

Review

Ancestry informative markers: A foundation for unravelling genetic ancestry and population structures

Akanksha Chaudhary^{1,2*}, Nidhishree N.S², Rakesh Kumar Pundir² and Amod Kumar²

¹Kurukshetra University, Kurukshetra - 136119, Haryana, India

²ICAR - National Bureau of Animal Genetic Resources, Karnal - 132001, Haryana, India

ABSTRACT

Deciphering the genetic ancestry in livestock species has important applications in population stratification and is used to explore the genetic basis for differences among the breeds within the populations. Genetic ancestry plays a crucial role in identifying population structure by determining the number of subpopulations within a population and assigning individuals to their respective groups. It is also instrumental in defining the number of ancestral populations in admixed populations and estimating the proportions of these ancestral populations in admixed individuals. The study of Ancestry Informative Markers (AIMs) has significantly advanced our understanding of genetic variation within populations, with applications ranging from anthropology to livestock management. AIMs are specific genetic markers, such as SNPs, that show significant allele frequency differences between populations, allowing researchers to trace lineage and analyse population structure. With the advent of next-generation sequencing technologies and SNP genotyping, AIMs have become invaluable for uncovering the biogeographical origins of species, aiding in conservation efforts, and improving livestock breeding strategies. In this review, we have briefly explained an overview of about the AIMs, methods of estimation and their importance in livestock management.

Key words: AIMs, Ancestry Informative Markers, Livestock species, SNPs

*Corresponding author: akankashachoudhary0@gmail.com

INTRODUCTION

The study of genetic variation within and across populations has been a basis of evolutionary biology and anthropology for many years. With the offset of high-throughput sequencing and Next-Generation Sequencing (NGS) technologies our ability to investigate these variations with unprecedented detail and accuracy has also advanced (Satam *et al.*, 2023). The development of genotyping methods and NGS has made it easier to study genetic variation and structure, which is critical to understanding our evolutionary history. The NGS analysis of biological components can lead to ancestry and phenotypic insights, encompassing results such as AIMs (Ancestry Informative Markers) and SNPs (Single Nucleotide Polymorphisms) as depicted in workflow given below (Fig. 1). AIMs emerged as a concept in early 2000s with advancements in genetic research and the growing availability of genome-wide data. They are a set of informative SNPs with significant differences in allele frequency between ancestral populations and have become essential tools for determining genomic ancestry (Santangelo *et al.*, 2017; Kehdy *et al.*, 2015; Vongpaisarnsin *et al.*, 2015). They provide insights into the biogeographic origins of individuals and populations by examining the frequency of specific alleles rather than their complete presence or absence (Das *et al.*, 2018). This approach is not

only crucial for anthropological research but has also found applications in other fields, including forensic science, personalized medicine, and, increasingly, livestock management. Initially, researchers sought to understand human population history and structure by identifying genetic variations that were significantly different between distinct populations. AIMs were identified as key markers that could distinguish these differences.

Also, population genetic structure analyses have demonstrated that continental population groups can be distinguished by examining allele frequency differences (Rosenberg *et al.*, 2002, 2005). In recent years, research has revealed that thousands of single nucleotide polymorphisms (SNPs) across the genome exhibit significant allele frequency disparities between two or more continental populations (Mao *et al.*, 2007; Price *et al.*, 2007; Smith *et al.*, 2004). These findings have laid the groundwork for admixture mapping and accounting for population genetic structure in association studies. The latter is crucial, as differences in population genetic structure between case and control groups can confound SNP-disease associations, potentially leading to false-positive or false-negative results (Campbell *et al.*, 2005; Clayton *et al.*, 2005; Freedman *et al.*, 2004). Methods to quantify and address population structure differences in association

testing have been developed (Epstein *et al.*, 2007; Hoggart *et al.*, 2003) and are particularly applicable in whole-genome association (WGA) scans. However, for subsequent association studies aiming to refine

critical candidate regions in larger population sets or to analyze additional populations, a compact set of AIMs is highly valuable.

