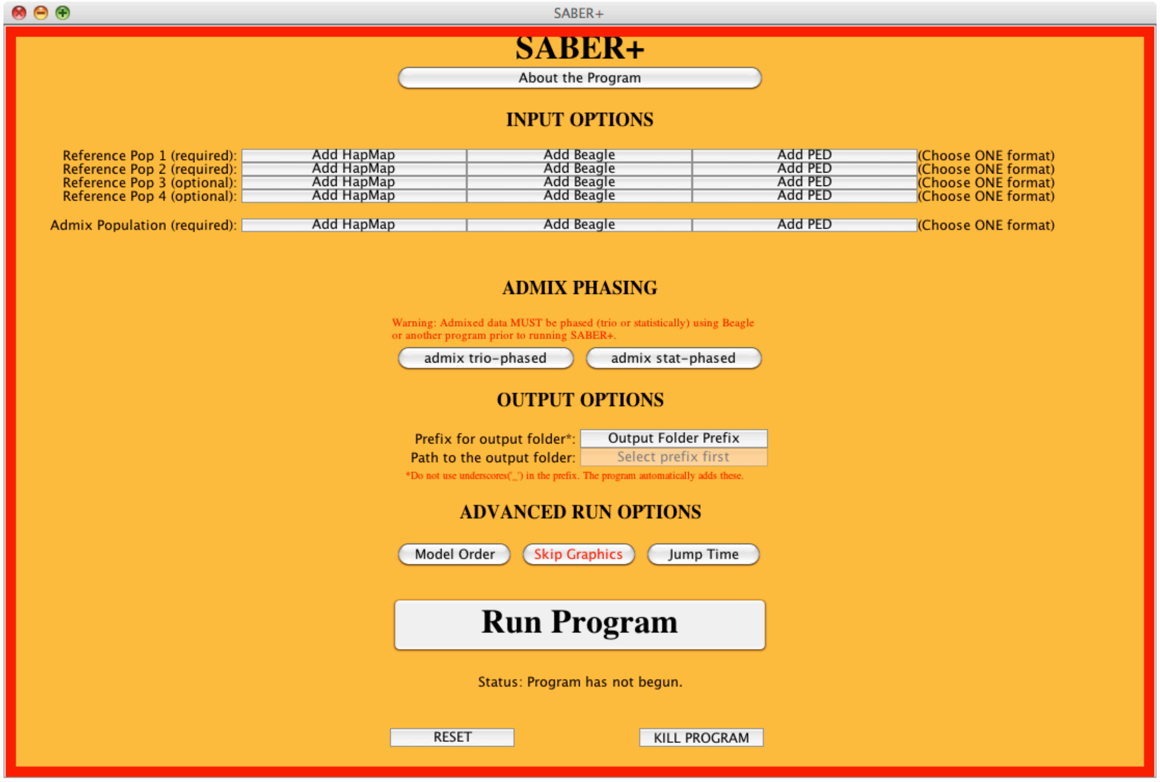
SABER+ Tutorial

SABER+ is a program designed to estimate (local) ancestry along a chromosome in admixed individuals, based on genotype data. This tutorial takes you through the analysis of a small example dataset. Before you begin, please make sure that you have R and a C/C++ compiler installed on your machine. Also, you may need an internet connection to download R packages. Please see the ReadMe file for more information on input and output formats, as well as advanced features.

Step 1: Install SABER+. Download (and unzip, if necessary) the entire SABER+ folder in your working directory. In this example, we will assume the path is /myhome/proj/saber+beta_0.1.

