

SUPPLEMENTARY MATERIALS

Text S1 – concepts used in the study

Admixed Population

A population of individuals with ancestors from two or more relatively distinct populations relatively recent in human history.

Admixture mapping

Mapping of susceptibility alleles for a phenotype that shows differential risk by ancestry with the aim of correlating the ancestry around the genomic region associated with the phenotype.

Bayesian clustering

Assignment of individuals to clusters based on genetic similarity without assuming predefined populations, using statistical methods that allow inferences to be drawn from the data and prior information.

Expectation-Maximisation (EM) algorithm

An iterative method to find maximum likelihood estimates (MLE) of parameters in statistical models, altering between an expectation step (creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters) and a maximisation step (computes parameters maximizing the output of the expectation step).

Genetic Relatedness Matrix (GRM)

The GRM represents the genomic similarities among all individuals. Each cell in the matrix measures the genotypic correlation between a pair of individuals (the rows and columns). The GRM can be used with a phenotypic distance matrix to estimate heritability without estimating the phenotypic effect of individual SNPs.

Hidden Markov Model (HMM)

A statistical Markov model is a randomly changing system assumed to consist of future states that only depend on current states, whereby states are unobservable (hidden).

Markov Chain Monte Carlo (MCMC)

A simulation method used in Bayesian calculations, incorporating a class of algorithms that can obtain a sample of the desired distribution by observing several steps of the Markov chain, which is a sequence of a probability of events that depend only on the state of the previous event and has the desired distribution at its equilibrium.

Maximum Likelihood Estimate (MLE)

A method for estimating the parameters of a probability distribution by maximising the likelihood function, where the observed data, under the assumed statistical model, are the most probable.

Maximum optimisation

Maximising a function by choosing input values from a specified set and using this to calculate the value of the function.

Mixed Linear Models (MLM) or Linear Mixed Models (LMMs)

Mixed linear models that incorporate both fixed and random effects where there is non-independence in the data. These models are used as a form of “global correction” as they account for both ancestry

and relatedness.

Multidimensional Scaling (MDS)

A type of multivariate analysis which allows multidimensional information to be displayed graphically (usually two dimensions) with minimum loss of information.

Principal Components Analysis (PCA)

A type of multivariate analysis that reduced the dimensionality of the dataset with minimum loss of information while preserving covariance of data. PCA is typically applied to genotype data as a form of “global correction” and the derived PCs are used for further analyses. PCA represents the reduced dataset as principal components (PCs) that can be displayed on scatter plots.

Table S1: A summary of the advantages and limitations of seven of the most popular global genomic ancestry tools, as discussed in this review.

TOOL	DESCRIPTION	SUMMARY
STRUCTURE (Pritchard, Stephens and Donnelly, 2000) Citations: 24,418 (1285.16)	<i>STRUCTURE</i> uses a Bayesian clustering approach and applies Markov Chain Monte Carlo (MCMC) estimation, which randomly assigns sampled individuals to k groups based on estimations of genetic variant frequencies (Pritchard, Stephens and Donnelly, 2000). In theory, samples belonging to the same cluster can be considered either belonging to the same population or at least sharing a close evolutionary history. <i>STRUCTURE</i> was extended later to accommodate LD, permitting the inclusion of weakly linked markers with some degree of non-independence (Falush, Stephens and Pritchard, 2003).	ADVANTAGES • User friendly (Pritchard, Stephens and Donnelly, 2000). • Applicable to most common genetic markers (Pritchard, Stephens and Donnelly, 2000). LIMITATIONS • Very computationally intensive and slower than FRAPPE, ipPCA, BAPS and LAMP (Tang <i>et al.</i>, 2005; Sankararaman <i>et al.</i>, 2008; Intarapanich <i>et al.</i>, 2009). • Difficulty resolving ancestry clusters containing a small number of individuals (Intarapanich <i>et al.</i>, 2009). • Interval and standard error estimation are costly (Pritchard, Stephens and Donnelly, 2000). • Less accurate than LAMP, ipPCA and BAPS (Sankararaman <i>et al.</i>, 2008; Intarapanich <i>et al.</i>, 2009). • High risk of misleading results when used in the context of introduction route inference (Lombaert, Guillemaud and Deleury, 2018). ASSUMPTIONS • Models LD (Pritchard, Stephens and Donnelly, 2000). • Hardy–Weinberg equilibrium (HWE) (Pritchard, Stephens and Donnelly, 2000). • Fixed number of populations (Pritchard, Stephens and Donnelly, 2000). • Marker loci are unlinked and at linkage equilibrium (Pritchard, Stephens and Donnelly, 2000). RUNNING TIME Analysing 307 individuals from 14 geographic locations using a 2.3-GHz processor took more than 3 days in <i>STRUCTURE</i> (Intarapanich <i>et al.</i>, 2009). Analysing 500 individuals from 2 ancestral populations using a 2-GHz processor took 5 hours in <i>STRUCTURE</i> versus 1 minute in <i>FRAPPE</i> (Tang <i>et al.</i>, 2005).
ADMIXTURE (Alexander, Novembre)	<i>ADMIXTURE</i> is based on a similar approach to <i>STRUCTURE</i>. Still, like <i>FRAPPE</i>, it uses a maximum likelihood estimation approach to	ADVANTAGES • Block relaxation algorithm that allows for fast convergence and highly accurate

