

OXFORD

GENETICS

# High resolution analysis of population structure using rare variants

Lei Huang<sup>1</sup>, Thisseas C. Lamnidis<sup>1</sup> and Stephan Schiffels<sup>1,\*</sup>

<sup>1</sup>Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

\*Correspondence should be addressed to [stephan\\_schiffels@eva.mpg.de](mailto:stephan_schiffels@eva.mpg.de).

## Abstract

Various statistical methods have been developed to identify population structure from genetic data, including  $F$ -statistics, which measure the average correlation in allele frequency differences between two pairs of populations. However, the SNPs analyzed with  $F$ -statistics are often limited to those found as part of microarrays or, in the case of ancient DNA, to SNP capture panels, which are those within the common allele frequency band. Recent advances in sequencing technology increasingly allow generating whole-genome sequencing data, both ancient and modern, which not only enable querying nearly every base of the genome, but also contain numerous rare variants. Rare variants, with their more population-specific distribution, allow detection of population structure with much finer resolution than common variants - an opportunity that has so far been under-exploited. Here, we develop a new statistical method, *RAS* (Rare Allele Sharing), for summarizing rare allele frequency correlations, similar to  $F$ -statistics but with flexible ascertainment on allele frequencies. We test *RAS* on both published and simulated data and find that *RAS* has better resolution in distinguishing populations, with appropriate ascertainment. Leveraging this, we further develop the use of *RAS* to compute ancestry proportions with higher accuracy than existing methods, in cases of closely-related source populations. We implemented the new statistical methods as an R package and a command line tool. In summary, our method can provide new perspectives to identify and model population structure, allowing us to understand more subtle relationships among populations in the recent human past.

**Keywords:** population structure; rare variation; ancient DNA

## Introduction

Human population structure is shaped by past demographic events, which in turn can be inferred using genomic data. For example, populations isolated from each other for an extended period of time will differ in their allele frequencies due to genetic drift (exacerbated in case of small population sizes). On the other hand, migrations and admixture tend to equalize allele frequencies and affect population structure. Therefore, by analyzing and modeling population structure we can infer demographic processes. To identify population structure and make inferences on past demographic events (such as isolation and migration), many statistical methods have been developed and established in the field.

One popular approach is  $F$ -statistics (Patterson *et al.* 2012; Peter 2016), subdivided into  $F_2$ ,  $F_3$  and  $F_4$  depending on the number of populations involved. All  $F$ -statistics can be formulated as  $F_4$ , and therefore can measure the average correlation in allele frequency differences between two pairs of populations (Reich *et al.* 2009; Lipson 2020), and reflect the overlap between two genetic drift paths in a demographic model of the relationships of all populations involved (Reich *et al.* 2009; Patterson *et al.* 2012).  $F$ -statistics were first proposed in Reich *et al.* (2009) to test for "tree-ness" and compute the admixture proportion of focal populations rejecting tree-ness. Hereafter,  $F$ -statistics have been widely applied in human archaeogenetics, such as testing genetic similarity of populations (Raghavan *et al.* 2014), determining ancestry components (Reich *et al.* 2012; Lazaridis

*et al.* 2014) and detecting past admixture events (Patterson *et al.* 2012).

One critical advantage of  $F$ -statistics, unlike many other methods relying on allele frequency modeling (e.g. momi (Kamm *et al.* 2020), fastsimcoal (Excoffier *et al.* 2013), dadi (Gutenkunst *et al.* 2009)), is their robustness to some forms of SNP ascertainment. Specifically, it was shown theoretically (and in simulations) that statistical tests based on  $F$ -statistics (e.g. for tests for admixture) are unbiased under certain outgroup-directed ascertainment schemes (Patterson *et al.* 2012). It turns out, empirically, that non-outgroup-directed ascertainment are also close to being unbiased (although see Flegontov *et al.* (2023)). The ability to analyze SNP-ascertained datasets was and still is critical, as Array-Genotyping (e.g. Illumina 650K (Li *et al.* 2008), Affymetrix Human Origins (Patterson *et al.* 2012), 1240K (Mathieson *et al.* 2015)) remain the primary tool for obtaining genome-wide variation for population structure analysis including  $F$ -statistics (Novembre and Stephens 2008; Li *et al.* 2008; Patterson *et al.* 2012). In the mean time sequencing cost dropped further and larger amounts of whole-genome sequencing datasets were generated in recent years (The 1000 Genomes Project Consortium 2015; Bergström *et al.* 2020). In the past several years we have also witnessed an explosion of ancient DNA data, in large part based on enrichment techniques on ascertained SNP sets (Haak *et al.* 2015; Mathieson *et al.* 2015), but increasingly also based on shotgun sequencing, due to improvements in extraction methods and sequencing technology (Orlando *et al.* 2021).

## 2 Rare variants and population structure

The ongoing shift to shotgun sequencing, both in modern and ancient DNA analysis, makes it possible to query nearly every base of the genome, allowing the application of advanced demographic inference methods, as well as addressing potential bias in  $F$ -statistics of samples genotyped via SNP arrays or in-solution target capture (Flegontov *et al.* 2023). More importantly, whole-genome sequencing data contain many rare variants (The 1000 Genomes Project Consortium 2015; Bergström *et al.* 2020), which are more likely to be recently derived and can lead to novel conclusions on recent demographic history. For example, signals of recent admixture can be emphasized, as different African populations share more doubletons (variants shared by only two individuals, i.e. with allele count two) with each other than with East Asians, which is also reflected in the significant positive value of  $D$ (Chimp, Yoruba; Han, Mbuti) at low derived allele frequency in Yoruba, contradicting the phylogeny (Mbuti, (Yoruba, Han)) inferred from genome-wide variation (Bergström *et al.* 2020). Rare variation has been recognized as a potential tool for identifying fine-scale population structure, especially for distinguishing closely related populations. When quantifying the affinity between pairs of individuals in the 1000 Genomes Project with doubleton sharing, rather than genotype covariance, boundaries among populations are more pronounced, and even subgroups in populations GBR (from Britain) and CHS (from Southern China) can be detected. That is because many more doubletons are shared within the same subgroup than between different subgroups of the same population (The 1000 Genomes Project Consortium 2015). Rare allele methods have also been successfully applied to ancestry estimation. When focusing on alleles with lower frequency, ancient British populations from the Iron Age and Anglo-Saxon period are better distinguished by the ratio of alleles they share with Dutch and Spanish, therefore allowing accurate estimation on the Anglo-Saxon British ancestry in present-day British populations (Schiffels *et al.* 2016). Similarly, among North American indigenous groups, present-day Athabaskans can be distinguished from other groups due to recent admixture from Paleo-Eskimos, and therefore can be modeled as being admixed between northern First Peoples and Paleo-Eskimos (Flegontov *et al.* 2019).

However, rare allele methods have not yet been coherently formalized in earlier publications. Considering the similarity between  $F$ -statistics and previous rare allele methods, in this article we incorporate rare allele analyses into the definition of RAS-statistics, and demonstrate through simulation and empirical data their ability to outperform ordinary  $F$ -statistics in detecting recent demographic events, even when the latter are applied to whole-genome data. We derive a RAS-based method for ancestry decomposition and show that it gives more accurate estimates than  $F$ -statistics based ancestry proportions.

## 50 Method

### 51 RAS: Rare allele sharing statistics

52 Here, we define a statistic summarizing rare allele frequency correlations, RAS, an acronym for "Rare Allele Sharing". RAS-  
53 statistics are computed on genome-wide biallelic SNPs, similar  
54 to  $F$ -statistics (Patterson *et al.* 2012) but with ascertainment on  
55 rare variants.

56 In order to define RAS-statistics, we first define the following  
57 concepts:

58 **Reference population  $R$**  A group of individuals that is used to  
59 ascertain variants within specific allele frequency ranges.

**Outgroup  $O$**  An individual or group, which is an outgroup to all other individuals/groups involved in the analysis. It is used to define the ancestral allele, and hence polarize alleles into ancestral and derived.

**Genome length  $L$**  The number of all positions in the genome considered for analysis. Usually this covers all biallelic SNPs in large panels such as 1000 Genomes (The 1000 Genomes Project Consortium 2015) and/or HGDP (Bergström *et al.* 2020).

**Sample allele frequencies  $\mathbf{x}_A$**  A vector of length  $L$  representing outgroup-directed derived allele frequencies in population  $A$ . We use expression  $x_{A,i}$  to refer to sample allele frequency  $A$  at position  $i$ .

