

# AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications

## Version 2.0

**Authors:** Anneleen Van Geystelen and Maarten Larmuseau

E-mail: [anneleen.vangeystelen@bio.kuleuven.be](mailto:anneleen.vangeystelen@bio.kuleuven.be) and [maarten.larmuseau@bio.kuleuven.be](mailto:maarten.larmuseau@bio.kuleuven.be)

Centre for Forensic Genetics and Molecular Archaeology  
Department of Forensic Medicine  
University Hospitals Leuven  
Kapucijnenvoer 33  
3000 Leuven  
Belgium

Laboratory for Socioecology and Social Evolution  
Department of Biology  
KU Leuven  
Naamsestraat 59  
3000 Leuven  
Belgium

## Changes

### V2.0

- A graphical user interface in java was added. This makes running the script with all its arguments much easier.
- The script for extra quality assessment is now integrated in the AMY-tree\_v2.0.pl script.
- An algorithm for analysis custom capture Y-chromosome sequencing added as an option.
- Only one fasta file and status file of reference genome needs to be given: hg18 or hg19.
- The status reference file can now be generated via the interface (and the underlying getStatusReference.pl script).

### V1.1

- The phylogenetic tree was updated.

- Extra controls and actions are implemented for haplogroups belonging to R1b (R-M343) when determined in the 'sufficient' mode and with the MCC lower than 0.95.
- The extra check of Z381 & L2 and Z381 & L20 is added to the quality control file and is also implemented in AMY-tree\_v1.1.pl.
- Extra quality assessment: Matthews correlation coefficient (MCC), accuracy, sensitivity, specificity, precision, recall and F1-score. AMY-tree V1.1 uses the script SHELly\_v1.0.pl to calculate the values of these measures. Make sure that SHELly\_v1.0.pl is in the same folder as AMY-tree\_v1.1.pl. If not you can change the path of SHELly\_v1.0.pl in line 1761.

## Content

Introduction .....	4
Getting started.....	5
Windows .....	5
Linux .....	5
Running AMY-tree.....	6
Grafical user interface (Java) .....	6
Input files .....	6
Y-SNP (or SNP calling data) file .....	6
Output directory .....	7
Phylogenetic tree file .....	7
Mutation conversion file.....	7
Hg18/Hg19 Y-chromosome fasta file .....	8
Status file of Hg18/Hg19 Y-chromosome.....	8
Quality control file .....	8
Version (hg18 or hg19).....	8
Regions custom capture file (optional).....	8
Output files .....	9
Analysis file. ....	9
SNP status file .....	9
ConvNotTree file .....	9

New SNPs file .....	10
Quality file .....	10
StatusSNPs file .....	10
Command (when not using graphical user interface).....	10
List of files included .....	11
How to cite? .....	11
Reporting bugs and comments .....	11

## **Introduction**

An explosion of human whole genome data will become available in the coming years. These data can be used to optimize and to increase the resolution of the phylogenetic Y chromosomal tree. A first step toward this increased resolution is the requirement of an automatic determination of the phylogenetic position of an individual based on whole genome SNP calling data independently from the NGS platform and SNP calling program, whereby mistakes in the SNP calling or phylogenetic Y chromosomal tree are taken into account.

AMY-tree is a useful tool to determine the Y lineage of a sample based on SNP calling, to identify Y-SNPs with yet unknown phylogenetic position and to optimize the Y chromosomal phylogenetic tree in the future. AMY-tree will not add lineages to the existing phylogenetic tree of the Y-chromosome but it is the first step to analyse whole genome SNP profiles in a phylogenetic framework.

We strongly recommend consulting the reference papers in which the AMY-tree is described:

<http://www.biomedcentral.com/1471-2164/14/101/abstract>

<http://dx.doi.org/10.1016/j.fsigen.2013.03.010>

## **Getting started**

To run the AMY-tree script you need Java (for the interface) and a Perl distribution. AMY-tree needs also A Perl module from the CPAN website: Bio::SeqIO.