and Lange, 2009) Citations: 2,111 (211.10)	optimise the likelihood for allele frequencies and group memberships, using slightly different algorithms. By default, <i>ADMIXTURE</i> uses a block relaxation algorithm that allows for fast convergence and highly accurate parameter estimates (Alexander, Novembre and Lange, 2009) and has an optional Expectation-Maximisation (EM) algorithm.	parameter estimations (Alexander, Novembre and Lange, 2009). • Faster and more accurate than <i>STRUCTURE</i> and <i>FRAPPE</i>, which allows analyzing thousands of individuals across thousands of genetic markers (Alexander, Novembre and Lange, 2009). LIMITATIONS • Bootstrap resampling to compute standard errors takes a long time (Alexander, Novembre and Lange, 2009). • Difficulty resolving ancestry clusters containing a small number of individuals (Alexander, Novembre and Lange, 2009). • Parameters of the <i>ADMIXTURE</i> model must satisfy linear constraints and bounds (Alexander, Novembre and Lange, 2009). • Limited when using samples of ancient DNA due to violation of HWE assumption (Joseph and Pe'er, 2018). ASSUMPTIONS • HWE (Alexander, Novembre and Lange, 2009). • Linkage equilibrium among markers (Alexander, Novembre and Lange, 2009). RUNNING TIME For 10 simulated datasets with 10,000 markers, the supervised analysis took an average of 5.15 seconds, while unsupervised analysis averaged 27.5 seconds (Alexander and Lange, 2011).
FRAPPE (Tang <i>et al.</i>, 2005) Citations: 476 (34.00)	<i>FRAPPE</i> is based on a similar approach to <i>STRUCTURE</i>, but like <i>ADMIXTURE</i>, it uses a maximum likelihood estimation approach to optimise the likelihood for allele frequencies and group memberships, using slightly different algorithms. <i>FRAPPE</i> uses solely the EM algorithm, which is used to optimise the likelihood for both allele frequencies and fractional group memberships (Tang <i>et al.</i>, 2005).	ADVANTAGES • More robust than <i>ADMIXTURE</i> (faster and more efficient (Yuan <i>et al.</i>, 2017) • Faster than <i>STRUCTURE</i> (Tang <i>et al.</i>, 2005). • Does not require phased ancestral data and is therefore not prone to phasing errors (Tang <i>et al.</i>, 2005). LIMITATIONS • The strict convergence criteria of its EM algorithm makes it computationally intensive and slower than <i>ADMIXTURE</i> (Alexander, Novembre and Lange, 2009). • Loosening the convergence criteria of the EM algorithm allows faster termination of the EM algorithm but jeopardises the accuracy of estimates (Alexander, Novembre and Lange, 2009). ASSUMPTIONS • Linkage equilibrium among markers (Tang <i>et al.</i>, 2005).

		RUNNING TIME Analysing 500 individuals from 2 ancestral populations using a 2-GHz processor took 1 minute in <i>FRAPPE</i> (Tang <i>et al.</i>, 2005).
GENELAND (Guillot, Mortier and Estoup, 2005) Citations: 984 (70.29)	Spatial approaches, exemplified by <i>GENELAND</i> (Guillot, Mortier and Estoup, 2005), <i>TESS</i> (Durand <i>et al.</i>, 2009) and <i>BAPS</i> (Corander, Waldmann and Sillanpää, 2003) are conceptually similar to <i>STRUCTURE</i>, but consider geographical coordinates in their prior distributions, allowing identification of the spatial location of genetic variants between populations. These software do not only group individuals genetically into clusters but can also estimate the spatial distribution of these clusters (Corander, Waldmann and Sillanpää, 2003; Guillot, Mortier and Estoup, 2005; Durand <i>et al.</i>, 2009). <i>GENELAND</i> makes use of georeferenced individual multilocus genotypes for the inference of the number of populations and of the spatial location of genetic discontinuities between those populations. <i>GENELAND</i> can be used to detect population structure in the form of systematic variation of allele frequency, which can be detected from departure from Hardy-Weinberg and linkage equilibrium. <i>GENELAND</i> requires individual multilocus genetic data that are optionally georeferenced, and implements several models that can make use of both geographic and genetic information to estimate the number of populations in a dataset and delineate their spatial organisation (Guillot, Mortier and Estoup, 2005).	ADVANTAGES • Can identify the spatial location of genetic variants between populations (Guillot, Mortier and Estoup, 2005). • User friendly (Guillot, Mortier and Estoup, 2005). • Flexible to extend (Guillot, Mortier and Estoup, 2005). LIMITATIONS • Outperformed by <i>TESS</i> and <i>BAPS</i> in terms of accuracy and misassignment rates (Latch <i>et al.</i>, 2006; Chen <i>et al.</i>, 2007). • Does not handle admixture (Guillot, Mortier and Estoup, 2005). • Computationally intensive (Guillot, Mortier and Estoup, 2005). ASSUMPTIONS • HWE (Guillot, Mortier and Estoup, 2005). • Linkage equilibrium among markers (Guillot, Mortier and Estoup, 2005). • Known allele frequencies (Guillot, Mortier and Estoup, 2005). RUNNING TIME <i>GENELAND</i> is computationally intensive with large datasets (Guillot, Mortier and Estoup, 2005), taking roughly 20 minutes to analyse 500 individuals with 20 loci using a 2-GHz processor (Corander <i>et al.</i>, 2003).
BAPS (Corander, Waldmann and Sillanpää, 2003) Citations: 582 (44.77)	Spatial approaches, exemplified by <i>GENELAND</i> (Guillot, Mortier and Estoup, 2005), <i>TESS</i> (Durand <i>et al.</i>, 2009) and <i>BAPS</i> (Corander, Waldmann and Sillanpää, 2003) are conceptually similar to <i>STRUCTURE</i>, but consider geographical coordinates in their prior distributions, allowing identification of the spatial location of genetic variants between populations. These software do not only group individuals genetically into clusters but are also able to estimate the spatial distribution of these clusters (Corander, Waldmann and Sillanpää, 2003; Guillot, Mortier and Estoup, 2005; Durand <i>et al.</i>, 2009).	ADVANTAGES • Can identify the spatial location of genetic variants between populations (Corander, Waldmann and Sillanpää, 2003). • Outperforms <i>GENELAND</i> in terms of accuracy and misassignment rates (Latch <i>et al.</i>, 2006; Chen <i>et al.</i>, 2007). LIMITATIONS • Memory intensive (Corander, Waldmann and Sillanpää, 2003). • Slower and less accurate than ipPCA (Intarapanich <i>et al.</i>, 2009). • Lower accuracy than <i>GENELAND</i> at low levels of genetic differentiation (Latch <i>et al.</i>,