**Ascertained SNP set  $M(O, R, f_{\min}, f_{\max})$**  The set of all positions at which the derived allele frequency (polarized via outgroup  $O$ ) in reference population  $R$  is between  $f_{\min}$  and  $f_{\max}$ .

**Ascertained non-missing overlap  $L_M(A, B)$**  the number of sites in  $M$  that are non-missing in population  $A$  and  $B$ .

We then define a simple statistic as the correlation of two frequencies (for brevity, we write  $M$  instead of  $M(O, R, f_{\min}, f_{\max})$ ):

$$RAS(A; B) = \frac{1}{L_M(A, B)} \mathbf{x}_A^T \mathbf{x}_B = \frac{1}{L_M(A, B)} \sum_{i \in M} x_{A,i} x_{B,i} \quad (1)$$

Intuitively, this statistic measures the average rate of allele sharing among any pair of haplotypes from groups  $A$  and  $B$  across all ascertained variants.

Indeed, there is a close correspondence of  $RAS(A; B)$  and so-called outgroup- $F_3$ -statistics, which are more generally defined as

$$F_3(A, B; O) = \frac{1}{L} (\mathbf{x}_O - \mathbf{x}_A)^T (\mathbf{x}_O - \mathbf{x}_B).$$

Using derived allele frequencies polarized by our ascertainment outgroup  $O$ , and using only monomorphic sites in  $O$ , this definition simplifies to  $\mathbf{x}_A^T \mathbf{x}_B$  as in equation 1.

From our basic form of RAS-Statistics (eq. 1), we derive the following RAS-differences, termed RASD:

$$\begin{aligned} RASD(A_1, A_2; B) &= RAS(A_1; B) - RAS(A_2; B) \\ RASD(A; B_1, B_2) &= RAS(A; B_1) - RAS(A; B_2) \\ RASD(A_1, A_2; B_1, B_2) &= RASD(A_1; B_1, B_2) - RASD(A_2; B_1, B_2) \\ &= RAS(A_1; B_1) - RAS(A_2; B_1) - \\ &\quad RAS(A_1; B_2) + RAS(A_2; B_2) \end{aligned} \quad (2)$$

Those derived RASD-statistics can be used to test symmetry or treeness, similar to the widely used  $F_4$  or  $D$  statistics (Patterson *et al.* 2012).  $RASD(A_1, A_2; B)$  corresponds to  $F_4(A_1, A_2; B, O)$ , which tests relative sharing between  $A_1$  and  $A_2$  with respect to  $B$ . Similarly,  $RASD(A; B_1, B_2)$  corresponds to  $F_4(A, O; B_1, B_2)$ . Ultimately, the difference of differences  $RASD(A_1, A_2; B_1, B_2)$  corresponds to  $F_4(A_1, A_2; B_1, B_2)$ .

Note that this correspondence to  $F_4$  statistics becomes an equivalence in the special case that all sites are non-missing in all considered groups, or only sites non-missing in all groups are considered (e.g. using `maxmiss = 1` in the Software `qpfstats`

from ADMIXTOOLS2 (Maier *et al.* 2023), and no allele frequency ascertainment is applied (i.e.  $f_{\min} = 0$  and  $f_{\max} = 1$ ). In the more general case relevant here, due to different patterns of missing SNPs in ancient samples, our linear combinations in the definition of  $RASD$  are not equal to  $F_4$  even without frequency ascertainment.

## Ancestry decomposition

We can use  $RAS$  and  $RASD$  statistics to compute ancestry proportions. Specifically, we model a given *target population*  $T$  as a linear sum of *source populations*  $\{S_1, S_2, \dots, S_n\}$ , with the coefficients  $\{\beta_1, \beta_2, \dots, \beta_n\}$  summing up to 1. Symbolically, we can write

$$\begin{aligned} T &= \sum_{i=1}^n \beta_i S_i \\ \sum_{i=1}^n \beta_i &= 1 \end{aligned} \quad (3)$$

The key idea is to represent the *target* and *sources* by their shared genetic drift (as estimated using  $RAS$ -statistics) with a selected group of *reference populations*. Following the nomenclature from the popular qpAdm Software (Patterson *et al.* 2012; Maier *et al.* 2023), we denote *target* and *sources* as *left populations*, and *references* as *right populations*.

Specifically, we choose a set of  $m$  *right populations*  $\{R_1, R_2, \dots, R_m\}$ . If  $T$  is then admixed as specified in equation 3, and if there was no gene flow going from the *left* into the *right populations* (only from *right* to *left*, see below for a discussion on relaxing this condition), then we can write:

$$\begin{pmatrix} RAS(T; R_1) \\ RAS(T; R_2) \\ \vdots \\ RAS(T; R_m) \end{pmatrix} = \sum_i \beta_i \begin{pmatrix} RAS(S_i; R_1) \\ RAS(S_i; R_2) \\ \vdots \\ RAS(S_i; R_m) \end{pmatrix} \quad (4)$$

Defining the vectors  $\mathbf{v} = \{v_j\} = \{RAS(T; R_j)\}$  and  $\mathbf{b} = \{\beta_i\}$ , and the matrix  $\mathbf{W} = \{W_{ij}\} = \{RAS(S_i; R_j)\}$ , we can write this as:

$$\mathbf{v} = \mathbf{W}^T \cdot \mathbf{b} \quad (5)$$

This would be a simple linear regression model, but we still have to satisfy the constraint  $\sum_{i=1}^n \beta_i = 1$ . We therefore first write the last element of  $\mathbf{b}$  as

$$\beta_n = 1 - \sum_{i=1}^{n-1} \beta_i$$

Restricting to a single row  $j$  of the equation, we then get

$$v_j = \sum_{i=1}^{n-1} \beta_i W_{ij} + \left(1 - \sum_{i=1}^{n-1} \beta_i\right) W_{nj} \quad (6)$$

which in turn becomes

$$v_j - W_{nj} = \sum_{i=1}^{n-1} \beta_i (W_{ij} - W_{nj}) \quad (7)$$

The differences on both sides of the equation are actually  $RASD$ -statistics. We define  $X_{ij} = W_{ij} - W_{nj} = RASD(S_i, S_n; R_j)$

for  $i = 1, \dots, (n-1)$ ,  $y_j = v_j - W_{nj} = RASD(T, S_n; R_j)$  and  $\mathbf{a} = \{\beta_1, \dots, \beta_{n-1}\}$  can then write

$$\mathbf{y} = \mathbf{X}^T \cdot \mathbf{a} \quad (8)$$

or

$$\begin{pmatrix} RASD(T, S_n; R_1) \\ RASD(T, S_n; R_2) \\ \vdots \\ RASD(T, S_n; R_m) \end{pmatrix} = \sum_{i=1}^{n-1} \beta_i \begin{pmatrix} RASD(S_i, S_n; R_1) \\ RASD(S_i, S_n; R_2) \\ \vdots \\ RASD(S_i, S_n; R_m) \end{pmatrix} \quad (9)$$

which is a simple linear regression equation. The least-square solution of this equation is (see Hastie *et al.* (2009)):

$$\hat{\mathbf{a}} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y} \quad (10)$$

which is a genome-wide point-estimate of admixture proportions. There are two sources of uncertainty/error to consider in this estimation: First, the sampling noise and the finite length of the genome, and second the standard error from the least-square fit itself. We estimate both of these errors using a genome-wide block-Jackknife procedure (Busing *et al.* 1999) to re-estimate the admixture proportions using equation 10 with each of the blocks removed, and derive the standard error of them. This approach is inspired by the Jackknife implemented in qpAdm and the ADMIXTOOLS (Patterson *et al.* 2012) and ADMIXTOOLS2 packages (Maier *et al.* 2023)

The basic decomposition (equation 4) relies on there being no gene flow from left to right populations, only vice versa. This is hardly ever true in real populations. However, in contrast to ordinary F-Statistics, which is the basis for ancestry decomposition in qpAdm,  $RAS$  values are only affected by reverse gene flow if it occurred very recently. This can often be ruled out. But even if not, violations of the no-left-to-right gene flow assumption may affect goodness-of-fit statistics more than the actual ancestry estimates which we focus on here.

## Implementation of the method

All scripts used to process and analyze data, exclusively in the R programming language (Ihaka and Gentleman 1996), are provided within a GitHub repository [https://github.com/huanglei-artificium/RAS\\_tools](https://github.com/huanglei-artificium/RAS_tools), including documentation.

