Review

Methodological Advances in the Analysis of Genetic Population Structure: Implications for Biodiversity Conservation

Karolína Pálešová, Nina Moravčíková*, Radovan Kasarda
Slovak University of Agriculture in Nitra, Faculty of Agrobiology and Food Resources, Institute of Nutrition and Genomic, Nitra, Slovakia

Article Details: Received: 2025-02-24 | Accepted: 2025-03-10 | Available online: 2025-03-31

https://doi.org/10.15414/afz.2025.28.01.28-39

This paper provides an overview of advances in the analysis of the genetic structure of populations, focusing on the evolution of statistical approaches and their applications in conservation genetics. Understanding genetic relationships among populations is crucial for assessing evolutionary processes such as gene flow, genetic drift, and selection, which fundamentally affect genetic diversity over time. Traditionally, studies relied on a limited number of genetic markers and summary statistics; however, the advent of high-throughput genomic technologies has dramatically enhanced both the resolution and accuracy of these analyses. Whole-genome sequencing and dense SNP arrays now provide unprecedented insights into neutral and adaptive variations, enabling fine-scale detection of population subdivisions and historical demographic trends. In parallel, the development of advanced statistical models has refined genetic analyses, allowing for more precise estimations of genetic differentiation, admixture, and ancestral relationships. These innovations are particularly valuable in conservation genetics, where robust assessments are essential for optimising strategies to maintain genetic diversity, identify populations at risk, and mitigate the effects of inbreeding and effective population size decline. Despite these improvements, challenges remain, including computational demands and the need to account for complex demographic histories and selection pressures. Given the continuous evolution of analytical techniques, selecting appropriate methods tailored to specific research questions is critical for producing reliable insights into population structure and effectively guiding conservation efforts. In conclusion, the continuous advancement of genomic analysis tools enhances the ability to study population dynamics in greater detail and supports more effective conservation planning.

Keywords: population structure, biodiversity, genomics, conservation

1 Introduction

The main objective of population genetics is to identify populations and elucidate their relationships, providing critical insights into biological diversity and informing conservation strategies. Population structure, defined as the distribution of genetic variation within and among populations, reflects evolutionary processes such as gene flow, genetic drift, and selection (Lehocká et al., 2020; Hohenlohe et al., 2020). Genetic structure encompasses allele frequencies, genotype distributions, and chromosomal variation, all influencing a population's resilience and adaptability (Herbers, 2010). However, habitat fragmentation, population declines, and the spread of invasive species can lead to a loss of genetic diversity, reducing fitness and limiting adaptive potential in response to environmental changes (Ceballos et al., 2017). Genetic diversity is fundamental to a population's ability to persist in dynamic environments, making its monitoring essential for sustainable conservation and management efforts (Kasarda et al., 2020; Moravčíková and Kasarda, 2020). Reduced variation increases susceptibility to inbreeding depression and local extinctions, whereas maintaining connectivity between populations can enhance genetic exchange and evolutionary potential (Frankham, 2018). Therefore, understanding genetic population structure is essential for designing effective conservation strategies, such as establishing habitat

*Corresponding Author: Nina Moravčíková, Slovak University of Agriculture in Nitra, Faculty of Agrobiology and Food Resources, Institute of Nutrition and Genomics, Tr. Andreja Hlinku 2, 949 76 Nitra, Slovakia
✉ nina.moravcikova@uniag.sk https://orcid.org/0000-0003-1898-8718

corridors, reinforcing small populations, or guiding reintroduction programs.

In addition to conservation efforts, optimising breeding strategies is crucial for maintaining genetic diversity and enhancing population resilience. Breeding programs have to balance genetic gain with the preservation of variability, ensuring long-term adaptability. Advances in genomic techniques have played a significant role in achieving these goals, allowing researchers to assess population structure with unprecedented resolution and infer demographic history by examining both neutral genetic variation, which arises through mutation and genetic drift without selective pressure, adaptive genetic changes, which arise through natural selection acting on alleles that confer a fitness advantage in a given environment (Funk et al., 2012). High-resolution genotyping methods, including whole-genome sequencing and SNP arrays, have largely replaced microsatellites due to their superior scalability and precision (Hauser et al., 2021; Leaché and Oaks, 2017). Accurately characterising population structure and evolutionary history relies on robust statistical frameworks capable of detecting patterns of allelic variation and demographic processes. These methods are fundamental for both conservation and breeding applications, as they provide critical insights into genetic drift, divergence, and admixture. Stochastic simulations have been used to design breeding strategies that maximise desired traits while conserving genetic diversity (Hassanpour et al., 2023). Furthermore, strategies such as minimising kinship and implementing optimal contribution selection can prevent inbreeding depression and maintain evolutionary potential, making breeding programs more effective in supporting both conservation and sustainable population management (Li et al., 2022). However, interpreting population genetic patterns requires careful consideration, as confounding factors such as historical gene flow, selection, and incomplete lineage sorting can obscure underlying evolutionary processes (Sul et al., 2018; Moorjani and Hellenthal, 2023).

1.1 Estimation and Visualisation of Genetic Relationships

Sewall Wright (1921; 1923) revolutionised population genetics by introducing F-statistics, a foundational framework in evolutionary theory that integrates concepts such as genetic variance, allele identity by descent (IBD), and genetic diversity. Among the indices derived from this framework, the fixation index (F_{ST}) is the most widely used metric for assessing genetic differentiation between populations (Subramanian, 2022). F_{ST} quantifies the extent to which allele frequencies differ between subpopulations relative to the total population, providing a measure of genetic structure and divergence. Values of F_{ST} range from 0 to 1, where 0 indicates no genetic differentiation and high gene flow, while 1 signifies complete genetic isolation and absence of gene flow (Wright, 1965). A widely used approach for calculating F_{ST} is the method developed by Weir and Cockerham (1984), which provides a weighted estimate of genetic differentiation based on allele variance within and between populations. This method accounts for unequal population sizes by assigning greater weight to larger populations, thereby enhancing the precision of the estimates. This method is implemented in various software tools, with the most commonly utilised being PLINK (Chang et al., 2015), VCFtools (Danecek et al., 2011) and the StAMPP package in the R programming language (Pembleton et al., 2013), which efficiently handle large genomic datasets. In contrast, GenAlEx (Peakall and Smouse, 2006) and Arlequin (Excoffier and Lischer, 2010) are more suited for smaller datasets due to limitations in the number of markers they can process. Given its broad applicability, F_{ST} has become a standard measure of genetic differentiation, widely used to quantify population structure and gene flow across taxa. Beyond its role in quantifying genetic differentiation, F_{ST} serves as a comparative metric in conservation genetics, aiding in identifying genetic barriers and population connectivity (Hedgecock et al., 2007). In evolutionary biology, it is instrumental in disentangling the relative contributions of selection and genetic drift to genetic variation (Whitlock and Guillaume, 2009). Despite its usefulness, F_{ST} has several limitations. It assumes Hardy-Weinberg Equilibrium (HWE), often violated in natural populations, leading to biased estimates (Guillot and Orlando, 2013). Additionally, F_{ST} can overestimate differentiation in weakly structured populations and fails to account for continuous population structure and admixture (Putman and Carbone, 2014; Moura and Eurico, 2010). Its application to polyploid organisms is also problematic due to its inability to handle allele dosage effects (Liu and Meirmans, 2018).