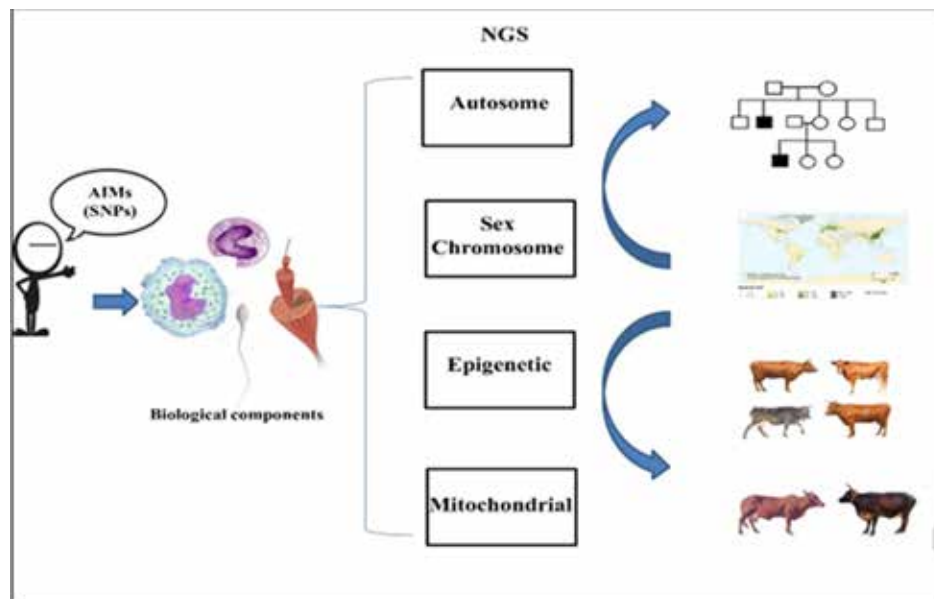


Fig. 1: Information generated from biological samples using NGS methods that include AIMs and SNPs.

The development of AIMs was driven by the need to trace lineage and understand the genetic diversity within and between populations. Early studies focused on human genetics, using AIMs to infer ancestral origins and migration patterns. These markers soon proved valuable in other fields, including livestock genetics, where they provided insights into the history and genetic composition of different breeds.

The significance of Ancestry Informative Markers

AIMs help in analysing the genetic structure of populations by revealing how different groups are related and how they have evolved (Das *et al.*, 2019). The importance of AIMs in livestock cannot be overstated. Livestock breeds have been developed over centuries through selective breeding, often influenced by geographical, cultural, and economic factors. Understanding the ancestry of these breeds helps in preserving genetic diversity, which is crucial for the long-term sustainability of livestock populations. Moreover, AIMs can assist in identifying the genetic basis of traits associated with productivity, evolutionary genetics, biomedical research, and forensic analyses (Mekhfi *et al.*, 2024).

One of the primary applications of AIMs in livestock is in the assessment of genetic diversity. Understanding the genetic variation within and among breeds allows for better management of breeding programs. For example, in sheep breeding, AIMs have been used to identify genetic markers associated with wool quality

and reproductive traits (Somenzi *et al.*, 2020; Getachew *et al.*, 2017). In goats, researchers have utilized AIMs to assess genetic diversity among indigenous breeds, contributing to conservation efforts and the development of sustainable breeding programs (Monau *et al.*, 2022).

In addition to conservation, AIMs play a crucial role in breeding programs. By identifying specific genetic markers associated with desirable traits, such as disease resistance, growth rate, and milk production, breeders can implement marker-assisted selection (MAS) (Rezende *et al.*, 2012). This approach allows for more efficient selection of animals that carry advantageous alleles, ultimately improving the overall performance of livestock populations. For instance, studies have shown that AIMs can be used to select for traits like heat tolerance in cattle (Macciotta *et al.*, 2017), which is increasingly important in the face of climate change.

Admixture is a common form of gene flow between populations. It refers to the process in which two or more genetically and phenotypically diverse populations with different allele frequencies mate and form a new, mixed or 'hybrid' population (Chakraborty, 1986). Modeling studies showed that in contrast to the million markers suggested to be necessary for genome-wide association studies (GWAS) (Hirschhorn *et al.*, 2005), 2000 and 5000 well-distributed ancestry informative markers (AIMs) distinguishing parental origins are sufficient for whole

genome scanning under the admixture mapping strategy (Tian *et al.*, 2007). Hence, it is important to identify and choose most ancestry informative markers across populations, the power of admixture mapping relies heavily on the ability of informative markers to infer ancestry along the chromosomes of admixed individuals (Shriver *et al.*, 2003).