Briefly, our tool computes  $RAS$  and  $RASD$  with flexible ascertainment on allele frequency. Besides .geno files, our tool also accepts allele frequency data consisting of two columns representing numerator (nr of alternative alleles) and denominator (nr of non-missing haplotypes) as input. Then we compute allele frequencies for each population and each site, which are then used to i) select sites that fulfill ascertainment conditions, and ii) to compute the actual  $RAS$ -statistics on those sites.

To compute optional uncertainties based on a blockwise jackknife estimate (Busing *et al.* 1999),  $RAS$  gets computed blockwise (typically by chromosome), which are then combined to obtain genome-wide statistical values.

Multiple statistics, cycling through several populations, and multiple ascertainment conditions are handled efficiently inside our tools, and can be computed in one run.

## 4 Rare variants and population structure

### Simulations

To illustrate how *RAS* and *RASD* perform in contrast to  $F_3$  and  $F_4$  statistics, we devised a simulation scheme that allows for varying levels of population structure by tuning migration rates. Specifically, we use msprime ([Kelleher et al. 2016](#)) to simulate a set of nine populations, located in a  $3 \times 3$  grid, each with an effective diploid population size  $N_e = 20,000$  individuals, with  $n = 50$  diploid samples drawn from each population. Each population is connected to its non-diagonal neighboring populations through a symmetric two-way per-generation migration rate. All the migration rates ( $m$ ) in the same simulation were uniform, and varied across different simulations to produce different degrees of population structure among the simulated populations, with the scaled migration rate ( $4mN_e$ ) being 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000 and 5000 respectively. Each sample has 20 chromosomes, each 100Mbp in length. We number our nine populations from 0 to 8 following a left-to-right, bottom-to-top order, as depicted in Figure 1A. No specific outgroup was simulated, instead we just used a genome consisting exclusively of ancestral alleles as outgroup.

To assess the population structure in the simulated data, we perform the statistical analysis across three different ascertainment schemes (i) Using all variants in the simulated data, (ii) Using only rare variants with different ascertainment conditions, (iii) Using only a subset of 1.2M variants with derived allele frequency between 0.05 and 0.95, to mimic the 1240K panel ([Mathieson et al. 2015](#)).

We chose 10 individuals from each population to act as "test individuals" (denoted by  $T_i$ ), whereas the other 40 individuals are then used as References  $R_i$ . The ascertainment is therefore based on  $9 \times 40 = 360$  reference individuals. For some analysis, in order to average sampling noise, we rotated test and reference individuals 5 times so that each individual served as a test individual once. The code for running the simulation is available at [https://github.com/Schiffels-Popgen/RAS\\_exploration](https://github.com/Schiffels-Popgen/RAS_exploration).

### Modern reference data

For our modern reference data, we chose as a starting point the recently released harmonized dataset of 1000 Genomes Project (1kGP) and Human Genome Diversity Project (HGDP) ([Koenig et al. 2024](#)), where a new genotype calling was made based on the raw sequencing data from 1kGP and HGDP, with more than 150 million high-quality variants identified, including a large number of rare variants. We chose to focus on the European populations in this dataset, which includes five from 1kGP and eight from HGDP, in which genetic outliers and relatives closer than second-degree were filtered out according to the analysis of [Koenig et al. \(2024\)](#) (see Table 1 for the number of individuals for each population). We further supplemented this basic dataset with three European public datasets with genome-wide allele count data: Danish from GenomeDenmark project ([Maretty et al. 2017](#)), Dutch from Genome of the Netherlands (GoNL) project ([The Genome of the Netherlands Consortium 2014](#)) and Swedish from SweGen project ([Ameur et al. 2017](#)) (Table 1).

Even though we focused on Europe, for all the *RAS* statistics regarding real data, we used all African groups in 1kGP and HGDP as outgroups, ascertaining to strictly fixed sites within Africa.

We screened 133 million biallelic SNPs from the harmonized dataset of 1kGP and HGDP for analysis. Variant sites in the Danish, Dutch and Swedish datasets were filtered to this 1kGP+HGDP SNP set, to ensure that our African outgroup is

available on all analyzed SNPs. We excluded sites with different alternative alleles when joining the datasets. Unless otherwise stated, we used all European populations in Table 1 for references, which consist of 16 populations and 2184 individuals (4368 sets of haploid chromosomes). Allele frequencies for each site are based on all non-missing individuals.

**Table 1** Number of haploid copies (2N) of present-day populations used for analysis

Population abbr. <sup>a</sup>	Population	Haploids	Dataset
CEU	Northern and Western European ancestry	242	1kGP
FIN	Finnish	196	1kGP
GBR	British	176	1kGP
IBS	Spanish	208	1kGP
TSI	Toscani in Italy	206	1kGP
-	Adygei	34	HGDP
-	Basque	46	HGDP
-	French	54	HGDP
-	Italian	22	HGDP
-	Orcadian	26	HGDP
-	Russian	50	HGDP
-	Sardinian	54	HGDP
-	Tuscan	16	HGDP
DK	Danish	40	GenomeDenmark
NL	Dutch	998	GoNL
SE	Swedish	2000	SweGen
AFR_all	All Africans	1978	1kGP+HGDP

<sup>a</sup> Where available, we use abbreviations in the figures.

### Ancient genomes

Our ancient dataset consists of 34 individuals from Great Britain with shotgun-sequencing data, 7 dating to the Late Iron Age (LIA) and 27 to the Early Middle Ages (EMA) ([Martiniano et al. 2016; Schiffels et al. 2016; Gretzinger et al. 2022](#)). We started with alignment files (.bam files) and called variants overlapping with our reference SNPs, using the Majority-Call method with a minimum coverage of 3 and downsampling in pileupCaller (article submitted). We then selected individuals with more than 1 million SNPs overlapping with our reference data set (see above). We used the ancestry decomposition published in [Gretzinger et al. \(2022\)](#), with two major components: CNE ("Continental North European") and WBI ("Western British-Irish"), and labeled individuals whose dominant ancestry (CNE or WBI) is more than 70% as "England\_CNE" ( $N = 17$ ) and "England\_WBI" ( $N = 7$ ).

## 1 Results

### 2 Exploring RAS with simulated data

3 We first explored, how our *RAS*-Statistics can distinguish populations from our simulated  $3 \times 3$  grid of connected populations (Methods). We chose two corner populations as references, namely population 6 (in the top left, see Figure 1A) and population 8 (in the top right) and computed  $RAS(x; R_6)$  and  $RAS(x; R_8)$  for all test individuals  $x$  across all populations. For comparison, we also computed  $F_3(O; x, R_6)$  and  $F_3(O; x, R_8)$ , without any allele frequency ascertainment.

11 For low migration rate, both *RAS* and  $F_3$  reveal clearly separated clusters of individuals, corresponding to the 9 populations (Figure 1A, with populations closer to the references being placed high on the respective axis, and populations more distant to the references being placed low on the axes (Figure 1A).

16 Increasing the migration rate shows less clearly defined clusters in the case of  $F_3$ , with no apparent structure being visible for the highest migration rate tested here ( $4mN_e = 2000$ ) (1B and C). In contrast, ascertaining SNPs to be rare with respect to the reference populations (which here comprise  $9 \times 40$  individuals, excluding the  $9 \times 10$  test individuals, see Methods) reveals structure being visible also at higher migration rates. For example, at  $4mN_e = 100$  (Figure 1E), *RAS* scatter plots reveal still well-separated clusters while  $F_3$  (Figure 1B) already shows considerable overlap between groups. Even at  $4mN_e = 2000$  (Figure 1F), *RAS* still shows some power to distinguish groups, whereas  $F_3$  appears random (Figure 1C). For our simulated "1240K" dataset with only 1.2 million common variants, structure is substantially less resolved (Supplementary Figure S1). With the highest simulated migration rate, even rare variation appears quite random (Supplementary Figure S1).

