# Phylogrowth User Guide

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

July 19, 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*.

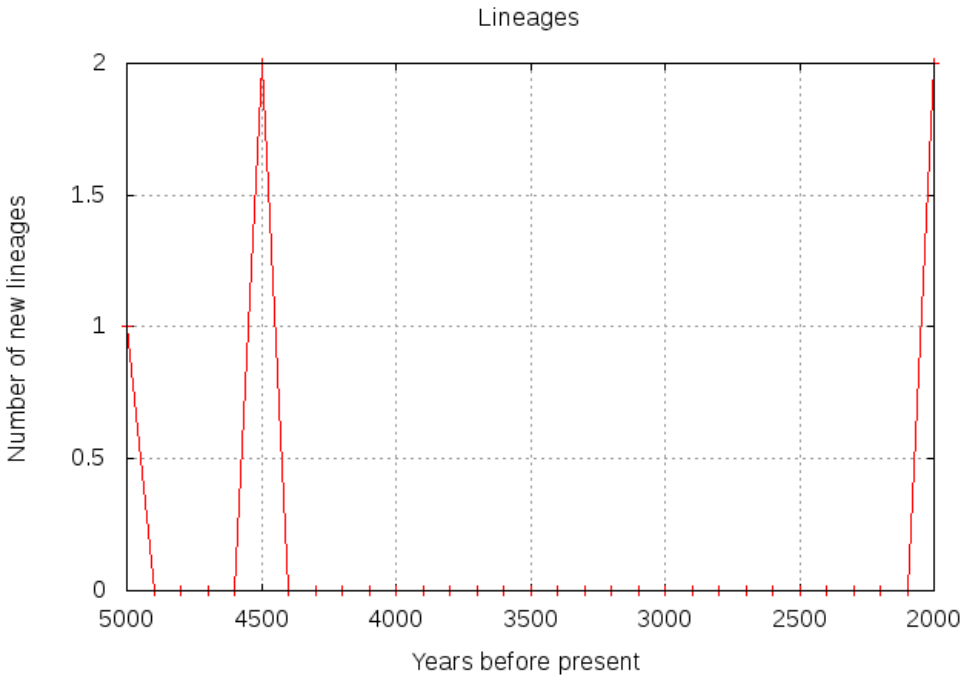


Figure 1: The first example showing time periods of haplogroup expansion.

## 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?

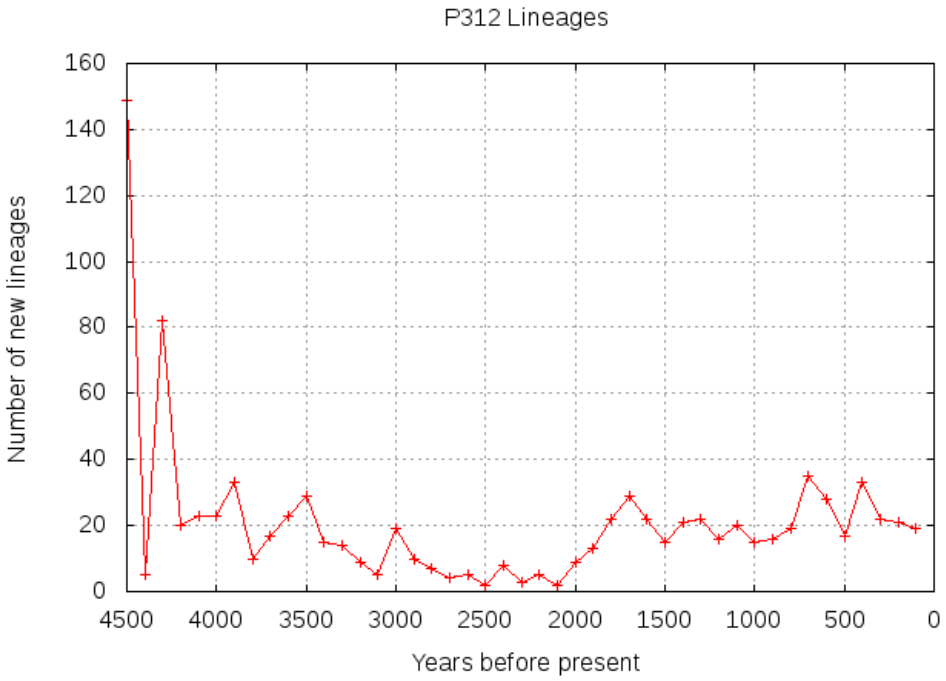


Figure 2: Number of new lineages of the P312 haplogroup.

# 5 Theory

## 5.1 Population Growth

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.

