.

Some experimental studies involving crossbred beef cattle (Dillard et al., 1980) and dairy cattle (Robison et al., 1981) which exclusively used purebred sires concluded that epistatic effects are of little or no importance. Moreover, exclusive use of purebred sires set up a relationship between the coefficient of dominance and the coefficient of epistatic loss.

We estimated the correlations of the dominance component of heterosis with the two definitions of epistatic loss. It was extremely high (0.97) for epistatic loss based on Kinghorn's definition (Figure 4-1), not allowing statistical separation. Confounding of these effects was also reported by (Fries et al., 2002). The correlation was lower (0.64), but still high, with our new parameter of epistatic loss.

The consequence of these dependencies is that much of the variation in epistatic effects is explained by the dominance component of a regression model fitting dominance alone. Thus having accounted for epistatic effect, it could equally be concluded that dominance or heterozygosity effects are of little or no importance (Kinghorn, 1983). We are not aware of any recent studies considering genomic predictors of genome-wide epistatic loss in crossbred livestock populations.

The concept of epistasis is important for understanding the genetic architecture of traits. Recent studies applying genetic models (e.g., Maki-Tanila & Hill, 2014) indicate only small contributions of epistatic effects to the genetic variance of traits in outbred populations. Our study did not estimate non-additive variances but estimated effects of components of heterosis in a crossbred population by regression. While epistatic loss was significant for two of the four traits investigated, it is hard to tell how this transforms into non-additive genetic variance.

4-5 | CONCLUSION

Crossbred populations provide the unique opportunity to study the effects of components of heterosis on traits potentially influenced by these effects. In this study, we used genomic information to replace the traditional probabilistic interpretation of non-additive genetic effects of heterosis in a crossbred population. For ejaculate volume, transformed number of spermatozoa and percentage of live sperm in admixed Swiss Fleckvieh bulls, estimates of the dominance component of heterosis were 1.24 ml, 0.28 and 1.40 %. We therefore expect more volume, number of spermatozoa and percentage of live sperm compared with bulls of the parental breeds. Epistatic effects, applying a new genome-wide definition, were significant and negative for volume (-7.56 ml) and transformed number of spermatozoa (-1.57). Separation of the dominance component of heterosis and genome-wide indicators of epistatic effects is difficult. Significant effects were found for a definition of proportion of pairs of loci derived from the opposite parental populations.

ACKNOWLEDGEMENTS

We would like to thank the Swissherdbook cooperative Zollikofen for providing genotypes and Swissgenetics for providing phenotypes for the analysis. Y. T. Utsunomiya was supported by Sao Paulo Research Foundation - FAPESP (process 2014/01095-8 and 2016/07531-0).

REFERENCES

Akaike, H. (1974). New Look at Statistical-Model Identification. *Ieee Transactions on Automatic Control, Ac19*(6), 716-723. doi: Doi 10.1109/Tac.1974.1100705

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. [Research Support, N.I.H., Extramural]. *Genome Research, 19*(9), 1655-1664. doi: 10.1101/gr.094052.109

Amuzu-Aweh, E. N., Bijma, P., Kinghorn, B. P., Vereijken, A., Visscher, J., van Arendonk, J. A., & Bovenhuis, H. (2013). Prediction of heterosis using genome-wide SNP-marker data: application to egg production traits in white Leghorn crosses. [Research Support, Non-U.S. Gov't]. *Heredity (Edinb), 111*(6), 530-538. doi: 10.1038/hdy.2013.77

Anderson, D. R. (2008). *Model based inference in the life sciences : a primer on evidence*. New York ; London: Springer.

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 26*(2), 211-252.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference : a practical information-theoretic approach* (2nd ed. ed.). New York ; London: Springer.

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. [Research Support, N.I.H., Intramural

Research Support, Non-U.S. Gov't]. *Gigascience, 4*, 7. doi: 10.1186/s13742-015-0047-8

Dillard, E. U., Rodriguez, O., & Robison, O. W. (1980). Estimation of additive and nonadditive direct and maternal genetic effects from crossbreeding beef cattle. *Journal of Animal Science, 50*(4), 653-663.

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics* (4th ed. ed.). Harlow: Longman.

Ferencakovic, M., Solkner, J., Kaps, M., & Curik, I. (2017). Genome-wide mapping and estimation of inbreeding depression of semen quality traits in a cattle population. *Journal of Dairy Science, 100*(6), 4721-4730. doi: 10.3168/jds.2016-12164

Freyer, G., Konig, S., Fischer, B., Bergfeld, U., & Cassell, B. G. (2008). Invited review: crossbreeding in dairy cattle from a German perspective of the past and today. [Review]. *Journal of Dairy Science, 91*(10), 3725-3743. doi: 10.3168/jds.2008-1287

Fries, L. A., Schenkel, F. S., Roso, V. M., Brito, F. V., Severo, J. L. P., & Piccoli, M. L. (2002). "Epistazygosity" and Epistatic Effects. [Proceeding]. *7th World Congress on Genetics Applied to Livestock Production, Monpellier, France*, 5.

Garoia, F., Guarniero, I., Grifoni, D., Marzola, S., & Tinti, F. (2007). Comparative analysis of AFLPs and SSRs efficiency in resolving population genetic structure of Mediterranean Solea vulgaris. [Comparative Study

Research Support, Non-U.S. Gov't]. *Molecular Ecology, 16*(7), 1377-1387. doi: 10.1111/j.1365-294X.2007.03247.x

Kargo, M., Madsen, P., & Norberg, E. (2012). Short communication: Is crossbreeding only beneficial in herds with low management level? *Journal of Dairy Science, 95*(2), 925-928. doi: 10.3168/jds.2011-4707

Khayatzadeh, N., Meszaros, G., Utsunomiya, Y. T., Garcia, J. F., Schnyder, U., Gredler, B., . . . Solkner, J. (2016). Locus-specific ancestry to detect recent response to selection in admixed Swiss Fleckvieh cattle. *Animal Genetics, 47*(6), 637-646. doi: 10.1111/age.12470

Kinghorn, B. (1980). The Expression of Recombination Loss in Quantitative Traits. *Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie-Journal of Animal Breeding and Genetics, 97*(2), 138-143.

Kinghorn, B. (1983). Genetic-Effects in Crossbreeding .3. Epistatic Loss in Crossbred Mice. *Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie-Journal of Animal Breeding and Genetics, 100*(3), 209-222.

Koch, R. M., Dickerson, G. E., Cundiff, L. V., & Gregory, K. E. (1985). Heterosis Retained in Advanced Generations of Crosses among Angus and Hereford Cattle. *Journal of Animal Science, 60*(5), 1117-1132.

Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sunderland, Ma.: Sinauer.

