Instruction Manual for "GLOBETROTTER: a program for identifying, dating and describing admixture events in population data"

Garrett Hellenthal

December 30, 2016

Contents

1 Introduction

GLOBETROTTER is an R program, originally described in [1], that can identify, date and describe admixture events occurring in the ancestral history of a given target population within the last ≈4,500 years. Unlike similar programs, it requires no *a priori* specification of surrogates for the original sources involved in the admixture event(s) in the ancestry of the target population. Instead you provide GLOBETROTTER with DNA information (as summarized by companion program CHROMOPAINTER; see below) on multiple sampled groups (or "surrogates") that may or may not be related to ancestral sources of the target population. GLOBETROTTER then uses these sampled groups to identify whether the target population descends from any admixture event(s), and – if so – it determines precisely when the event(s) occurred and the admixing source groups involved. To describe each original admixing source, GLOBETROTTER provides our "best-guess" of the single sampled surrogate group that genetically best represents the admixing source, but also represents the admixing source as a mixture of the the DNA of all sampled surrogate groups (often many of which are inferred not to contribute to the mixture at all), allowing for a richer characterization and subsequent interpretation.

Alternatively, you can use GLOBETROTTER to describe the DNA of a given target population as a mixture of that of given surrogate populations *without* attempting to infer any admixture events, using the technique described in [3]. To do so, in the *parameter_infile* simply specify **prop.ind** as "1" and **num.mixing.iterations** as "0" (see Section 4.1 for a detailed description of *parameter_infile*).

Throughout this document, we refer to the population you are testing for admixture as the "target" population. We refer to any other sampled populations used to describe the target population's admixture event(s) as "surrogate" populations. Finally, "donor" populations refer to groups used to paint both target and surrogate populations using CHROMOPAINTER (which may include the surrogate and/or target populations).

2 Running CHROMOPAINTER to make input files for GLOBETROTTER

GLOBETROTTER uses "copying vectors" and "painting samples" generated by a companion program CHROMOPAINTER" [2], which identifies haplotype sharing patterns among different groups by studying Single-Nucleotide-Polymorphism (SNP) data. Specifically, CHROMOPAINTER "paints" the DNA of a "recipient" population conditional on a set of "donor" populations. I.e. at each genetic region across the haplotypes of a set of recipient individuals, CHROMOPAINTER identifies the single best matching haplotype among the set of donor individuals, which is indicative that the recipient and donor share recent common ancestry

at the particular genetic locus.

In order to use GLOBETROTTER, you must first use CHROMOPAINTER to generate the following input files:

1. "copy vectors" for each of the target and surrogate populations, giving the total length of DNA segments that individuals of each population copy from a set of donors under CHROMOPAINTER

2. "painting samples" for haplotypes of the target population, giving CHROMOPAINTER's sample realisations of the individual donor haplotypes copied at each SNP of each recipient haplotype of the target population (**Note** this is not required if specifying only to describe the target as a mixture of the surrogate groups as in [3], i.e. by specifying **prop.ind** as "1" and **num.mixing.iterations** as "0" in the *parameter_infile*.)

(1) and (2) refer to the "XXX.chunklengths.out" and "XXX.samples.out" output files, respectively, from CHROMOPAINTER. To make the "copy vectors" of (1) comparable, the target and surrogate populations should copy from the same set of donors under the CHROMOPAINTER model. For example, this set of donors could contain all individuals from the target and recipient populations, which is akin to the main "full" analyses described in [1]. Or the set of donors might contain only the surrogate populations, as is the case in the "regional analyses" of [1]. However, the set of donors could also be entirely separate from the surrogate and donor groups, or only contain a subset of these groups and/or a subset of individuals from these groups. In Section 6 below, we walk through one recommended way of analysing of a set of populations from start to finish, using both CHROMOPAINTER and GLOBETROTTER (and optionally fineSTRUCTURE).

3 Getting Started

After extracting the .tar file, compile GLOBETROTTER in the following manner:

```
R CMD SHLIB -o GLOBETROTTERCompanion.so
GLOBETROTTERCompanion.c -lz
```

To compile, note that you must have "zlib" installed (e.g. sudo apt-get install zlib1g-dev). You must also have the package "nnls" installed in R (i.e. install.packages("nnls")). Note also that in order to run on e.g. clusters, you may need to change the line in GLOBETROTTER.R that reads dyn.load("GLOBETROTTERCompanion.so") to include the pathway directory; i.e. change this line to dyn.load("/directorypath/GLOBETROTTERCompanion.so").

The basic command line is as follows:

```
R < GLOBETROTTER.R [parameter_infile] [painting_samples_filelist_infile]
[recom_rate_filelist_infile] --no-save
```

or to direct screen output to another file "*screen_output*":

R < GLOBETROTTER.R [*parameter_infile*] [*painting_samples_filelist_infile*] [*recom_rate_filelist_infile*] --no-save > [*screen_output*]

There are three user-input files, all required unless specifying only to describe the target as a mixture of the surrogate groups as in [3], i.e. by specifying **prop.ind** as "1" and **num.mixing.iterations** as "0" in the *parameter_infile*:

1. *parameter_infile* (always required) provides a description of all the parameters to use in GLOBETROTTER, including specification of the target, surrogate and donor populations and the file of copy vectors to use.

2. *painting_samples_filelist_infile* (not required only in above exception) gives a list of the XXX.samples.out files from a CHROMOPAINTER analysis of the target population conditional on the donors. (When inferring confidence intervals around dates of admixture, GLOBETROTTER will bootstrap resample individuals independently for each file in *painting_samples_filelist_infile*, so that e.g. you might specify one XXX.samples.out file per chromosome.)

3. *recom_rate_filelist_infile* (not required only in above exception) provides a list of the recombination rate files corresponding to each file in *painting_samples_filelist_infile*. Each of these recombination rate files is in the same format as the files provided as input in CHROMOPAINTER, giving the genetic distance between each pair of contiguous SNPs that have been painted in each XXX.samples.out file.

See section "Input Format" below for details on the format of each of these three input files and section "Output" for details on the files output by GLOBETROTTER.

Type "R < GLOBETROTTER.R help --no-save" to get a brief description of this command line and the parameter input file options.

4 Input Format

There are three types of input files that must be specified from the command line, each of which we have provided examples for (given in parentheses):

1. *parameter_infile* (example files: **"example/BrahuiYorubaSimulation.paramfile.txt"**, **"example/BrahuiYorubaSimulation.paramfileII.txt"**)

2. *painting_samples_filelist_infile* (**"example/BrahuiYorubaSimulation.samplesfile.txt"**)

3. *recom_rates_filelist_infile* (**"example/BrahuiYorubaSimulation.recomfile.txt"**).

These example files reflect (partially) the simulated example in Figure 1 of [1].

