ADMIXFRQ

Version 1 user manual

Authors:
Rafael A. Nafikov
Ellen M. Wijsman

Division of Medical Genetics
Department of Medicine
University of Washington

Last modified on: June 4, 2018

Contents

1. Introduction

1.1 Citing ADMIXFRQ

If you use the ADMIXFRQ Perl script (ADMIXFRQ.pl) or any other scripts associated with the ADMIXFRQ in your data analysis and intend to publish your work, please cite the following reference:

> Nafikov, R. A., Nato, A. Q., Jr., Sohi, H., Wang, B., Brown, L., Horimoto, A. R., Vardarajan, B. N., Barral, S. M., Tosto, G., Mayeux, R. P., Thornton, T. A., Blue, E., & Wijsman, E. M. (2018). Analysis of Pedigree Data in Populations with Multiple Ancestries: Strategies for Dealing with Admixture in Caribbean Hispanic Families from the ADSP. *Genetic Epidemiology*, 1-16. doi.org/10.1002/gepi.22133

1.2 License

The ADMIXFRQ Perl script is licensed under the GNU General Public License, version 3 which can be found at the following URL: https://www.gnu.org/licenses/gpl-3.0.en.html.

1.3 Software URL

The ADMIXFRQ Perl script (ADMIXFRQ.pl) can be downloaded from: https://github.com/RafPrograms/ADMIXFRQ. The manual, an example-template parameter file, and supporting scripts for the ADMIXFRQ can be downloaded at the URL specified above as well.

1.4 General information about ADMIXFRQ

The ADMIXFRQ script can specify allele frequencies for markers listed in the '**chrN.geno**' files and used in multipoint analyses as described in Nafikov et al. (2018) and/or for markers provided by a user that can be used for any type of analysis where accurate allele frequency specification in admixed samples matters.

1.5 Computing hardware requirements

The ADMIXFRQ Perl script can be computer memory intensive depending on the size of your data. For example, the specification of allele frequencies for the whole genome sequence data and for over 5,000 linkage markers in one admixed family genome-wide using four different admixture models and all available options in the script requires about 19 GB of active computer memory. The run time for this analysis with just 1 CPU is slightly less than 3 hrs.

2. Running ADMIXFRQ

2.1 Command line to run ADMIXFRQ

The ADMIXFRQ Perl script can be run from a terminal using either:

perl ADMIXFRQ.pl parameters.txt

or

./ADMIXFRQ.pl parameters.txt

command line. A parameter file can have any name. To speed up computation, you can run this script in parallel on multiple CPUs by creating a separate parameter file for each of the runs.

2.2 Parameter file

A parameter file is a space or tab delimited text file which contains all the necessary information to run the ADMIXFRQ Perl script. An example-template parameter file, **parameters.txt,** is provided as a separate text file along with the script and can be used by users to create their own parameter files suited to their particular needs. In a parameter file empty lines or lines starting with a hash sign **#** are ignored. Any entries on a line with text after a hash **#** are ignored as well. Each line of a parameter file contains specifications for an option either alone or together with a few sub-options. Each option and sub-option has its corresponding entry separated by either equal sign **=** or colon **:** . On a single line sub-options are separated from an option or each other by a vertical slash **|** . Arrangement of an option and corresponding sub-options on a single line was dictated by their interrelatedness. The underlined capital letter **N** within an option or sub-option statement and inside file names is used to indicate a place for a number.

Please be consistent with how you name your populations and families across different options in a parameter file and across all input data files.

Parameter File Options:

- **CHROMOSOMES = [1 2 3 7 9]**

  This option specifies chromosome numbers for which a computation will be performed. Chromosome numbers separated by space can be entered inside square brackets. Alternatively, consecutive chromosome numbers can be specified inside square brackets as a range of numbers using two dots **..** and no spaces between the dots and numbers (e.g., **[1..5]** means chromosomes 1, 2, 3, 4, and 5). It is also allowed to use the two specification formats for chromosome numbers mentioned above simultaneously (e.g., **[1..5 9 13 17]**).