	<i>BAPS</i> combines a discrete parameterization of admixture proportions in genomes with a simulation framework that yields a clear biological interpretation of the estimation results and can be used to assess the statistical significance of putative admixture events. The number of genetically differentiated sources contributing to a dataset is inferred first using a mixture model. Thereafter, given such an estimate, admixture events can be learned on a more stable basis using a Monte Carlo simulation-based algorithm (Corander, Waldmann and Sillanpää, 2003). <i>BAPS</i> treats both allele frequencies of the molecular markers (or nucleotide frequencies for DNA sequence data) and the number of genetically diverged groups in the population as random variables. However, analyses and model comparisons can also be performed using a fixed number of genetically diverged groups or pre-specified population structures (Corander, Waldmann and Sillanpää, 2003).	2006). ASSUMPTIONS • HWE (Corander, Waldmann and Sillanpää, 2003). • Linkage equilibrium among markers (Corander, Waldmann and Sillanpää, 2003). RUNNING TIME The authors of <i>BAPS</i> note that their software took just 30 minutes to compute an entire sequential admixture estimation for complete human data (Corander, Waldmann and Sillanpää, 2003).
TESS (Durand <i>et al.</i>, 2009) Citations: 292 (22.46)	Spatial approaches, exemplified by <i>GENELAND</i> (Guillot, Mortier and Estoup, 2005), <i>TESS</i> (Durand <i>et al.</i>, 2009) and <i>BAPS</i> (Corander, Waldmann and Sillanpää, 2003) are conceptually similar to <i>STRUCTURE</i>, but consider geographical coordinates in their prior distributions, allowing identification of the spatial location of genetic variants between populations. These software do not only group individuals genetically into clusters but also estimate the spatial distribution of these clusters (Corander, Waldmann and Sillanpää, 2003; Guillot, Mortier and Estoup, 2005; Durand <i>et al.</i>, 2009). <i>TESS</i> implements ancestry estimation algorithms for spatial population genetic analyses. The program performs individual geographical assignment and admixture analysis and can run genome scans for selection. <i>TESS</i> is particularly suited to seeking genetic discontinuities in continuous populations and estimating spatially varying individual admixture proportions. <i>TESS</i> returns graphical displays of geographical cluster assignments or admixture	ADVANTAGES • Can identify the spatial location of genetic variants between populations (Durand <i>et al.</i>, 2009). • Outperforms <i>GENELAND</i> and <i>BAPS</i> in terms of accuracy and misassignment rates (Latch <i>et al.</i>, 2006; Chen <i>et al.</i>, 2007). • Provides admixture and none admixture models (Durand <i>et al.</i>, 2009). LIMITATIONS • Highest misassignment rates compared with <i>BAPS</i> and <i>GENELAND</i> (Basto <i>et al.</i>, 2016). ASSUMPTIONS • HWE (Durand <i>et al.</i>, 2009). • Parental populations are assumed to be in migration/drift equilibrium (admixture model) (Durand <i>et al.</i>, 2009).

	proportions (depending on the model used) and textual output of the admixture Q matrix (Durand <i>et al.</i> , 2009).	
EIGENSTRAT (Price <i>et al.</i> , 2006) Citations: 6,859 (527.62)	The PCA of <i>EIGENSTRAT</i> aims to reduce the dimensionality of the SNP dataset by reducing the genetic markers into principal components (PCs), the first two of which are typically plotted on a scatter plot. The output is used to identify outliers, assign individuals into populations, match cases and controls and more. PCA is one of the most commonly used tools in population genetics due to its high speed, relatively low computational demands, ease of use and appealing graphical results (Patterson <i>et al.</i>, 2004). <i>EIGENSTRAT</i> consists of three steps. First, PCA is applied to genotype data to infer continuous axes of genetic variation, which reduces the data to a small number of dimensions, describing as much variability as possible. In datasets with ancestry differences between samples, axes of variation often have a geographic interpretation. Second, genotypes and phenotypes are continuously adjusted by amounts attributable to ancestry along each axis via computing residuals of linear regressions, creating a virtual set of matched cases and controls. Finally, association statistics are computed using ancestry-adjusted genotypes and phenotypes (Price <i>et al.</i>, 2006).	ADVANTAGES • Faster than <i>STRUCTURE</i> (Price <i>et al.</i>, 2006). • Memory efficient and applicable to large genomic scale (Price <i>et al.</i>, 2006). • Output can be used to produce intuitive plots (Price <i>et al.</i>, 2006). LIMITATIONS • Difficulty resolving ancestry clusters containing a small number of individuals (Price <i>et al.</i>, 2006). • Ineffective when analysing closely related sub-populations (Intarapanich <i>et al.</i>, 2009). • Cannot account for admixture (Elhaik and Ryan, 2019) • Uneven sample size can bias results (McVean, 2009). • Most importantly, PCA results may not be reliable, robust, or replicable, as is commonly assumed. PCA should thereby not be used for population genomic analyses. (Elhaik 2021). ASSUMPTIONS NA. RUNNING TIME Less than 15 minutes to analyse a dataset with 1,000 samples and 100,000 makers (Price <i>et al.</i>, 2006).

Table S2: A summary of the advantages and limitations of the five of the most popular selected local ancestry inference tools.

TOOL	DESCRIPTION	SUMMARY
HAPMIX (Price <i>et al.</i>, 2009). Citations (normalised): 414 (41.40)	<i>HAPMIX</i> can be used to accurately infer chromosomal segments of distinct ancestry in admixed populations using dense genetic data. <i>HAPMIX</i> extends the model of Li and Stephens, estimating the likelihood of a haplotype (alleles at multiple SNPs) belonging to each reference population. The analysis of haplotypes accounts for LD and allows differentiation of very closely related populations (Price <i>et al.</i>, 2009). At each position in the genome, <i>HAPMIX</i> estimates the likelihood that a haplotype from an admixed individual is a better statistical match to one reference population or the other. A Hidden Markov Model (HMM) is used to combine these likelihoods with information from neighboring loci to provide a probabilistic estimate of ancestry at each locus (Price <i>et al.</i>, 2009).	ADVANTAGES • Does not require phasing of data (Price <i>et al.</i>, 2009). • Computationally tractable, even with large datasets (provided a cluster of computing nodes available) (Price <i>et al.</i>, 2009). • Works effectively with unphased data and allows for some miscopying of ancestry segments from the wrong population (Price <i>et al.</i>, 2009). LIMITATIONS • Slower than <i>LAMP-LD</i> (Price <i>et al.</i>, 2009). • Extremely computationally intensive. • Restricted to two reference populations, limiting accuracy and usefulness (Pugach <i>et al.</i>, 2011). • Requires specification of several biological parameters that may be difficult to obtain (Dias-Alves, Mairal and Blum, 2018). • Reduced accuracy when using haplotypes that have been computationally phased (Dias-Alves, Mairal and Blum, 2018). ASSUMPTIONS • Models LD (Price <i>et al.</i>, 2009). • Admixture tract lengths are exponentially distributed and independent (Price <i>et al.</i>, 2009). • Limited to two populations (Price <i>et al.</i>, 2009). RUNNING TIME For a dataset of roughly 1,000 admixed individuals on a cluster of 100 nodes, the running time is approximately five hours (Price <i>et al.</i>, 2009).
LAMP (Sankararaman <i>et al.</i>, 2008). Citations (normalised): 271 (24.64)	<i>LAMP</i> is a software package for the inference of locus-specific ancestry in recently admixed populations. <i>LAMP</i> uses a sliding-windows framework to calculate an optimal window length of contiguous SNPs and then uses a clustering algorithm to estimate ancestry from large reference haplotype data (Sankararaman <i>et al.</i>, 2008). <i>LAMP</i> firstly calculates an optimal window length, on which a clustering algorithm operates and estimates each individual's ancestry. A majority vote is	ADVANTAGES • Faster and more accurate than <i>SABER</i> and <i>STRUCTURE</i> (Sankararaman <i>et al.</i>, 2008). • SLA algorithm for optimal window size (Sankararaman <i>et al.</i>, 2008). • Robust to phasing errors (Sankararaman <i>et al.</i>, 2008). LIMITATIONS • Does not provide information about LD in ancestral populations (Sankararaman <i>et al.</i>, 2008). • Admixture of two populations only (<i>LAMP-ANC</i> >2) (Sankararaman <i>et al.</i>, 2008).

