Phylogrowth User Guide

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

June 27, 2016

Contents

1 Introduction

Phylogrowth is a program that calculates the expansion of a haplogroup by counting subclades and samples for given periods of time.

Periods of human growth and prosperity are often accompanied by strong population growth. For a society strong population growth simply means more surviving children. If more children survive there is also a higher chance of more mutations and thus more new branches on the phylogenetic tree. So a high number of new branches within a given period of time indicates strong population growth and good times.

Population growth usually continues until resources get scarce or the society is threatened by epidemics, natural disasters or war. In such cases we expect that the main haplogroups of a population would stop to grow or at least grow slower.

Given highly accurate phylogenetic trees with good time estimates it should be possible to detect different periods of human history and derive valuable conclusions.

Modern technologies like next generation sequencing have made it possible to create such phylogenetic trees and to utilize the important archeological evidence we all carry within our DNA.

Phylogrowth uses phylogenetic trees to compute changes in haplogroup growth. This should give us more insights into our history.

Have fun exploring human history!

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 Phylogrowth 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:

1. 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 Phylogrowth program with
   `go get github.com/yogischogi/phylogrowth`

4. Install the program with
   `go install github.com/yogischogi/phylogrowth`

Now the Phylogrowth program should be installed.

3 Command Line Options

Command line options may be given in arbitrary order. Parameters may be specified by using a space or equals sign. For example the following options are identical: `-treein=mytree`, `-treein mytree`.

-help Prints available program options.

-treein Input filename for phylogenetic tree (.txt).

-treeout Output filename for phylogenetic tree in TXT format.

-csvout Output filename for histogram data in CSV format.

-txtout Output filename for histogram data in TXT format.

-pngout Output filename for PNG image. Needs Gnuplot [2] to be installed.

-step Step length of histogram intervals in years. Default is 100.

-subclade Selects a specific branch of the tree.

4 Examples

4.1 First Steps

Phylogenetic Input Tree

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

```
// This is an example tree.
// Comments begin with //.

CTS4528, S1200 TMRCA 5000
    S11481 TMRCA 2000
        id:YF01234
        id:YF00301
        id:YF02016
    S14328 TMRCA 4500
        id:YF04242
        id:YF00101
        id:YF01010
```

Each line of the tree contains one or more SNPs or a sample ID. Subclades and samples are indented by using tabs or spaces. Each sample starts with `id:` followed by the ID.

If a subclade has a TMRCA estimate Phylogrowth counts the number of additional new lineages. As a first rough estimate we assume that the number of new lineages is proportional to population growth.

To create a graph that shows the new lineages from the example tree:

1. Save the example tree to a file, for example *tree.txt*.
2. Go to a command line and type:
   `phylogrowth -treein=tree.txt -pngout=example.png`
   For this to work you need Gnuplot [2] to be installed.
3. Done! Your first graph should be stored in the file *example.png*.

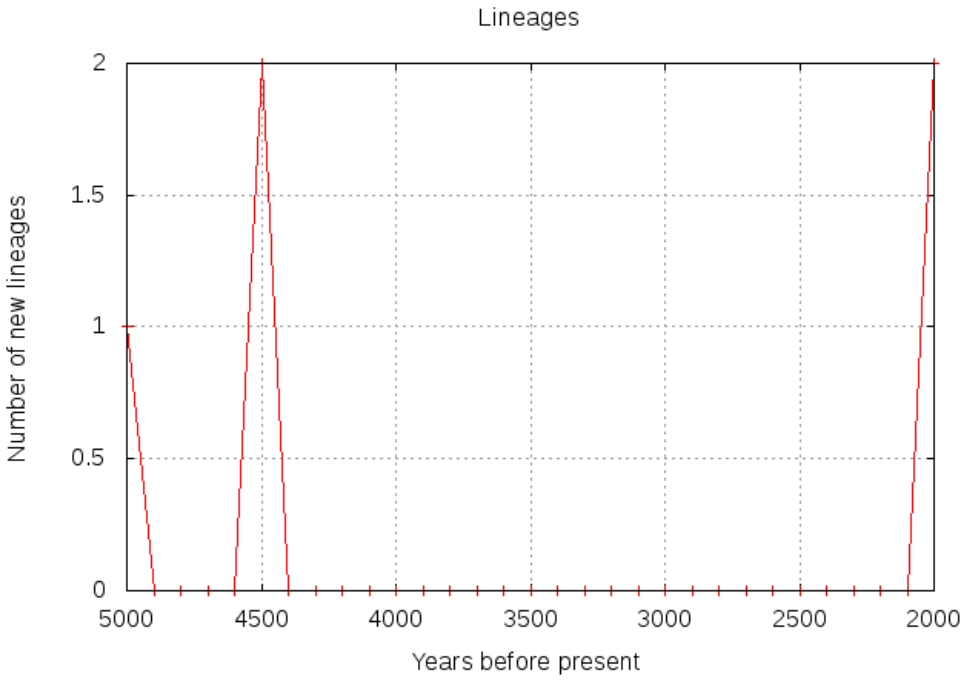

Figure 1: The first example showing time periods of haplogroup expansion.

4.2 Using the YFull Tree

Phylogrowth is especially well suited for the use of the YFull tree [3]. It's TMRCA estimates are based on SNP counting and calibrated by ancient and modern DNA samples [1].