Maki-Tanila, A., & Hill, W. G. (2014). Influence of gene interaction on complex trait variation with multilocus models. [Research Support, Non-U.S. Gov't]. *Genetics, 198*(1), 355-367. doi: 10.1534/genetics.114.165282

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. [Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't]. *American Journal of Human Genetics, 81*(3), 559-575. doi: 10.1086/519795

Rauw, W. M., Kanis, E., Noordhuizen-Stassen, E. N., & Grommers, F. J. (1998). Undesirable side effects of selection for high production efficiency in farm animals: a review. *Livestock Production Science, 56*(1), 15-33. doi: Doi 10.1016/S0301-6226(98)00147-X

Robison, O. W., McDaniel, B. T., & Rincon, E. J. (1981). Estimation of direct and maternal additive and heterotic effects from crossbreeding experiments in animals. *Journal of Animal Science, 52*(1), 44-50.

Sankararaman, S., Sridhar, S., Kimmel, G., & Halperin, E. (2008). Estimating local ancestry in admixed populations. *American Journal of Human Genetics, 82*(2), 290-303. doi: 10.1016/i.ajhg.2007.09.022

Shull, G. H. (1948). What Is "Heterosis"? *Genetics, 33*(5), 439-446.

Sorensen, M. K., Norberg, E., Pedersen, J., & Christensen, L. G. (2008). Invited review: crossbreeding in dairy cattle: a Danish perspective. [Review]. *Journal of Dairy Science, 91*(11), 4116-4128. doi: 10.3168/jds.2008-1273

VanRaden, P. M., & Sanders, A. H. (2003). Economic merit of crossbred and purebred US dairy cattle. *Journal of Dairy Science, 86*(3), 1036-1044. doi: 10.3168/jds.S0022-0302(03)73687-X

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics, 9*, 60-62. doi: DOI 10.1214/aoms/1177732360

Chapter 5

Genome-wide mapping of heterosis for percentage of live sperm in admixed Swiss Fleckvieh bulls

N. Khayatzadeh[1], G. Mészáros[1], M. Ferenčaković[2], Y. T. Utsunomiya[3], U. Schnyder[4], F. Seefried[4], I. Curik[2] & J. Sölkner[1]

[1] *University of Natural Resources and Life Sciences, Department of Sustainable Agricultural Systems, Division of Livestock Sciences, Gregor-Mendel-Strasse, 1180 Vienna, Austria*
Negar.khayatzadeh@students.boku.ac.at
[2] *University of Zagreb, Faculty of Agriculture, Department of Animal Science, Svetošimunska cesta 25, 10000 Zagreb, Croatia*
[3] *UNESP – Univ Estadual Paulista, Faculdade de Medicina Veterinária de Araçatuba, Departamento de Apoio, Saúde e Produção Animal, Araçatuba, São Paulo, Brazil*
[4] *Qualitas AG, Chamerstrasse 56, Ch-6300, Zug, Switzerland*

World Congress on Genetics Applied to Livestock Production (WCGALP) 2018
Accepted

Summary

Heterosis is the superiority of the crossbred offspring compared with the average of parental population due to favorable non-allelic gene interactions. Ignoring epistatic interactions, the extent of heterosis depends on degree of dominance and its direction, difference in allele frequencies of contributing loci. In this study we analyzed the possible association of heterosis with a sperm quality trait (percentage of live sperm) in bulls of admixed Swiss Fleckvieh, a composite of Simmental and Holstein Friesian, applying genome-wide mapping with genetic markers (SNP). Total of 41,749 phenotypic records of percentage of live sperm for both purebred and admixed bulls were used. After quality control of genotypes, 1169 bulls with 38,205 SNP remained for analyses. The model for single locus mapping consisted of genetic effect of bulls, fixed effects (age, contemporary groups, ejaculate intervals and semen collector), additive SNP effect, genomic breed percent and genomic breed heterosis. For percentage of live sperm 10 significant signals on chromosomes 3, 4, 5, 7, 13 and 14 were detected. Four of these regions contained genes related to spermatogenesis.

Keywords: local genetic ancestry, heterosis, genome-wide mapping, sperm quality, Swiss Fleckvieh

5-1 Introduction

Crossbreeding is a widely used mating system in livestock which is the result of interbreeding of purebred parental lines from at least two distinct breeds or lines. Systematic crossbreeding optimizes genetic merit of crossbred offspring by introducing favorable genes, decreasing inbreeding depression and benefits of gene interaction of heterosis. Heterosis or hybrid vigor is the complementary and opposite phenomenon of inbreeding depression, where the progeny of crossing inbred lines shows an increase of those characters suffering from inbreeding (Falconer & Mackay, 1996). The amount of heterosis depends on the degree of dominance and its direction in the contributed loci for the special trait, the difference of allele frequency between parental

populations for the special trait and epistatic interaction between loci (Falconer & Mackay, 1996).

Following the recent advances in genotyping technology and single nucleotide polymorphisms (SNP), the animal breeding research community has largely focused on using the genomic markers to study of population structure, genomic inbreeding and associated phenomena.

In this study, we aim to map heterosis effects with single locus models, using breed ancestry of the two alleles of a locus. Mixed ancestry is considered as dominant (single locus heterosis) gene action, while the additive breed effect on a locus as well as the additive of the actual SNP genotype are also included. The trait investigated is percentage of live sperm in pure Simmental and Red Holstein Friesian as well as the admixed Swiss Fleckvieh bulls from an AI station in Switzerland.

5-2 Materials and methods

5-2-1 Phenotypes and Genotypes

For 1298 Red Holstein Friesian (RHF), Simmental (SI) and admixed Swiss Fleckvieh bulls, 68,475 records on ejaculate volume (ml), concentration (10^9/ml) and percentage of live sperm from 2000 to 2015 were received from an artificial insemination (AI) station in Mülligen, Switzerland. For this study we used percentage of live sperm. The ejaculate records for percentage of live sperm which were in the range of the mean ± 3 standard deviation were kept for the analyses. After phenotypes filtering 41,749 records for 1296 bulls remained.

Both pure and admixed bulls were genotyped by Illumina® Bovine SNP 50k, 150k and 777k BeadChip. The imputed genotype data using *FImpute* software (Sargolzaei et al., 2014) to a subset of 44,999 SNP was received from Swissherdbook cooperative Zollikofen. The standard quality control was performed with PLINK (Chang et al., 2015; Purcell et al., 2007) to exclude monomorphic SNPs, those with call rate of < 0.95 and those deviated from Hardy-Weinberg equilibrium (*P*-value < 10^{-6}), and finally 38,205 SNPs for 1169 bulls remained for analyses.

