

RASpop - Rare variant analysis for population structure

Stephan Schiffels and Thiseas Lamnidis

March 2021

Introduction

One of the primary challenges in population genetics in general, and in archaeogenetics in particular, is analysing population structure. The challenge mostly consists of analysing extremely high-dimensional data (millions of markers in thousands of individuals) and making meaningful claims about structure between groups, admixture proportions, or the number of independent gene flow events in multiple populations.

With archaeogenetic, there is the additional challenge that ancient genomes have very specific batch effects caused by DNA damage, low coverage and missingness.

Most existing tools to study population structure are based on allele frequencies of variants that have been segregating in several human populations for a long time (potentially hundreds of thousands of years). This means that all genetic differences seen between populations are caused by genetic drift, which has a relatively slow time scale, limiting the ability to detect population structure to populations separated by at least several thousand years.

Here, we propose a new set of methods to analyze population structure based on rare genetic variants, and how they are shared between samples and populations. Specifically, below we set out to define statistical methods that mirror certain methods that have been developed for allele frequency statistics:

Based on allele-frequency differences	Based on rare allele sharing
MDS / PCA	RASmds / RAS_ReferenceMDS
D / F4 statistics	RAS-F4 (RASDA?) / RAS_ReferenceF4
qpWave	RASwave
qpAdm	RASadm

Rare Allele Sharing

Ascertainment

To ascertain variants as “rare,” we typically consider a high quality modern reference dataset, such as the 1000 Genomes dataset (1000 Genomes Project Consortium et al. 2015) or the HGDP dataset (Bergström et al. 2020). Within these datasets, we consider variants as “rare” if they occur with total allele count up to a specific maximal count, for example 2, 3, 4 or 5.

Without loss of generality, for rare alleles ascertained in this way, we consider the rare variant as the “derived” mutation, and the common variant as the “ancestral mutation,” even in cases where the human reference genome carries the derived variant.

Importantly, once we select such a reference dataset and ascertained rare alleles, we consider rare allele sharing between any two individuals or groups, even from outside the reference dataset. Particularly, if we

have ancient genomes that we study, we consider their sharing against any of the populations in the reference panel, even though the ancient genomes themselves are not included in the ascertainment. Furthermore, we can consider rare allele sharing between two ancient genomes in this way, by counting how many rare variants they share, even though those variants have been identified in the modern reference population.

An ascertainment of rare variants in a reference panel with maximum allele count m results in a set of genomic positions at which variants with allele count up to m occur. We denote that set as \mathcal{G} .

RAS(i, j)

The key quantity to measure then, is not the amount of genetic *differentiation* between individuals or groups (as done with allele frequency differences), but the amount of *similarity* based on sharing rare variants. Specifically, we define “Rare Allele Sharing” (termed RAS) as the core quantity between individuals or groups. Consider two individuals or populations i and j . We then define

$$\text{RAS}_m(i, j) = \langle x_i x_j \rangle_{\mathcal{G}}$$

where x_i denotes the allele frequency of individual or group i , and the average $\langle \cdot \rangle$ is defined as average across all sites in \mathcal{G} (note even though we use allele counts as integers, we still use allele frequencies as fractional numbers between 0 and 1). RAS will typically be very small, because it is the average of products of very low frequencies, but we are ultimately not interested in absolute but in relative numbers.

Missing Data

Note that the definition of RAS extends naturally towards missing data, either partially (in groups) or completely (as in low coverage individuals). Partial missingness simply results in less accurate allele frequency estimates x , which we consider as contributing unbiased noise. Complete missingness will simply reduce the set \mathcal{G} accordingly, with the overall estimate still being an unbiased estimate of the full RAS.

Symmetric analyses

We first consider the simplest case of analysing only populations or individuals within a reference dataset themselves. In other words, the individuals that serve for ascertainment of rare alleles are also the ones being studied. In the following, we consider n_R to be the number of individuals in the reference dataset.