Genetic distance metrics offer a quantitative framework for assessing genetic divergence, with specific methods tailored to different types of genetic markers and evolutionary contexts. Several widely utilised measures include the Cavalli-Sforza and Edwards chord distance (1967), which is particularly effective for microsatellite data, and Reynolds' genetic distance (1983), which is well-suited for recently diverged populations. Additionally, Rogers' genetic distance (1972) and Edwards' (1971) are frequently employed in phylogenetic studies to infer evolutionary relationships. The selection of an appropriate metric depends on various factors, including the genetic markers used (e.g., SNPs, microsatellites) and the underlying evolutionary model. Among these multiple approaches, Nei's genetic distance (Nei, 1972) has become the most widely applied metric in population genetics as it quantifies genetic differentiation based on allele frequency variation. This metric is particularly useful for inferring evolutionary relationships and reconstructing historical migration patterns (Takezaki and Nei, 1996; Nei and Kumar, 2000). Nei's genetic distance is widely used in population genetics, as it provides reliable estimates even when sample sizes are unequal (Sekino and Hara, 2001). It has been implemented in various software tools, with programs like GENEPOP (Rousset et al., 2008) and Arlequin (Excoffier and Lischer, 2010) suited for smaller datasets, while R packages such as *adegenet* (Jombart, 2008) and *poppr* (Kamvar, 2014) are more efficient for larger genomic analyses. MEGA (Kumar et al., 2018), on the other hand, is widely used for phylogenetic analysis and genetic distance estimation in evolutionary studies. Due to its sensitivity to genetic variation, Nei's genetic distance is widely used to detect evolutionary divergence, though it does not directly distinguish between neutral and adaptive changes (Fan et al., 2008; Nagai et al., 2007). It is also a key tool in phylogenetic analysis for reconstructing evolutionary relationships (Makrem et al., 2006) and is compatible with diverse genetic data types, including microsatellites and SNPs, making it versatile across research applications (Nam et al., 2016). Moreover, it integrates well with statistical frameworks like AMOVA, enhancing population structure analyses (Zhu et al., 2014). Although Nei's genetic distance does not explicitly assume HWE, deviations from it can influence allele frequency estimates, potentially affecting result accuracy (Chakraborty, 2010). Low genetic diversity within samples can further reduce its accuracy, leading to an underestimation of differentiation (Bublyket al., 2020). While valuable for assessing genetic divergence, Nei's distance does not directly estimate gene flow, limiting its effectiveness for highly dispersive species where migration plays a dominant role (Rosel et al., 2017).

Beyond traditional genetic distance measures, Identity by Descent (IBD) matrices and genomic relationship matrices offer alternative approaches for quantifying genetic relationships among individuals. IBD methods identify shared genomic segments inherited from a common ancestor without recombination events, offering a direct measure of both recent and ancient ancestry (Thompson, 2013). Closely related individuals share long IBD segments, while distant relatives exhibit shorter, fragmented regions that decrease over time. Due to shared ancestry, IBD segments persist over larger genomic distances in small, isolated populations, providing insights into demographic history, population structure, and bottleneck events (Browning and Browning, 2012). Beyond population structure, IBD-based methods aid parent selection in breeding and conservation by identifying individuals with minimal shared ancestry, helping to maintain genetic diversity and reduce inbreeding risks (Wellmann, 2019; Meuwissen et al., 2020). Several computational tools have been developed to estimate IBD matrices, each optimised for different data types and applications. PLINK (Chang et al., 2015) is widely used for IBD estimation, efficiently detecting shared genomic segments in large SNP datasets. For high-resolution IBD detection, BEAGLE (Browning and Browning, 2013) provides probabilistic phasing and IBD inference, making it particularly useful for recent ancestry analysis. GERMLINE (Gusev et al., 2009) is another widely used tool designed for identifying long IBD tracts, facilitating studies of kinship and demographic history. While methods like PCA and ADMIXTURE are more commonly used for broader population structure analysis, IBD-based approaches remain valuable for revealing fine-scale stratification and genetic differentiation. Additionally, IBD analyses help mitigate confounding in association studies and improve genetic variance estimates (Browning and Thompson, 2012). They also offer insights into evolutionary history, shedding light on past population structures and migration patterns (Palamara et al., 2012). However, IBD inference has limitations. Small sample sizes and low genetic diversity can reduce the accuracy and reliability of the inferred IBD segments (Henden et al., 2018). Additionally, false positives may arise due to genotyping errors, population structure effects, or repetitive genomic regions, complicating genetic association studies (Browning and Thompson, 2012).

Genomic relationship matrices (GRMs) were first introduced by VanRaden (2008) as a method to estimate genetic relatedness using dense SNP markers, providing an alternative to traditional pedigree-based matrices. By incorporating genome-wide information, GRMs capture both additive and non-additive genetic variances, offering a more comprehensive measure of relatedness, particularly when pedigree data is incomplete or unavailable (VanRaden, 2008; Su et al., 2012). Additionally, GRMs enhance genetic analyses by improving population structure control in statistical models, making genome-wide association studies (GWAS) and heritability estimates more reliable (Veerkamp et al., 2011; Villanueva et al., 2021). They also help identify population substructure, revealing genetic differentiation patterns that support conservation and breeding programs (Zapata-Valenzuela et al., 2013). Several computational tools are widely used for constructing and analysing GRMs, each optimised for different applications. GCTA (Yang et al., 2011) is commonly used for estimating GRMs and conducting heritability analysis in large datasets. PLINK (Chang et al., 2015) is a versatile tool that enables GRM computation from genotype data while also supporting genome-wide association studies. ASReml (Gilmour et al., 2009) integrates GRMs into mixed models, making it particularly useful for genetic evaluations in breeding programs. GRMs offer several advantages. They provide a more precise measure of genetic relatedness, improving genetic predictions and breeding value estimates (Makgahlela et al., 2014). However, GRMs also have limitations. Their computational complexity poses challenges for large datasets, while insufficient data in smaller datasets increases the risk of overfitting in predictive models (Wright et al., 2019; Aguilar et al., 2011).