5-2-2 Prediction of local ancestry

We used LAMP 2.5 (Sankararaman et al., 2008) in LAMPANC mode to infer locus-specific ancestry at each SNP position along the genome of admixed bulls. Configurations for LAMP were as in our previous study (Khayatzadeh et al., 2016). Ancestry origin of each SNP was estimated for each admixed bull with respect to two pure ancestral populations, representing the proportion of each ancestry (0, 0.5 and 1). 0 and 1 values indicate that two alleles at have the same origin and 0.5 indicates that each allele originate from different ancestral population at corresponding locus.

5-2-3 Genome-wide heterosis mapping model

A linear regression model for heterosis mapping similar to Ferencakovic et al. (2017) was used.

$$y_{ijklmnopqr} = \mu + \alpha_i + age_j + contemporarygroup_k + elapse_l + assistant_m + SNP_n + BreedPerc_o + BreedHet_q + \varepsilon_{ijklmnopqr}$$

This model was run for each SNP, using MIXED procedure and applying restricted maximum likelihood (REML) in SAS(*SAS/STAT user's guide*).

Where $y_{ijklmnopqr}$ denotes the phenotypic value, μ is the overall mean, $\alpha_i$is the random bull effect by $N(0, \boldsymbol{I}\sigma_a^2)$, where $\boldsymbol{I}$ is the identity matrix and σ_a^2 is the bull variance.

age_j, $contemporarygroup_k$, $elapse_l$ and $assistant_m$ are the fixed effects on bull age (< 16, 16-72 and > 72 months), year and season (Feb-May, June-Sep and Oct-Jan) of collection, ejaculate interval days between consecutive ejaculates (3-6, 7-9 and > 9 days) and semen collector, respectively.

SNP_n, $BreedPerc_o$ and $BreedHet_q$ were the regression coefficients for SNP additive genetic effect, locus-specific breed percent and breed heterosis according to LAMP results described above.

5-3 Results and discussion

Average percentage of live sperm was 86.28±3.38%. Breed difference and breed heterosis were estimated at global level by taking the average of breed ancestry and breed heterosis for all

incorporated loci for each individual. These breed percent estimates indicated +0.60% in favor of RHF compared to SI and a global heterosis effect of 1.40 %.

Single SNP analysis for mapping of genomic regions with effect on heterosis was performed as described above. Genome wide significance according to a threshold of $p = 10^{-6}$ was considered. A total of 10 regions on chromosomes 3, 4, 5, 7, 13 and 14 with positive effects of heterosis were identified (Table A5-1). Manhattan plot for breed heterosis on percentage of live sperm is shown in Figure A5-1.

The NCBI (https://www.ncbi.nlm.nih.gov/), Uniprot (http://www.uniprot.org) and GeneCards (http://www.genecards.org) were used for gene identification in the significant regions. Description of the genes with their physical positions on chromosomes and their function are given in the appendix.

PKDREJ is a protein coding gene on chromosome 5 which encodes a member of the polycystin protein family. This protein plays a role in human reproduction and fertilization by generating a calcium^{2+} transporting channel involved in initiating the acrosome reaction of the sperm (Fischer et al., 2015; Hamm et al., 2007). *THEG* gene is also a protein coding gene detected on chromosome 7 and is expressed in a nucleus of haploid male germ cells which involved possibly in spermatogenesis (Mannan et al., 2000). *ODF3L2* gene (Nuhrenberg et al., 2013) is also detected on chromosomes 7, which is paralog with *ODF3L1* and *SPAG4* is a protein coding gene on chromosome 13, involved in spermatogenesis and maintenance of the general polarity of the sperm head. It may also assist the organization and assembly of outer dense fibers (*ODFs*), a specific structure of the sperm tail (Fischer et al., 2015).

5-4 Conclusions

In this study we modeled single locus heterosis effects for percentage of live sperm along autosomes in admixed Swiss Fleckvieh bulls. Additive SNP effect, breed percent and breed heterosis were taken into account. Significant signals for heterosis were detected on chromosomes 3, 4, 5, 7, 13 and 14. Several of these regions hosted genes with function in spermatogenesis.

Acknowledgements

We would like to thank the Swissherdbook cooperative Zollikofen for providing genotypes and Swissgenetics for phenotypes for analyses.

List of References

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. doi: 10.1186/s13742-015-0047-8

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics* (4th ed. ed.). Harlow: Longman.

Ferencakovic, M., Solkner, J., Kaps, M., & Curik, I. (2017). Genome-wide mapping and estimation of inbreeding depression of semen quality traits in a cattle population. *Journal of Dairy Science, 100*(6), 4721-4730. doi: 10.3168/jds.2016-12164

Fischer, D., Laiho, A., Gyenesei, A., & Sironen, A. (2015). Identification of Reproduction-Related Gene Polymorphisms Using Whole Transcriptome Sequencing in the Large White Pig Population. *G3-Genes Genomes Genetics, 5*(7), 1351-1360. doi: 10.1534/g3.115.018382

Hamm, D., Mautz, B. S., Wolfner, M. F., Aquadro, C. F., & Swanson, W. J. (2007). Evidence of amino acid diversity-enhancing selection within humans and among primates at the candidate sperm-receptor gene PKDREJ. *American Journal of Human Genetics, 81*(1), 44-52. doi: 10.1086/518695

Khayatzadeh, N., Meszaros, G., Utsunomiya, Y. T., Garcia, J. F., Schnyder, U., Gredler, B., . . . Solkner, J. (2016). Locus-specific ancestry to detect recent response to selection in admixed Swiss Fleckvieh cattle. *Animal Genetics, 47*(6), 637-646. doi: 10.1111/age.12470

Mannan, A., Lucke, K., Dixkens, C., Neesen, J., Kamper, M., Engel, W., & Burfeind, P. (2000). Alternative splicing, chromosome assignment and subcellular localization of the

testicular haploid expressed gene (THEG). *Cytogenet Cell Genet, 91*(1-4), 171-179. doi: Doi 10.1159/000056840

Nuhrenberg, T. G., Langwieser, N., Binder, H., Kurz, T., Stratz, C., Kienzle, R. P., . . . Neumann, F. J. (2013). Transcriptome analysis in patients with progressive coronary artery disease: identification of differential gene expression in peripheral blood. [Comparative StudyResearch Support, Non-U.S. Gov't]. *J Cardiovasc Transl Res, 6*(1), 81-93. doi: 10.1007/s12265-012-9420-5

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses.]. *American Journal of Human Genetics, 81*(3), 559-575. doi: 10.1086/519795

Sankararaman, S., Sridhar, S., Kimmel, G., & Halperin, E. (2008). Estimating local ancestry in admixed populations. *American Journal of Human Genetics, 82*(2), 290-303. doi: 10.1016/i.ajhg.2007.09.022

Sargolzaei, M., Chesnais, J. P., & Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics, 15*, 478. doi: 10.1186/1471-2164-15-478

SAS/STAT user's guide. (6, Fourth edition. ed.): Cary N.C. : SAS Institute, 1990 (1994 [printing]).