Step 2: Start SABER+. Open the saber+beta_0.1 folder and double click on the SABER_PLUS.jar (a runnable jar folder). This should bring up a window like the one below:

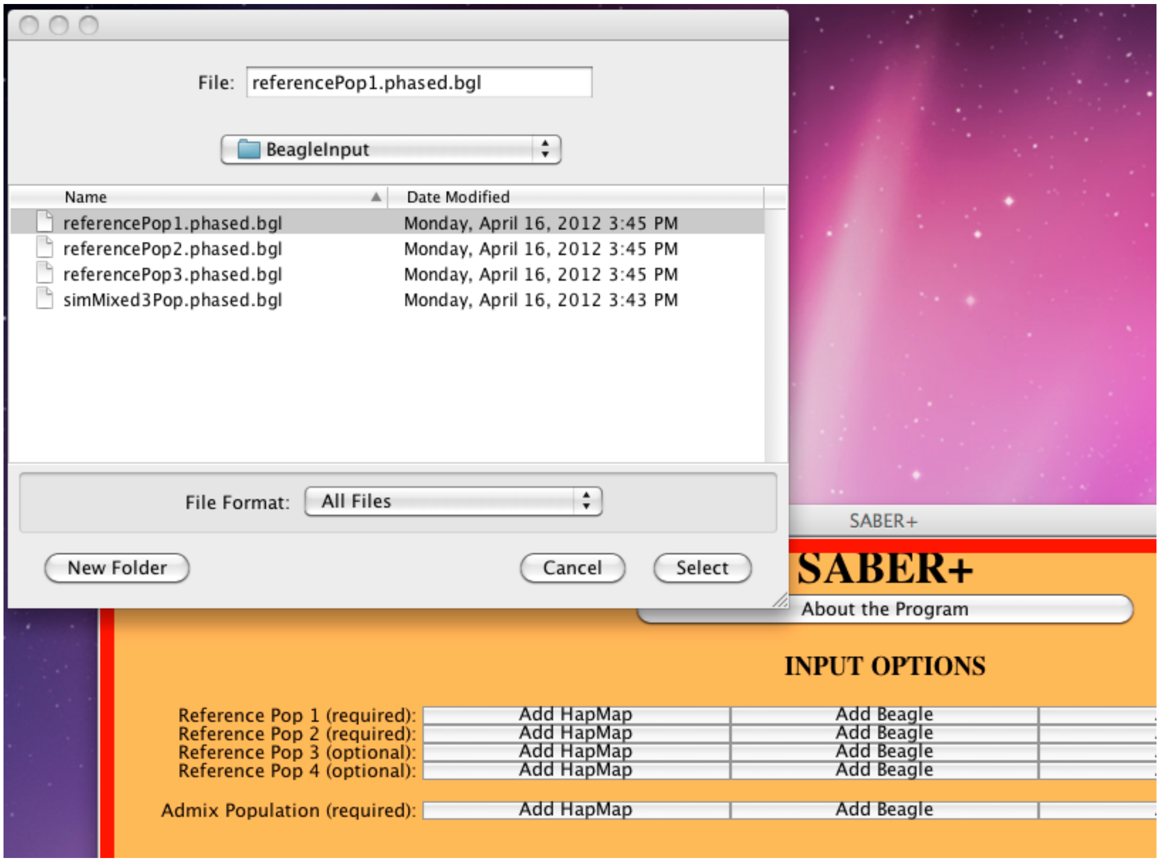


Warning: Admixed data MUST be phased (trio or statistically) using Beagle or another program prior to running SABER+.

*Do not use underscores('_') in the prefix. The program automatically adds these.

Status: Program has not begun.

Step 3: Specify input data files. Click on the “Add Beagle” button to add the reference 1, reference 2 and admixed datasets. These are found in the Examples/BeagleInput directory. SABER+ accepts two additional data format: HapMap phased haplotypes, and ped file used in programs such as PLINK. The Example directory provides the same data in all three formats. Each file contains haplotype data for one chromosome, and markers are assumed to be arranged in physical order. The precise physical location or recombination distance are not required. Please do not mix multiple chromosomes in one file.