Recent advances in machine learning provide a flexible, data-driven approach to estimating genetic relatedness. Semiparametric efficient estimators incorporating machine learning techniques improve genetic covariance and correlation estimation while reducing bias from model misspecification. These methods also enable the construction of valid confidence intervals, making them particularly effective for high-dimensional genomic data (Guo et al., 2023). Supervised learning techniques, including decision tree-based classifiers, have demonstrated high reliability in genetic classification tasks, particularly in distinguishing closely related populations (Kukučková et al., 2018). Among such approaches, ensemble learning methods like Random Forest (RF) (Breiman, 2001) have been widely used for genetic data analysis, leveraging multiple decision trees to enhance predictive accuracy and reduce variance in tree-based models (Kasarda et al., 2023). While RF is effective for feature selection and genotype-phenotype association studies, recent methodological developments have expanded beyond tree-based models to incorporate information-theoretic measures for genetic distance estimation. Mutual Information and Entropy H (MIH) employs an information matrix (IM) derived from genetic data, which encapsulates both positional heterogeneities, quantified through Shannon entropy, and coordinated substitutions among loci, assessed via mutual information (Campo et al., 2023). Mutual information measures the extent to which knowledge of one variable reduces uncertainty about another, making it a powerful tool for detecting complex dependencies between genetic sites. Unlike traditional correlation metrics, it does not assume linearity or other specific forms of dependence, allowing for the identification of both direct and indirect genetic interactions (Faith et al., 2007). Given its ability to capture nonlinear associations, mutual information is widely applicable in inferring various interaction networks, including biological, chemical, and social systems. In such networks, a high mutual information value indicates strong interdependence between components, whereas a value approaching zero suggests little to no relationship (Villaverde et al., 2014). However, MIH has notable limitations, including high computational complexity due to large-scale mutual information calculations and sensitivity to sequence length and data quality. Additionally, it does not directly measure allele frequency differences, making interpretation less intuitive for population geneticists. The method also requires large sample sizes to ensure statistical power, as small datasets may not provide robust entropy and mutual information estimates. Mutual Information Analyzer (MIA) (Lichtenstein et al., 2015) is a valuable tool for MIH analysis, but its suitability depends on the dataset and research objectives.

Genetic distance measures can be visualised using phylogenetic trees and network-based methods to interpret population structure and evolutionary relationships. These approaches reveal divergence patterns, clustering, and genetic connectivity, aiding in fine-scale genetic analysis. The Neighbor-Joining (NJ) method, a widely used distance-based technique, reconstructs evolutionary history by minimising total branch length without assuming a strict molecular clock (Saitou and Nei, 1987). NJ trees are particularly useful for inferring population relationships from genetic distance matrices such as Nei's genetic distance. However, while phylogenetic trees depict hierarchical relationships, network-based visualisations offer a more nuanced perspective, especially in populations exhibiting high levels of gene flow and admixture. Methods such as NetView and split decomposition networks capture reticulations and complex evolutionary histories, making them valuable for analysing fine-scale population structure (Morrison, 2010; Neuditschko et al., 2012). Applied to large genomic datasets, these approaches complement tree-based methods in population genetics studies (Al-Breiki et al., 2018). Beyond distance-based techniques, model-based approaches such as TreeMix provide an alternative for inferring population splits and mixtures from genome-wide allele frequency data (Pickrell and Pritchard, 2012). This software has been shown to accurately reconstruct known relationships among populations while identifying previously unrecognized connections, making it a powerful tool for clustering individuals into genetically homogeneous groups. However, TreeMix may not fully capture recent admixture events or complex demographic histories, where continuous patterns of genetic variation challenge traditional discrete models. To enhance population structure analyses, visualization techniques are often integrated with clustering methods like Principal Components Analysis (PCA) and model-based clustering approaches, providing a more comprehensive understanding of genetic relationships (Steinig et al., 2016). Advances in population genetics software have further refined visualization methodologies, increasing analytical precision. The latest version of STRAF 2 integrates multidimensional scaling (MDS) for population visualization and includes an R package, offering both interactive and offline tools for efficient genetic structure analysis (Gouy and Zieger, 2025).

Traditional methods for visualising genetic relationships, such as NetView and NJ trees, rely on precomputed genetic distance matrices, typically derived from F_{ST} or Nei's genetic distance. However, these approaches assume discrete population boundaries, which may not accurately represent continuous patterns of genetic variation. To address this limitation, Smith et al. (2024) developed a deep neural network framework that utilises geo-referenced SNP data to generate spatially heterogeneous maps of population density and dispersal rates. Trained on simulated datasets, mapNN integrates genotypic data and sampling locations to infer demographic parameters, offering a more precise representation of population structure. By simultaneously estimating both the magnitude and spatial variation of dispersal and density, this approach enhances fine-scale population visualisation and complements network-based methods like NetView. Its ability to incorporate both genetic and geographic information makes it particularly valuable for studying species with high gene flow, isolation-by-distance patterns, or complex population connectivity.

1.2 Admixture Analysis and Model-Based Clustering

Admixture refers to the genetic integration of two or more previously isolated populations, resulting in new genetic combinations. This process is a major driver of population genetic structure, offering critical insights into historical gene flow, demographic history, and patterns of genetic diversity (vonHoldt et al., 2011). Moreover, it is one of the most rapid evolutionary mechanisms, capable of significantly altering the genetic composition of populations within a few generations (Korunes and Goldberg, 2021). The detection and quantification of admixture are typically conducted using model-based clustering approaches, which estimate individual ancestry proportions based on multilocus genotype data (Lawson et al., 2018). These methods have become indispensable for inferring population structure and reconstructing genetic relationships, with clustering algorithms widely applied to genetic ancestry characterisation. Pritchard et al. (2000) introduced a Bayesian clustering algorithm, implemented in the STRUCTURE software, for defining genetic populations and assigning individuals to inferred clusters. While highly accurate, its computational intensity limits its use for large datasets. To improve scalability, fastSTRUCTURE (Raj et al., 2014), FRAPPE (Tang et al., 2005), and ADMIXTURE (Alexander et al., 2009) were developed, employing a similar inference model but optimised for large-scale genomic data. Despite its widespread use in population genetics, admixture analysis has inherent limitations that can affect the accuracy of inferred population structure. Bayesian clustering methods, such as those implemented in STRUCTURE, are computationally intensive, requiring extensive iterations for convergence. This makes them impractical for large genomic datasets without high-performance computing resources (Wang, 2022). A key drawback of model-based admixture analysis methods, such as STRUCTURE and ADMIXTURE, is their reliance on a predefined number of ancestral populations (K); if misestimated, this can introduce bias and misrepresent genetic structure, particularly in populations with complex admixture histories (Gopalan et al., 2022). Nevertheless, STRUCTURE remains favoured for small-scale analyses (Lawson et al., 2018). To improve the interpretation of STRUCTURE and ADMIXTURE outputs, several tools facilitate result processing and visualization. CLUMPP, developed by Jakobsson and Rosenberg (2007), addresses label switching and multimodality issues by aligning multiple replicate analyses, ensuring consistent cluster assignments. Similarly, DISTRUCT (Rosenberg, 2004), provides clear graphical representations of individual membership coefficients, making population structure patterns easier to interpret.