Appendix

Table A5-1 Physical positions on detected signals for percent of live sperm and annotated genes names with their functions

Chromosome	Signals positions (Mb)	Genes names	Functions
3	10.95-15.99 119.00-119.29 120.77 121.37	-	-
4	77.10-88.46	-	-
5	113.43-117.86	*PKDREJ* (117.25-117.27)	Polycystin family receptor for egg jelly
7	43.05-47.38	*THEG* (44.63-44.66) *ODF3L2* (44.75-44.77)	Spermatid protein Outer dense fibre of sperm tail 3like 2
13	32.06 62.49-73.90	*SPAG4* (65.49-65.51)	Sperm associated antigene 4
14	27.03-29.99		

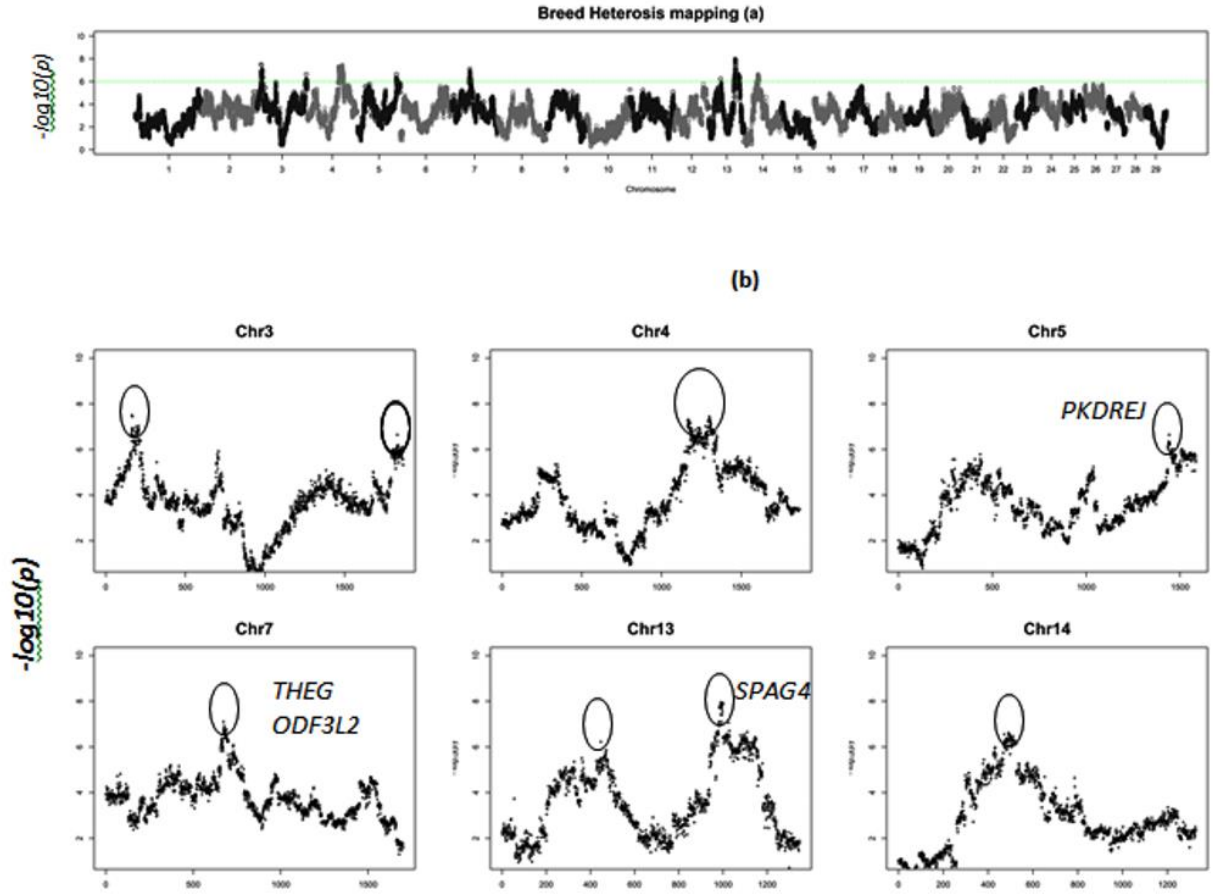


Figure A5-1. (a) Genome-wide heterosis mapping for percent of live sperm. (b)Significant signals together with gene name for chromosomes 3, 4, 5, 7, 13 and 14 for percent of live sperm

Chapter 6

General discussion and conclusions

6-1 Introduction

Identification of selection signature in admixed populations is one of the endeavors in evolutionary genetic studies. Genome of admixed animals in a recent admixed population is fragmented into ancestral segments with different proportions, due to limited number of recombination taking place each generation. Therefore, the ancestral contributions vary along the genome due to sampling error and genetic drift as the most prominent sources of random error and systematic biases caused by selection for or against some alleles.

In **Chapter 2**, we used LAMP program (Sankararaman *et al.* 2008) to estimate local genetic ancestries along autosome in admixed Swiss Fleckvieh cattle. We studied post- admixture selection signature using local genetic ancestries by three the results of three different approaches. We determined two genome-wide thresholds for signals of selection by 1) correction for multiple hypothesis testing and 2) permutation tests, where both provided very similar thresholds.

In **Chapter 3** we took advantage of the availability of various phasing algorithms and local ancestry estimation approaches. The same genotypes on two pure ancestral populations, RHF and SI, and admixed Swiss Fleckvieh bulls in **Chapter 2** were used for this study. Local ancestries in **Chapter 2** were estimated by LAMP. In this study we used two other approaches, implemented by LAMP-LD (Baran *et al.* 2012) and MULTIMIX (Churchhouse & Marchini 2013), to examine how inferences influence the local ancestry estimations. Reference haplotype panels and admixed bulls' haplotypes, as input files for LAMP-LD and MULTIMIX, were phased by *Shape-It* (Delaneau *et al.* 2012) and *AlphaPhase* (Hickey *et al.* 2011). Correlation of local ancestries, using different phasing algorithm were ~ > 0.91. Results on total 361 analyses, implemented by different parameter settings, showed inferences of local ancestries were susceptible to windows lengths and input data format (unphased or phased genotypes).