Estimation of genetic ancestry: Global and local

Global ancestry (GA) represents the proportion of genomic ancestry in each admixed individual that can

be attributed to the ancestral populations contributing to the recently admixed population (Fig. 2A). Various approaches can be used to estimate GA. Among the most widely used methods, probabilistic models that utilize genotype data is most preferred. These methods often assume Hardy–Weinberg equilibrium within populations and complete linkage equilibrium across all loci included in the estimation, such as STRUCTURE (Falush *et al.*, 2007) and ADMIXTURE (Alexander *et al.*, 2009).

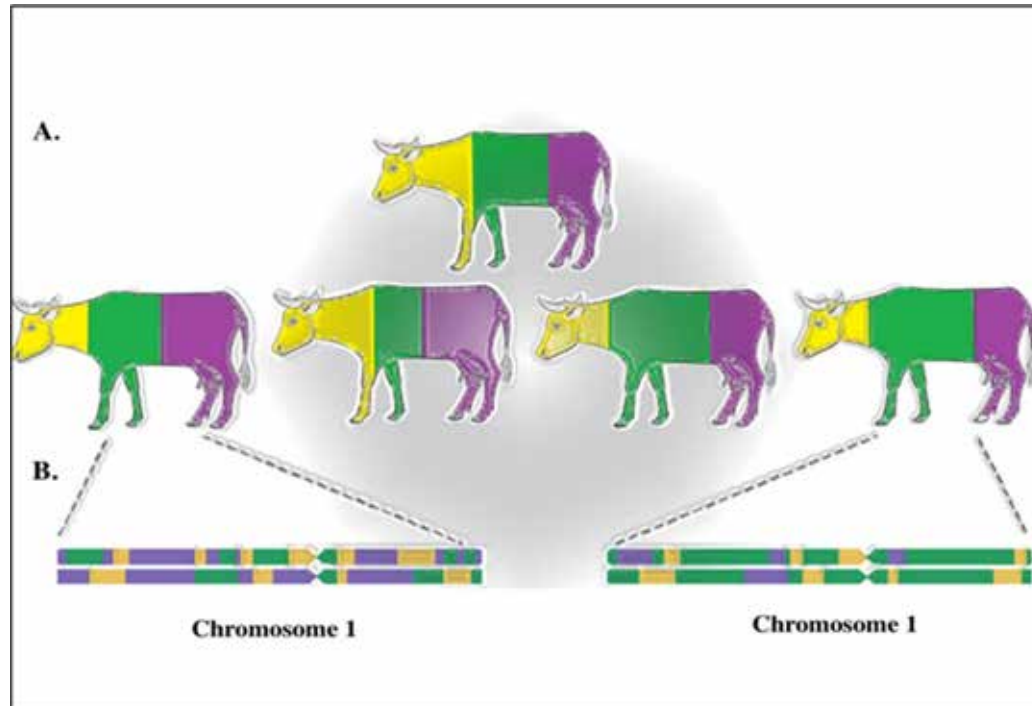


Fig. 2: Global (A) and local (B) genetic ancestries in a recently admixed population with three ancestral populations. The proportion of each of the ancestral populations is represented by the colours yellow, green, and purple.

Local ancestry (LA) refers to the ancestral origins of specific chromosome segments, known as ancestral tracts, within recently admixed individuals (Fig. 2B). For this, the number of copies derived of each ancestral population, in each genomic position, could be inferred per individual (from zero to two copies). Thus, GA can also be obtained by summarizing LA across the individual genomes. Briefly, the choice of the most suitable approach depends on the number and density of available markers, as well as the evolutionary history of the admixed population. Some models use haplotype data and require specific reference panels, which may not be available for all populations (Dias-Alves *et al.*, 2018; Maples *et al.*, 2013). Additionally, local ancestry inference becomes challenging in cases of admixture between populations with limited genetic divergence or in ancient admixture events (Winkler *et al.*, 2010). Some of the genetic characteristics of admixed population allow the estimation of ancestry with a relatively

small number of genetic markers. These markers are Ancestry Informative Markers. Their number will also depend on the assessed populations, the ancestral groups, and the time since the admixture event. To identify those markers that are useful and informative of ancestry, multiple measurements of population differentiation have been proposed (Rosenberg *et al.*, 2003; Ding *et al.*, 2011).