	then used for each SNP, over all the windows that overlap with the SNP, to decide the most likely ancestral populations at the SNP (Sankararaman <i>et al.</i>, 2008).	ASSUMPTIONS • Absence of recombination between windows (Sankararaman <i>et al.</i>, 2008). • Recombination rates across the genome are known (Sankararaman <i>et al.</i>, 2008). RUNNING TIME <i>LAMP</i> analysed 38,864 SNPs in 394 seconds (200 times faster than SABER and 10^4 times faster than STRUCTURE) ((Sankararaman <i>et al.</i>, 2008).
LAMP-LD (Baran <i>et al.</i>, 2012) Citations (normalised): 147 (21.00)	<i>LAMP-LD</i> is an extension of the model implemented in the original <i>LAMP</i> tool (Sankararaman <i>et al.</i>, 2008). <i>LAMP-LD</i> integrates a two-layer HMM into the <i>LAMP</i> window-based framework to leverage haplotype structure, in turn accounting for LD. <i>LAMP-LD</i> divides the genome into non-overlapping windows such that no transitions between ancestries are made within each window. Limiting the occurrence of ancestry transitions in this window-based framework improves the inference quality by eliminating extremely short, likely artifactual, ancestral segments (Baran <i>et al.</i>, 2012).	ADVANTAGES • Robust to phasing errors (Baran <i>et al.</i>, 2012). • Faster than HAPMIX (Baran <i>et al.</i>, 2012). • Works effectively with three-way admixture (Baran <i>et al.</i>, 2012). • Able to more efficiently handle large datasets than LAMP (Baran <i>et al.</i>, 2012). LIMITATIONS • Limited memory allocation and computationally intensive (Dias-Alves, Mairal and Blum, 2018). • Significant reductions in accuracy when analysing sequencing data rather than genotyping data (Brown and Pasaniuc, 2014). ASSUMPTIONS • Models LD (Baran <i>et al.</i>, 2012). • Constant recombination rate (Baran <i>et al.</i>, 2012). • No transitions between ancestries are made between windows (Baran <i>et al.</i>, 2012). RUNNING TIME <i>LAMP-LD</i> is faster than <i>HAPMIX</i>, analysing 200 admixed genotypes in roughly three days compared to <i>HAPMIX</i>'s 22 days (1.1-GHz processor) (Baran <i>et al.</i>, 2012).
ChromoPainter (Lawson <i>et al.</i>, 2012) Citations (normalised): 424 (60.57)	Chromosome painting is a related method that does not attempt to identify ancestry tracts but instead identifying the individual who is the most recent common ancestor for every individual at every location in the genome. It "paints" each individuals' genome as a combination of all others to produce a co-ancestry matrix, analogous to the GRM but counting recently shared ancestors instead of SNPs. <i>ChromoPainter</i> is part of the <i>fineSTRUCTURE</i> pipeline (Lawson <i>et al.</i>, 2012), which allows the identification of	ADVANTAGES • Visualisation of LAI estimates (Lawson <i>et al.</i>, 2012). • Can analyse more than two populations (Lawson <i>et al.</i>, 2012). LIMITATIONS • Less effective when the admixture is strong due to admixture tract assumption (Lawson <i>et al.</i>, 2012). • Susceptible to phasing errors (Lawson <i>et al.</i>, 2012). ASSUMPTIONS • Models LD (Lawson <i>et al.</i>, 2012).

	fine-scale population structure that cannot be identified by PCA or related approaches (Leslie <i>et al.</i>, 2015). <i>ChromoPainter</i> achieves this by utilising the tree relationship between haplotypes for different genes and considering the time to the most recent common ancestor, which changes along the genome due to recombination resulting in different trees for different genes. It then looks for the haplotypes in the genome that are closest to a haplotype in question, which is then used to paint said haplotype as a combination of all other haplotypes and to produce a co-ancestry matrix. This co-ancestry matrix can then be used for inference in <i>fineStructure</i>, which can identify the clusters that differ and assess the statistical differences between them, therefore identifying the potentially different populations (Lawson <i>et al.</i>, 2012).	• Admixture tracts are independent and exponentially distributed (Lawson <i>et al.</i>, 2012). • Computationally demanding (Lawson <i>et al.</i>, 2012). RUNNING TIME The <i>ChromoPainter</i> and <i>fineSTRUCTURE</i> pipeline analysed 500 individuals with 3,000-15,000 single-nucleotide variants in roughly 180-500 minutes, whilst taking 25,000-55,000 minutes to analyse 5,000 individuals with 3,000-15,000 SNVs (desktop computer with 32 GB RAM) (Pan <i>et al.</i>, 2017).
--	---	---

Table S3: A summary of the advantages and limitations of the six most popular of the selected relatedness inference tools, as discussed in this review.

TOOL	DESCRIPTION	SUMMARY
KING (Manichaikul <i>et al.</i>, 2010). Citations (normalised): 607 (67.44)	<i>KING</i> uses moment relatedness estimators to calculate the kinship coefficient, which uses the statistical method of moments to estimate the realised k coefficients; the proportion of genome at which two individuals share 0, 1, or 2 IBD genes (Wang, Sverdlov and Thompson, 2016). <i>KING</i> is a framework for relationship inference combined with a fast algorithm for relationship inference appropriate for use on samples with thousands of individuals genotyped at millions of SNPs from autosomes, consistent with a scale typically achieved in a GWAS. <i>KING</i> includes two different methods: (1) <i>KING</i>-homo, derived under the assumption of population homogeneity and (2) <i>KING</i>-robust, which provides robust relationship inference in the presence of population substructure (Manichaikul <i>et al.</i>, 2010).	ADVANTAGES • Produces reliable inference even on millions of unrelated pairs of individuals and thousands of relative pairs (Manichaikul <i>et al.</i>, 2010). • Performs well under extreme population stratification (Manichaikul <i>et al.</i>, 2010). • <i>KING</i> can produce reliable inference for large sample sizes (millions of unrelated and thousands of relative pairs) and is also significantly faster than <i>PLINK</i> (Manichaikul <i>et al.</i>, 2010). LIMITATIONS • Prone to biased estimates in admixed populations and the presence of population structure (Conomos <i>et al.</i>, 2016). • Cannot separate unrelated individuals from distantly related ones (Moltke and Albrechtsen, 2014). • Relatively high accuracy for first through third-degree classification; however, their accuracy decreased substantially to below 50% for fourth through seventh and unrelated classification (Ramstetter <i>et al.</i>, 2017). ASSUMPTIONS • HWE among SNPs with the same underlying allele frequencies (Manichaikul <i>et al.</i>, 2010). RUNNING TIME <i>KING</i> was able to analyse a dataset of 180,901 individuals in one minute (3.20-GHz processor) (Manichaikul <i>et al.</i>, 2010).
REAP (Thornton <i>et al.</i>, 2012). Citations (normalised): 135 (19.29)	<i>REAP</i> uses moment relatedness estimators to calculate the kinship coefficient, which uses the statistical method of moments to estimate the realised k coefficients; the proportion of genome at which two individuals share 0, 1, or 2 IBD genes (Wang, Sverdlov and Thompson, 2016). <i>REAP</i> appropriately accounts for population structure and ancestry-related assortative mating by using individual-specific allele	ADVANTAGES • Can work well with admixed samples (Thornton <i>et al.</i>, 2012). LIMITATIONS • Cannot separate unrelated individuals from distantly related ones (Moltke and Albrechtsen, 2014). • Relatively high accuracy for first through third-degree classification; however, their accuracy decreased substantially to below 50% for fourth through seventh and unrelated classification (Ramstetter <i>et al.</i>, 2017).