Crossbreeding is the mating of more distantly related animals, where sires of one breed or line mate with dams of another breed or line. The primary effect of crossbreeding is an increase in heterozygosity. The result of crossbreeding for polygenic traits is a gain in gene combination value, called hybrid vigor or heterosis, which is important to production with major effects on fertility and survivability. The gene combination value consists of favorable combinations of

dominance and unfavorable combinations caused by gradual breaking up of desirable epistatic blocks of linked loci in advanced generations of certain crosses. Different animal models were investigated for estimation of the non-additive genetic components of heterosis for semen traits in admixed Swiss Fleckvieh bulls in **Chapter 4**. The local genetic ancestry estimates for a bigger sample size (1179 bulls) in comparison with population studied in **Chapter 2** (485 bulls) were used to estimate local genetic ancestries. Local ancestry inferences were then used to define dominance and epistatic loss as the components of heterosis. Results of comparing model based on likelihood ratio test and Akaike Information Criteria (AICc) were presented in **Chapter 4**.

In **Chapter 5**, the result of genome-wide mapping of dominance component of heterosis (breed dominance) for percentage of live sperm (as one of the evaluated traits in **Chapter 4**) was presented. Genomic regions which contain genes associating with spermatogenesis and influence spermatogenesis were presented as well.

6-2 Local genetic ancestry to detect post-admixture signals

Availability of high density SNP markers provides the opportunity of estimation of ancestry proportions at genome-wide level. Moreover, various proposed statistical approaches exist for local ancestry estimates. Study of patterns and distribution of local ancestries at genome-wide level in admixed populations offers unique opportunities for the detection of selection signature. Extreme deviations of local genetic ancestries from genome-wide average ancestry (excess or deficiency) can be used to infer signals of post-admixture selection in crossbred populations (Tang *et al.* 2007) (**Chapter 2**).

$$\delta_k^m = \frac{1}{I}\sum_{i=1}^{I}\left(q_k^{i,m} - \bar{q}_k^i\right) = \tilde{q}_k^m - \bar{q}_k$$

where $q_k^{i,m}$ is the locus-specific ancestry of animal *i* at SNP *m*, estimated by LAMP, $\bar{q}_k^i$ is mean of locus-specific ancestries across the genome for individual *I*, $\tilde{q}_k^m$ is the mean of ancestry at SNP *m* averaged over all admixed animals; and $\bar{q}_k$ is the mean of locus-specific ancestry across the entire whole genome for admixed population *k*. This method of inferring selection signals has

been performed for Creole cattle (Gautier & Naves 2011a; Flori *et al.* 2014a), African cattle (Kim & Rothschild 2014; Bahbahani *et al.* 2015) and chickens (Qanbari *et al.* 2012) as well as human (Tang *et al.* 2007; Jin *et al.* 2012).

Using the most suitable significance threshold in genome-wide studies is a major challenge. It should account for multiple comparisons based on the number of the markers and provide a good balance of false positive and false negative results. A variety of statistical approaches accounting for multiple testing, such as Bonferroni correction and false discovery rate (Panagiotou *et al.* 2012).

Permutation tests (Doerge & Churchill 1996) in a specific way proposed by Tang *et al.* (2007), have also been used to define any significant deviations of local ancestries. According to the results of a simulations study (Tang *et al.* 2007), this type of permutation method maintains the original structure of dataset and is robust to demographic process of genetic drift. Finding very similar thresholds (Figure 2-4, **Chapter 2**) for detection of significant signals of selection in our study, using both permutation tests and extreme deviations of scaled local ancestries from normal distribution, correcting for multiple hypothesis tests, provided confidence in the thresholds chosen.

Results of 20,000 permutation tests, 5% genome-wide threshold and extreme deviations from normal distribution by correction for 1000 hypotheses (Bhatia *et al.* 2014), on the basis of higher admixture LD in admixed populations, identical regions on chromosomes 13 (46.3-47.3 Mb by permutation tests, 5% genome-wide threshold and 46.3-46.8 Mb by deviations from normal distribution, 1000 hypotheses) and on chromosome 18 (18.7-25.9 Mb by both genome-wide thresholds) detected as candidates of selection signature.

Width of detected signals in the recent crossbred population indicated that 1) selection did not have enough number of generations of crossbreeding to sharpen the signals; 2) finding similar pre- and post-admixture selection signature is not promising (F_{st} and Δ ancestries), and 3) vague candidate genes can be associated with selection signature in recent admixed Swiss Fleckvieh cattle, associated with fertility and reproduction traits, milk composition and morphology traits.

6-3 Different algorithms for the inference of local ancestry; opportunities and constraints

Recent advance in genotyping technology together with development of statistical methods facilitated the assessment of fine-scale ancestry along the genome of admixed individuals. In **Chapter 3** we took advantage of the availability of different phasing algorithms and local ancestry inference software tools to investigate how much the choice of different methods as well as parameter settings of the applied software tools can affect the estimations. In **Chapter 2**, we used LAMP for local ancestry estimations. In **Chapter 3** we applied two other different software tools (LAMP-LD and MULTIMIX) to estimate local ancestries. LAMP-LD and MULTIMIX model genetic ancestries by window length and genetic map which should be defined by users. Although all require some form of reference panel, LAMP uses clustering algorithm for local ancestry deconvolution, while the other two programs use an approximation to coalescent with recombination for inference of local ancestries (Li & Stephens 2003).

The key limitation of such programs is that they assume the admixture tract lengths can be modeled as independently and identically distributed and are considered as (iid) exponential random variables. Nevertheless, this assumption does not hold in recently admixed populations, leading false positives. Admixture tracts are stochastically and not evenly distributed along the genome of admixed individuals (Pool & Nielsen 2009; Liang & Nielsen 2014). Limitations of the existing models are: 1) need of large inference panels; 2) need to explicitly model linkage disequilibrium, and 3) computational time demand. Choosing small window lengths produced noise with LAMP-LD and MULTIMIX, since we expected larger windows in recent admixed population. The most comparable results were with window lengths 15-23 SNP, between LAMP-LD and MULTIMIX_MCMCgeno and between MULTIMIX_MCMCgeno and MULTIMIX_MCMC (**Chapter 3**, Table 3-3).

On the other hand, LAMP works based on breaking the genome into windows and decides for optimal window size internally based on information on the number of generations after admixture and admixture intensity and does not require assumption on parametric population genetic model. Most of studies of local ancestry patterns for selection signature use LAMP, since

it is faster, decides internally for optimum window size, does not need to phased data reference genotype panels (Qanbari *et al.* 2012; Flori *et al.* 2014b; Bahbahani *et al.* 2015).