32 Following this qualitative assessment of the ability to separate closely related groups, we devised a more quantitative assessment, by testing for each individual whether it is closest to the mean position of their own population or to some other population, in which case we consider it misclassified. The misclassification ratio is then the proportion of misclassified test individuals relative to all test individuals (using the rotation scheme described in Methods, this amounts to 450 tests). As expected, *RAS* performs better in distinguishing populations for medium and high migration rates, although at the highest migration rate  $4mN_e = 5000$ , the improvement is relatively weak (Supplementary Figure S2). Specifically, at  $4mN_e = 50, 100, 200$ , while  $F_3$  statistics (for both "all sites" and 1240K) start to misclassify, *RAS* can still distinguish the test individuals with no or little error. At low migration rates  $4mN_e = 1, 2$ , note that *RAS* estimates are noisier than the corresponding  $F_3$ -estimates, due to fewer shared rare variants: Low levels of migration make it difficult for variants to spread between populations, especially for rare variants between geographically distant populations (e.g., populations 0, 1 and 2 with respect to populations 6 and 8), so that the signal of rare allele sharing between them becomes particularly noisy.

### 54 Application to real data

55 Turning to real data, we first analyzed present-day European genetic diversity. Specifically, we used the five European populations from the 1000 Genomes Project (1kGP) as references (The 1000 Genomes Project Consortium 2015) and eight European populations from the HGDP project as test individuals (Bergström et al. 2020). We then quantified the affinity between 56 1kGP European populations and each European HGDP individ-

ual with outgroup- $F_3$  and *RAS* (Figure 2). FIN (Finnish) and IBS (Spanish) from 1KGP are selected as references because they are relatively different geographically and genetically. Both whole genome and 1240K SNP sets reveal differential affinities of the HGDP populations with respect to these references (Figure 2B and C). For example, Russian and Sardinian groups are closest to FIN or IBS references, respectively. Ascertaining on rare allele frequencies in the references with *RAS*, these differences become substantially larger (Figure 2A). In particular, Russian and Basque groups from the HGDP share substantially more rare alleles with either FIN or IBS, indicating recent shared ancestry between Russian and Finnish, and between Basque and Spanish. In addition, there are clearer boundaries for some isolated populations, such as Sardinian and Orcadian. All groups are more clearly separated with *RAS* than with un-ascertained variants.

57 We next turned to ancient DNA to explore the potential of rare variants to analyze ancient population structure. Specifically, we again used the present-day 1kGP reference groups and measured *RAS* and  $F_3$  to a set of ancient genomes from England (Gretzinger et al. 2022) for which whole-genome sequencing data is available. Without ascertainment, these two groups do not appear to separate clearly on either all sites or 1240K subset (Figure 2B,C). In particular, WBI individuals are distributed among the entire range of CNE individuals. In contrast, we find that in our *RAS* analysis (Figure 2A), the WBI individuals fall closer to present-day French compared to samples with CNE ancestry, with all WBI individuals being closer to IBS than to FIN, as indicated by the dashed line in Figure 2A.

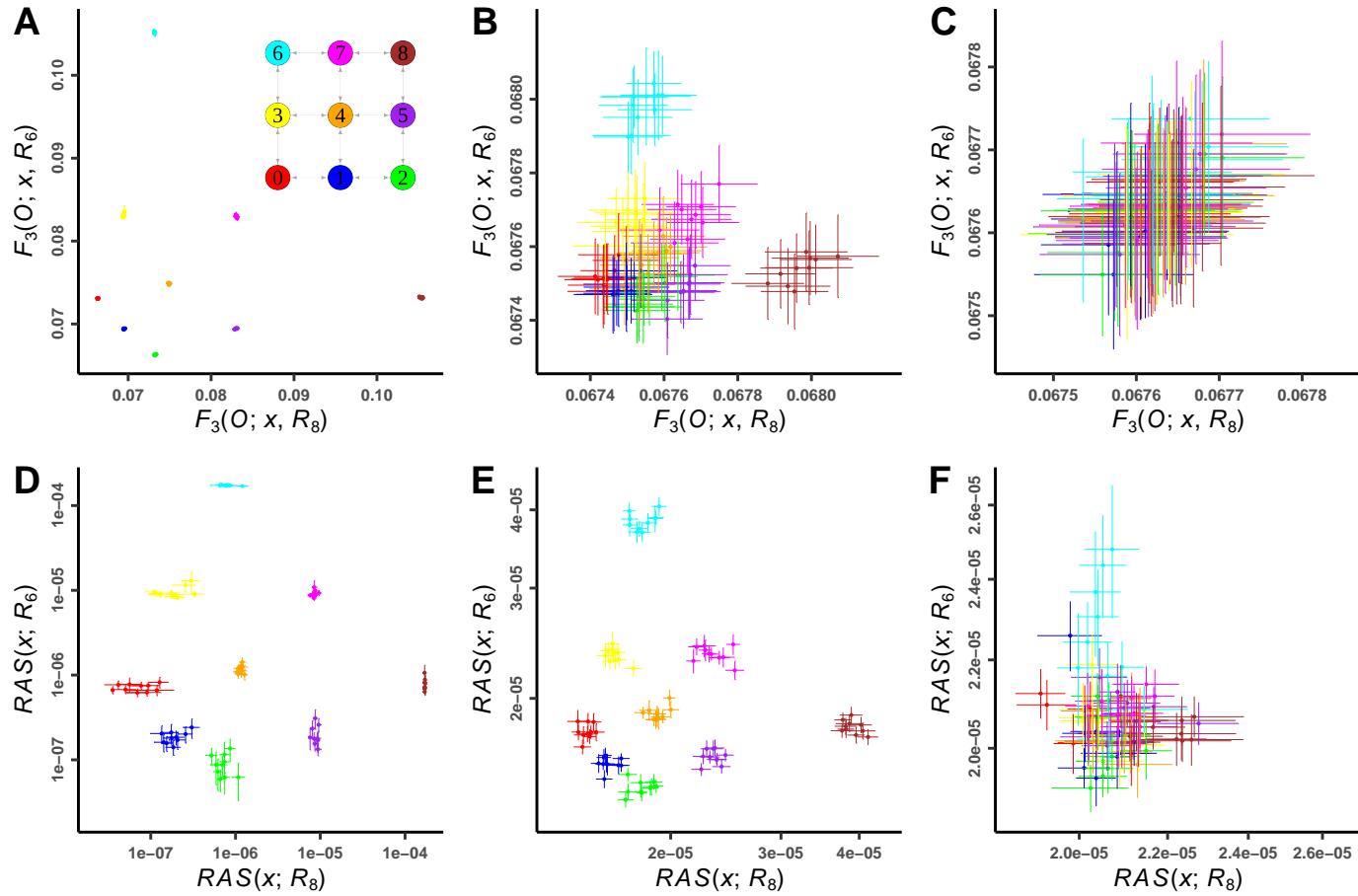
### 59 Testing for population structure with RASD

A formal test of population structure is a test for symmetry between two groups with respect to a reference group. For classical  $F$ -Statistics, this is done through  $F_4$ -Statistics, which are essentially differences of  $F_3$ -Statistics, to statistically quantify differential affinities as deviations from symmetry. Analogously, we defined *RASD*( $A_1, A_2; B_1, B_2$ )-Statistics (Methods), which tests whether  $A_1$  and  $A_2$  are differentially related to  $B_1$  and  $B_2$ .

Here, we explore the following form: *RASD*(England\_CNE, England\_WBI;  $R_1, R_2$ ), where the first two slots are cycling through ancient individuals from Britain with dominant CNE ( $n = 17$ ) or WBI ( $n = 7$ ) ancestry, respectively, and the last two slots are cycling through various present-day reference populations. We evaluated the results using the Z-Score (i.e. the statistical deviation from zero), using *RASD* at different ascertainment conditions and corresponding  $F_4$  at 1240K and all sites (Figure 3). When applying *RASD*, as long as it is not restricted to extremely rare alleles (up to 0.2%), the ability to distinguish between CNE and WBI is overall better than using  $F_4$ , with Z scores even greater than 10 for some combinations. When comparing different reference populations, using NL (Dutch) and SE (Swedish) generally resulted in better discrimination than using FIN (Finnish). This suggests that the immigrants into Britain from continental Europe during Early Middle Age were genetically closer to present-day Dutch or Swedish, consistent with the findings from Gretzinger et al. (2022).

In order to explore the genetic change in Early Middle Age Britain, we performed a systematic analysis by grouping individuals with CNE ( $n = 17$ ) and WBI ( $n = 7$ ) ancestry respectively and comparing them with present-day European populations, under different ascertainment conditions (Figure 4).