4.1 *parameter_infile*

This file contains all the parameter information for the GLOBETROTTER run, as well as the names of the files containing (a) all the copying vectors for the surrogate and target populations and (b) the labels and group assignments for all surrogate, target, and donor individuals. Many of these parameters deal with how to fit the "coancestry curves" described in [1]. These curves are constructed for all pairwise combinations of surrogate populations inferred to match >**props.cutoff** (see below) of the total ancestry of the target population, and e.g. are provided at http://admixturemap.paintmychromosomes.com/ for all analyses of [1].

The file *parameter_infile* contains 20 rows, formatted in the following manner (and order) shown in bold type, with brackets containing allowed values and a description provided in normal font provided to the right of each parameter:

- **prop.ind:** [**0,1**] – indicate whether ("**1**") or not ("**0**") to infer admixture proportions, dates and sources (if "**0**", this information will be read from previously made GLOBETROTTER files specified by **save.file.main**)
- **bootstrap.date.ind:** [**0,1**] – indicate whether to perform bootstrap resampling to infer confidence intervals around date estimates (see Section 4.2 for details)
- **null.ind:** [**0,1**] – indicate whether to standardize by a "NULL" individual when performing inference (e.g. used for inferring p-values for evidence of admixture in [1], also appropriate when "target" population has likely undergone bottleneck effects and in general testing for consistency – see [1])
- **input.file.ids:** [**input.filename1**] – pathway and name for file containing id labels for all samples (see Section 4.1.1 below)
- **input.file.copyvectors:** [**input.filename2**] – pathway and name for file containing copy vectors for all surrogate and target populations (see Section 4.1.2 below)
- **save.file.main:** [**output.filename1**] – pathway and name for main output file prefix (see Section 5)
- **save.file.bootstraps:** [**output.filename2**] – pathway and name for bootstrap output file prefix (see Section 5.4)
- **copyvector.popnames:** [**pop_1 pop_2 ... pop_k**] – names of all k populations used as donors; i.e. that both surrogate and target populations copied from when running CHROMOPAINTER (NOTE: Any painted segments that select the target population as a donor will be ignored, even if you include the target population in this line of the input file.)

- **surrogate.popnames: [pop_1 pop_2 ... pop_j]** – names of all j surrogate populations, i.e. used to describe admixture in **target.popname**
- **target.popname: [pop_rec]** – name of target population
- **num.mixing.iterations: [0,1,...,5,...]** – number of iterations of date and proportion/source estimation to perform; "0" specifies to only infer proportions of ancestry relating the target to each surrogate, and to not try and infer admixture events, as described in [3] (only used when **prop.ind: 1**)
- **props.cutoff: [0.0,...,1.0]** – at each iteration, remove any surrogates that contribute $\leq$ this value to the mixture describing the target population
- **bootstrap.num: [0,1,...]** – number of bootstrap re-samples (only used when **bootstrap.date.ind: 1**); each bootstrap re-sample is independent, so that this step can be run simultaneously on multiple computing cores (though be careful to specify a different **save.file.bootstraps** for each run to avoid recording over previous results)
- **num.admixdates.bootstrap: [1,2]** – number of dates to fit when performing bootstrapping (only used when **bootstrap.date.ind: 1**)
- **num.surrogatepops.perplot: [1,...]** – will plot this number squared of coancestry curves for each page of the curves output file (only used when **prop.ind: 1**)
- **curve.range: [lower.lim upper.lim]** – lower and upper bounds of x-axis (i.e. cM distance between DNA segments) to fit dates to when generating coancestry curves
- **bin.width: [e.g. 0.1]** – width of x-axis bins (in cM) when generating coancestry curves
- **xlim.plot: [lower.lim upper.lim]** – lower and upper bounds (in cM) of x-axis to plot for coancestry curves (only used when **prop.ind: 1**)
- **prop.continue.ind: [0,1]** – indicate whether you are continuing proportion estimation from those in a previous file (in which case the previous file will be read from **save.file.main** and output files will add the suffix "**_continue**")
- **haploid.ind: [0,1]** – indicate whether individuals are haploid ("**1**") or diploid ("**0**") (unfortunately these are the only two ploidies the present implementation of GLOBETROTTER can cope with)

See **"example/BrahuiYorubaSimulation.paramfile.txt"** and **"example/BrahuiYorubaSimulation.paramfileII.txt"** for examples. In both of these examples, the target population (called "BrahuiYorubaSimulation") is a population simulated as a mixture of Brahui and Yoruba individuals.

The admixing source groups will be described as mixtures of the populations "Adygei", "Armenian", ..., "Yi" (i.e. **surrogate.popnames**). All of these surrogates and the target population have been modeled using CHROMOPAINTER with **copyvector.popnames** as donors, with the copy vectors from this analysis stored in the **input.file.copyvectors** called "BrahuiYorubaSimulation.copyvectors.txt" (see Section 4.1.2) and the key to matching individuals from that file to each target, surrogate and donor label stored in the **input.file.ids** called "BrahuiYorubaSimulation.idfile.txt" (see Section 4.1.1).

For the provided example, the file **"example/BrahuiYorubaSimulation.paramfile.txt"** would be run initially (see Section 6), as it specifies to infer admixture proportions, dates and sources, specifically using five iterations (i.e. **num.mixing.iterations**) of date and proportion/source estimation. When estimating dates, only pairs of DNA segments separated by $\geq$ 1cM and $\leq$ 50cM (i.e. **curve.range**) are considered. (This is a sensible range to fit, in that segments separated by <1cM may be affected by within-population linkage disequilibrium that could distort signals, and only very recent admixture will have a signal beyond 50cM.) When tabulating the counts of DNA segment pairs that are copied (under CHROMOPAINTER) from each pairing of donor populations, the distance between segments within each pair are rounded to the nearest 0.1cM (i.e. **bin.width**), so that the total grid of x-axis values (genetic distances) fit for the coancestry curves is [1.0,1.1,1.2,....,49.8,49.9,50.0]cM in this example. This GLOBETROTTER run will generate estimates of the admixture proportions, dates and sources under (i) a single date of admixture with two admixing sources, (ii) a single date of admixture with more than 2 sources and (iii) two separate dates of admixture (possibly with different sources each time), written to a file with prefix "example/BrahuiYorubaSimulation.globetrotter.main" (i.e. **save.file.main**) (see Section 5.1). This run will also provide the coancestry curves for all pairwise combinations of surrogate populations inferred to match >0.001 of the total ancestry of the target population (i.e. **props.cutoff: 0.001**) in text format, and a pdf file that plots 9 coancestry curves per page (i.e. **num.surrogatepops.perplot: 3**) for an x-axis ranging from 0 to 50cM (i.e. **xlim.plot: 0 50**) (see Sections 5.2-5.3).