How does population size change with time? If we look after a short period of time we may take the old population size, add the number of births and subtract the number of deaths.

$$P(t + dt) = P(t) + b(t)dt - d(t)dt \tag{2}$$

$P$: population size
$b$: number of births per second
$d$: number of deaths per second
$t$: time

The probability of a mutation is the same for all newborn babies. This means the number of newly appearing mutations and lineages within a period of time scales with the number of births within that period. Thus we are interested in the number of births and rewrite the previous formula as

$$b(t)dt = P(t + dt) - P(t) + d(t)dt \tag{3}$$

or

$$b(t) = \frac{P(t + dt) - P(t)}{dt} + d(t) \tag{4}$$

$$= \frac{dP(t)}{dt} + d(t) \tag{5}$$

As an easy model let us assume that the number of deaths per second is proportional to the total population size. In this case the following equation holds:

$$d(t) = c_d P(t) \tag{6}$$

$d$: number of deaths per second
$P$: population size
$c_d$: constant

Under this assumption we can rewrite equation 5 as

$$b(t) = \frac{dP(t)}{dt} + c_d P(t) \tag{7}$$

This means that the number of births per second depends on two terms, the population growth and the population size.

## 5.2 Measurement of Phylogenetic Growth

We can not measure the number of births for every point in history directly. We can only look at a phylogenetic tree and examine how the number of lineages increases with time.

Let us assume that a fraction $c_l$ of all births develops into lineages that are still on the phylogenetic tree today. If we try to count these lineages we can count all new lineages on the phylogenetic tree between to points in time.

$$g(t) = c_l \int_{t-\frac{a}{2}}^{t+\frac{a}{2}} b(t)dt \tag{8}$$

If we substitute $b(t)$ with equation 7 we get

$$g(t) = c_l \int_{t-\frac{a}{2}}^{t+\frac{a}{2}} \frac{dP(t)}{dt} + c_d P(t)dt \tag{9}$$

$g$: growth of phylogenetic lineages between $t - \frac{a}{2}$ and $t + \frac{a}{2}$
$a$: time interval or step length used for counting
$c_l$: fraction of births that develops into long time lineages
$P$: population size
$c_d$: fraction of the population size that dies per second

For practical purposes we assume that the measurement time interval $a$ is small compared to the time it takes for any significant population changes. Then we may write the previous equation as

$$g(t) \approx c_l a \left( \frac{dP(t)}{dt} + c_d P(t) \right) \tag{10}$$

This means that the measured increase in phylogenetic lineages at any given time is roughly proportional to the number of births at that time and this depends on population size and growth.

Equation 10 is only valid for complete phylogenetic trees. In reality we have to take into account that only a fraction $c_t$ of the population has tested.

Furthermore the probability to detect a certain lineage depends on it's age. If we characterize a specific historic lineage by it's SNP mutation we will find this mutation whenever a person belonging to the SNP haplogroup has been tested. The probability to detect a certain SNP scales with the number of persons carrying it. For this reason it is much more likely to detect an older haplogroup than a younger subclade.

# 6 Appendix: Probability to Detect K Samples

Let us assume we are trying to detect a specific haplogroup within a population of N individuals. The haplogroup itself is represented by K individuals. How big is the probability to detect one of the K haplogroup members with S samples?

For a single sample the answer is easy:

$$p(1) = \frac{K}{N} \tag{11}$$

And the probability, not to detect one of the K individuals is

$$\tilde{p}(1) = \frac{N-K}{N} \tag{12}$$

| | |
|---|---|
| $p(S)$ | probability to detect K with S samples |
| $\tilde{p}(S)$ | probability not to detect K with S samples |
| $N$: | number of individuals |
| $K$: | number of persons to detect |
| $S$: | number of samples |

Suppose our single sample test did not yield a positive result and we try a second time. The previously tested individual does not count anymore. So we must subtract one from the number of individuals and one from the number of individuals who are not a member of our haplogroup. Thus the probability not to detect one of the K individuals the second time is

$$\tilde{p} = \frac{N-K-1}{N-1} \tag{13}$$

The maximum number of failures is $N-K$ because then only bearers of the haplogroup we are looking for are left and the next test will definitely a success.

To calculate the total probability of S failures we can multiply the probabilities of failure for each single try.

$$\tilde{p}(S) = \prod_{i=0}^{S-1} \tilde{p}_i \tag{14}$$

$$= \prod_{i=0}^{S-1} \frac{N-K-i}{N-i} \tag{15}$$

$$= \frac{(N-K)!/(N-K-S)!}{N!/(N-S)!} \tag{16}$$

$$= \frac{(N-K)!(N-S)!}{N!(N-K-S)!} \tag{17}$$

Because the total probability must always be one, the probability to detect K individuals with S samples is

$$p(S) = 1 - \tilde{p}(S) \tag{18}$$

$$= 1 - \frac{(N-K)!(N-S)!}{N!(N-K-S)!} \tag{19}$$

# 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.