Measures of Marker Informativeness for Ancestry

The various methods for estimating marker informativeness are summarized below:

1. **Fisher Information Content (FIC)** measures the amount of information a genetic marker provides about an individual's ancestry. It is particularly effective in distinguishing between different ancestral populations, making it valuable for enhancing the precision of ancestry estimations in genetic studies.

2. **Shannon Information Content (SIC)** derived from Shannon entropy, quantifies the uncertainty involved in predicting an individual's ancestry based on a specific genetic marker. Higher SIC values indicate that the marker is more informative, which is crucial for assessing the diversity of alleles in a population and understanding the degree of heterogeneity.
3. **F-Statistics (F_{ST})** is a metric that measures population differentiation due to genetic structure. It compares the genetic variability within sub populations relative to the total population. F_{ST} is widely used to quantify genetic differences between populations, helping to identify markers that are highly differentiated and thus particularly informative for ancestry analysis.
4. **Informativeness for Assignment Measure (In)** assesses how well genetic markers can assign

individuals to specific populations. It evaluates the probability that a given marker can correctly classify the ancestry of an individual, making it particularly useful in studies aiming for accurate population assignment.

5. **Absolute Allele Frequency Differences (δ)** calculates the absolute difference in allele frequencies between two populations. A high δ value indicates that the allele is differentially distributed between populations, making it informative for ancestry inference. This measure is commonly used to identify markers that distinguish between populations, aiding in understanding population structure and migration patterns (Ding *et al.*, 2011).

The process for identifying ancestry informative markers is illustrated in Fig. 3.

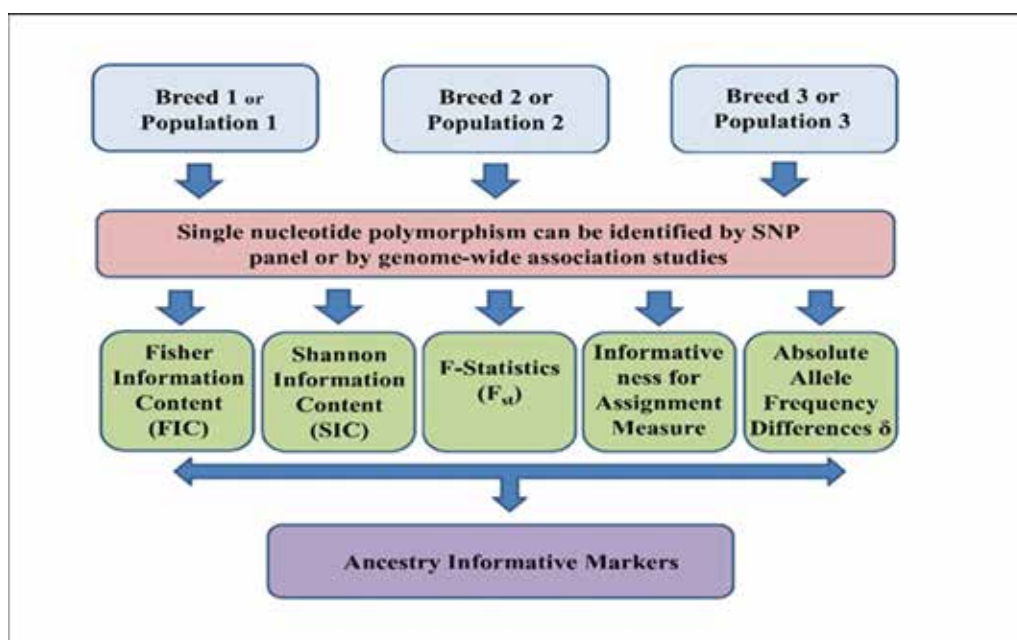


Fig. 3: Process of identification of Ancestry Informative Markers.