After inferring admixture proportions, dates and sources, the file **"example/BrahuiYorubaSimulation.paramfileII.txt"** would be run to bootstrap re-sample individuals (or individuals' chromosomes) in order to infer confidence intervals around the date(s) of admixture. By setting **prop.ind: 0** and **bootstrap.date.ind: 1**, the file **"example/BrahuiYorubaSimulation.paramfileII.txt"** specifies that GLOBETROTTER output files with prefix given in **save.file.main** (in this case "example/BrahuiYorubaSimulation.globetrotter.main") already exist and contain inferred admixture proportions, dates and sources. In this case, for bootstrapping we specify to perform 20 bootstrap re-samples (i.e. **bootstrap.num: 20**) assuming a single date of admixture (i.e. **num.admixdates.bootstrap: 1**). This GLOBETROTTER run will generate coancestry curves using the specified **curve.range** and **bin.width**, and output the bootstrap date estimates

to a file with prefix "example/BrahuiYorubaSimulation.globetrotter.boot" (i.e. **save.file.bootstraps**) (see Section 5.4).

4.1.1 input.file.ids

This file should match exactly the donor input file used in ChromoPainterv2 ('-t' switch), specifying the individual identifier and corresponding population label for each donor, target, and surrogate individual in the analysis. (For those using the old version of CHROMOPAINTER input format – see the comment at the end of this section – you must generate a new file in this manner.) The format of the file is one individual per row, with individuals ordered in the same manner as they are in the CHROMOPAINTER haplotype input file used in this analysis. There are three columns per row, with the first column giving the individual identifier, the second column giving the individual's population label and the third column an indicator for whether the individual is not included in the analysis (use "0" – i.e. "zero" – to specify NOT to include the given individual). Any individuals with a "0" in the third column will NOT be considered in any part of the GLOBETROTTER analysis.

For example, consider a file with the following 10 individuals:

IND1 Pop1 0
IND3 Pop1 1
IND2 Pop1 1
IND4 Pop2 1
IND5 Pop2 0
IND6 Pop2 1
IND7 Pop1 1
Pop4Ind1 Pop4 1
IND8 Pop3 1
IND9 Pop3 1

In the GLOBETROTTER analysis, as is at least recommended – though not strictly necessary – for the corresponding preliminary CHROMOPAINTER analyses, the individuals IND1 and IND5 will be ignored. This leaves IND3, IND2, IND7 as representing "Pop1", IND4, IND6 representing "Pop2", IND8, IND9 representing "Pop3" and Pop4Ind1 representing "Pop4". Therefore, GLOBETROTTER will search for these individual labels (i.e. column 1) when combining copy vector columns and rows from **input.file.copyvectors** for each of these four population labels "Pop1-Pop4", though when performing this concatonation for each of "Pop1-Pop4" it will also include any columns and rows from **input.file.copyvectors** labeled as "Pop1-Pop4", respectively. (Note that no value other than "0" in the third column specifies any action.)

An example of **input.file.ids** is provided in
"example/BrahuiYorubaSimulation.idfile.txt". Here column 3 has a "1"

in each row, specifying that no individuals are removed from analysis.

NOTE1: Again in general, each population label specified in **copyvector.popnames**, **surrogate.popnames** and **target.popname** of the file "*parameter_infile*" **MUST** be in column 2 of at least one row of the file **input.file.ids**. An exception, incorporated to make things more flexible for the user, is if all **surrogate.popnames** and **target.popname** labels missing from column 2 of **input.file.ids** are specified in the row labels of **input.file.copyvectors**, and similarly all **copyvector.popnames** missing from column 2 of **input.file.ids** are specified in the column labels of **input.file.copyvectors**. In other words, the column names and row names of **input.file.copyvectors** must contain the individual identifiers and/or the population labels (see Section 4.1.2).

NOTE2: It is critical that the order of individuals in **input.file.ids** corresponds to the donor indices used in the XXX.samples.out files used in the analysis (see Section 4.2). Each painting sample (row) of each XXX.samples.out file gives a number D for each SNP, corresponding to the row of the CHROMOPAINTER haplotype input file ('-g' switch) that contains the donor haplotype copied at that SNP (where the first haplotype in this input file is assigned label $D = 1$). The row containing the label for this donor individual in **input.file.ids** MUST be row D/p (with any decimal values of D/p rounded up to the nearest integer) where $p = 1, 2$ is the ploidy of the organism.

4.1.2 input.file.copyvectors

This file should contain the XXX.chunklengths.out files from the corresponding preliminary CHROMOPAINTER analyses, in the same format and combined across all chromosomes and individuals. I.e. each row is an individual (or population), and the columns give the total amount of genome-wide DNA that the given individual (or population) is inferred to copy from every "donor" individual (or population) in the corresponding preliminary CHROMOPAINTER analyses.

The first row of **input.file.copyvectors** lists the column labels reflecting the donor individuals and/or populations. The first column in this first row is "Recipient", and the remaining columns of this first row must contain for each label specified in **copyvector.popnames** of "*parameter_infile*" either (i) the label itself and/or (ii) the individual identifier of at least one individual assigned to that label, in any order (i.e. the order does NOT need to match that of **input.file.ids**). (The individual identifiers and their corresponding labels are specified in columns 1 and 2, respectively, of **input.file.ids** – see Section 4.1.1.) The remaining rows of **input.file.copyvectors** list the "recipient" individual (or population) label in the first column, with the remaining columns containing the total amount (or proportion) of genome-wide DNA that the given recipient individual (or population) copies from each donor label provided in the first row. Analogous to the first row, the first column of **input.file.copyvectors**,

which list the recipient labels, must contain for each label specified in **surrogate.popnames** and **target.popname** of "*parameter_infile*" either the label itself and/or the individual identifier of at least one individual assigned to that label, again in any order.

Operationally, within each row of **input.file.copyvectors**, first GLOBETROTTER will – for each donor label of **copyvector.popnames** – sum the values across all columns that either match or have an individual identifier assigned to the given donor label, i.e. calculate the total amount of genome-wide DNA each row copies from each donor label. Then GLOBETROTTER will – for each recipient label of **surrogate.popnames** and **target.popname** – average these resulting values across all rows that either match or have an individual identifier assigned to the given recipient label. In the end, this process calculates the average amount of genome-wide DNA that each surrogate and target population copies from each donor population. These values are standardized to sum to 1 within each surrogate and target, and then used to help infer admixture proportions, dates and sources.

An example of **input.file.copyvectors** is provided in **"example/BrahuiYorubaSimulation.copyvectors.txt"**. Here for column 1 we have used individual identifiers for all the recipients individuals (rows), while for row 1 we have simply used the population labels to specify each of the donors (columns). Therefore, for **target.popname** and each label specified in **surrogate.popnames** of "*parameter_infile*", GLOBETROTTER will first find all individual identifiers (column 1 of **input.file.ids**) for the given label (column 2 of **input.file.ids**), pull out all rows of **input.file.copyvectors** matching these individual identifiers, and average each column across these rows to get the final matrix of copying vectors (i.e. without having to do any summing over columns in this particular example).