- **FAMILY NAMES = [family1 family2 family3]**
  This option specifies family names for which a computation will be performed. Family names separated by space can be entered inside square brackets.

- **ADMIXTURE MODEL = Global FBCW FBGW Local | AVERAGE TYPE FOR GLOBAL AND FBGW = normal** or **weighted**
  This option is used to specify admixture models to be used in a computation. The available options for admixture model specification are: **Global FBCW FBGW Local**. Multiple models can be used if they are specified together and separated by space. A definition of each of the models can be found in Nafikov et al. (2018). Options for different admixture models were provided for exploratory purposes. ***For data analysis we strongly recommend using only the Local admixture model because the results from our paper along with past work on the effects of haplotype frequency misspecification indicate that this model provides the most appropriate results for the analysis.***

  The sub-option "**AVERAGE TYPE FOR GLOBAL AND FBGW**" is used to specify a type of averaging (normal or weighted) to be used in a computation of admixture proportions from local ancestry estimates for the Global and FBGW admixture models. Here, the "normal" averaging means computing mean admixture proportions without taking into consideration differences in number of markers across chromosomes for which local ancestry estimates were computed. ***We recommend using only the "weighted" averaging option.*** However, when local ancestry estimates for a genome-wide panel of dense markers are available, the choice of averaging type does not matter. The "normal" averaging option is only provided for exploratory purposes because this is how mean admixture proportions were computed for the Global and FBGW admixture models in our study (Nafikov et al., 2018).

- **EXTRAPOLATION OF LOCAL ADMIXTURE PROPORTIONS: yes** or **no**
  This option is used to specify whether extrapolation of admixture estimates beyond chromosomal boundaries defined by marker positions with admixture data will be performed or not. It is recommended to extrapolate if you have markers with admixture estimates spanning the entire chromosome. In this case admixture proportions at the first and last marker positions with admixture data will be assigned to the rest of your marker data below and above the defined boundaries. In situations, when markers with admixture estimates do not cover your entire chromosomes, the extrapolation of admixture estimates is not recommended.

- **GLOBAL ADMIXTURE MODEL PROPORTIONS: population1 = 0.64 | population2 = 0.27 | population3 = 0.09**
  This option is ***optional*** but ***becomes required*** if the Global admixture model is used and the "**COMPUTE GLOBAL ADMIXTURE MODEL PROPORTIONS**" option is not specified. The option above can be used to specify admixture proportions for reference populations assumed to be involved in admixture under the Global admixture model. The specified admixture proportions will be used in calculation of marker allele frequencies under the Global admixture model. If however the "**COMPUTE GLOBAL ADMIXTURE MODEL PROPORTIONS: yes**" option is specified, it will override the "**GLOBAL ADMIXTURE MODEL PROPORTIONS**" option.

- **COMPUTE GLOBAL ADMIXTURE MODEL PROPORTIONS: yes** or **no**
  This option can be used to force computation of admixture proportions under the Global admixture model. This computation requires genome-wide local ancestry estimates for all the families under consideration. The specification of this option will override the statements made with the "**GLOBAL ADMIXTURE MODEL PROPORTIONS**" option.

- **ADMIXTURE PROPORTIONS PRINTOUT: yes**
  This option is used to output admixture proportions for specified admixture models into separate files for plotting or exploratory purposes. By default, the ADMIXFRQ script will not perform this step. The admixture proportions for the Global admixture model will be printed out only if the "**COMPUTE GLOBAL ADMIXTURE MODEL PROPORTIONS: yes**" option is specified.