Phasing haplotypes using *Shape-It* by different window length does not considerably change the phasing results, since the important parameters are genetic map or effective population size. *AlphaPhase* as another phasing algorithm is robust to small variation in terms of core- and tail-length. Therefore, choosing for different algorithms of phasing haplotypes and different setting did not change drastically the local ancestries estimations. The highest correlations between *shape-It* and *AlphaPhase* results were observed, when total core- and tail-length were in the range of 300 SNP for *AlphaPhase*.

In **Chapter 2** significant signals were detected for local genetic ancestries estimated with LAMP. These signals were not confirmed by local ancestry deviation with LAMP-LD and MULTIMIX and no other significant signals were found with those two approaches. The results suggest that care should be taken when interpreting selection signature based on local ancestry detected by a single method and confirmation with alternative approaches is necessary and advisable.

6-4 Heterosis components for semen traits in a composite cattle breed

Crossbreeding is an important tool to increase the efficiency of livestock productions through heterosis and complementarity of breeds. Additive (breed proportion) and non-additive genetic (breed heterosis) coefficients need to be considered in a model statistically analyzing the effects of crossbreeding. Non-additive genetic coefficients consist of intra locus interaction (dominance) and non-allelic gene interaction (epistasis). If heterosis is primarily due to dominance with no epistasis, then it is proportional to heterozygosity (proportion of heterozygotes at individual loci). Confounding of the genetic effects (Figure 4-1, **Chapter 4**) complicates the estimation of dominance effects separately from epistatic effects, such that most of the animal models for multiple breed evaluations are only based on dominance effects (Kinghorn 1983). Accurate prediction of crossbred cattle performance is key-factor for running successful breeding programs which are very important to produce protein rich food. The accurate predictions require

to estimate the breeding values of crossbred animals considering all possible genetic effects (non-additive as well as additive genetic effects (Cardoso & Tempelman 2004).

For our study in **Chapter 4**, we defined heterosis components based on genomic data, where according to our knowledge no previous study used local ancestries to define these effects. In order to study the contribution of non-additive genetic effects of heterosis in admixed Swiss Fleckvieh bulls using genomic data, we applied different mixed models by including dominance and epistatic loss step-wise. Rather than genotypes at SNP level, we used locus-specific ancestries estimated by LAMP to define the dominance and epistatic loss.

Epistasis effect is important for understanding the genetic architecture of different traits. Recent studies applying genetic models (Maki-Tanila & Hill 2014) indicate only small contributions of epistatic effect to the genetic variance in outbred populations. Epistatic loss was significant for two important sperm quality traits, number of spermatozoa and percentage of live sperm (Table 4-4, **Chapter 4**). The current study estimated regression coefficients for epistatic effects, not genetic variances, and it is not clear how these results translate into non-additive variance.

6-5 genome-wide mapping of heterosis

In previous chapter (**Chapter 4**), we estimated the global effects of heterosis components. In **Chapter 5**, we focused on monitoring local heterosis across the genome, to use benefits of genomic data to find regions along genome associated with heterosis for one of the traits evaluated in **Chapter 4** (percentage of live sperm). A linear regression model for heterosis mapping, similar to the one applied by Ferencakovic *et al.* (2017) for mapping of inbreeding depression, was used.

$$y_{ijklmnopqr} = \mu + \alpha_i + age_j + contemporarygroup_k + elapse_l + assistant_m + SNP_n + BreedPerc_o + BreedHet_q + \varepsilon_{ijklmnopqr}$$

Where $y_{ijklmnopqr}$ denotes the phenotypic value, μ is the overall mean, $\alpha_i$is the random bull effect by $N(0, \boldsymbol{I}\sigma_a^2)$, where $\boldsymbol{I}$ is the identity matrix and σ_a^2 is the bull variance.

age_j, $contemporarygroup_k$, $elapse_l$ and $assistant_m$ are the fixed effects on bull age, year and season of collection, ejaculate interval days between consecutive ejaculates and semen collector, respectively. SNP_n, $BreedPerc_o$ and $BreedHet_q$ are the regression coefficients for SNP additive genetic effect, locus-specific breed percent and breed dominance as a component of heterosis. Modeling the dominance component of heterosis for percentage of live sperm gave several significant signals on chromosomes 3, 4, 5, 7, 13 and 14. Genes, responsible for sperm fertilization, spermatogenesis and structure of sperm tail, were identified on the detected regions. We did not include genomic relationship matrix and any correction for population stratification and polygenic effects as well, because it was assumed that corrections for population stratification can smooth heterosis effects.

Since considering epistatic effects between all SNP genome-wide demands a lot of computation, we did not consider this effect for heterosis mapping. Heterosis mapping has been performed in plant (sorghum, cotton and rice) and epistatic effects using two or three loci theory as well as dominance components of heterosis were defined (Ben-Israel *et al.* 2012; Guo *et al.* 2013). Because we found significant levels for epistatic loss, according to our new definition in **Chapter** 4, future work will include such epistasis mapping for the semen data included in this thesis.

6-6 Conclusions and recommendations

1) One of the key findings of this thesis was the detection of significant signals of selection post-admixture, based on local differences in average ancestry of a population of admixed individuals (**Chapter 2**). In a subsequent step (**Chapter 3**), we tested how sensitive these results are to choice of tools for phasing and estimation of local admixture. Surprisingly, differences were big and the significant signals of the previous study were not confirmed with alternative tools and settings. While simulation may bring light into this conundrum, the advice derived from the results of this thesis is to study local genetic ancestry with at least two different software tools. Also, new approaches, particularly machine learning, should be tested and compared.

2) This thesis has used high density genetic marker data to estimate components of heterosis that were previously derived from expected values of heterozygosity and levels of ancestry (**Chapter**

4). Effects of the dominance component of heterosis were positive, as expected and epistatic effects -when significant- were negative, also as expected. This approach should be used in other studies of crossbred populations because it provides valuable information about the crossbreeding system implemented. The models may be expanded to include not only additive × additive, but also the other combinations of epistatic effects (additive × dominance and dominance × dominance) to fully accommodate the definition of epistatic loss.

3) Search for regions of the genome responsible for heterosis is a very exciting endeavor that has only been explored for crops, including maize and rice, so far (Ben-Israel *et al.* 2012; Guo *et al.* 2013). In this thesis, one sperm quality trait was investigated, considering only the dominance component of heterosis. It is strongly suggested to include epistatic effects with two or three locus theory. This may be done for many crossbred populations with animals that are routinely SNP-Chip genotyped.

References

Bahbahani H., Clifford H., Wragg D., Mbole-Kariuki M.N., Van Tassell C., Sonstegard T., Woolhouse M. & Hanotte O. (2015) Signatures of positive selection in East African Shorthorn Zebu: A genome-wide single nucleotide polymorphism analysis. *Sci Rep* **5**, 11729.