RASmds - Multidimensional Scaling

We consider the matrix $r_{ij} = \text{RAS}_5(i, j)$ of RAS between all individuals within the reference dataset, with $i, j = 1 \dots n_R$. Note that this in principle is fast to compute, since for every SNP, we only have to consider at most 10 different pairs of individuals with an allele frequency other than zero, to contribute to the overall sharing matrix (see section “Implementation Details”).

Matrix r_{ij} is a *similarity matrix*, and in order to use multidimensional scaling to visualise it in a two dimensional scatter plot, we convert it to a distance matrix $d_{ij} = 1 - r_{ij}$. This distance matrix can then be used to compute principal components, for example using the R function `stats::cmdscale`.

We call this approach “RASmds,” although it’s really not a new invention, but simply a MDS visualization of pairwise rare allele sharing.

Projection of additional samples into an existing RASmds

Let's consider the case where we have n_x extra samples outside of the reference dataset. For example, one can imagine ancient genomes for this. For each of these additional samples, we have $RAS(i, k)$ with $i = 1 \dots n_R$ and $k = 1 \dots n_x$. In principle, it should be possible to "project" a sample k into the existing RASmds plot, by assigning it coordinates in a two-dimensional space such that the euclidian distances to all samples i in that space match most closely the relative similarities represented by $RAS(i, k)$. It is not entirely clear to me how to do that, but I believe it must be a solved problem. See for example [this link](#). **To be explored!**

RASf4 - Cladality Tests

One of the most widely used statistics used to analyse population structure is the D-test, or F4-statistics. It essentially tests whether one individual or group A is closer to an individual or group C than some other group or individuals B. Originally, it is defined as

$$F4(A, B; C, O) = \langle (x_A - x_B)(x_C - x_O) \rangle$$

where O is an outgroup, needed for technical reasons, and x_A, x_B etc. are allele frequencies. It is illuminating to express this statistics in terms of so-called F3 statistics:

$$F4(A, B; C, O) = F3(A, C; O) - F3(B, C; O)$$

with

$$F3(A, C; O) = \langle (x_A - x_O)(x_C - x_O) \rangle = \langle x_A x_C \rangle + \dots$$

where the dots represent terms that depend on allele sharing with the outgroup, which - if it is a true outgroup - will be the same for all groups A, B and C, so we ignore it here. We recognise now that $F3(A, C; O)$ is in its definition equivalent to $RAS(A, C)$, except for the fact that RAS is only measured on rare variants, while $F3$ is typically computed across common variants.

So F4 statistics are simply differences of allele sharing. They are expected to be close to zero for cases in which populations A and B are equally distantly related to population C, while positive values indicate that A is closer to C than B is to C, indicating for example gene flow from C into A after the split between A and B.

Based on these considerations, we now define $RASf4$ as the difference between two RAS-statistics:

$$RASf4(A, B, C) = RAS(A, C) - RAS(B, C)$$

with a similar interpretation of zero, negative and positive values as for the standard F4. Arguably, however, $RASf4$ will be more sensitive towards *recent* gene flow events. Arguably, $RASf4$ might even switch signs compared to standard F4 if the population history is complex, with ancient and recent gene flow pushing into different directions. **To be explored!**

Exploration via Simulations

Exploration via real data

Asymmetric analyses

RAS_ReferenceMDS - Multidimensional Scaling

RAS_ReferenceF4 - Cladality Tests

RASwave - Admixture rank analyses

RASadm - Admixture decomposition

Implementation Details

File Formats

Tools

References

1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, Coriell Institute for Medical Research, Max Planck Institute for Molecular Genetics, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. <http://www.nature.com/doifinder/10.1038/nature15393>.

Bergström, Anders, Shane A McCarthy, Ruoyun Hui, Mohamed A Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, et al. 2020. “Insights into Human Genetic Variation and Population History from 929 Diverse Genomes.” *Science* 367 (6484). <https://doi.org/10.1126/science.aay5012>.