

## Instruction for *frappe*

### Introduction

*frappe* is a program for estimating individual ancestry and admixture proportions using high-density SNP data. The statistical method is described in Tang et al. 2005.

### Download

The program can be downloaded from <http://tanglab.stanford.edu/software.html>. Compiled executables under Mac OS X, linux and Windows are available. We are still improving the algorithm, so please come back and check for the update frequently.

### Command Line

To run *frappe*, simply type

```
> frappe parm.txt
```

where *parm.txt* is the input parameter file (see below).

### Input Data

You can run *frappe* with either two or three input files.

#### Parameter file

This is the “*parm.txt*” used in the command line. It contains the following parameters:

MaxIter: maximum iteration of EM to run

K: number of (ancestral) populations. Note: currently, *frappe* does not provide measures to choose K. In practice, people often try different K, and choose the K that makes most biological sense.

M: number of markers in the genotype data

I: number of individuals

Nout and step: these two parameters specify how often you would like to output intermediate results. The intermediate results have the same format as the final output file (see below), but records the ancestry estimates in the current EM iteration. Looking at the intermediate file can help you to decide whether the program is running properly, and whether the estimates are converging. Nout specifies the total number of output file you would like; alternatively, you can specify that an intermediate file should be generated every n steps of EM iterations. Barring rounding errors,  $Nout * step = MaxIter$ . If both

Nout and step are specified, Nout overrides step. Caution: if you choose a large Nout, please make sure you have plenty disk space.

GenotypeFile: the name of the genotype data file. See below for format.

IndividualFile: name of an optional input file. By default, it is NONE. This file will be useful for some advanced applications (to be described later).

### **Genotype Data file**

The (large) genotype file in plink ped file format. The two alleles of a SNP are coded as 1 and 2, and missing alleles are coded as 0. Please consult PLINK website for details of the ped file format (<http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#ped>). PLINK provides a nice venue to convert from many different formats. For example if your dataset is coded as A/C/G/T or 1/2/3/4 for the four alleles, you can generate the desired frappe input file by:

```
plink --ped filename --recode12
```

Note: frappe does not take transposed file. If you have transposed input file, first convert it to the standard ped file:

```
plink --tped filename --recode12 --transpose
```

Note: It is very important that you remove markers or individuals with excess missing-values. Removing SNPs or excluding individuals can be done using PLINK.

### **Output file**

The output file consists of the ancestry proportion estimates for each individual. If your input genetic data consists of I individuals, and you specify K ancestral populations, the output file will have I lines, each line following the format:

```
FID IID      :      q1 q2 ... qK
```

where FID and IID are the family ID and individual ID, identical to the first two columns of the genotype input file. The ancestry attributions are in (q1, ..., qK).

### **Tips**

- Please make sure that the input and output file names are compatible with your platform.
- Please make sure that you have converted the genotype data to 0/1/2 coding
- Please make sure that the length of FID and IID are no longer than 25 characters in total.

### **Contributors**

Jie Peng, Pei Wang, Zhiyu Ma, Hong Gao, Marc Coram, Hua Tang.

### **Contact us**

For questions and bug reports, please contact us as [frappe.help@gmail.com](mailto:frappe.help@gmail.com). We welcome your suggestions and feedback!

### **Reference**

Tang H, Peng J., Wang P., and Risch N. (2005) Estimation of Individual Admixture: Analytical and Study Design Considerations. *Genet Epidemiol.* 28:289-301.