Note that you may include labels in row 1 and/or column 1 of **input.file.copyvectors** that are neither contained in **input.file.ids** nor correspond to any of **copyvector.popnames**, **surrogate.popnames** or **target.popname** of "*parameter_infile*". These rows/columns will be ignored in the GLOBETROTTER analysis. In addition, any rows/columns in **input.file.copyvectors** labeled with identifiers that have been specified to be excluded from analysis (i.e. by having a "0" in the third column of **input.file.ids**) will also be ignored.

4.2 *painting_samples_filelist_infile*

This file contains a list of file locations and names of XXX.samples.out output files from CHROMOPAINTER, specifying one such file per line. These XXX.samples.out files should be for a CHROMOPAINTER analysis that paints the target population specified in **target.popname** using the donor populations specified in **copyvector.popnames** of "*parameter_infile*" (though carefully considered deviations may technically be possible, and non-target indi-

viduals can be included – these individuals will be ignored). Each of the XXX.samples.out files specified in "*painting_samples_filelist_infile*" should contain the same number of ≥ 1 painting samples from CHROMOPAINTER for each individual, and all files should contain the same number of individuals. When inferring admixture proportions, dates and sources, the weighted counts of DNA segments copied from every pairwise combination of donor populations will be summed across all files specified in "*painting_samples_filelist_infile*" and used to generate the coancestry curves for the target population.

Note that for each bootstrap re-sample calculated when **bootstrap.date.ind: 1** in "*parameter_infile*", N target individuals will be sampled with replacement from each file listed in "*painting_samples_filelist_infile*", and combined across all files to generate each new (bootstrap) date estimate. For this reason, we recommend specifying one XXX.samples.out file in "*painting_samples_filelist_infile*" per chromosome, as each chromosome of an individual can act as an independent unit of information, though smaller regions (e.g. 5 or 10Mb) can be used instead.

Each XXX.samples.out file contains an initial row that gives details of the CHROMOPAINTER run (note that the number of samples MUST be listed in the 21st column of this file, exactly as it is by CHROMOPAINTER). The remaining rows give the labels and the painting samples inferred for each haplotype of each painted recipient individual, including those from the target population, giving the sample number in the first column and the index of the donor haplotype copied at each SNP in the remaining columns. **Note:** As noted in Section 4.1.1, it is critical that the same **input.file.ids** used in GLOBETROTTER is also used for each CHROMPAINTER run used to generate the XXX.samples.out files for a given analysis, as these donor indices correspond to the order of individuals in this **input.file.ids** (which in turn should correspond to the order of individuals in the haplotype file input into CHROMOPAINTER using the '-g' switch). Also, the labels in the XXX.samples.out files must correspond to the individual labels in the first column of **input.file.ids** (though in any order), in particular for the target individuals.

An example for "*painting_samples_filelist_infile*" is provided in **"example/BrahuiYorubaSimulation.samplesfile.txt"**, which lists three XXX.samples.out files: **"example/BrahuiYorubaSimulationChrom20.samples.out.gz"**, **"example/BrahuiYorubaSimulationChrom21.samples.out.gz"** and **"example/BrahuiYorubaSimulationChrom22.samples.out.gz"**. (Note that there is one file per chromosome, though for simplicity we include only 3 of the 22 autosomes for this example. Note also that these files are gzipped, though they need not be. **In fact, users have reported that some c compilers prefer the *samples.out files to NOT be zipped, so that GLOBETROTTER will not work if they are.**) Therefore if bootstrap re-sampling is specified (i.e. **bootstrap.date.ind: 1** in "*parameter_infile*"), individuals will be sampled with replacement independently within each of these three files and then combined to generate bootstrapped date estimates. Note that these three

files contain only painted target individuals (i.e. from the "BrahuiYorubaSimulation" population), but could have contained other painted individuals as well (though only the individuals with labels matching those for the target population as provided in **input.file.ids** will be used).

4.3 *recom_rates_filelist_infile*

This file contains a list of file locations and names of input files used in the CHROMOPAINTER analyses corresponding to each of the XXX.samples.out files listed in "*painting_samples_filelist_infile*". Specifically each line should point to a file that was input (using the '-r' switch) into CHROMOPAINTER and that contains the recombination rates used for each SNP position. The number of SNPs (rows) specified in each of these files should exactly match the number of SNPs (columns) in the painting samples of the corresponding XXX.samples.out file specified in "*painting_samples_filelist_infile*".

The format of each file listed in "*recom_rates_filelist_infile*" is as follows. There should be a header line followed by one line for each SNP. Each line should contain two columns, with the first column denoting the basepair position value, in increasing order. The second column should give the genetic distance per basepair between the SNP at the position in the first column of the same row and the SNP at the position in the first column of the subsequent row. The last row should have a "0" in the second column (though this is not required – this value is simply ignored by the program). Genetic distance should be given in Morgans, or at least the relevant output files assume this value is in Morgans.

For example, to specify a chromosome with four SNPs at basepair positions 100, 250, 335 and 450, and with corresponding recombination rates between contiguous SNPs of 0.01, 0.02 and 0.05 Morgan per basepair, respectively, the recombination rate input file should look as follows:

```
start.pos recom.rate.perbp
100 0.01
250 0.02
335 0.05
450 0
```

(Note that, in order to allow numerical stability in C, **the minimum allowed recombination rate is 1×10^{-15} Morgan per basepair**. Any values below this in the second column of *recom_rate_infile* will be fixed automatically to this value in the corresponding CHROMOPAINTER analysis that used the given file, though it will be treated as 0 in the GLOBETROTTER analysis.)

An example for "*recom_rates_filelist_infile*" is provided in **"example/BrahuiYorubaSimulation.recomfile.txt"**, which lists three XXX.samples.out files: **"example/BrahuiYorubaSimulationChrom20.recomrates"**, **"example/BrahuiYorubaSimulationChrom21.recomrates"** and

"example/BrahuiYorubaSimulationChrom22.recomrates". Note that there is one file corresponding to each of the files specified in "*painting_samples_filelist_infile*", and that the number of SNPs match between the corresponding files.

5 Output

There are four possible output files for GLOBETROTTER:

5.1 [output.filename1].txt

This main output file, with prefix specified by **save.file.main** in "*parameter_infile*" and suffix **".txt"**, summarizes the inferred admixture proportions, dates and sources. It is generated only if **prop.ind: 1** in "*parameter_infile*".

