GTmix: A Program for Maximum Likelihood Inference of Admixture Networks from Population Genetics Data

User Manual

Version 1.3.0

October 28, 2019

Yufeng Wu
CSE Department, University of Connecticut Storrs, CT 06269, U.S.A.
Email: yufeng.wu@uconn.edu

Yufeng Wu, Inference of Population Admixture Network from Local Gene Genealogies: a Coalescent-based Maximum Likelihood Approach, manuscript, 2019.

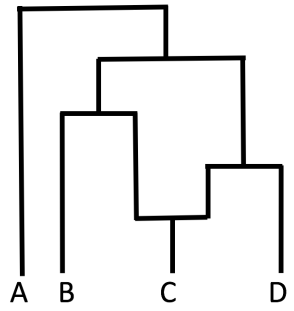
1 Getting Started with GTmix

1.1 Program availability

GTmix is written in C++. Executables for popular platforms such as Linux 32 bits or 64 bits and MacOS are downloadable from GitHub:
`https://github.com/yufengwudcs/xxxx`. Files can be downloaded using "Save Link/Target As..." After downloading the softwares, you may need to change file access permissions (e.g. chmod u+x gtmix-linux64). In case that you want to compile the code yourself, source code is also available for download at the above URL. To compile the code, first put the gzip file in the directory youd like and unzip it: use gunzip and tar commands such as:
▷ gunzip ⟨gtmix-src.tar.gz⟩
▷ tar -xvf ⟨gtmix-src.tar⟩
Then type:
▷ make at the prompt. This creates an executable called gtmix, which can be run by typing
▷ ./gtmix at the prompt. You will need to specify some input options - see below.

1.2 What is GTmix?

First, where does the name GTmix come from? It stands for {G}ene {t}ree based maximum likelihood inference of ad{mix}ture network.



(a) Admixture network

	s_1	s_2	s_3	s_4	s_5	s_6	s_7
a_1	1	0	0	0	1	0	0
a_2	0	0	0	0	1	0	0
b_1	0	0	0	0	0	1	0
b_2	0	1	1	0	0	1	0
c_1	0	1	1	0	0	0	0
c_2	0	0	1	0	0	0	1
d_1	0	0	0	1	0	0	1
d_2	0	0	0	1	0	0	1

(b) Haplotypes

Figure 1: *Illustration of admixture network. Network is shown in part 1(a). Four populations.* C *is admixed. Part 1(b): haplotypes for the two gene trees. Two lineages are sampled from each population:* a_1 *and* a_2 *are from population* A, b_1 *and* b_2 *are from population* B *and so on.*

GTmix is designed to infer population admixture network (the demographic history of population with admixture) from population genetic data. Admixture network is a model for population demographic history, which explicitly models admixture. One example is shown in Figure 1(a). Here, there are four extant populations, and the population C is admixed. The main objective of GTmix is inferring the population admixture network from population genetic data.

GTmix woks with population haplotypes. Haplotypes are the phased population genetic variation data at multiple variation sites (usually single nucleotide polymorphisms or SNPs). One or multiple haplotypes are sampled from each population. GTmix assumes the haplotypes are in the form of binary sequences. See Figure 1(b) for an illustration of haplotypes. Here, we sample two haplotypes from each population. For example, two haplotypes a_1 and a_2 are from the population A at seven SNP sites. Note that GTmix doesn't directly work with haplotypes. Instead, GTmix takes input as local gene genealogies, rather than haplotypes. Local gene genealogies can be inferred from population haplotypes (see below).

1.3 How does GTmix work?

GTmix chooses a subset of local gene genealogies and finds the maximum likelihood estimate (MLE) of admixture network for the chosen genealogies. The likelihood is based on the multispecies coalescent model. Refer to the paper for more details on the methodology of GTmix.

2 Functionalities and Usage of GTmix

2.1 Preparing inputs

2.1.1 Haplotypes

To run GTmix, the user needs to first have population haplotypes. These haplotypes should be in the form of binary sequences. You should have one or multiple haplotypes from each population under study. When you have multiple loci, you should have an individual haplotype file for each locus. At a single locus, the format of the haplotype file should be: the first line species the SNP positions. I would recommend to use fractional positions between 0 and 1. Here is an example with 11 SNPs and eight haplotypes (two haplotypes for each of four populations).

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
10001111111
10010101010
00101010101
01111100110
11001111010
10110011111
00011110000
11010011111

Here, each haplotype is at a different row. Don't leave space between SNP alleles. For more guideline on haplotypes, check out the instructions provided by the program RENT+.

2.1.2 Inferring local genealogies

GTmix takes inferred local genealogies as input, not haplotypes. To infer local genealogies, I recommend to use the program RENT+. The executable of RENT+ is distributed together with GTmix, for your convenience. To run RENT+, you run for haplotypes from a single locus:

java -jar RentPlus.jar test-mat-1.hap

Here, test-mat-1.hap contains the haplotypes. The inferred local genealogies are stored in a file called test-mat-1.hap.trees.

2.1.3 Picking trees for inference

Usually there are too many local genealogies inferred by RENT+, which makes GTmix too slow. I recommend to choose a subset of local genealogies for inference. A simple tool called TreePicker is distributed as part of GTmix. To run TreePicker, do:

./treepicker test-mat-1.hap.trees test-mat-1.hap 5 > test-mat-1.hap.trees.chosen

The above command samples five trees and store in a file called test-mat-1.hap.trees.chosen.