File: referencePop1.phased.bgl

BeagleInput

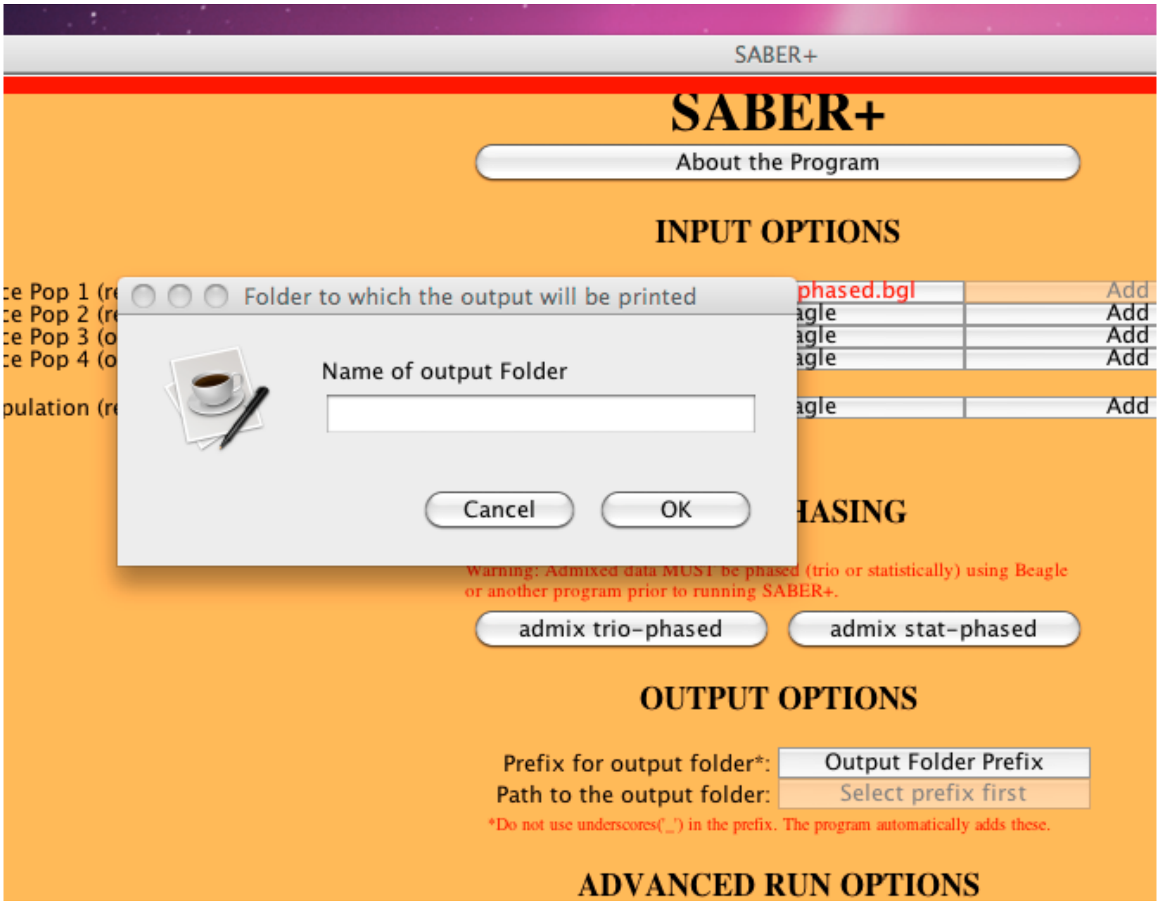
Name	Date Modified
referencePop1.phased.bgl	Monday, April 16, 2012 3:45 PM
referencePop2.phased.bgl	Monday, April 16, 2012 3:45 PM
referencePop3.phased.bgl	Monday, April 16, 2012 3:45 PM
simMixed3Pop.phased.bgl	Monday, April 16, 2012 3:43 PM

File Format: All Files

New Folder
Cancel
Select
SABER+
SABER+
About the Program
INPUT OPTIONS
Reference Pop 1 (required):
Reference Pop 2 (required):
Reference Pop 3 (optional):
Reference Pop 4 (optional):
Admix Population (required):
Add HapMap
Add Beagle

Step 4: Required program parameters. SABER+ assumes that the reference population datasets consist of haplotype data, which are phased using trios (i.e. no phasing error). The admixed data can either be trio-phased or statistically phased, using programs such as Beagle. ***Importantly, you must first analyze your admixed individuals through a phasing algorithm.*** If your admixed dataset consists of unrelated individuals, you can use the output from Beagle (or other phasing programs of your choice), and specify “stats-phased;” SABER+ tries to correct occasional phasing error. If your admixed individuals come in parent-offspring trios, you should take advantage of the family structure in phasing, and specify “trio-phased” in SABER+; in this case, SABER+ assumes that the admixed haplotypes are perfectly phased. *When in doubt, use “stats-phased.”* The data in the example are true haplotypes, so you could use either option.