These methods offer complementary insights into genetic ancestry by evaluating the informativeness of genetic markers from different angles. FIC and SIC focus on the information content, F_{ST} measures population differentiation, evaluates assignment accuracy, and δ highlights allele frequency differences. Together, they form a comprehensive toolkit for ancestry analysis in population genetics.

Research and Advances in Ancestry Informative Markers

Recent research in livestock genetics has seen significant advancements in the use of informative markers. One of the key developments is the refinement of SNP panels, which now offer greater resolution and accuracy in genetic analysis. Modern

SNP panels can cover millions of SNPs across the genome, providing a comprehensive view of genetic variation and ancestry.

High-Density SNP Panels

High-density SNP panels have revolutionized livestock genetics by enabling more detailed and accurate genetic assessments. These panels include a large number of SNPs distributed across the genome, allowing researchers to pinpoint genetic variations associated with specific traits and diseases. For example, high-density SNP panels have been used to identify markers linked to resistance to diseases such as mastitis in dairy cattle or heat tolerance in beef cattle (Cardoso *et al.*, 2020; Macciotta *et al.*, 2017).

Genome-Wide Association Studies (GWAS)

Genome-wide association studies have become a powerful tool for identifying genetic markers associated with complex traits. By analysing the entire genome of a large population of livestock, researchers can find correlations between SNPs and traits of interest. Recent GWAS have provided valuable insights into the genetic basis of traits such as meat quality, reproductive performance, and feed efficiency (Sbardella *et al.*, 2021; Zhang *et al.*, 2020).

Admixture Analysis

Admixture analysis is another area where informative markers have made a significant impact. Admixture occurs when individuals from different genetic backgrounds interbreed, resulting in a new population with mixed ancestry. Understanding admixture patterns can help researchers identify the genetic contributions of different breeds or populations, which is important for breeding programs and conservation efforts (VonHoldt *et al.*, 2018). Recent studies have used SNP panels to assess admixture in livestock populations. For example, admixture analysis has been used to trace the influence of exotic breeds on local livestock populations, helping breeders make informed decisions about improving genetic diversity and performance (Berthouly-Salazar *et al.*, 2012; Edea *et al.*, 2015).

Also, the ancestry informative markers research in livestock have technological improvements in genotyping and sequencing, leading to significant breakthroughs in understanding breed development, hybridization, and selection. AIMS have been integrated into genomic selection programs, helping identify markers linked to key traits such as milk production, growth rate, and disease resistance in cattle, thereby enhancing breeding strategies and improving livestock performance (Lewis *et al.*, 2011). These markers have also uncovered historical hybridization events, revealing the genetic contributions of ancestral breeds to modern commercial livestock, which is crucial for maintaining genetic diversity and managing inbreeding (Yaro *et al.*, 2017). Additionally, AIMS have played a vital role in the conservation of endangered breeds by identifying unique genetic variants that need preservation, helping guide efforts to protect populations at risk (Supple and Shapiro, 2018). Furthermore, AIMS have provided insights into the genetic basis of disease resistance, such as in sheep, where markers linked to resistance against parasites like gastrointestinal nematodes have been identified (Alvarez *et al.*, 2019). This knowledge allows for the development of targeted breeding strategies to enhance the overall resilience and health of livestock populations, ensuring their sustainability in the face of environmental and disease challenges.

In conclusion, ancestry informative markers are powerful tools that have revolutionized the study of livestock genetics. By providing insights into the origins and genetic composition of livestock breeds, AIMS have become indispensable for breeding programs, conservation efforts, and research into disease resistance and productivity traits. The continued advancement of AIMS research, combined with new genomic technologies, will undoubtedly lead to even greater improvements in livestock management and sustainability in the years to come.

ACKNOWLEDGEMENT

Authors would like to thank Director, ICAR-NBAGR, Karnal and Principal, Institute of Integrated and Honors Studies, Kurukshetra University Kurukshetra for providing necessary help.