The first line lists our "best-guess" conclusion for admixture in the target population. The choices are listed below, as well as a brief description in parenthesis of how each is chosen (a more complete description of these choices is given in [1]):

1. **uncertain** – admixture is detected but difficult to describe (combined fit quality for two events "fit.quality.2events" <0.985)
2. **one-date** – a single date of admixture between two sources (combined fit quality for two events ≥ 0.985; two-date score "maxScore.2events" <0.35; fit-quality for a single event "fit.quality.1event" ≥ 0.975)
3. **one-date-multiway** – a single date of admixture between more than two sources (combined fit quality for two events ≥ 0.985; two-date score <0.35; fit-quality for a single event <0.975)
4. **multiple-dates** – two (or more) distinct dates of admixture between two or more sources (combined fit quality for two events ≥ 0.985; two-date score ≥ 0.35)
5. **unclear signal – check curves/bootstraps** – when fitting two dates of admixture, no coancestry curve (red line) provided a very good fit to the data (black lines), suggesting the admixture signal – if any – is very unclear or nonexistent (the maximum two-date R^2 fit across all curves is <0.3, so "maxR2fit.1date" should also be low). In these cases, date estimates when bootstrapping may contain 1 or ≥ 400 when specifying **null.ind: 1**, indicating no clear evidence of admixture.

We note that these values and conclusions are based on a particular set of simulations, with specific sample sizes and population combinations, described in [1]. They at best can be used as a guide, though we recommend careful visual exploration of the inferred coancestry curves provided in the .pdf file to see how well

e.g. one versus two (versus 0) events fit the data (see Section 5.3). For example, we also advise performing admixture proportion, date and source inferrence specifying **null.ind: 1** in "*parameter_infile*" and performing some number (e.g. 100) of bootstrap re-samples to see if any re-samples have a "non-sensical" date (e.g. equal to 1 or $\geq$400 generations), suggesting GLOBETROTTER cannot reliably detect admixture. See Section 6 for a detailed example.

The next lines ("1-DATE FIT EVIDENCE, DATE ESTIMATE, SINGLE BEST-FITTING DONORS") give GLOBETROTTER's inferred date, proportions and "best-guess" sources of admixture for a single event or multiway admixture when assuming only a single date of admixture (i.e. particularly appropriate when the "best-guess" conclusion is "one-date" or "one-date-multiway"), as well as measures of "goodness-of-fit" for these events. In particular:

gen.1date – inferred date of admixture (in generations from present), when assuming only a single date

proportion.source1 – inferred proportion of admixture from the minority contributing source (for the strongest signaled event, in the case of multiway admixture) when assuming a single date

maxR2fit.1date – the goodness-of-fit (R^2) for a single date of admixture, taking the maximum value across all inferred coancestry curves

fit.quality.1event – the fit of a single admixture event (i.e. the first principal component, reflecting admixture involving two sources)

fit.quality.2events – the fit of the first two principal components capturing the admixture event(s) (the second component might be thought of as capturing a second, less strongly-signaled event)

bestmatch.event1.source1 – the single "best-guess" surrogate population that matches the inferred minority contributing source (for the strongest signaled event, in the case of multiway admixture) when assuming a single date

bestmatch.event1.source2 – the single "best-guess" surrogate population that matches the inferred majority contributing source (for the strongest signaled event, in the case of multiway admixture) when assuming a single date

proportion.event2.source1 – inferred proportion of admixture from the minority contributing source for the second, less strongly signaled event (when assuming a single date; particularly appropriate when the "best-guess" conclusion is "one-date-multiway")

bestmatch.event2.source1 – the single "best-guess" surrogate population that matches the inferred minority contributing source for the second, less strongly signaled event (when assuming a single date; particularly appropriate when the "best-guess" conclusion is "one-date-multiway")

bestmatch.event2.source2 – the single "best-guess" surrogate population that matches the inferred majority contributing source for the second, less strongly signaled event (when assuming a single date; particularly appropriate when the "best-guess" conclusion is "one-date-multiway")

The next lines of output ("2-DATE FIT EVIDENCE, DATE ESTIMATES, SINGLE BEST-FITTING DONORS") give GLOBETROTTER's inferred dates, proportions and "best-guess" sources of admixture when assuming two distinct dates of admixture (i.e. particularly appropriate when the "best-guess" conclusion is "multiple-dates"), as well as measures of "goodness-of-fit" for these events. In particular:

gen.2dates.date1 – inferred date of admixture (in generations from present) for the first event, when assuming two dates

gen.2dates.date2 – inferred date of admixture (in generations from present) for the second event, when assuming two dates

maxScore.2events – the additional goodness-of-fit (R^2) explained by adding a second date versus assuming only a single date of admixture, taking the maximum such value across all inferred coancestry curves

proportion.date1.source1 – inferred proportion of admixture from the minority contributing source for the first date's event (when assuming two dates)

bestmatch.date1.source1 – the single "best-guess" surrogate population that matches the inferred minority contributing source for the first date's event (when assuming two dates)

bestmatch.date1.source2 – the single "best-guess" surrogate population that matches the inferred majority contributing source for the first date's event (when assuming two dates)

proportion.date2.source1 – inferred proportion of admixture from the minority contributing source for the second date's event (when assuming two dates)

bestmatch.date2.source1 – the single "best-guess" surrogate population that matches the inferred minority contributing source for the second date's event (when assuming two dates)

bestmatch.date2.source2 – the single "best-guess" surrogate population that matches the inferred majority contributing source for the second date's event (when assuming two dates)

The next lines of output ("1-DATE FIT SOURCES, PC1") give GLOBETROTTER's inferred composition of each admixing source in the most strongly signaled event, when assuming only a single date of admixture. In particular every two consecutive rows describe the inferred genetic composition of one admixing source, giving both the proportion of DNA contributed by that source (first

column), and GLOBETROTTER's inferred mixture coefficients to describe each source (remaining columns – these should sum to 1 for each source). For example, the following output:

```
########################################
### 1-DATE FIT SOURCES, PC1
proportion Pop1 Pop5 Pop2
0.4 0.1 0.25 0.65
proportion Pop1 Pop6
0.6 0.23 0.77
########################################
```

implies that GLOBETROTTER has inferred the first admixing source, which contributes 40% of the DNA of the target population, to be best represented genetically as a mixture of (0.1,0.25,0.65) times the copy vectors of surrogate labels (Pop1,Pop5,Pop2), respectively. And GLOBETROTTER has inferred the second admixing source, which contributes 60% of the DNA of the target population, to be best represented genetically as a mixture of (0.23,0.77) times the copy vectors of surrogate labels (Pop1,Pop6), respectively.

Analogous source proportion and mixing coefficient inference is given next for the less strongly signaled event when assuming a single date of admixture (i.e. "1-DATE FIT SOURCES, PC2"), which is particularly appropriate when the "best-guess" conclusion is "one-date-multiway". And following this are the analogous inference for the first date's event when assuming two dates of admixture ("2-DATE FIT SOURCES, DATE1-PC1"), and the second date's event when assuming two dates of admixture ("2-DATE FIT SOURCES, DATE2-PC1"), which is particularly appropriate when the "best-guess" conclusion is "multiple-dates".