	frequencies at SNPs that are calculated based on ancestry derived from whole-genome analysis (Thornton <i>et al.</i>, 2012).	Bias in the individual-specific allele frequencies used to account for the different ancestries of admixed individuals can significantly impact relatedness estimates (Conomos <i>et al.</i>, 2016). ASSUMPTIONS NA. RUNNING TIME REAP took approximately 1 hr and 31 min to analyse 2,560 individuals and 100,000 SNPs. The REAP analysis of the HapMap MXL sample with 86 individuals and 150,872 SNPs took around 20 seconds (Intel Xeon quad-core 2.66 GHz processors with 16 GB of random-access memory) (Thornton <i>et al.</i>, 2012).
PC-relate (Conomos <i>et al.</i>, 2016). Citations (normalised): 79 (26.33)	<i>PC-Relate</i> uses moment relatedness estimators to calculate the kinship coefficient, which uses the statistical method of moments to estimate the realised k coefficients; the proportion of genome at which two individuals share 0, 1, or 2 IBD genes (Wang, Sverdlov and Thompson, 2016). <i>PC-Relate</i> uses principal components calculated to separate genetic correlations among sampled individuals into two separate components: a component for the allele sharing as a result of inherited IBD from recent common ancestors, which represents familial relatedness, and another component for allele sharing due to more distant common ancestry, which represents population structure (Conomos <i>et al.</i>, 2016).	ADVANTAGES - Can work well with admixed samples (Conomos <i>et al.</i>, 2016). - More accurate than REAP and RelateAdmix (Conomos <i>et al.</i>, 2016). - Able to separate unrelated individuals from more distantly related ones (Moltke and Albrechtsen, 2014). LIMITATIONS - Bias in the individual-specific allele frequencies used to account for the different ancestries of admixed individuals can lead to significantly biased relatedness estimates (Conomos <i>et al.</i>, 2016). - Likely suffers from the same limitation of PCA (Elhaik 2021). ASSUMPTIONS NA. RUNNING TIME <i>PC-Relate</i> analysed all 3,587 individuals in the WHI-SHARE Hispanic cohort with 87,180 SNPs in 12.1 minutes (Conomos <i>et al.</i>, 2016).
PLINK (Purcell <i>et al.</i>, 2007). Citations (normalised): 16,675 (1389.58)	Methods for IBD detection identify similarities between haplotypes that are statistically unlikely to occur in the absence of IBD sharing (Durand, Eriksson and Mclean, 2014). <i>PLINK</i> incorporates a method of moments approach using an HMM to infer underlying IBD in chromosomal segments based on observed IBS states.	ADVANTAGES - More accurate and has shown lower rates of false-positive results compared with <i>GERMLINE</i> (Purcell <i>et al.</i>, 2007). LIMITATIONS - Slower than <i>GERMLINE</i> (Gusev <i>et al.</i>, 2009). - Has shown higher rates of false-positive results than <i>GERMLINE</i> (Gusev <i>et al.</i>, 2009). - Produces high levels of false positives for second-degree relationships (Stevens <i>et al.</i>, 2011).

		ASSUMPTIONS Homogeneous, random-mating population (Purcell <i>et al.</i>, 2007). RUNNING TIME PLINK was able to analyse 2,000 individuals in 556 hours (2x2.4-GHz processors) (Gusev <i>et al.</i>, 2009).
fastIBD (Browning and Browning, 2011). Citations (normalised): 270 (33.75)	Methods for IBD detection identify similarities between haplotypes that are statistically unlikely to occur in the absence of IBD sharing (Durand, Eriksson and Mclean, 2014). <i>fastIBD</i> detects “seeds” of identical short haplotype matches and extends them to nearby sites. The <i>fastIBD</i> algorithm samples a fixed number of haplotype pairs for each individual from the posterior haplotype distribution. Each sampled haplotype corresponds to a sequence of hidden Markov model (HMM) states. <i>fastIBD</i> searches for pairs of sampled haplotypes sharing the same sequence of HMM states for a set of consecutive markers. For each pair of individuals, overlapping shared haplotype tracts are merged and a score is calculated for each merged tract. If this score is below a user-specified threshold, the tract is printed to an output file (Gusev <i>et al.</i>, 2009).	ADVANTAGES - Can be applied to large sample sizes across genome-wide SNP data (Browning and Browning, 2011). - High power and low false-discovery rate compared to PLINK and GERMLINE (Browning and Browning, 2011). LIMITATIONS - Slower and less accurate than GERMLINE and PLINK (Gusev <i>et al.</i>, 2009; Browning and Browning, 2011). - Requires phased data which is prone to phasing errors and increases computation time (Browning and Browning, 2011). - Unfeasible to analyse over 100,000 individuals due to the phasing requirement which limits the number of individuals that can be phased at one time (Browning and Browning, 2011). ASSUMPTIONS NA. RUNNING TIME To analyse 4806 individuals and 37,645 SNPs, for which IBD was estimated for all possible pairs of individuals across the chromosome, the running time for <i>fastIBD</i> was approximately 17 hours (single core of an Intel Xeon E5620 Quad-Core compute node running at 2.40GHz) (Browning and Browning, 2011).