REFERENCES

- Alexander DH, Novembre J, and Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655-1664.
- Alvarez I, Fernández I, Soudré A, Traoré A, Pérez-Pardal L, Sanou M, Tapsoba SA, Menéndez-Arias NA, and Goyache F. 2019. Identification of genomic regions and candidate genes of functional importance for gastrointestinal parasite resistance traits in Djallonké sheep of Burkina Faso. *Archives Animal Breeding*, 62(1):313-323.
- Berthouly-Salazar C, Thevenon S, Van TN, Nguyen BT, Pham LD, Chi CV, and Maillard JC. 2012. Uncontrolled admixture and loss of genetic diversity in a local Vietnamese pig breed. *Ecology and Evolution*, 2(5):962-975.
- Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, and Hirschhorn JN. 2005. Demonstrating stratification in a European American population. *Nature Genetics*, 37(8):868-872.
- Cardoso DF, Fernandes Junior GA, Scaletz DCB, Alves AAC, Magalhães AFB, Bresolin T, Ventura RV, Li C, de Sena Oliveira MC, Porto-Neto LR, and Carvalheiro R. 2020. Uncovering sub-structure and genomic profiles in across-countries subpopulations of Angus cattle. *Scientific Reports*, 10(1):8770.
- Chakraborty R. 1986. Gene admixture in human populations: models and predictions. *American Journal of Physical Anthropology*, 29(S7):1-43.
- Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, and Nutland S. 2005. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, 37(11):1243-1246.
- Das R, and Upadhyai P. 2018. An ancestry informative marker set which recapitulates the known fine structure

- of populations in South Asia. *Genome Biology and Evolution*, 10(9):2408-2416.
- Das R, Roy R, and Venkatesh N. 2019. Using ancestry informative markers (AIMs) to detect fine structures within Gorilla populations. *Frontiers in Genetics*, 10, 43.
- Dias-Alves T, Mairal J, and Blum MG. 2018. Loter: a software package to infer local ancestry for a wide range of species. *Molecular Biology and Evolution*, 35(9):2318-2326.
- Ding L, Wiener H, Abebe T, Altaye M, Go RC, Kercsmar C, Grabowski G, Martin LJ, Khurana Hershey GK, Chakorborty R, and Baye TM. 2011. Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics*, 12:1-18.
- Edea Z, Bhuiyan MSA, Dessie T, Rothschild MF, Dadi H, and Kim KS. 2015. Genome-wide genetic diversity, population structure and admixture analysis in African and Asian cattle breeds. *Animal*, 9(2):218-226.
- Epstein MP, Allen AS, and Satten GA. 2007. A simple and improved correction for population stratification in case-control studies. *The American Journal of Human Genetics*, 80(5):921-930.
- Falush D, Stephens M, and Pritchard, JK. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, 7(4):574-578.
- Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, and Pato MT. 2004. Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, 36(4):388-393.
- Getachew T, Huson HJ, Wurzinger M, Burgstaller J, Gizaw S, Haile A, Rischkowsky B, Brem G, Boison SA, Mészáros G, and Mwai AO. 2017. Identifying highly informative genetic markers for quantification of ancestry proportions in crossbred sheep populations: implications for choosing optimum levels of admixture. *BMC Genetics*, 18:1-14.
- Hirschhorn JN, and Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nature Review Genetics*, 6(2):95-108.
- Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, and McKeigue PM. 2003. Control of confounding of genetic associations in stratified populations. *The American Journal of Human Genetics*, 72(6):1492-1504.
- Kehdy FS, Gouveia MH, Machado M, Magalhães WC, Horimoto AR, Horta BL, Moreira RG, Leal TP, Scliar MO, Soares-Souza GB, and Rodrigues-Soares F. 2015. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proceedings of the National Academy of Sciences*, 112(28):8696-8701.
- Lewis J, Abas Z, Dadousis C, Lykidis D, Paschou P, and Drineas P. 2011. Tracing cattle breeds with principal components analysis ancestry informative SNPs. *PLoS One*, 6(4): e18007.
- Macciotta NPP, Biffani S, Bernabucci U, Lacetera N, Vitali A, Ajmone-Marsan P, and Nardone A. 2017. Derivation and genome-wide association study of a principal component-based measure of heat tolerance in dairy cattle. *Journal of Dairy Science*, 100(6): 4683-4697.
- Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert, TD, Leon-Velarde F, Moore LG, Vargas E, McKeigue PM, and Shriver MD. 2007. A genomewide admixture mapping panel for Hispanic/Latino populations. *The American Journal of Human Genetics*, 80(6):1171-1178.
- Maples BK, Gravel S, Kenny EE, and Bustamante CD. 2013. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2):278-288.
- Mekhfi L, El Khalfi B, Saile R, Yahia H, and Soukri A. 2024. The interest of informative ancestry markers (AIM) and their fields of application. In BIO Web of Conferences (Vol. 115, p. 07003). *EDP Sciences*.
- Monau PI, Raphaka K, and Nsoso SJ. 2022. Adoption of Genomics and Breeding Strategies to Improve Goat Productivity in Southern Africa. In Food Security and Safety Volume 2: African Perspectives (pp. 471-479). Cham: Springer International Publishing.
- Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, Tandon A, Schirmer C, Neubauer J, Bedoya G, and Duque C. 2007. A genome wide admixture map for Latino populations. *The American Journal of Human Genetics*, 80(6):1024-1036.
- Rezende FMD, Ferraz JBS, Eler JP, Silva RCGD, Mattos EC, and Ibanez-Escriche N. 2012. Study of using marker assisted selection on a beef cattle breeding program by model comparison. *Livestock Science*, 147(1-3):40-48.
- Rosenberg NA, Li LM, Ward R, and Pritchard, JK. 2003. Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics*, 73(6):1402-1422.
- Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, and Feldman MW. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, 1(6):e70.
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, and Feldman MW. 2002. Genetic structure of human populations. *Science*, 298(5602):2381-2385.
- Santangelo R, González-Andrade F, Børsting C, Torroni A, Pereira V, and Morling N. 2017. Analysis of ancestry informative markers in three main ethnic groups from Ecuador supports a trihybrid origin