5.2 [output.filename1]_curves.txt

This output file, with prefix specified by **save.file.main** in "*parameter_infile*" and suffix **"_curves.txt"**, gives the coancestry curves for every pairwise combination of surrogate populations inferred to match >**props.cutoff** (see "*parameter_infile*") of the total ancestry of the target population, as well as information related to these curves. It is generated only if **prop.ind: 1** in "*parameter_infile*".

The first line of **[output.filename1]_curves.txt** gives our "best-guess" conclusion for admixture in the target population (see Section 5.1). The second line gives a header key, and the genetic distance (in cM) corresponding to the x-axis of each coancesty curve (see Section 4.1 and [1] for a description of how the bins and range for these cM distance values are specified).

The first two columns ("surrogate1", "surrogate2") denote the surrogate populations in a given pair.

The third column ("curve.description") describes each curve; of which there are the following four types per surrogate population pairing (output in consecutive rows):

1. **scaled.data** – the re-weighted counts of DNA segment pairs inferred to copy from surrogate populations donor1 and donor2; i.e. the (re-weighted) "data"

2. **gen.fit.1date** – GLOBETROTTER's inferred fit for a single date of admixture

3. **source.fit.1date** – GLOBETROTTER's inferred fit for a single admixture event (demonstrates reliability of source and proportion estimation when assuming a simple admixture event at a single time between only two sources)

4. **gen.fit.2date** – GLOBETROTTER's inferred fit for two distinct dates of admixture

The fourth column ("rsquared.date.fit") gives the goodness-of-fit (R^2) for a single admixture date for **gen.fit.1date** and **source.fit.1date** (note this column has identical values for these two rows within a given surrogate pair) or for two dates for **gen.fit.2date**.

The fifth column ("intercept.fit") and sixth column ("intercept.fit.date2") give the coefficients from fitting **scaled.data** using as predictors one or two exponential distributions with rates equal to the cM bins given in the second line scaled by the inferred dates of admixture. For **gen.fit.1date** and **gen.fit.2date**, this fit is accomplished using linear regression, while for **source.fit.1date** the coefficient is determined using the inferred admixture proportions and source mixing coefficients for the most strongly signaled event assuming a single date of admixture (see [1] for details). (Note that the sixth column is "NA" for **gen.fit.1date** and **source.fit.1date**, as they only have a single inferred date and hence a single predictor and coefficient.)

The remaining columns for each row give the y-axis values, corresponding to each x-axis cM bin value given in the second line, for each respective curve. In particular these y-axis values give the (scaled) probability of copying surrogate1 and surrogate2 at a pair of DNA segments separated by the corresponding x-axis (i.e. cM distance) value, for the raw data or one of the fitted models. For each of 1-4 above, these values are plotted for each pairing of surrogate populations in the pdf file described in Section 5.3.

5.3 [output.filename1].pdf

This output file, with prefix specified by **save.file.main** in "*parameter_infile*" and suffix **".pdf"**, plots the coancestry curves for every pairwise combination of

surrogate populations inferred to match >**props.cutoff** (see "*parameter_infile*") of the total ancestry of the target population. The given surrogate pairing is specified in each plot's title. The x-axis gives genetic distance in cM (see Section 4.1 and [1] for a description of how the bins and range for these cM distance values are specified). The y-axis gives the weighted and symmetrized probability of copying from the first and second surrogate populations listed in the title at a pair of DNA segments separated by the corresponding x-axis (i.e. cM distance) value. For each surrogate population pair, four such probabilities (lines) are shown, corresponding to 1-4 in Section 5.2 with the colors black, green, blue and red, respectively. It is generated only if **prop.ind: 1** in "*parameter_infile*".

5.4 [output.filename2].txt

This output file, with prefix specified by **save.file.bootstraps** in "*parameter_infile*" and suffix **".txt"**, gives the inferred dates and goodness-of-fit (R^2) values for bootstrap re-samples of individuals' DNA. It has 4-5 columns (depending on whether **num.admixdates.bootstrap** in "*parameter_infile*" equals **"1"** or **"2"**), denoting the bootstrap index ("bootstrap.num"), and for this given bootstrap the inferred date(s) of admixture ("date1.est.boot", "date2.est.boot"), the maximum goodness-of-fit (R^2) across all coancestry curves when fitting a single date of admixture ("maxR2fit.1date.boot"), and the maximum additional goodness-of-fit (R^2) across all coancestry curves explained by adding a second date versus assuming only a single date of admixture ("maxScore.2events.boot"). This file can be used to generate e.g. 95% confidence intervals around date estimates. It is generated only if **bootstrap.date.ind: 1** in "*parameter_infile*".

6 Detailed (recommended usage) example

As throughout the document, here we refer to the population you are testing for admixture as the "target" population. We refer to any other sampled populations used to describe the target population's admixture event(s) as "surrogate" populations. Finally, "donor" populations refer to groups used to paint both target and surrogate populations using CHROMOPAINTER.

As a first step, sampled individuals must be classified into population labels. Rather than strictly relying on provided labels based on e.g. self-identification, we recommend using the program fineSTRUCTURE [2] to cluster individuals into genetically homogeneous groups, which can then be used as "surrogate", "target" and "donor" populations. Though we note it can still often be helpful to use provided labels in order to interpret admixture signals (e.g. to use geographic information on where individuals were sampled from), signals may be confusing or misleading if some individuals assigned to the same label are genetically rather different, so that we recommend this is assessed *a priori*. We suggest doing this using fineSTRUCTURE or principal components of the fineSTRUCTURE coancestry matrix, since these are based on similar or identical copy vectors to

those used by GLOBETROTTER. It is also possible to do this step using other software, e.g. ADMIXTURE [4], EIGENSOFT [5] or similar programs.

6.1 CHROMOPAINTER (and possibly fineSTRUCTURE) analyses

There are two initial CHROMOPAINTER analyses to run:

1. generate "copy-vectors" – paint all surrogate and target individuals conditional on a set of donor individuals
2. generate "painting samples" – paint all target individuals conditional on a set of donor individuals

For (1), the set of "donor" individuals might (and often likely will) contain the set of surrogate (and possibly target) individuals, though this need not be the case. Ideally each target/surrogate individual should be painted using the same number of individuals from each donor (e.g. population) label, though this number need not be the same for all donor labels. For example, if you use as donors your K surrogate (and possibly target) groups, with group $k \in [1, ..., K]$ containing n_k individuals, you might paint each surrogate and target individual using $n_k - 1$ individuals from each donor population $k \in [1, ..., K]$ to account for the fact that each individual from population k cannot copy themself (and hence only copy $n_k - 1$ individuals from population k). However, for simplicity users e.g. may want to define target, surrogate, and/or donor labels using a fineSTRUCTURE [2] analyis. (I.e. the program fineSTRUCTURE clusters a set of individuals into genetically homogeneous groups, and these groups can be used as the "population labels" when testing for admixture.) To perform fineSTRUCTURE analysis in order to cluster a set of N individuals, CHROMPAINTER must be run allowing each of the N individuals to copy from all others (i.e. using the '-a' switch in CHROMOPAINTER). The XXX.chunklengths.out files from this analysis can subsequently be used (once appropriately combined across chromosomes and into a single file) as the **input.file.copyvectors** of "*parameter_infile*" (see Section 4.1) in the GLOBETROTTER analysis (though see **WARNING** at end of this section).