Step 5: Output path. Add the name of the output folder, which must not already exists. This folder is created in the parent directory of saber+beta_0.1. In our example, a directory will be created in /myhome/proj/. The name of the output folder also serves as the prefix of the output file. Once you have entered the prefix, you can modify the path to this folder by click on the button next to "Path to the output folder." ***Note:*** *Currently we require that each run write output to a new directory.*



Warning: Admixed data MUST be phased (trio or statistically) using Beagle or another program prior to running SABER+.

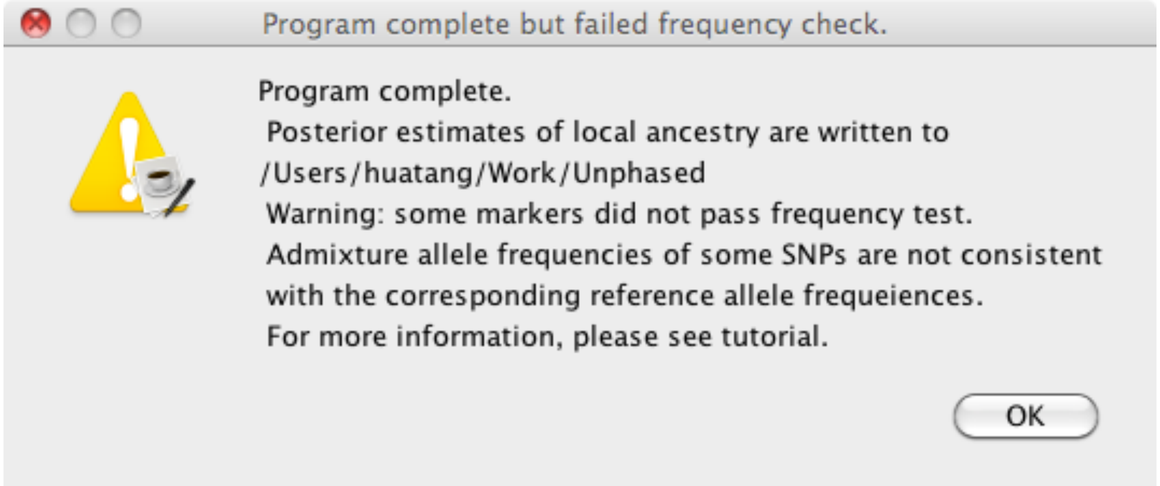
*Do not use underscores('_') in the prefix. The program automatically adds these.

Step 6: Optional parameters. When "Skip Graphics" is in black, a pdf figure will be generated corresponding to each individual. The default is to NOT generating these figures, since we anticipate that many users will be analyzing large datasets and these figures will take a large amount of disk space. The other two optional parameters, "Model Order" and "Jump Time" do not need to be changed.

Step7: Output. When Program is complete, you will get a message indicating that program is finished and the "Program Running" button will say "Program Done!" You may find output in the specified directory, and the output should look like that given in the Examples directory. If you specified that the admixed individual is trio-phased, then the posterior estimate is produced corresponding to each haplotype; otherwise, the posterior estimate is produced for each individual. In either case, each row of the output is a SNP, and the posteriors corresponding to Reference Populations 1, 2, and 3 are estimated.

Summary files combine the local ancestry determinations for all individuals, listing one individual per row and one SNP per column. For example, if there were 2 reference populations, two files are created *_ RefPop0_combinedFile.txt and *_ RefPop1_combinedFile.txt, that combine the local ancestry estimates for all individuals for reference population 1 and 2 respectively. *_snpIDs_combinedFile.txt lists the SNPs ID used in order.