- of Ecuadorians. *Forensic Science International: Genetics*, 31:29-33.
- Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, Thakare RP, Banday S, Mishra, AK, Das G, and Malonia SK. 2023. Next-generation sequencing technology: current trends and advancements. *Biology*, 12(7):997.
- Sbardella AP, Watanabe RN, da Costa RM, Bernardes PA, Braga LG, Baldi Rey FS, Lôbo RB and Munari DP. 2021. Genome-wide association study provides insights into important genes for reproductive traits in Nelore cattle. *Animals*, 11(5):1386.
- Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N, and Baron A. 2003. Skin pigmentation, biogeographical ancestry and admixture mapping. *Human genetics*, 112:387-399.
- Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, Malasky MJ, Scafe C, Le E, and De Jager PL. 2004. A high-density admixture map for disease gene discovery in African Americans. *The American Journal of Human Genetics*, 74(5):1001-1013.
- Somenzi E, Ajmone-Marsan P, and Barbato M. 2020. Identification of ancestry informative marker (AIM) panels to assess hybridisation between feral and domestic sheep. *Animals*, 10(4):582.
- Supple MA, and Shapiro B. 2018. Conservation of biodiversity in the genomics era. *Genome Biology*, 19:1-12.
- Tian C, Hinds DA, Shigeta R, Adler SG, Lee A, Pahl MV, Silva G, Belmont JW, Hanson RL, Knowler WC, and Gregersen PK. 2007. A genome wide single-nucleotide-polymorphism panel for Mexican American admixture mapping. *The American Journal of Human Genetics*, 80(6):1014-1023.
- Vongpaisarnsin K, Listman JB, Malison RT, and Gelernter, J. 2015. Ancestry informative markers for distinguishing between Thai populations based on genome-wide association datasets. *Legal Medicine*, 17(4):245-250.
- VonHoldt BM, Brzeski KE, Wilcove DS, and Rutledge LY. 2018. Redefining the role of admixture and genomics in species conservation. *Conservation Letters*, 11(2):e12371.
- Winkler CA, Nelson GW, and Smith MW. 2010. Admixture mapping comes of age. *Annual Review of Genomics and Human Genetics*, 11(1):65-89.
- Yaro M, Munyard KA, Stear MJ, and Groth, DM. 2017. Molecular identification of livestock breeds: a tool for modern conservation biology. *Biological Reviews*, 92(2):993-1010.
- Zhang F, Wang Y, Mukibi R, Chen L, Vinsky M, Plastow G, Basarab J, Stothard P, and Li C. 2020. Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: I: feed efficiency and component traits. *BMC Genomics*, 21:1-22.