  <u>These are a few general sub-options used multiple times with different option statements described below:</u>
  - **FILE PATH** – specifies a location for a particular file on your computer. A file path can contain the plus sign **+** and the asterisk ***** used as space holders for a family name or a chromosome number, respectively. During a run, the ADMIXFRQ script will loop through family and chromosome names specified by the "**CHROMOSOMES"** and **"FAMILY NAMES"** options in a parameter file and substitute each of the above symbols with the corresponding name and/or number.
  - **HEADER** - can take "**F**" or "**T**" letters which stand for the absence (**F**alse) or presence (**T**rue) of a header in your file.
  - **BP_POSITION**, **CM_POSITION**, **RS_NUMBER** – specify a column number in a file where data associated with a particular sub-option is located.

- **GENO FILE FOR GL_AUTO: FILE PATH = /your_directory/+/chr*/chr*.geno | DETAILED ALLELE FREQUENCY FILE = yes** or **no**
  1.6 This option requests marker allele frequency modifications in the '**chr<u>N</u>.geno**' (**<u>N</u>** stands for a chromosome number) files subsequently used in multipoint analysis. The sub-option "**DETAILED ALLELE FREQUENCY FILE"** is used to request an output of a '**chr<u>N</u>.frq**' file which contains the additional information for markers used in the corresponding '**chr<u>N</u>.geno**' file. The default is to not output this file.

- **INFORMATION FOR GENO FILE: FILE PATH = /your_directory/chr*/chr*.tmap | HEADER = T** or **F | BP_POSITION = <u>N</u> | CM_POSITION = <u>N</u> | RS_NUMBER = <u>N</u>**
  This option is used to provide base pair positions for markers listed in '**chr<u>N</u>.geno**' files when the "**GENO FILE FOR GL_AUTO"** option in a parameter file is invoked. The required information is contained in '**chr<u>N</u>.tmap**' files generated by PBAP together with '**chr<u>N</u>.geno**' files (Nato et al., 2015).

- **MODIFY YOUR BP ALLELE FREQUENCIES: FILE PATH = /your_directory/chr*.txt | HEADER = T** or **F | BP_POSITION = <u>N</u> | RS_NUMBER = <u>N</u>**
  This option is used to provide base pair positions of markers for which allele frequency specification is desired considering an admixed nature of a sample. The "**RS_NUMBER**" sub-

option is optional. If a column containing rs numbers is specified, the script will check whether rs number for your sample variant at a particular base pair position is the same as the one for a variant at the same base pair position in a reference population sample. If rs numbers cannot be matched, it will be reported in a file with the 'rs_num_missmatch' prefix. In a column for rs numbers if a particular variant does not have an rs number, use "**na**" entry to indicate that.

- **YOUR SAMPLE ALLELE FREQUENCIES: FILE PATH = /your_directory/chr*.txt | HEADER = T** or **F | BP_POSITION = N | REF_ALELLE = N | ALT_ALELLE = N | REF_ALLELE_FRQ = N**
  This option is used to provide your sample allele frequencies computed by simple averaging ignoring the fact that there is relatedness (non-independence) among your samples. These allele frequencies are only used to determine which allele has a frequency ≥ 0.5 if data for a particular variant is not present in a reference population. The sub-options "**REF_ALELLE**", "**ALT_ALELLE**", and "**REF_ALLELE_FRQ**" are used to specify column numbers in input files containing the corresponding data entries.

- **OUTPUT DIRECTORY = your_output_directory**
  This option specifies a location where all output files will be created.

- **REFERENCE POPULATION NAME = population1 | FILE PATH = /your_directory/chr*.txt | HEADER = T** or **F | BP_POSITION = N | RS_NUMBER = N | REF_ALLELE = N | ALT_ALLELE = N | REF_ALLELE_FRQ = N**
  This option is used to provide reference population marker allele frequencies separately for each reference population used in the analysis. ***In a reference population each variant with allele frequency data must have rs number.*** The sub-options "**REF_ALLELE**", "**ALT_ALLELE**", and "**REF_ALLELE_FRQ**" are used to specify column numbers in input files containing corresponding data entries.