## 6 Rare variants and population structure



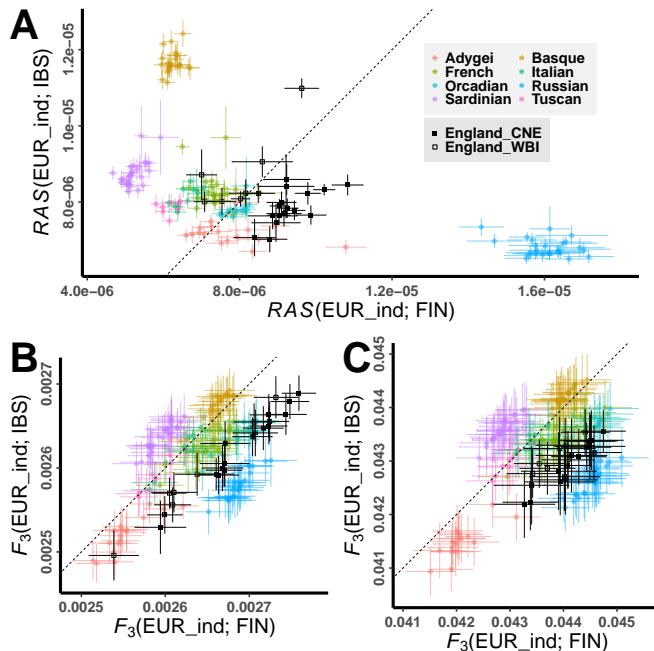
**Figure 1** Outgroup- $F_3$  and  $RAS$  statistics on test individual  $x$  and a specific reference population ( $R_6$  or  $R_8$ ) at different migration rates. The SNP panels used are all sites (A, B and C) and sites with derived allele frequency less than 1% in all reference individuals (D, E and F). The migration rates ( $4mN_e$ ) used in the simulation are 1 (A and D), 100 (B and E) and 2000 (C and F). Test individuals  $x$  are distinguished by colors representing different populations, shown in legend in (A), which also includes the schematic of simulated population migration. The coordinates of values of  $RAS$ -statistics are logarithmic.

For the 1240K panel and all sites, we observe non-significant positive Z scores for all European reference populations, although some Northern European populations, such as SE (Swedish), NL (Dutch) and FIN (Finnish) have relatively better ability to distinguish CNE and WBI ancestry (Figure 4). In contrast, with  $RASD$ -statistics, the difference between CNE and WBI, represented by Z score, becomes much more pronounced at low derived allele frequency. Among those present-day European populations, SE (Swedish), NL (Dutch), CEU (Northern and Western European ancestry) and FIN (Finnish) have better ability to distinguish CNE and WBI, suggesting that these present-day populations are most closely related to the actual source that migrated into early medieval Britain, in line with previous conclusion that the immigrants were from Northern Europe (Gretzinger *et al.* 2022).

Rare alleles can even provide more information about population history at different points in the past, which is reflected by the results of  $RASD$ -statistics at different derived allele frequency cutoffs. For French, TSI and IBS, their Z scores are much higher at very low frequency 0 - 0.2% than at 0 - 0.6%, which is due to recent low-level gene flow within the European continent making present-day Southern Europeans share increasing number of rare alleles with CNE ancestry, rather than WBI ancestry.

The Z scores of French and TSI at 0 - 0.2% are even slightly higher than those of FIN, reflecting that the magnitude of recent gene flow may be highly dependent on geography, since Finnish and the estimated location of CNE ancestry are on the opposite sides of the Baltic Sea. However, the Z score of FIN reaches its peak at maximum cutoff 1.4% and still remains significant until maximum cutoff 3.8%, suggesting that the gene flow between the CNE ancestry and the ancestor of Finnish occurred in the more distant past, probably mediated through a population closely related to present-day Swedish.

We have compared the ancient British individuals with mostly CNE or WBI ancestry. However, the actual CNE admixture in EMA Britain was like a spectrum, with varying proportions among different individuals (Gretzinger *et al.* 2022). As  $F$ -statistics are linear under a gradient of admixture components, so are  $RAS$  and  $RASD$ . When applied to real data, there is a strong correlation between  $RAS$  and  $RASD$ -estimates, and the actual ancestry proportion of a sample. Therefore, we performed  $RASD$  and  $F$ -statistics on each EMA British individual by comparing it to a pair of present-day European reference populations that are able to distinguish EMA British individuals. We then computed the correlation between our estimates and the high-resolution CNE ancestry estimated by supervised ad-



**Figure 2** *RAS* and outgroup- $F_3$  statistics on European individual  $x$  and FIN/IBS with error bars of  $\pm 1$  standard deviation (SD). Modern European individuals are distinguished by colors representing different populations in HGDP; ancient British individuals are marked black and distinguished by shapes representing different ancestries. SNP panels are based on the harmonized dataset of 1kGP and HGDP with the following filtering: monomorphic in all 1kGP Africans and derived allele frequency less than 2% in all 1kGP Europeans (A); all sites (B); 1240K sites (C). The dash line indicates the equal relationship to FIN and IBS. FIN: Finnish; IBS: Spanish.

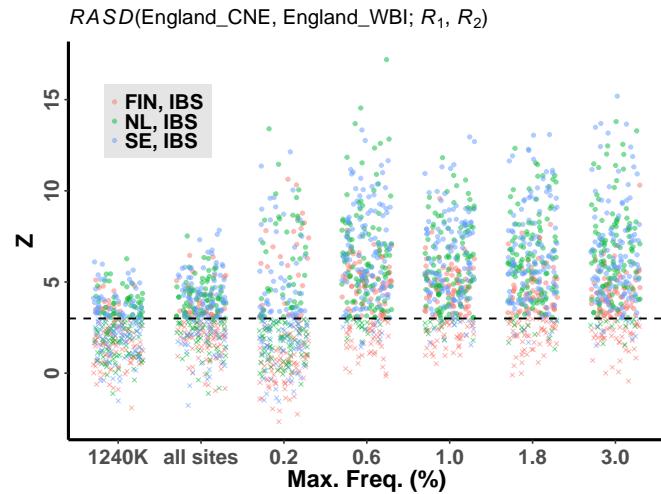
1 mixture using thousands of present-day Europeans as reported  
 2 in [Gretzinger et al. \(2022\)](#) (Figure 5; Supplementary Figure S3).  
 3 At specific  $f_{\max}$ , such as 0.6% and 1%, our *RASD* estimates using  
 4 any of Finnish (FIN), Netherlands (NL) or Swedish (SE) in contrast to Spanish (IBS), show a higher correlation to the reported  
 5 CNE ancestry estimates than  $F_4$ -values on 1240K, indicating a  
 6 better resolution in distinguishing EMA British individuals with  
 7 rare alleles, and potentially a more accurate estimation on CNE  
 8 ancestry from new samples provided the *RASD* statistical values.  
 9 Those results suggest a potential method for more accurate  
 10 ancestry decomposition using *RAS*-statistics with appropriate  
 11 ascertainment on rare alleles.  
 12

### 13 Decomposing ancestries using linear combinations of 14 *RAS*

15 Motivated by the correlation between *RAS*-Statistics and ancestry components, we devised a new method to decompose  
 16 ancestry components based on *RAS* (see Methods).

17 Briefly, every left population (i.e. the target and the sources)  
 18 has a specific profile of rare allele sharing with each of the right  
 19 populations, represented by a multi-dimensional vector. We  
 20 then model the target profile as a linear combination of source  
 21 profiles, with the coefficient reflecting the admixture proportion.

22 We tested this new method on our simulated data, focusing  
 23 on two-component models. According to the simulation scheme  
 24 (Method), some populations can be represented as admixtures



**Figure 3**  $Z$  score distribution of  $RASD(\text{England\_CNE}, \text{England\_WBI}; R_1, R_2)$  using different SNP panels, shown in each bar with jitter. The ascertained cases are represented in numbers, which are  $p_{\max}$  (in percentage) in 1kGP and HGDP European populations plus the Dutch, Danish, and Swedish populations. Different colors represent different reference population pairs ( $R_1, R_2$ ). Results with  $Z < 3$  are indicated as crosses and with  $Z > 3$  with circles.

26 of other populations. Specifically, we aim to model population  
 27 4 (in the middle of the grid) as an admixture of population 0  
 28 (bottom-left) and population 8 (top-right). We expect for all  
 29 migration rates a 50%/50% decomposition, due to symmetry.