Java can be downloaded via this link:

[http://www.java.com/en/download/help/index\\_installing.xml](http://www.java.com/en/download/help/index_installing.xml)

## **Windows**

Both Strawberry (strawberryperl.com) as ActiveState (www.activestate.com) provide a Perl distribution for Windows. The Bio::SeqIO module can be installed via the cpan.pl program. Open the Command Prompt (cmd.exe) and run the next command

```
cpan -fi Bio::SeqIO
```

## **Linux**

Most Linux operation systems include a Perl distribution. ActiveState provides a Perl distribution for Linux (www.activestate.com).

To install the Bio::SeqIO module under Linux, the next steps need to be taken:

1. Download the module from the CPAN website (www.cpan.org)
2. Open new Shell and run the next commands:
3. Decompress the file

```
gzip -d /yourpath/yourmodule.tar.gz
```

4. Unpack the file

```
tar -xof /yourpath/yourmodule.tar
```

5. Build the module

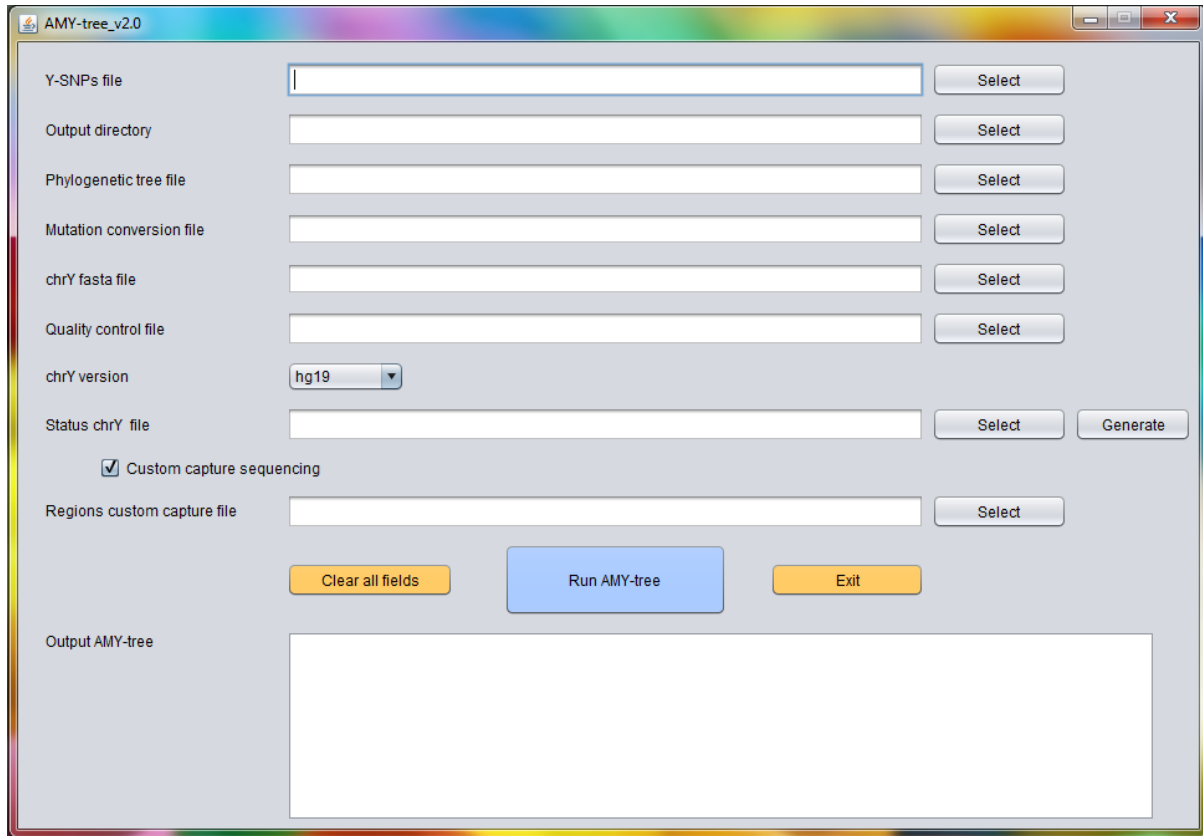
```
cd /yourpath/  
perl Makefile.PL  
make test
```