- **LOCAL ADMIXTURE PER FAMILY BY POPULATION = population1 | FILE PATH = /your_directory/chr*.txt | HEADER = T** or **F | VARIANT IDENTIFIER = rs_number** or **bp_position | VARIANT IDENTIFIER COLUMN = N | FAMILIES = [family1 family2 family3] | COLUMN NUMBERS = [N N N]**
  This option is used to provide local admixture estimates computed separately for each family. Separate files must be created for each combination of population name and chromosome number and this option has to be specified separately for each population in the analysis. The sub-option "**VARIANT IDENTIFIER**" could be either "**rs_number**" or "**bp_position**" and specifies what would be used by the script to combine together same data located in different input files. It is strongly recommended to use the "**rs_number**" entry for the "**VARIANT IDENTIFIER**" sub-option unless you are confident that base pair positions for variants in your reference populations and your samples used to estimate local ancestry are based on the same build and patch. ***If the "rs_number" entry for the "VARIANT IDENTIFIER" sub-option is used, each variant with admixture data must have rs number.*** The sub-option "**VARIANT IDENTIFIER COLUMN**" specifies a column number in the file where data for this sub-option are located. The sub-options "**FAMILIES**" and "**COLUMN NUMBERS**" list all the family names and column numbers in input files where data for these families are located. ***The order of***

the family names and their corresponding column numbers in the last two sub-options described above must be the same.

2.3 Data Input Files

Depending on the options specified in a parameter file, the following input files are needed for the ADMIXFRQ script to run. The information below is presented to describe the sources and correct formatting of input files. Instructions to use a particular input file to run the script are defined within the option statements for a parameter file.

Information about variant alleles and their frequencies in all input files must be grouped based on whether an allele is reference or alternative.

- **Your sample allele frequency files** – These files must be prepared separately for each chromosome and have separate columns with base pair positions, reference allele, reference allele frequency, and alternative allele data for each variant in your data set. Allele frequencies can be computed by simple averaging ignoring the fact that there is relatedness among samples in your data set. These files are used with the "**YOUR SAMPLE ALLELE FREQUENCIES**" option in a parameter file.

- **'chrN.geno' files** – These files must be prepared separately for each combination of the chromosome numbers and family names present in your data set. The '**chrN.geno**' file must be created first using PBAP (Nato et al., 2015). The format of these files follows guidelines described in a manual to run the gl_auto program of the Morgan package version 3.3 (Tong & Thompson, 2007; https://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml). These files are used with the "**GENO FILE FOR GL_AUTO**" option in a parameter file.

- **'chrN.tmap' files** – These files formatted as PLINK tmap files are created separately for each chromosome by PBAP (Nato et al., 2015) when you run it to generate '**chrN.geno**' files. The **'chrN.tmap'** files contain information about base pair and cM positions and rs numbers for all markers used to generate '**chrN.geno**' files by PBAP. These files are used with the "**INFORMATION FOR GENO FILE**" option in a parameter file.

- **Reference population allele frequency files** – These files have to be prepared by a user for each reference population intended to be used in the analysis. The choice of a reference population is left at a user's discretion. If a user has no data for a specific reference population used in the analysis, he can use existing general reference populations such as the 1000 Genomes Project data (Genomes Project, 2012). Reference population allele frequency files must have separate columns for base pair positions, rs numbers, reference and alternative alleles and reference allele frequencies. ***In a reference population each variant entry with allele frequency data must have rs number.*** These files are used with the "**REFERENCE POPULATION NAME**" option in a parameter file.

- **Individual local admixture estimates averaged within a family by population files** – These files have to be prepared by a user as described in the Local ancestry estimation subsection of the Methods section of Nafikov et al. (2018). A separate file for each combination of the reference populations and chromosomes has to be created with columns

containing variant-specific information about their base pair positions or rs numbers or both and about their admixture proportions for a particular reference population with family-specific entries placed in a separate column. ***If the "rs_number" entry for the "VARIANT IDENTIFIER" sub-option is used, each variant with admixture data must have rs number.*** These files are used with the "**LOCAL ADMIXTURE PER FAMILY BY POPULATION**" option in the parameter file.