30 We compared our ancestry estimates based on *RAS* with  
 31 the expected values and defined two types of error: (A) the  
 32 absolute difference between our estimate and the theoretical  
 33 value of 0.5, denoted as "true error"; (B) the standard error of our  
 34 estimate based on a chromosome-wise jackknife ([Busing et al. 1999](#)), denoted as "self error".

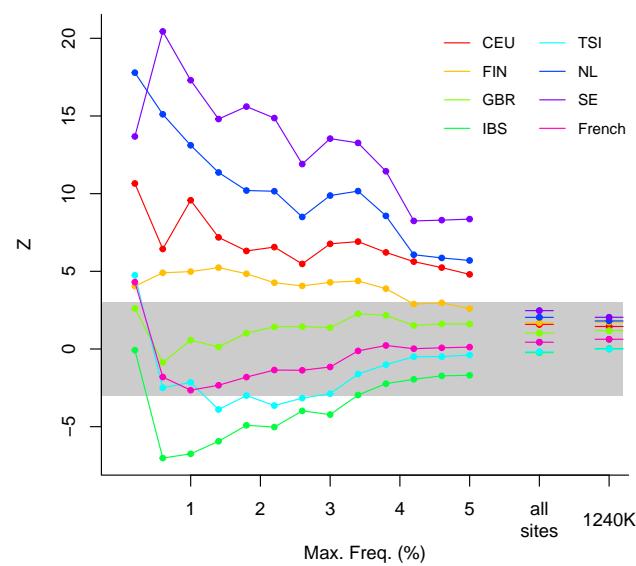
35 The results reveal that at high migration rates, both types  
 36 of errors are substantially lower for our *RAS*-based ancestry  
 37 estimate compared to  $F_3$ -based estimates on 1240K or even on  
 38 all sites (Figure 6). Note that towards lower migration rates,  
 39 we observe a turning point (around  $4Nm=100$ , Supplementary  
 40 Figure S4), where all sites and even 1240K are performing subtly  
 41 better than rare variants, although at a very low level of error,  
 42 which we attribute to a lower number of shared rare variants  
 43 for low migration rates. Therefore in real cases, it's appropriate  
 44 to use  $F$ -statistics when populations are highly differentiated,  
 45 while *RAS* fills the gap where  $F$ -statistics lose resolution.

### 46 Discussion

47 We have defined *RAS*, a statistical method based on rare  
 48 allele sharing between populations, and demonstrated that rare  
 49 variation provides powerful means of identifying fine-scale pop-  
 50 ulation structure and revealing unique population histories that  
 51 common alleles may not capture. For both simulated and em-  
 52 pirical data, we observe a signal reinforcement of recent demo-  
 53 graphic events, reflected in the much stronger allele sharing for  
 54 rare alleles (Figure 1D,E,F; 2A; 3; 4), compared to common alleles  
 55 and even the whole genome.

56 Investigating population structure is one of the primary goals

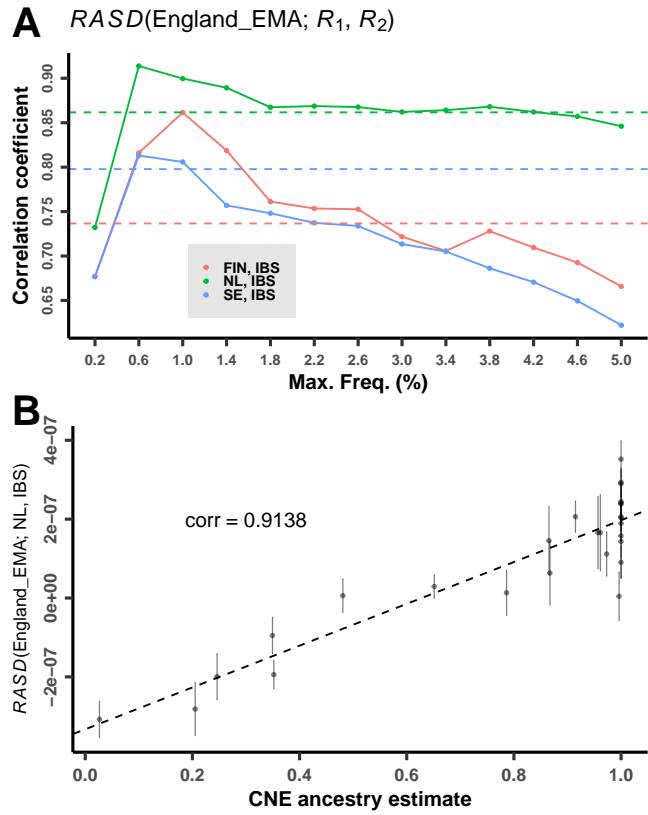
## 8 Rare variants and population structure



**Figure 4** Distinguishing CNE and WBI ancestry using *RASD* and  $F_4$  represented in Z scores. The colored lines are Z scores of  $RASD(\text{England\_CNE}, \text{England\_WBI}; x)$ , where  $x$  is a selected present-day European population, under different  $p_{\max}$  (in percentage) in present-day European references. Two categories on the right cover ordinary  $F_4(\text{England\_CNE}, \text{England\_WBI}; x, \text{AFR\_all})$  on all sites and on 1240K, respectively. The gray area represents  $|Z| < 3$ .

of population genetic analyses involving ancient and modern genomes. Such studies have greatly advanced our understanding of historical human populations, and their migration and admixture (Nielsen *et al.* 2017; Skoglund and Mathieson 2018; Liu *et al.* 2021; Stoneking *et al.* 2023). Especially in recent years, studies involving larger sample sizes have revealed more detailed historical demographic events (Lazaridis *et al.* 2022; Al-lentoft *et al.* 2024; Antonio *et al.* 2024; McColl *et al.* 2025). In these studies, 1240K remains the primary ascertainment scheme for exploring genome-wide population structure. While increasing sample sizes help with existing methods, we here show that different ascertainment approaches and in particular a focus on rare variants can more dramatically improve resolution, such as the clearly demarcated Russian and Basque populations (Figure 2A). Even for ancient DNA, by embracing an ascertainment scheme strictly in present-day data, the ability to distinguish between populations is also significantly increased, as exemplified here by differentiating between CNE and WBI ancestries (Figure 3; 4; 5). Surprisingly, the resolution for rare variants is even higher than using all variants (Figure 2; 3), which further emphasizes the importance of ascertaining rare alleles for population structure analysis.

Larger sample sizes have a more direct impact on the resolution of our method, compared to most traditional tools based on common alleles such as  $F$ -statistics and PCA. The rare alleles ascertained from reference populations generally also have low frequencies in the test populations. Therefore, the relative error of their allele frequency estimates is larger than that of common alleles. This highlights the role of increasing sample sizes in

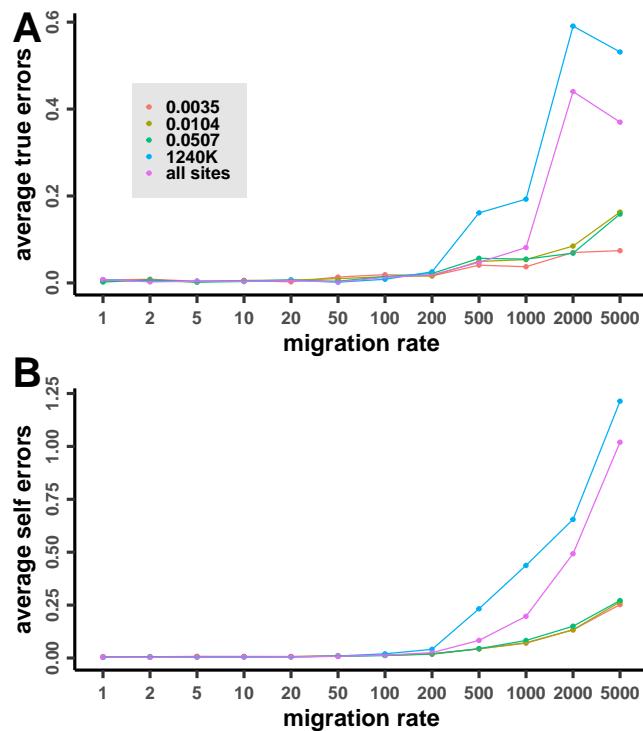


**Figure 5** (A) The colored lines are the correlations between  $RASD(\text{England\_EMA}; x_1, x_2)$  and estimates of CNE ancestry of each England\_EMA individual (from Gretzinger *et al.* (2022)), under different  $p_{\max}$  (in percentage) in present-day Europeans. The dashed lines are the correlations of corresponding  $F_4$ -statistics on 1240K. (B)  $RASD(\text{England\_EMA}; \text{NL, IBS})$  with error bars of  $\pm 1$  SD, ascertained on sites with 0 – 0.6% derived allele frequency in present-day Europeans.

improving the accuracy of rare allele frequency estimates, and therefore the resolution of *RAS* statistics. Here, we have compared CNE and WBI ancestries and observed higher Z scores when grouping individuals (Figure 4) rather than analyzing them individually (Figure 3). For simulated data, we also group individuals to reduce noise (Figure 6).