Wang (2024) developed PopCluster, a likelihood-based framework that integrates mixture and admixture models for high-resolution population analyses. This method demonstrates superior accuracy, particularly when analysing weakly differentiated populations, intricate population structures, or datasets with small or highly unbalanced sample sizes. Additionally, by utilising parallel computing frameworks such as MPI (Message Passing Interface) and OpenMP (Open Multi-Processing), PopCluster enhances computational efficiency, making it well-suited for large-scale genomic datasets. Its capacity to handle both multiallelic markers (e.g., microsatellites) and millions of biallelic markers (e.g., SNPs) further underscores its versatility in population genetics (Wang, 2022). Another major advancement in admixture analysis is Neural ADMIXTURE, introduced by Dominguez Mantes et al. (2023), a deep learning-based autoencoder designed to improve the efficiency of genomic clustering. Unlike traditional methods, which require extensive computational resources, Neural

ADMIXTURE leverages neural networks to achieve orders-of-magnitude speedup while maintaining high accuracy. Its multi-head architecture enables the simultaneous inference of multiple clustering solutions, reducing the need for separate runs at different values of K. This makes it particularly suitable for large-scale genomic datasets, such as biobanks, where traditional approaches become computationally prohibitive (Bycroft et al., 2018). Unlike statistical approaches, deep learning models can infer population structure without strong parametric assumptions, making them more adaptable to complex demographic histories and continuous genetic variation.

While clustering methods directly infer population structure, simulations serve as a complementary approach by generating artificial genetic data to test the accuracy and robustness of these inferences. One example is the Gametes Simulator, which generates multilocus genotypes based on observed allele frequencies and is particularly useful for analysing genetic structure in outbreeding diploid species (Porta et al., 2020). However, its reliance on accurate allele frequency data can be a limitation. Another useful tool is ADAM (Pedersen et al., 2009), a stochastic simulation program designed to model selective breeding schemes in animal populations. By simulating genetic changes under different selection strategies, mating designs, and population structures, ADAM provides valuable insights into how breeding practices influence genetic variation and population structure over time. Similarly, DYMEX (Li et al., 2015) models population dynamics in ecological contexts, enabling researchers to assess environmental influences on population structure over time. While primarily applied in ecological and conservation studies, its use varies depending on specific modelling needs. SFS_CODE (Sinha et al., 2011) is a forward-time simulation tool designed for modelling genetic data under complex selection and demographic scenarios, allowing researchers to study selective sweeps, evolutionary pressures, and the effects of mutation and recombination.

1.3 Multivariate Techniques

Principal component analysis (PCA) and Discriminant Analysis of Principal Components (DAPC) are widely used multivariate methods in population genetics for assessing genetic structure and variation with predefined population assignments. PCA reduces high-dimensional genetic data into a smaller set of uncorrelated variables (principal components), maximising the variance captured from the dataset. DAPC, in contrast, combines PCA with discriminant analysis (DA) to optimise population differentiation by maximising between-group variation while minimising within-group variation, making it particularly effective for assigning individuals to populations (Jombart et al., 2010; Karamizadeh et al., 2013). These methods help visualise population structure, identify genetic clustering, and detect admixture patterns, with PCA being an unsupervised approach, while DAPC requires predefined groups (Patterson et al., 2006; Qin et al., 2021). A key mathematical foundation of PCA lies in eigenvalues and eigenvectors, which are derived from the covariance matrix of the dataset. Eigenvalues represent the amount of variance explained by each principal component, with higher eigenvalues indicating components that capture more genetic variation. Eigenvectors, on the other hand, define the directions in which the data points vary the most, forming the principal axes of the PCA plot (Jolliffe and Kadima, 2016). Building on these principles, DAPC maximises between-group variance, enhancing its effectiveness for classifying individuals into populations (Chhotaray et al., 2019). PCA and DAPC can be performed using *adegenet* (Jombart, 2008), which handles datasets of various sizes, or GenAlEx (Peakall and Smouse, 2012), which is more commonly used for smaller datasets. For large-scale genomic data, PLINK efficiently computes PCA on genome-wide SNPs (Chang et al., 2015), while TASSEL integrates PCA for GWAS applications (Bradbury et al., 2007). However, PCA is sensitive to missing data and outliers, which can distort population clustering (Serneels and Verdonck, 2008), and DAPC requires predefined groups, making it less effective for detecting unknown structures (Miller et al., 2020). Additionally, PCA may fail to distinguish populations with high gene flow, as it only captures dominant variance components (Elhaik, 2022), while DAPC can be computationally intensive for large datasets due to the multiple steps of PCA transformation and discriminant function selection (Thia, 2023). To improve population structure and genetic relatedness analysis, PSReliP integrates PCA, multidimensional scaling (MDS), and clustering methods, providing a comprehensive framework for multivariate genetic analysis. Combining PLINK-based computations with an interactive Shiny web interface, enables dynamic visualisations, enhancing accessibility. These features make PSReliP a valuable tool for interpreting genetic variation in GWAS, genomic selection, and evolutionary studies, strengthening population genetics methodologies (Solovieva and Sakai, 2023).

A recent advancement in multivariate methods for population structure analysis is Population-Based Hierarchical Non-negative Matrix Factorization (PHNMF), introduced by Ding et al. (2022). PHNMF extends Hierarchical Non-negative Matrix Factorization (HNMF) to identify hierarchical population structures in complex genomic datasets. Unlike conventional techniques such as Principal Component Analysis (PCA)

