1000 Genomes Population Allele Frequency Difference File Specification (v1)

Andrew Wood
A.R.Wood@exeter.ac.uk

June 12, 2024

1 Motivation

This document describes a file format generated to hold chromosome-specific allele frequency differences across different genetic ancestry groups for 78,122,255 variants captured in 2,548 individuals from the 1000 Genomes Project (build 38), available here. Given the number of data points in the triangular matrix, this file format does not store differences in allele frequency = 0 between pairs of ancestry groups. Furthermore, data related to frequency differences have been stored in binary format to further reduce file sizes to assist in the sharing of this data. This data should be decompressed prior to reading if downloaded as a .gz file.

2 File set

2.1 Ancestry-pair index file (.pop)

This file is a text file with no header line, and one line per ancestry-pair for which differences in allele frequencies have been calculated. Ancestry-pairs have been derivied based on:

- 5 broad genetic ancestry groups (AFR, AMR, EAS, EUR, SAS)
- 26 sub-ancestries spanning the 5 broad ancestries.

The columns represent:

1. Ancestry-pair index
2. Ancestry-pair label

Note the order of the ancestry labels within a pair reflect the order of terms in the subtraction of frequencies. For example, `gbr_tsi` represents frequency differences based on subtracting the allele frequency in `TSI` from the allele frequency in `GBR`.

2.2 Variant list file (.var)

This file is a text file with no header line, and one line per variant for which allele frequency differences have been calculated. This file comprises of the following columns:

1. Variant ID
2. Allele 1 - the allele used to calculate frequency differences between ancestry pairs
3. Allele 2 - the other allele
4. Byte address of first byte related to the variant in the `.fd` file
5. Number of non-zero frequency differences calculated

Note, a variant may share the same first byte address as another variant if all frequency differences between ancestry groups = 0. This is necessary so that all variants can be looked up and accounted for. The logic implemented for reading the file will account for the number of non-zero frequency differences for a variant when cycling through the bytes associated with it (see below).

2.3 Population frequency difference file (.bin)

This is a non-human readable binary file storing data related to differences in allele frequencies between different genetic ancestry groups in the 1000 Genomes project. Bytes are in little-endian order.

2.3.1 The first 7 bytes

The first 7 bytes contains a 'magic number' and a label to enable existing programs developed to utilse this data to check expected input prior to use.

Byte address	No. bytes	Description
0	6	Magic Number `'ARWOOD'`
6	1	File format (value = 1)

2.3.2 The remaining bytes in the file

The remaining bytes consist of a repeating sequence of a 2-byte unsigned integer that encodes the index of the ancestry-pair for which an allele frequency difference is >0, and a 4-byte signed float that encodes the calculated allele frequency difference. The starting byte for a given variant can be looked up using the `.var` file. The ancestry pair can be looked up using the `.pop` file. The total number of bytes representing allele-frequency differences can be calculated as `6*N non-zero frequency differences`, available from the `.var` file.

Byte address	No. bytes	Description
Starting byte address for variant X	2	Unsigned integer: ancestry pair index
Starting byte address for variant X+2	4	Signed float: allele frequency difference
...	...	...

3 Notes

The human-readable version of the data as a full triangular matrix containing all for ancestral-group pairwise allele frequency differences is 124Gb (and .gz compressed). By contrast, the total size of the data stored using the described file specification is 45Gb (and 9.5Gb .gz compressed).

An example Python script (`GetFreqDiffs.py`) and command is provided here to enable you to extract allelel frequency differences among individuals across ancestry-group pairs. The input file containing the list of variants to extract data for is a human readable tab-delimited text file (without header), with the following columns:

- chromosome
- base-pair position (build 38)
- allele to align frequency differences to (this can be random or, for example, trait raising)
- other allele

Example command to run Python script for variants in the subset of 12111 GIANT height SNP list on chromosome 1. Note the first allele in the file is the height raising allele:

```
python GetFreqDiffs.py \
   --freq-diffs 1000g_build38_pop_freq_diffs_1 \
   --vars 12111_height_snps.txt \
   --out 12111_height_snps_1000g_freq_diffs.txt
```