2.4 Output files

A log file '**admixfrq.date.hr.min.sec.yr.random_number.log**' will be created in the specified output directory where 'date' has the following format: "day.month.date_of_the_month". Another file '**rs_num_missmatch.date.hr.min.sec.yr.random_number.log**' will be created with the information about mismatched rs numbers.

Depending on the options specified in a parameter file, an output directory might contain the following subfolders: '**gl_auto**', '**gl_auto_local_adm**', '**allele_frequencies**', and '**admixture_proportions**'.

The '**gl_auto**' subfolder will be created if the "**GENO FILE FOR GL_AUTO**" option in a parameter file is invoked. For each combination of admixture models, family names, and chromosome numbers, two text files '**chrN.geno**' and '**chrN.frq**' will be generated where **N** will be substituted by a chromosome number. The '**chrN.geno**' file will be in its usual format (Tong & Thompson, 2007; https://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml) and will contain modified allele frequencies according to the specified admixture model. The '**chrN.frq**' file is an optional file containing base pair and cM positions, alleles and their frequencies for all the markers listed in the '**chrN.geno**' file.

The '**gl_auto_local_adm**' subfolder will be created if the '**ADMIXTURE PROPORTIONS PRINTOUT**' and '**GENO FILE FOR GL_AUTO**' options in a parameter file are invoked together and the '**Local**' admixture model is selected. For each combination of family names and chromosome numbers a file '**chrN.local.adm.txt**', where **N** stands for a chromosome number, will be generated. Within the file, base pair and cM positions along with admixture proportions for all populations involved in admixture will be reported for all the markers contained in the '**chrN.geno**' file for the same family and chromosome.

The '**allele_frequencies**' subfolder will be created if the '**MODIFY ALLELE FREQUENCIES**' option in a parameter file is invoked. For each combination of admixture models, family names, and chromosome numbers a file with the name '**admixture_model.frq.txt**' will be generated where the expression 'admixture_model' will be substituted for the actual admixture model name (Global, FBGW, FBCW, or Local). The resulting file will contain base pair position, reference and alternative alleles and their frequency calculated using a specified admixture model.

The '**admixture_proportions**' subfolder will be created if the '**ADMIXTURE PROPORTIONS PRINTOUT**' and '**MODIFY ALLELE FREQUENCIES**' options in a parameter file are invoked together. For each combination of admixture models, family names, and chromosome numbers a file with the name '**chrN.family_name.admixture_model.admixture.txt**' will be generated where N, 'family_name", and "admixture_model' will be substituted by corresponding

chromosome number, family name, and admixture model used in the analysis. The resulting file will contain base pair positions and admixture proportions for all populations assumed to be contributing to admixture in a particular study sample.

3.0 References

Genomes Project, C., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., . . . McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65. doi:10.1038/nature11632

Nafikov, R. A., Nato, A. Q., Jr., Sohi, H., Wang, B., Brown, L., Horimoto, A. R., Vardarajan, B. N., Barral, S. M., Tosto, G., Mayeux, R. P., Thornton, T. A., Blue, E., & Wijsman, E. M. (2018). Analysis of Pedigree Data in Populations with Multiple Ancestries: Strategies for Dealing with Admixture in Caribbean Hispanic Families from the ADSP. *Genetic Epidemiology*, 1-16. doi.org/10.1002/gepi.22133

Nato, A. Q., Jr., Chapman, N. H., Sohi, H. K., Nguyen, H. D., Brkanac, Z., & Wijsman, E. M. (2015). PBAP: a pipeline for file processing and quality control of pedigree data with dense genetic markers. *Bioinformatics, 31*(23), 3790-3798. doi:10.1093/bioinformatics/btv444

Tong, L., & Thompson, E. (2007). Multilocus lod scores in large pedigrees: combination of exact and approximate calculations. *Human Heredity, 65*(3), 142-153. doi:10.1159/000109731