Table 1 Overview of commonly used software for population structure analysis

Category	Software	Author(s)	Operation system
Genetic distance and relatedness estimation	VCFtools	Danecek et al., 2011	Linux, macOS
	StAMPP	Pembleton et al., 2013	Windows, macOS, Linux (R package)
	GenAlEx	Peakall and Smouse, 2006	Windows, macOS (Excel add-on)
	Arlequin	Excoffier and Lischer, 2010	Windows, Linux (GUI for Windows, command-line for Linux)
	GENEPOP	Rousset, 2008	Windows, macOS, Linux (Web-based)
	PLINK	Chang et al., 2015	Windows, macOS, Linux
	MEGA	Kumar et al., 2018	Windows, Linux
	GCTA	Yang et al., 2011	Windows, Linux
	ASReml	Gilmour et al., 2009	Windows, Linux
	BEAGLE	Browning and Browning, 2013	Windows, macOS, Linux (Java-based)
	GERMLINE	Gusev et al., 2009	macOS, Linux
	adegenet	Jombart, 2008	Windows, macOS, Linux (R package)
	poppr	Kamvar et al., 2014	Windows, macOS, Linux (R package)
Genetic relationships visualization	TreeMix	Pickrell and Pritchard, 2012	macOS, Linux
	NetView	Neuditschko et al., 2012	Windows, macOS, Linux (Python-based)
	straf	Gouy and Zieger, 2025	Windows, macOS, Linux (R package)
	mapNN	Smith et al., 2024	Linux
Admixture analysis	STRUCTURE	Pritchard et al., 2000	Windows, macOS, Linux
	ADMIXTURE	Alexander et al., 2009	Linux, macOS
	fastSTRUCTURE	Raj et al., 2014	Linux, macOS
	FRAPPE	Tang et al., 2005	Linux
	PopCluster	Wang, 2024	Windows, macOS, Linux (R package)
	Neural ADMIXTURE	Dominguez Mantes et al., 2023	Linux, macOS (Python-based)
Interpretation of structure and admixture outputs	CLUMPP	Jakobsson and Rosenberg, 2007	Windows, macOS, Linux
	DISTRUCT	Rosenberg, 2004	Windows, macOS, Linux
Multivariate statistical approaches	TASSEL	Bradbury et al., 2007	Windows, macOS, Linux (Java-based)
	PSReliP	Solovieva and Sakai, 2023	Linux
	PHNMF	Ding et al., 2022	Windows, macOS, Linux (Python-based)
Simulations	Gametes Simulator	Porta et al., 2020	Windows, macOS, Linux (Python-based)
	ADAM	Pedersen et al., 2009	Windows (C++-based)
	DYMEX	Li et al., 2015	Windows
	SFS_CODE	Sinha et al., 2011	Linux, macOS
Information-theoretic approaches	MIA	Lichtenstein et al., 2015	Windows (Python-based)

and Discriminant Analysis of Principal Components (DAPC), which rely on linear transformations, PHNMF employs non-negative matrix factorisation to uncover latent population structures based on feature similarity. This approach automatically assigns subpopulations while preserving a hierarchical clustering framework, enhancing both interpretability and scalability for large genomic datasets. A key advantage of PHNMF is its ability to analyse large-scale genomic data with minimal assumptions about underlying genetic distributions, making it particularly effective for genetic population structure analysis. Numerical evaluations demonstrate that PHNMF accurately reconstructs latent hierarchical subpopulations, outperforming established methods such as Latent Class Analysis (LCA) and Latent Profile Analysis (LPA). LCA is a probabilistic modelling approach used to infer latent categorical groupings under the assumption of conditional independence among variables within each group (Haughton et al., 2009; Linzer and Lewis, 2011). LPA extends this framework to continuous data, representing a form of multivariate mixture modelling that maintains the same conditional independence assumption (Peugh and Fan, 2013; Oberski, 2016). Despite its advantages, PHNMF has some limitations. It is computationally intensive, requiring substantial memory for large datasets. Additionally, it is sensitive to data sparsity, which can affect accuracy when dealing with missing values or low-frequency variants.

An overview of commonly used software for analysing population structure is presented in Table 1. Many widely used tools, such as StAMPP, PLINK, TASSEL, and *adegenet*, function across Windows, macOS, and Linux due to their implementation in R or Java. MEGA is available for Windows, macOS, and Linux, while *straf*, designed for forensic genetics, is available as an R package. However, some tools have platform-specific requirements. PSReliP is restricted to Linux due to its reliance on bash shell scripts, while ASReml is compatible with Windows, macOS, and Linux. FRAPPE is designed for Linux, whereas ADMIXTURE supports both Linux and macOS but lacks a native Windows version. Software for IBD and kinship analysis, such as BEAGLE, is Java-based and cross-platform, while GERMLINE primarily supports Linux and macOS. Network-based and phylogenetic visualization tools, including NetView and TreeMix, are compatible with Linux and macOS, whereas mapNN, a deep learning-based tool for spatial demographic inference, is primarily available for Linux. Admixture and model-based clustering programs, including STRUCTURE, fastSTRUCTURE, PopCluster, and Neural ADMIXTURE, provide genetic ancestry inference, with most being optimized for Linux and macOS. Post-processing and visualization tools such as CLUMPP and DISTRUCT assist in interpreting clustering results and are compatible with Windows, macOS, and Linux. Software designed for multivariate and latent structure analysis, such as PHNMF, supports Windows, macOS, and Linux. For population genetic simulations, these tools facilitate forward-time and coalescent-based modelling: ADAM and DYMEX are available for Windows, Gametes Simulator supports Windows, macOS, and Linux, while SFS_CODE is restricted to Linux and macOS.

2 Conclusions

The continuous advancement of statistical methodologies has significantly improved the resolution and accuracy of genetic population structure analyses. By integrating high-throughput genomic data with sophisticated analytical approaches, researchers can now detect fine-scale genetic patterns, infer demographic history, and assess evolutionary processes more precisely. These improvements have profound implications for conservation genetics, where robust population assessments are essential for maintaining genetic diversity and mitigating inbreeding risks. However, challenges remain, including computational limitations and the need for methods that account for complex demographic histories and selection pressures. Future research should focus on refining these analytical tools, with an emphasis on incorporating artificial intelligence and machine learning to enhance automation, scalability, and accuracy in genetic data analysis. Such innovations will improve the efficiency and reliability of genetic analysis, ultimately strengthening evidence-based strategies for biodiversity conservation.

Acknowledgements

This research was funded by the Slovak Agency for Research and Development (grant number APVV-20-0161, APVV-17-0060 and DS-FR-22-0016).

References

Al-Breiki, H., Kennington, W. J., Brown, C., Burt, J. A., & DiBattista, J. D. (2018). Genome-wide SNP data reveal cryptic population structure in the scalloped spiny lobster (*Panulirus homarus*) along the Omani coastline. *BMC Genomics*, 19, 690. https://doi.org/10.1186/s12864-018-5044-8

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

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Browning, B. L., & Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2), 459–471. https://doi.org/10.1534/genetics.113.150029

Browning, S. R., & Browning, B. L. (2012). Identity by descent between distant relatives: Detection and applications. *Annual Review of Genetics*, 46, 617–633. https://doi.org/10.1146/annurev-genet-110711-155534

Browning, S. R., & Thompson, E. A. (2012). Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, 190(4), 1521–1531. https://doi.org/10.1534/genetics.111.136937

Bublyk, O., Andreev, I., Parnikoza, I., & Kunakh, V. (2020). Population genetic structure of *Iris pumila* L. in Ukraine: Effects of habitat fragmentation. *Acta Biologica Cracoviensia Series Botanica*, 62(1), 51–61. https://doi.org/10.24425/abcsb.2020.131665

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Cavalli-Sforza, L. L., & Edwards, A. W. F. (1967). Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics*, 19(3 Pt 1), 233–257.