Baran Y., Pasaniuc B., Sankararaman S., Torgerson D.G., Gignoux C., Eng C., Rodriguez-Cintron W., Chapela R., Ford J.G., Avila P.C., Rodriguez-Santana J., Burchard E.G. & Halperin E. (2012) Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **28**, 1359-67.

Ben-Israel I., Kilian B., Nida H. & Fridman E. (2012) Heterotic trait locus (HTL) mapping identifies intra-locus interactions that underlie reproductive hybrid vigor in Sorghum bicolor. *PLoS One* **7**, e38993.

Bhatia G., Tandon A., Patterson N., Aldrich M.C., Ambrosone C.B., Amos C., Bandera E.V., Berndt S.I., Bernstein L., Blot W.J., Bock C.H., Caporaso N., Casey G., Deming S.L., Diver W.R., Gapstur S.M., Gillanders E.M., Harris C.C., Henderson B.E., Ingles S.A., Isaacs W., De Jager P.L., John E.M., Kittles R.A., Larkin E., McNeill L.H., Millikan

R.C., Murphy A., Neslund-Dudas C., Nyante S., Press M.F., Rodriguez-Gil J.L., Rybicki B.A., Schwartz A.G., Signorello L.B., Spitz M., Strom S.S., Tucker M.A., Wiencke J.K., Witte J.S., Wu X., Yamamura Y., Zanetti K.A., Zheng W., Ziegler R.G., Chanock S.J., Haiman C.A., Reich D. & Price A.L. (2014) Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. *American Journal of Human Genetics* **95**, 437-44.

Cardoso F. & Tempelman R. (2004) *Hierarchical Bayes multiple-breed inference with an application to genetic evaluation of a Nelore-Hereford population*.

Churchhouse C. & Marchini J. (2013) Multiway admixture deconvolution using phased or unphased ancestral panels. *Genet Epidemiol* **37**, 1-12.

Delaneau O., Marchini J. & Zagury J.F. (2012) A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179-81.

Doerge R.W. & Churchill G.A. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285-94.

Ferencakovic M., Solkner J., Kaps M. & Curik I. (2017) Genome-wide mapping and estimation of inbreeding depression of semen quality traits in a cattle population. *Journal of Dairy Science* **100**, 4721-30.

Flori L., Thevenon S., Dayo G.K., Senou M., Sylla S., Berthier D., Moazami-Goudarzi K. & Gautier M. (2014a) Adaptive admixture in the West African bovine hybrid zone: insight from the Borgou population. *Molecular Ecology* **23**, 3241-57.

Flori L., Thevenon S., Dayo G.K., Senou M., Sylla S., Berthier D., Moazami-Goudarzi K. & Gautier M. (2014b) Adaptive admixture in the West African bovine hybrid zone: insight from the Borgou population. *Mol Ecol* **23**, 3241-57.

Gautier M. & Naves M. (2011a) Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Molecular Ecology* **20**, 3128-43.

Gautier M. & Naves M. (2011b) Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Mol Ecol* **20**, 3128-43.

Guo X., Guo Y., Ma J., Wang F., Sun M., Gui L., Zhou J., Song X., Sun X. & Zhang T. (2013) Mapping heterotic loci for yield and agronomic traits using chromosome segment introgression lines in cotton. *J Integr Plant Biol* **55**, 759-74.

Hickey J.M., Kinghorn B.P., Tier B., Wilson J.F., Dunstan N. & van der Werf J.H. (2011) A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet Sel Evol* **43**, 12.

Jin W.F., Xu S.H., Wang H.F., Yu Y.G., Shen Y.P., Wu B.L. & Jin L. (2012) Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Research* **22**, 519-27.

Kim E.S. & Rothschild M.F. (2014) Genomic adaptation of admixed dairy cattle in East Africa. *Front Genet* **5**, 443.

Kinghorn B. (1983) Genetic-Effects in Crossbreeding .3. Epistatic Loss in Crossbred Mice. *Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie-Journal of Animal Breeding and Genetics* **100**, 209-22.

Li N. & Stephens M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213-33.

Liang M. & Nielsen R. (2014) The lengths of admixture tracts. *Genetics* **197**, 953-67.

Maki-Tanila A. & Hill W.G. (2014) Influence of gene interaction on complex trait variation with multilocus models. *Genetics* **198**, 355-67.

Panagiotou O.A., Ioannidis J.P. & Genome-Wide Significance P. (2012) What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int J Epidemiol* **41**, 273-86.

Pool J.E. & Nielsen R. (2009) Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711-9.

Qanbari S., Strom T.M., Haberer G., Weigend S., Gheyas A.A., Turner F., Burt D.W., Preisinger R., Gianola D. & Simianer H. (2012) A high resolution genome-wide scan for significant selective sweeps: an application to pooled sequence data in laying chickens. *PLoS One* **7**, e49525.

Sankararaman S., Sridhar S., Kimmel G. & Halperin E. (2008) Estimating local ancestry in admixed populations. *American Journal of Human Genetics* **82**, 290-303.

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**, 626-33.

Zusammenfassung

Die Identifikation des Abstammungsursprungs von chromosomalen Segmenten in Kreuzungs-Populationen, lokale genetische Abstammung genannt, wurde bislang in Assoziationsstudien für Erbkrankheiten und quantitative Merkmale sowie zur Aufklärung von Populationsstrukturen untersucht. Eine genom-weite Perspektive von Kreuzungspopulationen zeigt auf, wie sich die Abstammungsbeiträge entlang des Genoms ändern. Die wichtigsten Quellen der Variation im Genom von gekreuzten Populationen sind Selektion und evolutionäre Fluktuationen durch genetische Drift. Anders als genetische Drift betrifft Selektion spezifische Regionen des Genoms. Die extremsten Abweichungen der lokalen genetischen Abstammungen von der durchschnittlichen genom-weiten Abstammung können als Selektion-Signaturen entdeckt werden. Kürzlich entwickelte Kreuzungspopulationen bieten eine ausgezeichnete Gelegenheit, um solche Selektions-Signaturen post Admixtur zu studieren. Das Ziel dieser Dissertation war, Abstammungsbeiträge auf lokalem Niveau zu studieren, um Selektions-Signaturen zu entdecken und um Effekte der Heterosis einzuschätzen. Swiss Fleckvieh (SWF), eine Schweizer Rinderrasse, die durch Kreuzung der lokalen Rasse Simmental (SI) mit Red Holstein Friesian (RHF) seit 1970 entstand, wurde dazu untersucht.