GERMLINE (Gusev <i>et al.</i>, 2009). Citations (normalised): 331 (33.10)	Methods for IBD detection identify similarities between haplotypes that are statistically unlikely to occur in the absence of IBD sharing (Durand, Eriksson and Mclean, 2014). <i>GERMLINE</i> detects “seeds” of identical short haplotype matches and extends them to nearby sites. <i>GERMLINE</i> firstly detects completely identical match-seeds of potentially shared segments by creating a dictionary of allele combination words across the population observed at different slices along the genome. <i>GERMLINE</i> then uses a dynamic programming algorithm along different slices to extend these candidate matches and resolve the likelihood of IBD (Gusev <i>et al.</i>, 2009).	ADVANTAGES • Faster than <i>fastIBD</i> and <i>PLINK</i> (Gusev <i>et al.</i>, 2009; Browning and Browning, 2011). • More accurate than <i>PLINK</i> (Gusev <i>et al.</i>, 2009). • Has shown lower rates of false-positive results than <i>PLINK</i> (Gusev <i>et al.</i>, 2009). • Does not require phased data (Gusev <i>et al.</i>, 2009). LIMITATIONS • Has shown higher rates of false-positive results than <i>PLINK</i> (Browning and Browning, 2013). ASSUMPTIONS NA. RUNNING TIME <i>GERMLINE</i> was able to analyse 2,000 individuals in 40.3 hours compared to <i>PLINK</i>’s 556 hours (2x2.4-GHz processors) (Browning and Browning, 2013).
---	---	---

Figure S1: The popularity (normalised number of citations) of the different GAI software, separated into model-based (red) and non-parametric (blue) tools. The tools were found using existing review papers and free search, using the search engines 'Google', 'Google Scholar' and the journal 'Bioinformatics' to search for keywords including: 'software', 'tools', 'inference', 'biogeographic', 'ancestry', 'kinship', and 'haplotype'. The number of citations for the paper proposing the tool, taken from 'Google Scholar', were compared for each tool within each domain. To account for the differences in the number of years since publication, the number of citations was normalised by dividing the number of citations by the number of years since publication.

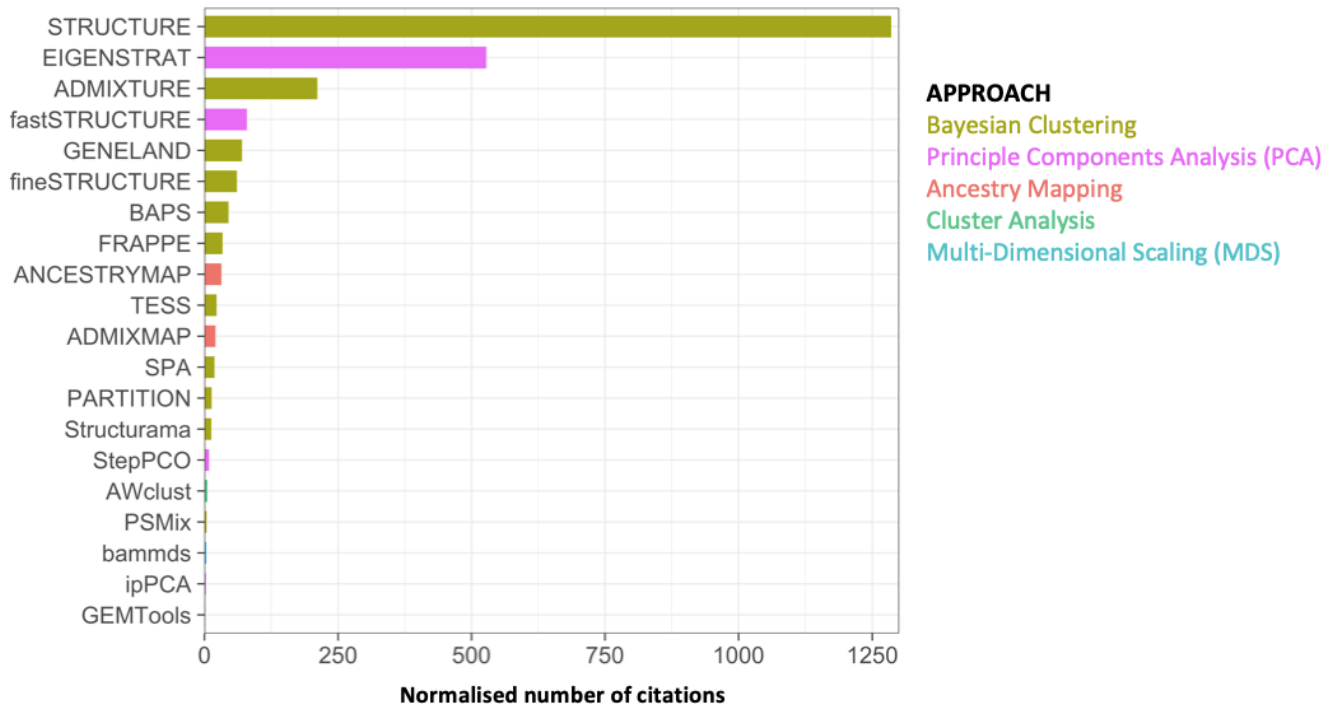


Figure S2: The popularity (normalised number of citations) of the different LAI software, separated into their technologies: Hidden Markov Model (HMM) (green), Chromosome Painting (red) and Statistical Learning Algorithm (SLA) (blue). Finding the tools and calculating the normalised citation number was done as in Figure S1.