In order to get a more complete picture of rare genetic variation, we have used shotgun data for ancient samples. We did not use capture data, due to the far lower amount and uneven distribution of rare alleles, although the raw sequence data still cover some neighboring rare allele sites. Fortunately, recent advances in sequencing technology have made shotgun sequencing more efficient and cost-effective, and shotgun data are becoming increasingly available for ancient DNA studies (Maisano Delser *et al.* 2021; Mallick *et al.* 2024), which coincides with the requirement of *RAS* for a larger sample size. The increasing availability of shotgun data is transforming the field of archaeogenetics as a whole by offering more detailed insights into human population history.

Western Eurasia serves as a prime example: Through the spread of early European farmers from the Near East (Lazaridis *et al.* 2014), and the movement of Indo-European speaking groups from the Eurasian steppe (Haak *et al.* 2015), populations



**Figure 6** Error estimate for modeling population 4 as an admixture of population 0 and 8. We rotated test individuals for 5 times and calculated the average of errors for 5 parallel tests.

that might have been more genetically distinct have gradually become more similar, leading to an overall more homogeneous genetic structure across Europe, which may have remained stable since the Iron Age (Antonio *et al.* 2024). Nonetheless, regional differences still persist, which are shaped by local history: during the Roman period, while Northern provinces maintained higher levels of local continuity, Southern sites absorbed the influences from Northern Africa, the Near East, and Eastern European Slavic groups, displaying more genetic variability (Antonio *et al.* 2019; Olalde *et al.* 2023); Celtic and Germanic tribes occupying different regions of Europe had different genetic profiles due to their different migration routes and interactions with different neighboring groups (McColl *et al.* 2025). Method development has so far relied on haplotype based analysis, such as IBD-based inference (McColl *et al.* 2025) or ancestral recombination graph inference (Speidel *et al.* 2025). In the future, using RAS, we will be able to additionally study rare variants that may be overlooked by other methods and expect more insights on subtle demographic events from these comprehensive datasets.

In our final demonstration, we have implemented a new way of estimating ancestry proportions, showing a better performance of RAS-based compared to ordinary *F*-statistics-based estimates (Figure 6). In the framework of *F*-statistics, there are extensions based on  $F_4$  matrices: qpWave for testing symmetry or external sources, and qpAdm for testing hypothetical admixture modeling (Patterson *et al.* 2012). In future work, RAS-statistics may be used similarly to develop formal tests for symmetry and admixture.

## Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 851511).

## Data Availability

All data used in this manuscript can be accessed from public resources. All modern genomes or allele frequency data (.vcf files) can be obtained from the following websites: the harmonized 1kGP+HGDP dataset at <https://gnomad.broadinstitute.org/downloads#v3-hgdp-1kg> (Koenig *et al.* 2024), allele frequency data from the GenomeDenmark project at <https://ega-archive.org/studies/EGAS00001002108> (Marety *et al.* 2017), allele frequency data from the Genome of the Netherlands (GoNL) project at <https://www.nlgenome.nl/menu/main/app-go-nl/download-data> (The Genome of the Netherlands Consortium 2014), allele frequency data from the SweGen project at <https://ega-archive.org/studies/EGAS50000000906> (Ameur *et al.* 2017). All ancient genomic data from Great Britain (.bam files or raw read data from the European Nucleotide Archive) from Martiniano *et al.* (2016); Schiffels *et al.* (2016); Gretzinger *et al.* (2022) as described in these publications.

## Code Availability

Software to estimate RAS, RASD and ancestry decomposition, as well as to prepare the datasets, can be found at [https://github.com/huanglei-artificium/RAS\\_tools](https://github.com/huanglei-artificium/RAS_tools). Documentation on how to use the software is available in the accompanying README file. The scripts used for data simulation are available at [https://github.com/Schiffels-Popgen/RAS\\_exploration](https://github.com/Schiffels-Popgen/RAS_exploration).

## Literature cited

- Allentoft ME, Sikora M, Refoyo-Martínez A, Irving-Pease EK, Fischer A, Barrie W, Ingason A, Stenderup J, Sjögren KG, Pearson A *et al.* 2024. Population genomics of post-glacial western Eurasia. *Nature*. 625:301–311. Publisher: Nature Publishing Group.
- Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, Viklund J, Kähäri AK, Lundin P, Che H *et al.* 2017. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *European Journal of Human Genetics*. 25:1253–1260. Publisher: Nature Publishing Group.
- Antonio ML, Gao Z, Moots HM, Lucci M, Candilio F, Sawyer S, Oberreiter V, Calderon D, Devitofranceschi K, Aikens RC *et al.* 2019. Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science*. 366:708–714. Publisher: American Association for the Advancement of Science.
- Antonio ML, Weiß CL, Gao Z, Sawyer S, Oberreiter V, Moots HM, Spence JP, Cheronet O, Zagorac B, Praxmarer E *et al.* 2024. Stable population structure in Europe since the Iron Age, despite high mobility. *eLife*. 13:e79714. Publisher: eLife Sciences Publications, Ltd.
- Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J *et al.* 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science*. 367. Publisher: American Association for the Advancement of Science Section: Research Article.