For (2), the set of "donor" individuals should ideally be the same set as that used in (1), though it is recommended that you do NOT use members of the target population as donors, as this will mask admixure signals (i.e. the target individuals will predominantly be painted using other individuals from their own population label, rather than identifying more distantly related ancestors). In such a case, when dating we ignore any segments in the target population that were painted using other individuals from their own population label, potentially leading to a substantial decrease in power.

In practice we have found GLOBETROTTER analyses to be robust to the exact procedure used, with no obvious "correct" pipeline. For example, in the

"full analysis" and some simulations described in [1], for step (1) all target and surrogate individuals were painted using both target and surrogate individuals as donors, while for step (2) target individuals were painted using only surrogate individuals as donors. In contrast, for the "regional analyses" and other simulations in [1], including the "BrahuiYorubaSimulation" example provided here, for step (1) all target and surrogate individuals were painted using only surrogate individuals as donors and for step (2) all target individuals were painted using only surrogate individuals as donors.

WARNING: If using the '-a' switch, specifying all included individuals copy from all other included individuals, bear in mind that each individual cannot copy from themself. As a result, if population k contains n_k individuals, each individual in that population will copy from $n_k - 1$ individuals of population k. In contrast, individuals from all other populations will copy from all n_k individuals of population k. GLOBETROTTER assumes the copying vectors across all populations are generated in the same manner, e.g. that each individual copies from precisely the same number of individuals from each group. It is not clear how to correct for this "$n_k - 1$" discrepancy. (Indeed this is why we performed a "leave-one-out" analysis – which excluded one individual from each population in the set of donors used to generate copying vectors – in our "full analysis" results reported in [1].) For example, in practice we have found that if you exclude an individual from population k, the amount of genome-wide DNA that would have been copied from that excluded individual under CHROMOPAINTER largely is distributed amongst the remaining $n_k - 1$ individuals, i.e. so that the resulting copy vectors are not noticeably different from the ideal scenario where everyone copies from the *exact* same set of donors. This suggests that the slight shortcut from using results taken when specifying the '-a' switch will not greatly alter GLOBETROTTER's analysis. However, it seems possible that this effect might be more pronounced for populations where n_k is particularly small (e.g. <5-10 individuals or so) and/or the surrogate/donor populations are relatively genetically similar (e.g. inferring admixture among different European groups). This has not been explored extensively.

6.2 GLOBETROTTER analyses

When running GLOBETROTTER, we recommend you first test for *any* evidence of admixture by setting **null.ind: 1** in "*parameter_infile*" (see Section 4.1). This accounts for any "unusual" patterns of linkage disequilibrium that may lead to a false signal of admixture. Specifically, you want to first run GLOBETROTTER setting **null.ind: 1** and **prop.ind: 1** in "*parameter_infile*" to infer admixture proportions, dates and sources (we recommend setting **num.mixing.iterations: 5** when doing so). Once this finishes, with **null.ind: 1** still, you then want to set **save.file.main** to be the prefix of the output from this initial GLOBETROTTER run, and then perform a second GLOBETROTTER run setting **prop.ind: 0**, **bootstrap.date.ind: 1** and **bootstrap.num: 100**. (Note for this last GLOBETROTTER run, if you have multiple computing cores at your disposal,

you may want to e.g. set **bootstrap.num: 20** and repeat 5 times, specifying a different **save.file.bootstraps** each time, in order to complete these bootstraps 5 times faster.) In the resulting **save.file.bootstrap** file(s), the proportion of inferred date(s) that are $\leq$1 or $\geq$400 give you the p-value for *any* evidence of detectable admixture.

Assuming there is detectable admixture, set **null.ind: 0** in "*parameter_infile*" and re-run GLOBETROTTER setting **prop.ind: 1** in "*parameter_infile*" to infer admixture proportions, dates and sources (again we recommend setting **num.mixing.iterations: 5** when doing so). Once this finishes, you then want to set **save.file.main** to be the prefix of the output from this initial GLOBETROTTER run, and then perform a second GLOBETROTTER run (with **null.ind: 0**) setting **prop.ind: 0**, **bootstrap.date.ind: 1** and **bootstrap.num: 100**. (Again note that if you have multiple computing cores at your disposal, you can run multiple GLOBETROTTER runs in parallel to infer bootstrapped dates.) Use the results in the **save.file.main** and **save.file.bootstrap** output files to assess the evidence for simple admixture involving two admixing sources intermixing at a single date, versus complex admixture involving multiple sources and/or multiple dates of admixture, and to describe the event(s) including the confidence intervals around the inferred date(s) of admixture. You should also check for consistency of all inferred values (i.e. dates, sources, and proportions) with the **null.ind: 1** analysis.

Note that the two **null.ind** procedures described above can be run in parallel.

6.3 analysis summary

To summarize, given a set of sampled "target" individuals you wish to explore for admixture, we recommend the following approach (though note that there are other valid approaches!):

1. Use CHROMOPAINTER to paint each individual conditional on every other individual, combining the resulting XXX.chunklengths.out files across all chromosomes (this combined file will be used as **input.file.copyvectors** of "*parameter_infile*" – see Section 4.1).

2. Using the XXX.chunkcounts.out files from step 1 (you could also use the XXX.chunklengths.out files, though fineSTRUCTURE has been optimized for counts rather than lengths), combined across chromosomes, use fineSTRUCTURE to cluster individuals into K genetically homogeneous groups. (See [2] for details – note that you may not want to use fineSTRUCTURE's final inferred number of clusters, but instead use a smaller number of groups K by e.g. extracting results from fineSTRUCTURE's inferred tree at an "appropriate" level.) These K groups can be used to label column 2 of **input.file.ids** of "*parameter_infile*", and used as e.g. **copyvector.popnames**, **surrogate.popnames**, and (one at at time as) **target.popname**.

3. Separately for each group $k \in [1, ..., K]$ defined in step 2, use CHROMOPAINTER to paint each individual in the group conditional on individuals from all *other* groups (i.e. excluding the other individuals from k) to generate one XXX.samples.out file per chromosome. These XXX.samples.out files will be listed in "*painting_samples_filelist_infile*" (see Section 4.2), and the corresponding recombination files input into CHROMOPAINTER for this analysis will be listed in "*recom_rates_filelist_infile*" (see Section 4.3).