As a more sophisticated example let us calculate the number of new lineages for the P312 haplogroup. P312 is widespread in Western Europe and most common in Spain, France and on the British Isles.

1. Go to the YFull R1b tree at https://yfull.com/tree/R1b/.
2. Copy the tree directly from the Web page and save it to a text file named *yfull-tree.txt*. Be careful that the indentations are preserved.
3. Go to a command line and type:
   `phylogrowth -treein=yfull-tree.txt -subclade=P312 -pngout=P312.png`
4. Done! Now watch *P312.png*. Can you spot the Bronze Age expansion, the Roman Empire and the Migration Period?

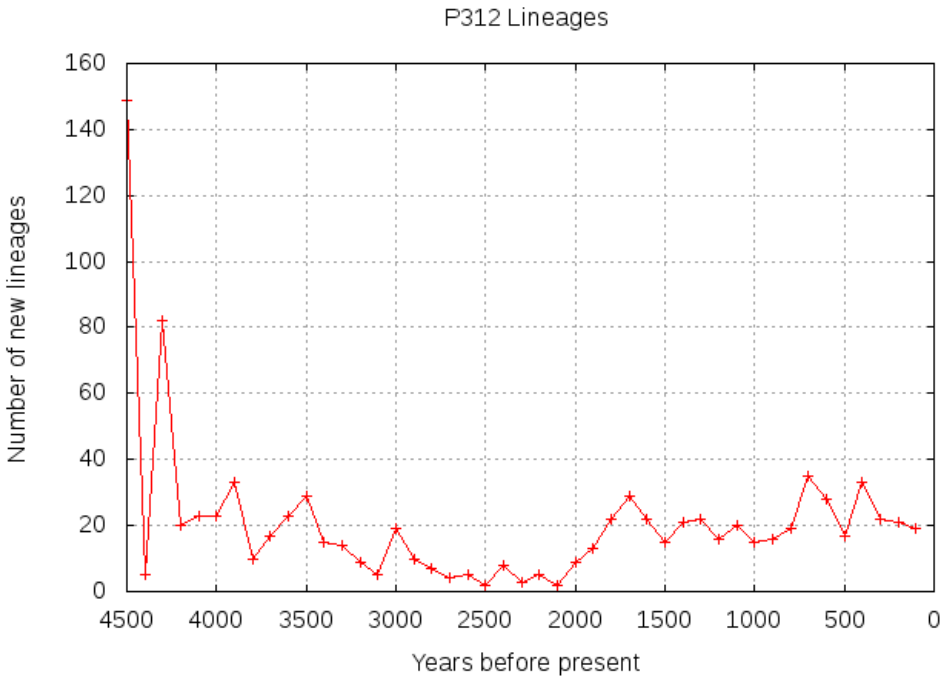

Figure 2: Number of new lineages of the P312 haplogroup.

5 Theory

Let us assume that we have a closed population and take a number of genetic samples. We analyse the samples and get different genetic lineages as a result.

The number of genetic lineages is proportional to the total population size.

$$L \sim P \tag{1}$$

L: number of detected lineages
P: population size

If we detect all male lineages, for example, the population size would be about twice as much.

We want to know how the number of lineages changes with time. For this we pinpoint the number of lineages at a specific time t_0 and look at infinitesimal changes.

$$L(t_0 + dt) = L(t_0) + dL \tag{2}$$

L: number of detected lineages
t_0: arbitrary point in time
dt: infinitesimal change in time
dL: infinitesimal change in the number of lineages

What can we say about the change in the number of lineages dL? We know for sure that at every given time some lineages disappear because they die out and others appear because of newborn babies. Thus we may write

$$dL = dL_{new} - dL_{dis} \tag{3}$$

dL: change in the number of lineages
dL_{new}: newly appearing lineages
dL_{dis}: disappearing lineages

This is exactly the same as

$$dL_{new} = dL + dL_{dis} \tag{4}$$

As an easy model let us assume that the number of disappearing lineages is proportional to the total number of lineages or in other words: The fraction of disappearing lineages is always the same. In this case the following equation holds:

$$dL_{dis} = c_{dis} L \, dt \tag{5}$$

dL_{dis}:	infinitesimal amount of disappearing lineages
L:	total number of lineages
dt:	infinitesimal amount of time
c_{dis}:	constant

With this we can rewrite 4 as

$$dL_{new} = dL + c_{dis} L \, dt \tag{6}$$

or

$$\frac{dL_{new}}{dt} = \frac{dL}{dt} + c_{dis} L \tag{7}$$

$\frac{dL_{new}}{dt}$:	change in newly appearing lineages
$\frac{dL}{dt}$:	change in in the total number of lineages
L:	total number of lineages
c_{dis}:	constant

We already know that the number of lineages is proportional to the population size 1. Thus we can interpret the previous formula in the following way:

The change in newly appearing lineages depends on two terms: the population growth rate and the population size.

Phylogrowth measures the change in newly appearing lineages (actually it measures a fraction of it, but I skip the details for now).

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] Thomas Williams, Colin Kelley, *Gnuplot*. Thomas Williams, Colin Kelley, 1986–1993, 1998, 2004, Date visited: 2016-06-23.

[3] YFull, *YFull Phylogenetic Tree*. Date visited: 2016-06-23.