10 Rare variants and population structure

- 1 Busing FMTA, Meijer E, Leeden RVD. 1999. Delete-m Jackknife  
2 for Unequal m. *Statistics and Computing*. 9:3–8.
- 3 Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M.  
4 2013. Robust Demographic Inference from Genomic and SNP  
5 Data. *PLOS Genetics*. 9:e1003905. Publisher: Public Library of  
6 Science.
- 7 Flegontov P, Altınışik NE, Changmai P, Rohland N, Mallick S,  
8 Adamski N, Bolnick DA, Broomandkhoshbacht N, Candilio F,  
9 Culleton BJ et al. 2019. Palaeo-Eskimo genetic ancestry and the  
10 peopling of Chukotka and North America. *Nature*. 570:236–  
11 240. Number: 7760 Publisher: Nature Publishing Group.
- 12 Flegontov P, İşıldak U, Maier R, Yüncü E, Changmai P, Reich D.  
13 2023. Modeling of African population history using f-statistics  
14 is biased when applying all previously proposed SNP ascer-  
15 tainment schemes. *PLOS Genetics*. 19:e1010931. Publisher:  
16 Public Library of Science.
- 17 Gretzinger J, Sayer D, Justeau P, Altena E, Pala M, Dulias K,  
18 Edwards CJ, Jodoin S, Lacher L, Sabin S et al. 2022. The  
19 Anglo-Saxon migration and the formation of the early English  
20 gene pool. *Nature*. 610:112–119. Publisher: Nature Publishing  
21 Group.
- 22 Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante  
23 CD. 2009. Inferring the Joint Demographic History of Multi-  
24 ple Populations from Multidimensional SNP Frequency Data.  
25 *PLOS Genetics*. 5:e1000695. Publisher: Public Library of Sci-  
26 ence.
- 27 Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas  
28 B, Brandt G, Nordenfelt S, Harney E, Stewardson K et al. 2015.  
29 Massive migration from the steppe was a source for Indo-  
30 European languages in Europe. *Nature*. 522:207–211. Pub-  
31 lisher: Nature Publishing Group.
- 32 Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical  
33 Learning*. Springer Series in Statistics. Springer. New York, NY.
- 34 Ihaka R, Gentleman R. 1996. R: A Language for Data Analy-  
35 sis and Graphics. *Journal of Computational and Graphical  
36 Statistics*. 5:299–314.
- 37 Kamm J, , Jonathan T, , Richard D, , Song YS. 2020. Efficiently  
38 Inferring the Demographic History of Many Populations With  
39 Allele Count Data. *Journal of the American Statistical As-  
40 sociation*. 115:1472–1487. Publisher: ASA Website \_eprint:  
41 <https://doi.org/10.1080/01621459.2019.1635482>.
- 42 Kelleher J, Etheridge AM, McVean G. 2016. Efficient Coalescent  
43 Simulation and Genealogical Analysis for Large Sample Sizes.  
44 *PLOS Computational Biology*. 12:e1004842. Publisher: Public  
45 Library of Science.
- 46 Koenig Z, Yohannes MT, Nkambule LL, Zhao X, Goodrich JK,  
47 Kim HA, Wilson MW, Tiao G, Hao SP, Sahakian N et al. 2024.  
48 A harmonized public resource of deeply sequenced diverse  
49 human genomes. *Genome Research*. 34:796–809. Company:  
50 Cold Spring Harbor Laboratory Press Distributor: Cold Spring  
51 Harbor Laboratory Press Institution: Cold Spring Harbor Lab-  
52 oratory Press Label: Cold Spring Harbor Laboratory Press  
53 Publisher: Cold Spring Harbor Lab.
- 54 Lazaridis I, Alpaslan-Roodenberg S, Acar A, Açıkkol A, Age-  
55 larakis A, Aghikyan L, Akyüz U, Andreeva D, Andrijašević  
56 G, Antonović D et al. 2022. The genetic history of the South-  
57 ern Arc: A bridge between West Asia and Europe. *Science*.  
58 377:eabm4247. Publisher: American Association for the  
59 Advancement of Science.
- 60 Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kir-  
61 sanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M  
62 et al. 2014. Ancient human genomes suggest three ancestral  
populations for present-day Europeans. *Nature*. 513:409–413.  
Number: 7518 Publisher: Nature Publishing Group.
- 63 Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ra-  
64 machandran S, Cann HM, Barsh GS, Feldman M, Cavalli-  
65 Sforza LL et al. 2008. Worldwide Human Relationships In-  
66 ferred from Genome-Wide Patterns of Variation. *Science*.  
67 319:1100–1104. Publisher: American Association for the Ad-  
68 vancement of Science.
- 69 Lipson M. 2020. Applying f4-statistics and ad-  
70 mixture graphs: Theory and examples. *Molec-  
71 ular Ecology Resources*. 20:1658–1667. \_eprint:  
72 <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13230>.
- 73 Liu Y, Mao X, Krause J, Fu Q. 2021. Insights into human history  
74 from the first decade of ancient human genomics. *Science*.  
75 373:1479–1484. Publisher: American Association for the Ad-  
76 vancement of Science.
- 77 Maier R, Flegontov P, Flegontova O, İşıldak U, Changmai P,  
78 Reich D. 2023. On the limits of fitting complex models of  
79 population history to f-statistics. *eLife*. 12:e85492. Publisher:  
80 eLife Sciences Publications, Ltd.
- 81 Maisano Delser P, Jones ER, Hovhannisyan A, Cassidy L, Pinhasi  
82 R, Manica A. 2021. A curated dataset of modern and ancient  
83 high-coverage shotgun human genomes. *Scientific Data*. 8:202.  
84 Publisher: Nature Publishing Group.
- 85 Mallick S, Micco A, Mah M, Ringbauer H, Lazaridis I, Olalde I,  
86 Patterson N, Reich D. 2024. The Allen Ancient DNA Resource  
87 (AADR) a curated compendium of ancient human genomes.  
88 *Scientific Data*. 11:182. Publisher: Nature Publishing Group.
- 89 Marety L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P,  
90 Skov L, Belling K, Theil Have C, Izarzugaza JMG et al. 2017.  
91 Sequencing and de novo assembly of 150 genomes from Den-  
92 mark as a population reference. *Nature*. 548:87–91. Publisher:  
93 Nature Publishing Group.
- 94 Martiniano R, Caffell A, Holst M, Hunter-Mann K, Montgomery  
95 J, Müldner G, McLaughlin RL, Teasdale MD, van Rheenen  
96 W, Veldink JH et al. 2016. Genomic signals of migration and  
97 continuity in Britain before the Anglo-Saxons. *Nature Com-  
98 munications*. 7:10326. Publisher: Nature Publishing Group.
- 99 Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N,  
100 Roodenberg SA, Harney E, Stewardson K, Fernandes D, Nov-  
101 ak M et al. 2015. Genome-wide patterns of selection in 230  
102 ancient eurasians. *Nature*. .
- 103 McColl H, Kroonen G, Moreno-Mayar JV, Seersholm FV, Scor-  
104 rano G, Pinotti T, Vimala T, Sindbæk SM, Ethelberg P, Fyfe R  
105 et al. 2025. Steppe Ancestry in Western Eurasia and the Spread  
106 of the Germanic Languages. Pages: 2024.03.13.584607 Section:  
107 New Results.
- 108 Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S,  
109 Willerslev E. 2017. Tracing the peopling of the world through  
110 genomics. *Nature*. 541:302–310. Publisher: Nature Publishing  
111 Group.
- 112 Nobembre J, Stephens M. 2008. Interpreting principal com-  
113 ponent analyses of spatial population genetic variation. *Nature  
114 Genetics*. 40:646–649. Publisher: Nature Publishing Group.
- 115 Olalde I, Carrión P, Mikić I, Rohland N, Mallick S, Lazaridis  
116 I, Mah M, Korać M, Golubović S, Petković S et al. 2023. A  
117 genetic history of the Balkans from Roman frontier to Slavic  
118 migrations. *Cell*. 186:5472–5485.e9. Publisher: Elsevier.
- 119 Orlando L, Allaby R, Skoglund P, Der Sarkissian C, Stockham-  
120 mer PW, Ávila Arcos MC, Fu Q, Krause J, Willerslev E, Stone  
121 AC et al. 2021. Ancient DNA analysis. *Nature Reviews Meth-*
- 122 123 124

- ods Primers. 1:14. Publisher: Nature Publishing Group.
- Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient Admixture in Human History. *Genetics*. 192:1065–1093.
- Peter BM. 2016. Admixture, Population Structure, and F-Statistics. *Genetics*. 202:1485–1501.
- Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford Jr TW, Orlando L, Metspalu E *et al.* 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*. 505:87–91. Number: 7481 Publisher: Nature Publishing Group.
- Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N *et al.* 2012. Reconstructing Native American population history. *Nature*. 488:370–374. Number: 7411 Publisher: Nature Publishing Group.
- Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. *Nature*. 461:489–494. Number: 7263 Publisher: Nature Publishing Group.
- Schiffels S, Haak W, Paajanen P, Llamas B, Popescu E, Loe L, Clarke R, Lyons A, Mortimer R, Sayer D *et al.* 2016. Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nature Communications*. 7:10408. Publisher: Nature Publishing Group.
- Skoglund P, Mathieson I. 2018. Ancient Genomics of Modern Humans: The First Decade. *Annual Review of Genomics and Human Genetics*. 19:381–404. Publisher: Annual Reviews.
- Speidel L, Silva M, Booth T, Raffield B, Anastasiadou K, Barrington C, Götherström A, Heather P, Skoglund P. 2025. High-resolution genomic history of early medieval Europe. *Nature*. 637:118–126. Publisher: Nature Publishing Group.
- Stoneking M, Arias L, Liu D, Oliveira S, Pugach I, Rodriguez JJRB. 2023. Genomic perspectives on human dispersals during the Holocene. *Proceedings of the National Academy of Sciences*. 120:e2209475119. Publisher: Proceedings of the National Academy of Sciences.
- The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature*. 526:68–74. Number: 7571 Publisher: Nature Publishing Group.
- The Genome of the Netherlands Consortium. 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*. 46:818–825. Publisher: Nature Publishing Group.

### Supplementary Figure Captions

Figure S1: Outgroup- $F_3$  statistics on test individual  $x$  and a specific reference population ( $R_6$  or  $R_8$ ) at different migration rates on 1240K SNP panel. The migration rates ( $4mN_e$ ) used in the simulation are 1 (A), 100 (B) and 2000 (C). Test individuals  $x$  are distinguished by colors representing different populations, shown in legend in (A), which also includes the schematic of simulated population migration.

Figure S2: The misclassification rate for different migration rates.

Figure S3: *RASD(England\_EMA; FIN, IBS)* (A) and *RASD(England\_EMA; SE, IBS)* (B) with error bars of  $\pm 1$  SD, ascertained on sites with 0 – 0.6% derived allele frequency in present-day Europeans.

Figure S4: Error estimate for modeling population 4 as an admixture of population 0 and 8. We rotated test individuals for 5 times and did 5 parallel tests. The coordinates of errors are logarithmic.