Phyloage User Guide

Dirk Struve phylofriend at projectory.de https://github.com/yogischogi/phyloage/

March 1, 2016

Contents

1	Introduction	3
2	Installation	5
3	Command Line Options	6
4	First example 4.1 Input Tree	7 7
\mathbf{R}	eferences	8

1 Introduction

Phyloage is an experimental program for genetic genealogy and history research. It combines SNP and Y-STR results to calculate TMRCA (Time To Most Recent Common Ancestor) estimates from STR mutational differences.

Since the emergence of next generation sequencing SNP based phylogenetic trees are rapidly evolving and less attention is paid to Y-STR mutations. But if we look at the properties of SNP mutations in detail, we see that they are not without drawbacks.

- SNP mutations are extremely stable. Phylogenetic trees based on SNPs are highly reliable.
- Not all SNPs are useful for genealogical purposes. This is problematic for TMRCA calculations because wrong SNPs lead to false results.
- For TMRCA calculations the margins of error are rather large. YFull has tried to include only genealgical relevant SNPs into their phylogenetic tree [8] and estimates one mutation every 140 years [1].

What we want is more precise estimates. This could be achieved by using another kind of mutation, the well known Y-STRs. Their properties are rather different from SNPs.

- STRs have been used for genetic genealogy for a long time. Some specific marker sets are well known and many people have tested.
- STRs mutate back and forth. It is impossible to create reliable phylogenetic trees that go back deep into history. Furthermore the back and forth mutations lead to a saturation effect that invalidates time estimates for long time spans (thousands of years).
- Next generation sequencing reveals results for up to 500 markers. This should give us enhanced precision for TMRCA estimates.

So what does Phyloage do? The basic idea is to start with an SNP based phylogenetic tree, insert the Y-STR results and calculate modal haplotypes for each node of the tree.

In theory this should reduce saturation effects and give better time estimates for both deep history and genealogical time frames.

If you use this program, please remember that it is still highly experimental. At the time I am writing this, there is no person on earth who has a long time experience with 500 Y-STR markers.

Have fun experimenting!

Dirk

2 Installation

This guide is mainly targeted towards persons who use Linux Mint or other Linux versions of the Debian family. Some familiarity with the use of Linux commands is assumed.

Currently there are no binary distributions available for Windows or the Mac. Users of these operating systems can use Phyloage as well, but they will experience some laborious installation work. The best way is to follow the instructions provided on the Go home page.

The following list applies to Linux users only:

- Make sure that the Go programming language is installed. If not it can be installed by typing sudo apt-get install golang
- 2. Read the Go Getting Started guide. Make sure to set your GOPATH variable and include it in your PATH so that Go programs can be found.
- 3. Fetch the Phyloage program with go get github.com/yogischogi/phyloage
- 4. Install the program with go install github.com/yogischogi/phyloage

3 Command Line Options

Command line options may be given in arbitrary order.

- -help Prints available program options.
- **-treein** Filename of the SNP based phylogenetic tree.
- **-treeout** Filename of the results tree in text format.
- -personsin Filename or directory of files containing the persons' Y-STR values. If this is a single file it must contain results for multiple persons. The input file format is CSV (comma separated values) or text format. If a directory is provided for input it must contain multiple files in

YFull format, each file containing the results for a single person. The person's ID is extracted from the filename.

personsin supports multiple file names separated by commas.

- **-mrin** Filename of the mutation rates to use.
- **-gentime** Generation time.
- -cal Calibration factor.

4 First example

4.1 Input Tree

Before you can start you need to create a phylogenetic tree, that contains SNPs and the IDs of the genetic samples. The file format is text based and looks like this:

```
// This is an example tree.
// Comments begin with //.
L51
   P310
       P311
           P312
               id:YF00009
               id:YF00301
               id:YF02016
           U106
               id:YF00815
               id:YF01234
               id:YF05678
           S1194
               CTS4528
                  id:YF04242
                  id:YF00101
                  id:YF01010
```

The tree is indented by using tabs. Each sample starts with *id:* followed the ID. In our case these are typical YFull IDs but Phyloage supports Family Tree DNA data as well. Phyloage uses the Phylofriend [6] program for data import and many calculations as well. See the Phylofriend User Guide [7] for the details of the supported input formats.

You can safe the input tree as *tree.txt*. Next you need to put the files containing the sample data into a directory for example *allsamples*. Now you can invoke the program by

```
phyloage -treein tree.txt -treeout results.txt
-personsin allsamples -mrin 111-average.txt -gentime 32
```

The results will be stored in a file named *results.txt*. We have used average mutation rates for 111 markers and a generation time of 32 years.

References

- [1] Dmitry Adamov, Vladimir Guryanov, Sergey Karzhavin, Vladimir Tagankin, Vadim Urasin. *Defining a New Rate Constant for Y-Chromosome SNPs based on Full Sequencing Data*. The Russian Journal of Genetic Genealogy (Русская версия), Vol 6, No 2 (2014)/Vol 7, No 1 (2015).
- [2] David Hamilton, *An accurate genetic clock*, bioRxiv preprint, first posted online June 15, 2015, doi: 10.1101/020933.
- [3] ISOGG, *Y-DNA Haplogroup R and its Subclades*. Date visited: 2015-10-05.
- [4] Anatole A. Klyosov, DNA Genealogy, Mutation Rates, and Some Historical Evidence Written in Y-Chromosome, Part I: Basic Principles and the Method. Journal of Genetic Genealogy, 5(2):186-216, 2009.
- [5] Sergey Malyshev, R1b-M269 (P312- U106-) DNA Project (aka ht35 Project) Phylogenetic Tree. R1b-M269 (P312- U106-) DNA Project, Date visited: 2015-10-05.
- [6] Dirk Struve, *Phylofriend*, a program to calculate genetic distances. Google Project Hosting, 2014; GitHub, 2015.
- [7] Dirk Struve, *Phylofriend User Guide*. Google Project Hosting, 2014; GitHub, 2015.
- [8] YFull, YFull Phylogenetic Tree. Date visited: 2016-03-01.