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences*, 114(30), E6089–E6096. https://doi.org/10.1073/pnas.1704949114

Chakraborty, S. (2010). Comparative study of various genetic distance measures between populations for the ABO gene. *Nature and Science of Sleep*, 2(4). https://doi.org/10.15835/nsb245447

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

Chhotaray, S., Panigrahi, M., Pal, D., Ahmad, S. F., Bhushan, D., Gaur, G. K., Mishra, B. P., & Singh, R. K. (2019). Ancestry informative markers derived from discriminant analysis of principal components provide important insights into the composition of crossbred cattle. *Genomics*. https://doi.org/10.1016/j.ygeno.2019.10.008

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Ding, X., Dong, X., McGough, O., Shen, C., Ulichney, A., Xu, R., Swartworth, W., Chi, J. T., & Needell, D. (2022). Population-based hierarchical non-negative matrix factorization for survey data. *arXiv*. https://doi.org/10.48550/arXiv.2209.04968

Dominguez Mantes, A., Mas Montserrat, D., Bustamante, C. D., Giró-i-Nieto, X., & Ioannidis, A. G. (2023). Neural ADMIXTURE for rapid genomic clustering. *Nature Computational Science*, 3, 621–629. https://doi.org/10.1038/s43588-023-00482-7

Edwards, A. W. F. (1971). Distances between populations on the basis of gene frequencies. *Biometrics*, 27(4), 873–881.

Elhaik, E. (2022). Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports*, 12, Article 14683. https://doi.org/10.1038/s41598-022-14395-4

Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10(3), 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G. et al. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1), e8. https://doi.org/10.1371/journal.pbio.0050008

Fan, C.-C., Pecchioni, N., & Chen, L.-Q. (2008). Genetic structure and proposed conservation strategy for natural populations of *Calycanthus chinensis* Cheng et S.Y. Chang (Calycanthaceae). *Canadian Journal of Plant Science*, 88(1), 179–186. https://doi.org/10.4141/CJPS07015

Frankham, R., Ballou, J. D., & Briscoe, D. A. (2018). *Conservation genetics*.

Funk, W. C., McKay, J. K., Hohenlohe, P. A., & Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, 27(9), 489–496. https://doi.org/10.1016/j.tree.2012.05.012

Gilmour, A. R., Gogel, B. J., Cullis, B. R., & Thompson, R. (2009). *ASReml user guide release 3.0*. VSN International Ltd.

Gopalan, S., Smith, S. P., Korunes, K., Hamid, I., Ramachandran, S., & Goldberg, A. (2022). Human genetic admixture through the lens of population genomics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1852). https://doi.org/10.1098/rstb.2020.0410

Gouy, A., & Zieger, M. (2025). STRAF 2: New features and improvements of the STR population data analysis software. *Forensic Science International: Genetics*, 76, 103207. https://doi.org/10.1016/j.fsigen.2024.103207

Guillot, G., & Orlando, L. (2013). *Population Structure*.

Guo, X., Qian, Y., Shi, H., Yang, W., & Zhou, N. (2023). Semiparametric efficient estimation of genetic relatedness with machine learning methods. *arXiv preprint arXiv:2304.01849*. https://doi.org/10.48550/arXiv.2304.01849

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., & Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, 19(2), 318–326. https://doi.org/10.1101/gr.081398.108

Hassanpour, A., Geibel, J., Simianer, H., & Pook, T. (2023). Optimization of breeding program design through stochastic simulation with kernel regression. *G3: Genes|Genomes|Genetics*, 13(12). https://doi.org/10.1093/g3journal/jkad217

Haughton, D., Legrand, P., & Woolford, S. (2009). Review of three latent class cluster analysis packages: Latent Gold, poLCA, and mclust. *The American Statistician*, 63, 81–91. http://dx.doi.org/10.1198/tast.2009.0016

Hauser, S. S., Athrey, G., & Leberg, P. L. (2021). Waste not, want not: Microsatellites remain an economical and informative technology for conservation genetics. *Ecology and Evolution*, 11(22), 15800–15814. https://doi.org/10.1002/ece3.8250

Hedgecock, D., Barber, P. H., & Edmands, S. (2007). Genetic approaches to measuring connectivity. *Oceanography*, 20(3), 70–79.
https://tos.org/oceanography/assets/docs/20-3_hedgecock.pdf

Henden, L., Lee, S., Mueller, I., Barry, A., & Bahlo, M. (2018). Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLoS Genetics*, 14(5), e1007279.
https://doi.org/10.1371/journal.pgen.1007279

Herbers, J. M. (2010). Evolution: Fundamentals. In *Encyclopedia of Animal Behavior* (pp. 670–678). Elsevier. https://doi.org/10.1016/B978-0-08-045337-8.00110-8

Hohenlohe, P. A., Funk, W. C., & Rajora, O. P. (2020). Population genomics for wildlife conservation and management. *Molecular Ecology*, 30(1), 62–82. https://doi.org/10.1111/mec.15720

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14), 1801–1806.
https://doi.org/10.1093/bioinformatics/btm233

Jombart, T. (2008). *adegenet*: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, 11.
https://doi.org/10.1186/1471-2156-11-94

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2, e281. https://doi.org/10.7717/peerj.281

Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., & Hooman, A. (2013). An overview of principal component analysis. *Journal of Signal and Information Processing*, 4(3B), 173–175. https://doi.org/10.4236/jsip.2013.43B031

Kasarda, R., Jamborová, Ľ., & Moravčíková, N. (2020). Genetic diversity and production potential of animal food resources. *Acta Fytotechnica et Zootechnica*, 23(2), 102–108. https://doi.org/10.15414/afz.2020.23.02.102-108

Kasarda, R., Moravčíková, N., Mészáros, G., Simčič, M., & Zaborski, D. (2023). Classification of cattle breeds based on the random forest approach. *Livestock Science*, 267, 105143. https://doi.org/10.1016/j.livsci.2022.105143

Korunes, K. L., & Goldberg, A. (2021). Human genetic admixture. *PLOS Genetics*, 17(3), e1009374.
https://doi.org/10.1371/journal.pgen.1009374

Kukučková, V., Kasarda, R., Žitný, J., & Moravčíková, N. (2018). Genetic markers and biostatistical methods as appropriate tools to preserve genetic resources. *AGROFOR International Journal*, 3(2), 41–48. https://doi.org/10.7251/AGRENG1802041K

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, 35(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret Structure and Admixture bar plots. *Nature Communications*, 9, Article 3258.
https://doi.org/10.1038/s41467-018-05257-7

Leaché, A. D., & Oaks, J. R. (2017). The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 48, 69–84.
https://doi.org/10.1146/annurev-ecolsys-110316-022645

Lehocká, K., Kasarda, R., Olšanská, B., & Moravčíková, N. (2020). Assessment of genetic drift and migration in six cattle breeds. *Acta Fytotechnica et Zootechnica*, 23 (Monothematic Issue: Future Perspectives in Animal Production), 46–51.
https://doi.org/10.15414/afz.2020.23.mi-fpap.46-51