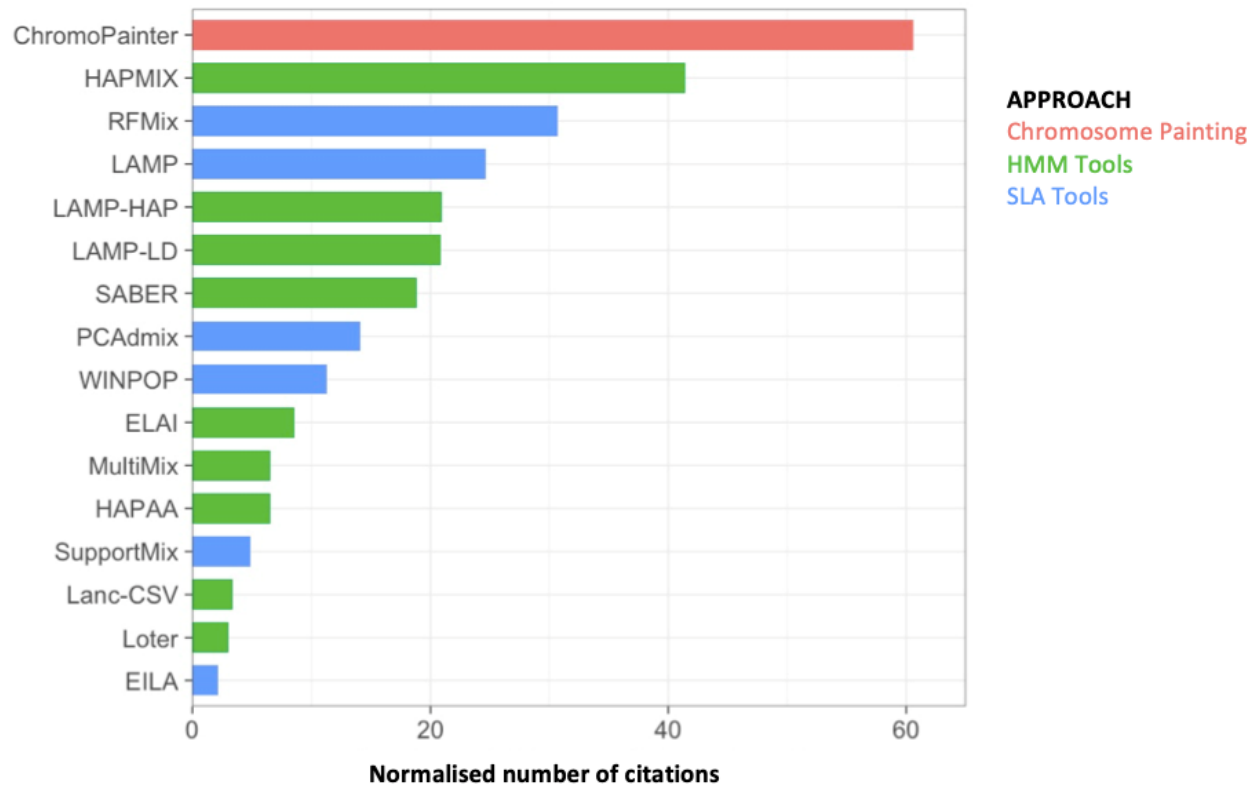
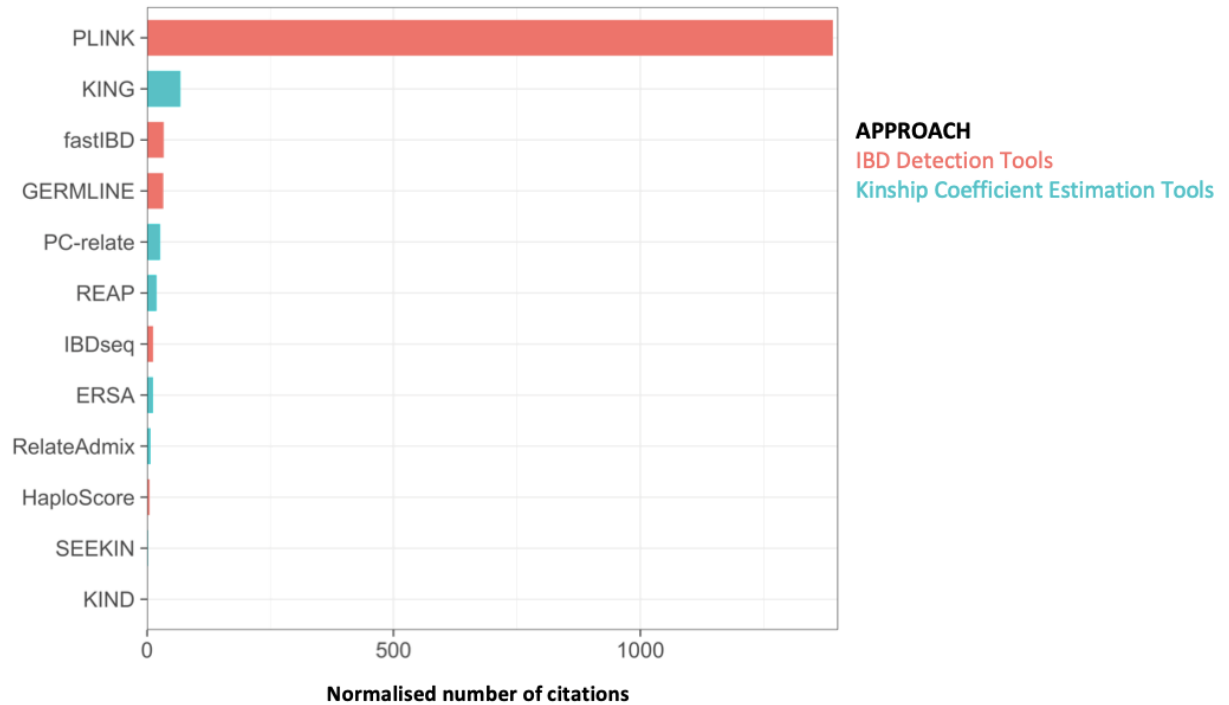


Figure S3: The popularity (normalised number of citations) of the different kinship inference software, separated into their software strategies: Identity-By-Descent (IBD) detection (red) and Kinship Coefficient Estimation (blue). Finding the tools and calculating the normalised citation number was done as in Figure S1.



REFERENCES

Alexander, D. H. and Lange, K. (2011) 'Enhancements to the ADMIXTURE algorithm for individual ancestry estimation', *BMC Bioinformatics*, 12. doi: 10.1186/1471-2105-12-246.

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

Baran, Y. *et al.* (2012) 'Fast and accurate inference of local ancestry in Latino populations', *Bioinformatics*, 28(10), pp. 1359–1367. doi: 10.1093/bioinformatics/bts144.

Basto, M. P. *et al.* (2016) 'Assessing genetic structure in common but ecologically distinct carnivores: The stone marten and red fox', *PLoS ONE*, 11(1). doi: 10.1371/journal.pone.0145165.

Brown, R. and Pasaniuc, B. (2014) 'Enhanced Methods for Local Ancestry Assignment in Sequenced Admixed Individuals', *PLoS Computational Biology*, 10(4). doi: 10.1371/journal.pcbi.1003555.

Browning, B. L. and Browning, S. R. (2011) 'A fast, powerful method for detecting identity by descent', *American Journal of Human Genetics*, 88(2), pp. 173–182. doi: 10.1016/j.ajhg.2011.01.010.

Browning, B. L. and Browning, S. R. (2013) 'Detecting identity by descent and estimating genotype error rates in sequence data', *American Journal of Human Genetics*, 93(5), pp. 840–851. doi: 10.1016/j.ajhg.2013.09.014.

Chen, C. *et al.* (2007) 'Bayesian clustering algorithms ascertaining spatial population structure: A new computer program and a comparison study', *Molecular Ecology Notes*, 7(5). doi: 10.1111/j.1471-8286.2007.01769.x.