For the tutorial example, you will get a warning message like this.

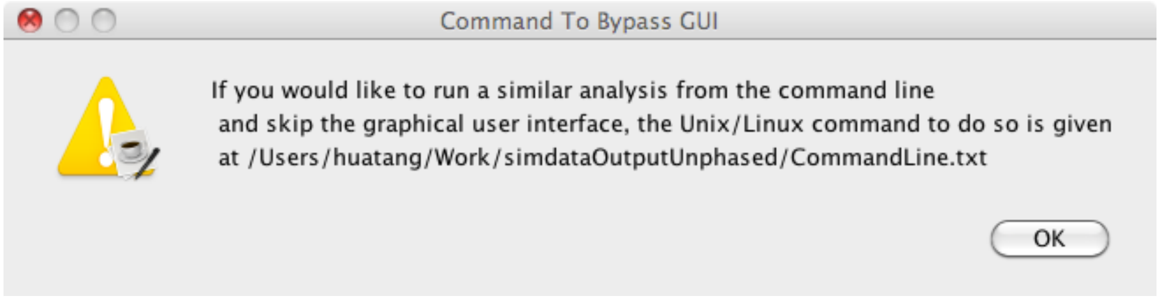


Program complete.
Posterior estimates of local ancestry are written to
/Users/huatang/Work/Unphased
Warning: some markers did not pass frequency test.
Admixture allele frequencies of some SNPs are not consistent
with the corresponding reference allele frequeiences.
For more information, please see tutorial.

SABER+ implements some diagnostics that detect obvious data error. [This section will be expanded later.]

Running SABER+ through command line.

At the end of the analysis, we get this message:



If you would like to run a similar analysis from the command line
and skip the graphical user interface, the Unix/Linux command to do so is given
at /Users/huatang/Work/simdataOutputUnphased/CommandLine.txt

The CommandLine that does what we have done in Step 2-6 is:

```
java -jar -Xms1024m -Xmx2048m SABER_PLUS.jar NO_GUI=true
OUTPUT_PREFIX=simdataOutputUnphased
OUTPUT_PATH=/Users/huatang/Work/simdataOutputUnphased
```

```
REF_FILE_1=/Users/huatang/Work/SABER+_beta/Examples/BeagleInput/referenc
ePop1.phased.bgl
REF_FILE_2=/Users/huatang/Work/SABER+_beta/Examples/BeagleInput/referenc
ePop2.phased.bgl
REF_FILE_3=/Users/huatang/Work/SABER+_beta/Examples/BeagleInput/referenc
ePop3.phased.bgl
ADMIX_FILE=/Users/huatang/Work/SABER+_beta/Examples/BeagleInput/simMix
ed3Pop.phased.bgl REFERENCE_1_FORMAT_INDEX=3
REFERENCE_2_FORMAT_INDEX=3 REFERENCE_3_FORMAT_INDEX=3
REFERENCE_4_FORMAT_INDEX=0 ADMIX_FORMAT_INDEX=3 STATS_PHASED=true
GRAPHICS=false
```

Our intention is that you will run a small test dataset using the graphic interface, and then modify the command line to batch process other chromosomes and subsets of individuals. Note that since SABER+ analyze each individual separately, you can take advantage of multiple processors by dividing data into subsets of individuals.

To run this command line again, you must change the OUTPUT_PATH to a non-existing directory. Of course, you should change the file name of the admixed individuals.

In this example, we assume the admixed individuals are statistically phased (STATS_PHASED=true); if the haplotypes were derived from trios, then use TRIO_PHASED=true. To produce graphic output of the posterior ancestry for each individual (may generate a lot of large .pdf files), change GRAPHICS=true.

References:

*We are in the process of write a paper describing SABER+. In the meanwhile, you can cite:

Johnson NA, Coram MA, Shriver MD, Romieu I, Barsh GS, London SJ, Tang H. (2011) Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genet.* 7:e1002410.