Zunächst wurde die Abstammung auf globalem und lokalem Niveau von 485 Stieren aller drei Rassen, davon 300 SWF, durch Analyse von Illumina ® BovineSNP50k Genotypen ermittelt. Die globalen RHF und SI Abstammungen der SWF Tiere wurden 0.70:0.30 geschätzt. Die Schätzungen der lokalen genetischen Abstammungen wurden verwendet, um Selektions-Signaturen zu entdecken. Zwei Methoden basierend auf Permutationen und Bonferroni-Korrektur für extreme Abweichungen von Normalverteilung wurden eingesetzt um die Schwelle der Signifikanz von starken Abweichungen der lokalen von globalen Herkunftsniveaus zu ermitteln. Beide Methoden führten zu ähnlichen Schwellen. Zwei bemerkenswerte Spitzen, eine auf Chromosom 13 (46.3-47.3 Mb) und ein anderes Gebiet auf Chromosom 18 (18.7-25.9) wurden als Selektion-Signaturen identifiziert. Erweiterte Haplotypen-Homozygotie (*EHH*), die Signale des Prä-und Post-Admixtur zu untersucht, offenbarte ein Signal auf Chromosom 18 (25.5-26.4 Mb) durch *iHS* Statistik für RHF und ein breites Gebiet auf Chromosom 18 (6.6-24.6) durch *Rsb* Statistik zwischen SWF und SI. Die breiten Post-Admixtur Selektion-Signaturen zeigten, dass 1) die beschränkte Zahl von Generationen nachdem Kreuzung (~ 10-15) nicht

genug war, um Signale zu schärfen; 2) Vergleich der Prä-und Post-Admixtur Signalen nicht vielversprechend war, und 3) vage Kandidaten von kausalen Genen gefunden wurden.

In einer nächsten Untersuchung wurden die lokalen Abstammungen zusätzlich zur in der vorigen Studie angewandten Methode LAMP auch durch zwei andere Software-Tools (LAMP-LD und MULTIMIX), welche die Annahme eines parametrischen genetischen Modells benötigen, ermittelt. Verschiedene Parameter-Einstellungen zum Phasing der Daten und Fensterlängen wurden definiert. Relativ niedrigen Korrelationen zwischen den Resultaten von verschiedenen Software-Tools zeigten, dass die Wahl der Methode zur Ermittlung lokalen Abstammungen einen sehr starken Einfluss auf die Ergebnisse hat und folglich Selektions-Signale nach Untersuchung mit einer Methode vorsichtig betrachtet werden müssen. Bestätigung mit alternativen Annäherungen wird empfohlen.

Schließlich wurden die lokalen Abstammungsschätzungen für eine große Zahl von 1179 Stieren verwendet, um die Effekte von Dominanz und epistatischem Verlust (zwei Definitionen) als Bestandteile von Heterosis für Sperma-Eigenschaften bei gekreuzten Stieren von Swiss Fleckvieh zu schätzen. Der Dominanz-Bestandteil von Heterosis war sehr wichtig und verbesserte die Genauigkeiten des Modells von drei von vier untersuchten Sperma-Eigenschaften. Effekte der Gen-Interaktion (Epistasie) wurden für zwei der vier untersuchten Sperma-Eigenschaften gefunden, wenn diese mit einer hier erstmals verwendeten Definition ermittelt wurde. Im letzten Teil der Arbeit wurde genom-weites Mapping des von Dominanz verursachten Anteils von Heterosis für Prozentsatz der lebenden Spermien im Ejakulat durchgeführt Das statistische Modell enthielt neben Umwelteffekten und dem globalen Effekt des Stieres für jeden SNP separat den additiven Effekt, den Rasse-Anteil und den Effekt der Dominanz aufgrund von gemischter Rasse-Herkunft. Signifikante Signale wurden auf Chromosomen 5, 7 und 13 ermittelt und es wurden in diesen Regionen Gene mit Bezug zu Spermiogenese gefunden.

Stichwörter: Admixtur, Kreuzung, Heterosis, Dominanz, Epistasie, genetische Abstammung, genomweites Mapping, Haplotyp, Permutation, Selektions-Signatur

Personal list of publications

Scientific articles in (peer-) reviewed journals

2015

Khayatzadeh N., Mészáros G., Gredler B., Schnyder U., Curik I. and Sölkner J. (2015). Prediction of global and local Simmental and Red Holstein Friesian admixture levels in Swiss Fleckvieh cattle. Poljoprivreda 21 63-67.

2016

Khayatzadeh N., Mészáros G., Utsunomiya Y. T., Garcia J. F., Schnyder U., Gredler B., Curik I. and Sölkner J. (2016). Locus-specific ancestry to detect recent response to selection in admixed Swiss Fleckvieh cattle. Animal Genetics 47 637-646.

2016

Khayatzadeh N., Mészáros G., Gredler B., Schnyder U., Curik I. and Sölkner J. (2016). Estimation of local genetic ancestry in an admixed cattle population applying different methods. ACTA agriculturae Slovenica 5 31-36.

2016

Ferenčaković M., Banadinović M., Mercvajler M., **Khayatzadeh N.,** Mészáros G., Cubric-Curik V., Curik I. and Sölkner J. Mapping of heterozygosity rich regions in Austrian Pinzgauer cattle. ACTA agriculturae Slovenica 5 41-44.

2017

Khayatzadeh N., Mészáros G., Utsunomiya Y. T., Schmitz-Hsu F., Gredler B., Schnyder U., Ferenčaković M., Curik I. and Sölkner J. (2017). Estimation of breed composition, breed heterosis and epistatic loss for percent of live spermatozoa in admixed Swiss Fleckvieh bulls. ACS Agriculturae Conspectus Scientificus.

International congress presentations with published in proceedings

2006

Khayatzadeh N., Nejati-Javaremi A., VaezTorshizi R. and Parvandi M. (2006). Effect of model adequacy on reliability of estimation of genetic parameters and breeding values of early growth traits in sheep. In Proceeding of the 8th World Congress on Genetic Applied to livestock Production 13-18 August 2006 Belo Horizonte, MG, Brazil.

International congress presentations with published abstracts

2016

Khayatzadeh N., Mészáros G., Utsunomiya Y. T., Garcia J. F., Schnyder U., Gredler B., Curik I. and Sölkner J. (2016). Local vs global ancestry: regions deviating from genome-wide admixture in a composite cattle breed. EAAP 67th Annual Meeting of the European Federation of Animal Science, 29 August – 2 September 2016 Belfast, Northern Ireland, United Kingdom.

Sölkner J., Milanesi M., **Khayatzadeh N.,** Utsunomiya A.T., Ferenčaković M., Curik I., Ajmone Marsan P., Garcia J. F. and Utsunomiya U.T. Predicting autozygosity via runs of homozygosity from NGS versus SNP chip data for Nellore bulls. EAAP 67th Annual Meeting of the European Federation of Animal Science, 29 August – 2 September 2016 Belfast, Northern Ireland, United Kingdom.