4. Separately for each group $k \in [1, ..., K]$, run GLOBETROTTER with group k as the target population setting **null.ind: 1** in "*parameter_infile*", inferring admixture proportions, dates and sources over 5 iterations (i.e. **num.mixing.iterations: 5**) and performing 100 bootstrap re-samples to generate confidence intervals around the estimated date.

5. Separately for each group $k \in [1, ..., K]$, run GLOBETROTTER with group k as the target population setting **null.ind: 0** in "*parameter_infile*", inferring admixture proportions, dates and sources over 5 iterations and performing 100 bootstrap re-samples to generate confidence intervals around the estimated date.

6. Calculate the proportion of date estimates (including all bootstrap re-samples) from step 4 that are $\leq$1 or $\geq$400 – this gives the p-value for evidence of "any detectable admixture" for group k.

7. If the p-value from step 6 is "significantly small", use the results from step 5 to describe the admixture in group k, assessing the evidence for simple admixture involving two admixing sources intermixing at a single date, versus complex admixture involving multiple sources and/or multiple dates of admixture. Check for consistency with the inferred results from step 4.

6.3.1 included example files

To describe admixture in the attached simulated example, which consists of 20 individuals simulated as descendants of an admixture event occuring 30 generations ago with 80% of the DNA contributed by the Brahui and 20% contributed by the Yoruba (this is the simulation described in Figure 1 of [1]), type the following:

```
R < GLOBETROTTER.R example/BrahuiYorubaSimulation.paramfile.txt
example/BrahuiYorubaSimulation.samplesfile.txt
example/BrahuiYorubaSimulation.recomfile.txt --no-save > output.out
```

As **prop.ind: 1** in "example/BrahuiYorubaSimulation.paramfile.txt", this will infer admixture proportions, dates and sources using five iterations (i.e. **num.mixing.iterations: 5**). Note that only data from chromosomes 20-22 are included, so that results may not be consistent with that of Figure 1 of [1].

This will generate the following output files describing the admixture event and showing the inferred coancestry curves: "example/BrahuiYorubaSimulation.globetrotter.main.txt", "example/BrahuiYorubaSimulation.globetrotter.main_curves.txt" and "example/BrahuiYorubaSimulation.globetrotter.pdf".

To next perform bootstrap re-samples for inferring confidence intervals around the inferred date of admixture in this simulation, type the following:

```
R < GLOBETROTTER.R example/BrahuiYorubaSimulation.paramfileII.txt
example/BrahuiYorubaSimulation.samplesfile.txt
example/BrahuiYorubaSimulation.recomfile.txt --no-save > output.out
```

Note that "example/BrahuiYorubaSimulation.paramfileII.txt" differs from the input file "example/BrahuiYorubaSimulation.paramfile.txt" from the previous GLOBETROTTER run only in that **prop.ind: 0** and **bootstrap.date.ind: 1**, so that now 20 bootstrap re-samples will be output to a new file "example/BrahuiYorubaSimulation.globetrotter.boot.txt".

Alternatively, to describe the DNA of a given target population as a mixture of that of given surrogate populations *without* attempting to infer any admixture events, i.e. using the technique described in [3], change **num.mixing.iterations** to "0" in "example/BrahuiYorubaSimulation.paramfile.txt" and run using:

```
R < GLOBETROTTER.R example/BrahuiYorubaSimulation.paramfile.txt
--no-save > output.out
```

This will give a single output file, "example/BrahuiYorubaSimulation.globetrotter.main.txt", which will contain the inferred proportions of ancestry relating the target group to each surrogate group with inferred proportion above the threshold specified by **props.cutoff**.

7 Computational Complexity

The computational complexity of GLOBETROTTER is $o[NC(B+M)(SL+J^2I+J^2I_j^2+GJ^2K^2)+C[\min(N,100)]^2(L+I_j^2)]$ for N target population individuals, C chromosomes, B bootstrap re-samples, M mixing iterations, S painting samples, L SNPs (maximum across all chromosomes), J donor populations, K surrogate groups, $I(\leq SL)$ total "chunks" (i.e. the maximum number of "chunks" across chromosomes and individuals), $I_j(\leq I)$ "chunks" copied from donor population j (the maximum number of "chunks" across chromosomes and individuals copied from a single donor population) and G the number of cM-scaled grid points over which the coancestry curve is evaluated (i.e. $G =$**(curve.range[2]-curve.range[1])/bin.width**). Here a "chunk" is a contiguous segment of DNA copied from a single donor population. The number I of such "chunks" refers to the total number of chunks across all S samples in a single target individual and

a single chromosome, and specifically the maximum such value across all target individuals and chromosomes. The exact I is data-dependent, but $I \leq SL$ and is often orders of magnitude less than SL depending on linkage disequilbrium levels.

As an example, for the Brahui-Yoruba simulations in Figure 1 of [1], when using N=20, C=22 (i.e. whole genome chip data), B=20, M=0, S=10, L=39655 and J=93, it took GLOBETROTTER <24 hours to run on a 3.16GHz Intel Core 2 Duo with 3.7Gb RAM.

This updated version of GLOBETROTTER may increase the speed of the program from previous versions, by now storing information from target individuals (specifically the coancestry curves of each individual) as it goes along. (You may not notice the increase unless your XXX.samples.out files contain lots of individuals.) But this potential speed-up comes with a memory cost. While the memory of the previous GLOBETROTTER was fairly trivial, this version of GLOBETROTTER will store $O(\max\{NL, SL\})$ integers plus $O(NGK^2)$ doubles.

8 Citation

When making use of GLOBETROTTER, please cite the following paper:

Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D. and Myers, S. (2014) "A Genetic Atlas of Human Admixture History" *Science* **343**:747-751

When making use of ChromoPainter and/or fineSTRUCTURE, please cite the following paper:

Lawson, D., Hellenthal, G., Myers, S., and Falush, D (2012) "Inference of population structure using dense haplotype data" *PLoS Genet* **8(1)**:e1002453

Questions? Bugs? Please contact Garrett Hellenthal at ghellenthal@gmail.com.

References

[1] G. Hellenthal, G.B.J. Busby, G. Band, J.F. Wilson, C. Capelli, D. Falush, and S. Myers. A genetic atlas of human admixture history. *Science*, 343:747–751, 2014.

[2] D.J. Lawson, G. Hellenthal, S. Myers, and D. Falush. Inference of population structure using dense haplotype data. *PLoS Genet*, 8(1):e1002453, 2012.

[3] S. Leslie, B. Winney, G. Hellenthal, D. Davison, A. Boumertit, T. Day, K. Hutnik, E.C. Royrvik, B. Cunliffe, Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, D.J. Lawson, D. Falush, C. Freeman, M. Pirinen, S. Myers, M. Robinson, P. Donnelly, and W. Bodmer. The fine scale genetic structure of the British population. *Nature*, 519:309–314, 2015.

[4] D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664, 2009.

[5] N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Gen*, 2(12):e190, 2006.