Li, Y., Kaur, S., Pembleton, L. W., Valipour-Kahrood, H., Rosewarne, G. M., & Daetwyler, H. D. (2022). Strategies of preserving genetic diversity while maximizing genetic response from implementing genomic selection in pulse breeding programs. *Theoretical and Applied Genetics*, 135(6), 1813–1828. https://doi.org/10.1007/s00122-022-04071-6

Li, Z., Zalucki, M. P., Yonow, T., Kriticos, D. J., Bao, H., Chen, H., Hu, Z., Feng, X., & Furlong, M. J. (2015). Population dynamics and management of diamondback moth (*Plutella xylostella*) in China: The relative contributions of climate, natural enemies, and cropping patterns. *Bulletin of Entomological Research*, 106(2), 197–214. https://doi.org/10.1017/S0007485315000853

Lichtenstein, F., Antoneli, F., & Briones, M. R. S. (2015). MIA: Mutual Information Analyzer, a graphic user interface program that calculates entropy, vertical and horizontal mutual information of molecular sequence sets. *BMC Bioinformatics*, 16, 409. https://doi.org/10.1186/s12859-015-0837-0

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(10), 1–29.
https://www.jstatsoft.org/index.php/jss/article/view/v042i10

Makgahlela, M. L., Strandén, I., Nielsen, U. S., Sillanpää, M. J., & Mäntysaari, E. A. (2014). Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population. *Journal of Dairy Science*, 97(2), 1117–1127. https://doi.org/10.3168/jds.2013-7167

Makrem, A., Najeh, B. F., Laarbi, K. M., & Mohamed, B. (2006). Genetic diversity in Tunisian *Ceratonia siliqua* L. (Caesalpinioideae) natural populations. *Genetic Resources and Crop Evolution*, 53, 1501–1511.
https://doi.org/10.1007/s10722-005-7761-5

Meirmans, P. G., Liu, S., & Van Tienderen, P. H. (2018). The analysis of polyploid genetic data. *Journal of Heredity*, 109(3), 283–296. https://doi.org/10.1093/jhered/esy006

Meuwissen, T. H. E., Sonesson, A. K., Gebregiwergis, G., & Woolliams, J. A. (2020). Management of genetic diversity in the era of genomics. *Frontiers in Genetics*, 11.
https://doi.org/10.3389/fgene.2020.00880

Miller, J. M., Cullingham, C. I., & Peery, R. M. (2020). The influence of a priori grouping on inference of genetic clusters: Simulation study and literature review of the DAPC method. *Heredity*, 125, 269–280.
https://doi.org/10.1038/s41437-020-0348-2

Moorjani, P., & Hellenthal, G. (2023). Methods for assessing population relationships and history using genomic data. *Annual Review of Genomics and Human Genetics*, 24, 305–332.
https://doi.org/10.1146/annurev-genom-111422-025117

Moravčíková, N., & Kasarda, R. (2020). Use of high-density SNP analyses to develop a long-term strategy for conventional populations to prevent loss of diversity – Review. *Acta Fytotechnica et Zootechnica*, 23(4), 236–240. https://doi.org/10.15414/afz.2020.23.04.236-240

Morrison, D. A. (2010). Using data-display networks for exploratory data analysis in phylogenetic studies. *Molecular Biology and Evolution*, 27(5), 1044–1057. https://doi.org/10.1093/molbev/msp309

Moura, A. and Eurico,V. (2010) Investigating the relative inuence of genetic drift and natural selection in shaping patterns of population structure in Delphinids (*Delphinus delphis*; *Tursiops* spp.), Durham theses, Durham University. Available at Durham E-Theses Online:

http://etheses.dur.ac.uk/755/

Nagai, S., Lian, C., Yamaguchi, S., Hamaguchi, M., Matsuyama, Y., Itakura, S., Shimada, H., Kaga, S., Yamauchi, H., Sonda, Y., Nishikawa, T., Kim, C.-H., & Hogetsu, T. (2007). Microsatellite markers reveal population genetic structure of the toxic dinoflagellate *Alexandrium tamarense* (Dinophyceae) in Japanese coastal waters. *Journal of Phycology*, 43(1), 43–54. https://doi.org/10.1111/j.1529-8817.2006.00304.x

Nam, B. E., Nam, J. M., & Kim, J. G. (2016). Effects of habitat differences on the genetic diversity of *Persicaria thunbergii*. *Journal of Ecology and Environment*, 40(11). https://doi.org/10.1186/s41610-016-0012-1

Nei, M. (1972). Genetic distance between populations. *The American Naturalist*, 106(949), 283–292. https://doi.org/10.1086/282771

Nei, M., & Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press.

Neuditschko, M., Khatkar, M. S., & Raadsma, H. W. (2012). NetView: A high-definition network-visualization approach to detect fine-scale population structures from genome-wide patterns of variation. *PLoS ONE*, 7(10), e48375. https://doi.org/10.1371/journal.pone.0048375

Oberski, D. (2016). *Mixture models: Latent profile and latent class analysis* (pp. 275–287). Springer Nature.

Palamara, P. F., Lencz, T., Darvasi, A., & Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, 91(5), 809–822. https://doi.org/10.1016/j.ajhg.2012.08.030

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

Peakall, R., & Smouse, P. E. (2006). GENALEX 6: Genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, 6(1), 288–295. https://doi.org/10.1111/j.1471-8286.2005.01155.x

Pedersen, L. D., Sørensen, C., Henryon, M. M., Mahyari, S. A., & Berg, P. (2009). ADAM: A computer program to simulate selective breeding schemes for animals. *Livestock Science*, 121(2), 343–344. https://doi.org/10.1016/j.livsci.2008.06.028

Pembleton, L. W., Cogan, N. O. I., & Forster, J. W. (2013). StAMPP: An R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular Ecology Resources*, 13(5), 946–952. https://doi.org/10.1111/1755-0998.12129

Peugh, J., & Fan, X. (2013). Modeling unobserved heterogeneity using latent profile analysis: A Monte Carlo simulation. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(4), 616–639. https://doi.org/10.1080/10705511.2013.824780

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

Porta, B., Fernández, P., Galván, G., & Condón, F. (2020). Gametes simulator: A multilocus genotype simulator to analyze genetic structure in outbreeding diploid species. *Crop Breeding and Applied Biotechnology*, 20(1). https://doi.org/10.1590/1984-70332020v20n1s9

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Putman, A. I., & Carbone, I. (2014). Challenges in analysis and interpretation of microsatellite data for population genetic studies. *Ecology and Evolution*, 4(22), 4399–4428. https://doi.org/10.1002/ece3.1305

Qin, X., Lock, T. R., & Kallenbach, R. L. (2022). DA: Population structure inference using discriminant analysis. *Methods in Ecology and Evolution*, 13(2), 485–499. https://doi.org/10.1111/2041-210X.13748

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, 197(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Reynolds, J., Weir, B. S., & Cockerham, C. C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, 105(3), 767–779.