Conomos, M. P. *et al.* (2016) 'Model-free Estimation of Recent Genetic Relatedness', *American Journal of Human Genetics*, 98(1), pp. 127–148. doi: 10.1016/j.ajhg.2015.11.022.

Corander, J., Waldmann, P. and Sillanpää, M. J. (2003) 'Bayesian analysis of genetic differentiation between populations.', *Genetics*, 163(1), pp. 367–374.

Dias-Alves, T., Mairal, J. and Blum, M. G. B. (2018) 'Loter: A software package to infer local ancestry for a wide range of species', *Molecular Biology and Evolution*, 35(9), pp. 2318–2326. doi: 10.1093/molbev/msy126.

Durand, E. *et al.* (2009) 'Spatial inference of admixture proportions and secondary contact zones', *Molecular Biology and Evolution*, 26(9), pp. 1963–1973. doi: 10.1093/molbev/msp106.

Durand, E. Y., Eriksson, N. and Mclean, C. Y. (2014) 'Reducing pervasive false-positive

identical-by-descent segments detected by large-scale pedigree analysis', *Molecular Biology and Evolution*, 31(8), pp. 2212–2222. doi: 10.1093/molbev/msu151.

Elhaik E. 2021. (2021) Why most Principal Component Analyses (PCA) in population genetic studies are wrong. bioRxiv.2021.2004.2011.439381. doi: 10.1101/2021.04.11.439381

Elhaik, E. and Ryan, D. M. (2019) 'Pair Matcher (PaM): fast model-based optimization of treatment/case-control matches', *Bioinformatics*, 35(13), pp. 2243–2250. doi: 10.1093/bioinformatics/bty946.

Falush, D., Stephens, M. and Pritchard, J. K. (2003) 'Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.', *Genetics*, 164(4), pp. 1567–1587.

Guillot, G., Mortier, F. and Estoup, A. (2005) 'GENELAND: A computer package for landscape genetics', *Molecular Ecology Notes*, 5(2), pp. 712–715. doi: 10.1111/j.1471-8286.2005.01031.x.

Gusev, A. et al. (2009) 'Whole population, genome-wide mapping of hidden relatedness', *Genome Research*, 19, pp. 318–326. doi: 10.1101/gr.081398.108.

Intarapanich, A. et al. (2009) 'Iterative pruning PCA improves resolution of highly structured populations', *BMC Bioinformatics*, 10(382). doi: 10.1186/1471-2105-10-382.

Joseph, T. A. and Pe'er, I. (2018) 'Inference of Population Structure from Ancient DNA', *bioRxiv*, pp. 90–104. doi: 10.1007/978-3-319-89929-9_6.

Latch, E. K. et al. (2006) 'Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation', *Conservation Genetics*, 7, pp. 295–302. doi: 10.1007/s10592-005-9098-1.

Lawson, D. J. et al. (2012) 'Inference of population structure using dense haplotype data', *PLoS Genetics*, 8(1). doi: 10.1371/journal.pgen.1002453.

Leslie, S. et al. (2015) 'The fine-scale genetic structure of the British population', *Nature*, 519(7543), pp. 309–314. doi: 10.1038/nature14230.

Lombaert, E., Guillemaud, T. and Deleury, E. (2018) 'Biases of STRUCTURE software when exploring introduction routes of invasive species', *Heredity*, 120(6), pp. 485–499. doi: 10.1038/s41437-017-0042-1.

Manichaikul, A. et al. (2010) 'Robust relationship inference in genome-wide association studies', *Bioinformatics*, 26(22), pp. 2867–2873. doi: 10.1093/bioinformatics/btq559.

McVean, G. (2009) 'A genealogical interpretation of principal components analysis',

PLoS Genetics, 5(10). doi: 10.1371/journal.pgen.1000686.

Moltke, I. and Albrechtsen, A. (2014) 'RelateAdmix: A software tool for estimating relatedness between admixed individuals', *Bioinformatics*, 30(7), pp. 1027–1028. doi: 10.1093/bioinformatics/btt652.

Pan, X. *et al.* (2017) 'Fine population structure analysis method for genomes of many', *Scientific Reports*, 7(1). doi: 10.1038/s41598-017-12319-1.

Patterson, N. *et al.* (2004) 'Methods for High-Density Admixture Mapping of Disease Genes', *The American Journal of Human Genetics*, 74(5), pp. 979–1000. doi: 10.1086/420871.

Price, A. L. *et al.* (2006) 'Principal components analysis corrects for stratification in genome-wide association studies', *Nature Genetics*, 38, pp. 904–909. doi: 10.1038/ng1847.

Price, A. L. *et al.* (2009) 'Sensitive detection of chromosomal segments of distinct ancestry in admixed populations', *PLoS Genetics*, 5(6). doi: 10.1371/journal.pgen.1000519.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000) 'Inference of population structure using multilocus genotype data', *Genetics*, 55(2), pp. 945–959.

Purcell, S. *et al.* (2007) 'PLINK: a tool set for whole-genome association and population-based linkage analyses.', *American journal of human genetics*, 81(3), pp. 559–75. doi: 10.1086/519795.

Ramstetter, M. D. *et al.* (2017) 'Benchmarking relatedness inference methods with genome-wide data from thousands of relatives', *Genetics*, 207(1), pp. 75–82. doi: 10.1534/genetics.117.1122.

Sankararaman, S. *et al.* (2008) 'Estimating Local Ancestry in Admixed Populations', *American Journal of Human Genetics*, 82(2), pp. 290–303. doi: 10.1016/j.ajhg.2007.09.022.

Stevens, E. L. *et al.* (2011) 'Inference of relationships in population data using identity-by-descent and identity-by-state', *PLoS Genetics*, 7(9). doi: 10.1371/journal.pgen.1002287.

Tang, H. *et al.* (2005) 'Estimation of individual admixture: Analytical and study design considerations', *Genetic Epidemiology*, 28(4), pp. 289–301. doi: 10.1002/gepi.20064.

Thornton, T. *et al.* (2012) 'Estimating kinship in admixed populations', *American Journal of Human Genetics*, 91(1), pp. 122–138. doi: 10.1016/j.ajhg.2012.05.024.

Wang, B., Sverdlov, S. and Thompson, E. (2016) 'Efficient estimation of realized kinship from SNP genotypes', *Genetics*, 205(3), pp. 1–23. doi: 10.1534/genetics.116.197004.

Yuan, K. *et al.* (2017) 'Models, methods and tools for ancestry inference and admixture analysis', *Quantitative Biology*, 5(3), pp. 236–250. doi: 10.1007/s40484-017-0117-2.