Note: GTmix takes a file containing local gene genealogies. In this file, each line contains a local genealogy in Newick format. Don't include any other information such as site positions. The genealogies should have numerical taxa from 1 to n (the number of haplotypes).

Here, the numerical taxa corresponds to the index of the haplotype row in the haplotype matrix. For example, the following is a list of trees to be given to GTmix (there are four populations; each has two haplotypes):
((((2,6),8),(1,5)),((4,7),3))
((((((1,7),3),(2,6)),5),8),4)
(((((1,5),7),(2,8)),6),(3,4))
(((((1,5),3),7),((6,8),2)),4)
(((((1,5),3),7),((6,8),2)),4)
(((((1,5),(2,4)),7),(6,8)),3)
((((1,5),7),(2,4)),((6,8),3))
((((1,5),2),((4,7),3)),(6,8))
(((((6,8),1),4),((2,5),7)),3)
((((6,8),1),(3,7)),((2,5),4))
((((2,5),4),7),(((6,8),1),3))

2.1.4 Population information file

GTmix requires a text-based file for specifying the population information. In particular, GTmix needs to know which haplotypes are for which population. The format of the file is: each line is for a different population; the format of a line is:

<population name> <number of haplotypes for this population> <haplotype row index 1> <haplotype row index 2> ...

Haplotype row index starts from 1 (i.e. the first haplotype is row 1). For example, suppose there are four populations and each population has two haplotypes. Assume haplotypes 1 and 2 are from the population A, haplotypes 3 and 3 are from the population B and so on. Then the following specifies such a settings:
A 2 1 2
B 2 3 4
C 2 5 6
D 2 7 8

2.2 Usage

To run GTmix, you must provide an input file with the chosen gene genealogies (using the approach as specified above). You also need a file to specify the population.

▷ ./gtmix -P <population-spec-file> <list-of-genealogies-file>

By default, GTmix infers a network with a single admixture event. In order to infer a network with more admixture events, you should specify the "-n" option:

▷ ./gtmix -n <num-of-admixture> -P <population-spec-file> <list-of-genealogies-file>

GTmix outputs the inferred admixture networks in two ways. First, it outputs all the population trees contained in the network. These trees are obtained by keeping exactly one

incoming edge at each admixture node. For example, there are two population trees in the network in Figure 1(a). The population trees are in the well-known Newick format. Second, the network is output to a file in the GML format. The default file name for the output is "optimal-network.gml". The GML format is a format for graphs. There are open source GML format graph viewers. Also, GML is in plain text and so it is easy to understand and do the format conversion yourself.

▷ ./gtmix -o <output-network-file> -P <population-spec-file> <list-of-genealogies-file>

Sometimes there is a population that is the outgroup. You may specify the outgroup using the "-r" option. For example, if the population A is the output, use the following command:

▷ ./gtmix -r A -P listPops.txt listtrees.txt

By default, GTmix will use up to K trees to infer networks, if there are more than K trees given. By default, $\mathsf{K} = 500$. This is to ensure the scalability of the inference when the number of trees is very large. There is a trade-off between accuracy and efficiency here: using more trees may increase the inference accuracy but will also slow down the inference. To change the value of K, use:

▷ ./gtmix -T <number-of-trees> -P <population-spec-file> <list-of-genealogies-file>

2.3 Command line options

For ease of reference, I now provide the list of (optional) command line options.

1. -n ⟨number-of-admixture⟩ by default, the number of admixtures is one. **Note**: inference will be slower if the number of admixtures increases.
2. -r ⟨outgroup-population⟩ specify the outgroup population; by default there is no outgroup.
3. -o ⟨output-file-name⟩ specify the name of output network file (in GML format). By default, the output file name is: optimal-network.gml.
4. -T ⟨number-of-trees-to-use⟩ specify the maximum number of trees to use in inference; by default, use at most 500 trees.

3 Remarks on GTmix

3.1 Multiple loci

When you have more than one loci (say you have 100 loci), I would suggest you to infer the local genealogies for each locus separately and then use the "TreePicker" utility to choose a fixed number of trees to use from each locus (say each locus, you choose 50). Then, put all these chosen files in one file (in the example here, you would have $100 \times 50 = 5,000$ trees; recall to put each tree in a separate line). Then you run GTmix with this concatenated tree

file. Recall that GTmix would again choose a subset of trees in inference (by default, 500). You can change this by the "-T" option.

3.2 Efficiency

GTmix scales well with the number of trees. However, GTmix doesn't scale very well with regarding to the number of populations, and the number of haplotypes (alleles) per population. As a rule of thumb, GTmix is likely to be slow when the total number of haplotypes of all populations is 50 or more. I would suggest you to experiment with data of difference sizes. My simulation shows that even with small number of haplotypes, GTmix can still infer reasonably accurate admixture network. In fact, GTmix can infer more accurate networks than existing methods such as TreeMix even when existing methods are given much larger data.

4 Revision History

1. 10/28/2019: Release of v.1.3.0. Include basic functionality of admixture network inference.