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.

In archaeogenetics, 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_ReferencePCA
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 k results in a set of genomic positions at which variants with allele count up to k 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

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

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.

Jackknifing

For our basic statistics $\mathrm{RAS}_k(i,j)$ we can naturally estimate error bars, by evaluating the statistics for genomic blocks (e.g. chromosomes) and then use weighted block jackknifing to estimate errors (Busing, Meijer, and Van Der Leeden 1999).

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 $r_ij = RAS_5(i,j)$ of RAS between all individuals within the reference dataset, with $i,j = 1 \ldots n_R$. Note that this in principle is fast to compute, since for every SNP, we only have to consider at most $\binom{5}{2} = 10$ 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 (exploratory)

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" another sample l 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 $RAS(i,l)$. 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!

UMAP

Just to mention it, of course given an existing RASmds analysis, one can always compute many more components, say 10, and feed them to a UMAP algorithm to gain ideas about more global structure.

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

$$F3(A,C;O) = \langle (x_A - x_O)(x_C - x_O) \rangle = \langle x_A x_C \rangle + \text{Outgroup-dependent terms}$$

which shows that Outgroup-F3 statistics are essentially again allele frequency products, plus some additional terms which depend on the relationships of our populations to the outgroup. As long as we can assume that our outgroup is a true outgroup (such as - say - Chimpanzee for humans), we can ignore these additional terms for how F3 is used in F4. 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:

$$RASf4_k(A,B,C) = RAS_k(A,C) - RAS_k(B,C)$$

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 (tunable by the maximum allele count m). 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

This would be a good place to add your current results in, Thiseas.

Exploration via real data

We can explore rare allele sharing in both the 1000 Genomes and the HGDP dataset. I propose the following analyses:

1) RASmds/UMAP for all $\text{RAS}_k(i, j)$ statistics, with (i, j) running over all possible pairs of individuals.
2) RASmds/UMAP for selected groups, such as Western-Eurasian, Eastern-Eurasian, American and African
3) Group all individuals into populations, either based on context information or even PCA/RASmds clustering (take note of outliers). Then, pick a focal population n and list all $\text{RAS}_k(n, m)$ for all other populations m, ordered by value, including error bars. Then, compare with ordinary F3. The idea here is to find out whether the RAS_k gives stronger differences in the ordering of "closeness" to other populations than ordinary F3. I suggest to do this analysis for several focal groups, perhaps one from each continent.
4) Based on 3), pick pairs of very closely related populations n and m, and explore $\text{RASf4}_k(n, m, l)$ with all other populations l. Then compare to ordinary F4. The idea here is to see whether cladality tests with other groups are more sensitive for RASf4_k than for ordinary F4.

Asymmetric analyses

In the asymmetric case, we also make use of the large reference datasets analysed above (1000 Genomes and HDGP), but this time we focus on a list of so-called test individuals or populations *outside* of this reference set. This foremost affects the ascertainment, which will now be based only on the reference dataset, but not the test individuals outside.

Most use cases will involve ancient genomes as test individuals, for which we have insufficient coverage or are worried about calling biases with respect to the variant calls from 1000 Genomes or HGDP. However, note that - importantly - most use cases might also include genomes that are *taken out* of the reference datasets to be included as test individuals. This will then affect the ascertainment, which should exclude all test individuals.

In general, we denote the number of reference individuals as n_R as before, and the number of test individuals as n_T. We now also explicitly consider a grouping of the reference dataset into n_P "reference populations," which are created based on context and/or genetic clustering.

The starting point for all asymmetric analysis is again a matrix $\text{RAS}_k(i, j)$, where $i = 1 \dots n_T$ loops over the test individuals, and $j = 1 \dots n_R$ loops over the reference individuals. In many cases, we will use the grouped version of the reference datasets, so we'd then switch to a matrix $\text{RAS}_k(i, m)$ where $m = 1 \dots n_P$.

Because the matrix $\text{RAS}_k(i, m)$ is such a central starting point for most analyses, we give it a more prominent name and call it "RAS table." Each row of this table contains an individual, and each column a specific reference populations:

	French	Yoruba	Chukchi	...
Test Individual #1	0.0003	0.00001	0.000002	...
Test Individual #2	0.0001	0.00002	0.000004	...
Test Individual #3	0.0005	0.00003	0.000001	...
...	...	...	...	...

where we've omitted error bars for simplicity, which of course exist for every numeric entry in the table.

RAS_ReferencePCA - PCA based on RAS with reference populations

The first thing we can do with a RAS-table is to use principal components analysis to plot all test individuals along the two primary axis of variation in RAS values with the reference populations. This should be straight forward and follow standard procedures for multivariate analyses. There is no conceptual difference between a PCA based on 1.2 million genetic variants per individual or a PCA based on 20 or so RAS values with reference populations.

Note that this type of analysis would ignore the error bars for RAS values in the table.

RAS_ReferenceF4 - Cladality Tests based on RAS with reference populations

Another useful analysis could be to test for cladality of two test individuals (i and j) with respect to a single reference population m, which we denote again as $\mathrm{RASf4}_k(i,j,m)$. Note that it is critical in this test that i and j are either *both excluded* from the ascertainment (so in the "test individuals") or *both included*, as above in the symmetric case.

RASwave - Generalization of RAS_ReferenceF4

As a generalization of RAS_ReferenceF4 we can consider the entire RAS-table from above to perform a cladality test via a Chi-squared test. Specifically, we can ask the question whether two rows in the table share the same "histogram" of RAS values with all the reference populations. So consider two test samples i and j, and their entire "rows" of RAS values $\mathrm{RAS}_k(i,m)$ and $\mathrm{RAS}_k(i,m)$, with $m = 1 \ldots n_P$, then one can construct a chi-square test to compare those two rows. Specifically, one would calculate

$$\chi^2 = \sum_m^{n_P} \left(\frac{\mathrm{RAS}_k(i,m) - \mathrm{RAS}_k(j,m)}{\sqrt{\Delta\mathrm{RAS}_k(i,m)^2 + \Delta\mathrm{RAS}_k(j,m)^2}} \right)^2$$

where $\Delta\mathrm{RAS}_k(\cdot,m)$ is the jackknife standard error, and the denominator essentially rescales the observed differences between the two statistics in the numerator with the expected standard error of that difference.

Under the null-hypothesis, we expect χ^2 to be chi-squared distributed with n_P degrees of freedom, so a standard tail test will yield a p-value for that null-hypothesis.

As an even further generalization, we can try to test for more complex cases with more test individuals, similar to qpWave (ADMIXTOOLS package), but that will involve a bit more involved linear algebra which I haven't yet written up.

RASadm - Admixture decomposition based on RAS with reference populations

Exploration via real data

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.

Busing, Frank M T A, Erik Meijer, and Rien Van Der Leeden. 1999. "Delete-m Jackknife for Unequal m." *Statistics and Computing* 9 (1): 3–8. http://link.springer.com/article/10.1023/A:1008800423698.