6. Install the module

```
make install
```

## Running AMY-tree

### Grafical user interface (Java)



### Input files

#### Y-SNP (or SNP calling data) file

This file contains all Y-SNPs of a single individual which are called by comparing the whole genome sequences with the reference genome. The format of the SNP calling file is a simple tab separated values format. We created software, the so-called 'WHY conversion tool', which can convert the formats of vcf, cga, tsv to our own defined tsv format. This format needs the chromosome in first field. The next field holds the position of the SNP on that chromosome (1-based). Fields #3 and #4 contain the reference base and the called base. A last field can be used to put the dbSNP name when available. An example of such a SNP calling data file can be found below.

chrY	10056145	G	A	x
chrY	10057061	T	G	x
chrY	10058354	A	G	rs35567891
chrY	10082860	T	G	rs35368665

### Output directory

This directory will contain the output files of AMY-tree.

### Phylogenetic tree file

This file contains the Y chromosomal phylogenetic data. The format of this file is also a simple tab separated values format that is easy to understand and created by the user. The first field contains the name of the node and the second field holds its alternative name. The third field contains its parent node and from field #4 on the Y-SNPs of that node are listed. An example of such phylogenetic tree file can be found below.

Root	Root	-	-				
A1b	A-V148	Root	V148	V149	V150	V151	V153
A1a-T	A1a-T-V168*	Root	V168	V171	V174	V203	V238
A1a	A-V4	A1a-T	V4	V14	V15	V25	V26
A2-T	A2-T-V221*	A1a-T	V221				
A2	A-V50	A2-T	V50	V61	V70	V72	V79
A3	A-V1	A2-T	V1	V10	V51	V56	V66

### Mutation conversion file

This file contains all the essential data of the Y-SNPs for which mutation conversion is scientifically reported. The format of this file is again a simple tab separated value format. The first field contains the name of the Y-SNP. Fields #2 and #3 contain the position on the Y-chromosome of Hg18 (NCBI36) and Hg19 (GRCh37). The fourth field contains the mutation conversion in the form of “ancestral base”->”mutant base”. The fifth field holds the type of mutation (SNP, indel or unknown) and this is linked to the next field which keeps the information if the mutation needs to be ignored or not. Only Y-SNPs are not ignored at the moment. The seventh field contains the possible phylogenetic position of the mutation according to ISOGG (International Society of Genetic Genealogy). Fields #8 and #9 contain the synonymous mutations and the dbSNP reference. An example of such a mutation conversion file can be found below.

L162	14528466	16019072	G->C	SNP	no	T1a	Page21	rs34179999
L164	14211253	15701859	G->A	SNP	no	R1b1a2a1a1a5c1a1a		rs13305517
L165	20389703	21930315	C->T	SNP	no	R1b1a2a1a1b2b2	S68	
L166	22399272	23989884	C->A	SNP	no	G (Investigation)		
L175	20355469	21896081	CTGT->del	indel	yes	R1a1a1g3a1a		
L176.1	20238645	21779257	AAAAC->del	Unknown	yes	R1a1a1g3a1	S179.1	

### Hg18/Hg19 Y-chromosome fasta file

This file just contains the sequence of the Hg18 (NCBI36) or Hg19 (GRCh37) Y-chromosome in fasta format.

### Status file of Hg18/Hg19 Y-chromosome

This file just contains the status for each SNP in the reference genome Hg18 of Hg19. The format of this file is again a simple tab separated value format. The first field contains the name of the Y-SNP and the second field contains the status of the Y-SNP: 0 means ancestral, 1 means mutant and -1 another base. An example of such a status file can be found below.