Rogers, J. S. (1972). Measures of genetic similarity and genetic distance. *Studies in Genetics VII, University of Texas Publication 7213* (pp. 145–153).

Rosel, P. E., Hancock-Hanser, B. L., Archer, F. I., Robertson, K. M., Martien, K. K., Leslie, M. S., Berta, A., Cipriano, F., Viricel, A., Viaud-Martinez, K. A., & Taylor, B. L. (2017). Examining metrics and magnitudes of molecular genetic differentiation used to delimit cetacean subspecies based on mitochondrial DNA control region sequences. *Marine Mammal* Science, 33(S1), 76–100. https://doi.org/10.1111/mms.12410

Rosenberg,N.A.(2004).DISTRUCT:aprogramforthegraphical display of population structure. *Molecular Ecology Notes*, 4(1), 137–138. https://doi.org/10.1046/j.1471-8286.2003.00566.x

Rousset, F. (2008). GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, 8(1), 103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Sekino, M., Hara, M. Application of Microsatellite Markers to Population Genetics Studies of Japanese Flounder Paralichthys olivaceus. *Marine Biotechnology*, 3, 572–589. https://doi.org/10.1007/s10126-001-0064-8

Serneels, S., & Verdonck, T. (2008). Principal component analysis for data containing outliers and missing elements. *Computational Statistics & Data Analysis*, 52(3), 1712–1727. https://doi.org/10.1016/j.csda.2007.05.024

Sinha, P., Dincer, A., Virgil, D., Xu, G., Poh, Y.-P., & Jensen, J. D. (2011). On detecting selective sweeps using single genomes. *Frontiers in Genetics*, 2, 85. https://doi.org/10.3389/fgene.2011.00085

Smith, C. C. R., Patterson, G., Ralph, P. L., & Kern, A. D. (2024). Estimation of spatial demographic maps from polymorphism data using a neural network. *bioRxiv*. https://doi.org/10.1101/2024.03.15.585300

Solovieva, E., & Sakai, H. (2023). PSReliP: An integrated pipeline for analysis and visualization of population structure and relatedness based on genome-wide genetic variant data. *BMC Bioinformatics*, 24, 135. https://doi.org/10.1186/s12859-023-05169-4

Steinig, E. J., Neuditschko, M., Khatkar, M. S., Raadsma, H. W., & Zenger, K. R. (2016). NetView P: A network visualization tool to unravel complex population structure using genome-wide SNPs. *Molecular Ecology Resources*, 16(1), 216–227. https://doi.org/10.1111/1755-0998.12442

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., & Lund, M. S. (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS ONE*, 7(9). https://doi.org/10.1371/journal.pone.0045293

Subramanian, S. (2022). The difference in the proportions of deleterious variations within and between populations influences the estimation of F_{ST}. *Genes*, 13(2), 194. https://doi.org/10.3390/genes13020194

Sul, J. H., Martin, L. S., & Eskin, E. (2018). Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genetics*, 14(12). https://doi.org/10.1371/journal.pgen.1007309

Takezaki, N., & Nei, M. (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics*, 144(1), 389–399. https://doi.org/10.1093/genetics/144.1.389

Tang, H., Peng, J., Wang, P., & Risch, N. J. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, 28(4), 289–301. https://doi.org/10.1002/gepi.20064

Thia, J. A. (2023). Guidelines for standardizing the application of discriminant analysis of principal components to genotype data. *Molecular Ecology Resources*, 23(3), 523–538. https://doi.org/10.1111/1755-0998.13706

Thompson, E. A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, 194(2), 301–326. https://doi.org/10.1534/genetics.112.148825

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Veerkamp, R. F., Mulder, H. A., Thompson, R., & Calus, M. P. L. (2011). Genomic and pedigree-based genetic parameters for scarcely recorded traits when some animals are genotyped. *Journal of Dairy Science*, 94(8), 4189–4197. https://doi.org/10.3168/jds.2011-4223

Villanueva, B., Fernández, A., Saura, M., Caballero, A., Fernández, J., Morales-González, E., Toro, M. A., & Pong-Wong, R. (2021). The value of genomic relationship matrices to estimate levels of inbreeding. *Genetics Selection Evolution*, 53, 42. https://doi.org/10.1186/s12711-021-00647-5

Villaverde, A. F., Ross, J., Morán, F., & Banga, J. R. (2014). MIDER: Network inference with mutual information distance and entropy reduction. *PLoS ONE*, 9(5). https://doi.org/10.1371/journal.pone.0096732

vonHoldt, B. M., Pollinger, J. P., Earl, D. A., Knowles, J. C., Boyko, A. R., Parker, H., Geffen, E., Pilot, M., Jedrzejewski, W., Jedrzejewska, B., Sidorovich, V., Greco, C., & Wayne, R. K. (2011). A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Research*, 21(8), 1294–1305. https://doi.org/10.1101/gr.116301.110

Wang, J. (2022). Fast and accurate population admixture inference from genotype data from a few microsatellites to millions of SNPs. *Heredity*, 129, 79–92. https://doi.org/10.1038/s41437-022-00535-z

Wang, J. 2024. PopCluster: A Population Genetics Model-Based Toolset for Simulating, Inferring and Visualising Individual Admixture and Population Structure. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.14058

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6), 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Wellmann, R. Optimum contribution selection for animal breeding and conservation: the R package optiSel. *BMC Bioinformatics,* 20, 25. https://doi.org/10.1186/s12859-018-2450-5

Whitlock, M. C., & Guillaume, F. (2009). Testing for spatially divergent selection: Comparing QST to F_{ST}. *Genetics*, 183(3), 1055–1063. https://doi.org/10.1534/genetics.108.099812

Wright, B., Farquharson, K. A., McLennan, E. A., Belov, K., Hogg, C. J., & Grueber, C. E. (2019). From reference genomes to population genomics: Comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species. *BMC Genomics*, 20, 453. https://doi.org/10.1186/s12864-019-5806-y

Wright, S. (1921). Systems of mating. I. The biometric relations between parent and offspring. *Genetics*, 6(2), 111–123. https://doi.org/10.1093/genetics/6.2.111

Wright, S. (1923). The theory of gene frequencies. *American Naturalist*, 57(649), 289–295. https://doi.org/10.1086/279872

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, 19(3), 395–420. https://doi.org/10.1111/j.1558-5646.1965.tb01731.x

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Zapata-Valenzuela, J., Whetten, R. W., Neale, D., McKeand, S., & Isik, F. (2013). Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. *G3: Genes|Genomes|Genetics*, 3(5), 909–916. https://doi.org/10.1534/g3.113.005975

Zhu, Z. H., Li, H. Y., Qin, Y., & Wang, R. X. (2014). Genetic diversity and population structure in *Harpadon nehereus* based on sequence-related amplified polymorphism markers. Genetics and Molecular Research, 13(3), 5974–5981.

https://doi.org/10.4238/2014.August.7.13