L16	1
L160	0
L166	0
L167	-1
L168	0

### Quality control file

This file contains all Y-SNPs which will be used to determine the Call Quality test score. Again, the format of this file is a tab separated value format. The first field contains the name of the haplogroup and all other fields contain the name of the used SNPs.

### Version (hg18 or hg19)

### Regions custom capture file (optional)

This files contains the regions that are sequenced via custom capture sequencing. Again, the format of this file is a tab separated value format. This format needs the chromosome in first field. The next fields hold the start position of the region and the end position of the region. The last field can hold a comment.

chrY	2650042	2650882	x
chrY	2651354	2652881	Region 1
chrY	2653469	2662342	x
chrY	2663962	2665162	x



## Output files

### Analysis file.

This file reports the 'Call quality test' score, the results of the vertical, horizontal, combinatorial and specific methods and the determined sub-haplogroup of each sample. An example of such an analysis file can be found below.

```
QUALITY
-----
F - R
0.588235294117647 (insufficient)

VERTICAL
-----
# P* [P-92R7*]          0 0
# R1* [R-M173*]         0 0
# R1b1b2a1a1* [R-U106*] 0 0

HORIZONTAL
-----
! Root

COMBI
-----
% P* [P-92R7*]
% R1* [R-M173*]
% R1b1b2a1a1* [R-U106*]

SPECIFIC
-----
@ R1b1b2a1a1* [R-U106*]      2
@ R1* [R-M173*]              1

RESULTS
-----
> R1b1b2a1a1* [R-U106*]
```

### SNP status file

This file shows the state of all Y-SNPs which are given in the Y-SNP mutation conversion file. The state can be ancestral (0), mutant (1) or other (-1). It also reports if the mentioned state is obtained by calling or by derivation from the reference. An example of such a SNP status file can be found below.

S186	0	reference
P227	1	reference
M405	1	called
S144	0	called
M83	0	reference

### ConvNotTree file

This file contains all Y-SNPs which have the mutant state and which are mentioned in the Y-SNP conversion file but are not present in the phylogenetic tree that was used. All synonyms of these SNPs are as well mentioned in this file.

### New SNPs file

This file contains all the Y-SNPs are listed for which the phylogenetic position is not yet determined and which are not yet reported in the Y-SNP conversion file.

### Quality file

This file contains all the values for the measures: Matthews correlation coefficient (MCC), accuracy, sensitivity, specificity, precision, recall and F1-score.

### StatusSNPs file

This file shows the status of all Y-SNPs which are given in the Y-SNP mutation conversion file. The state can be true negative (TN), true positive (TP), false negative (FN) or false positive (FP) An example of such a SNP status file can be found below.

S186	TN
P227	TP
M405	FN
S144	FP

### **Command (when not using graphical user interface)**

Type in the Command Prompt (Windows) or in the Shell (Linux)

- Whole genome sequencing:

```
perl /yourpath/AMY-tree_v2.0.pl /yourpath/SNPcallDataFile.txt  
/yourpath/OutputDirectory/ /yourpath/PhylogeneticTreeFile.txt  
/yourpath/MutationConversationFile.txt /yourpath/hg19.fa  
/yourpath/hg19_status.txt /yourpath/QualityControl.txt hg19
```

- Custom capture sequencing:

```
perl /yourpath/AMY-tree_v2.0.pl /yourpath/SNPcallDataFile.txt  
/yourpath/OutputDirectory/ /yourpath/PhylogeneticTreeFile.txt  
/yourpath/MutationConversationFile.txt /yourpath/hg19.fa  
/yourpath/hg19_status.txt /yourpath/QualityControl.txt hg19  
/yourpath/regionsCustomCapture.txt
```

### **List of files included**

The ZIP file AMY-tree.zip includes besides the Java executable AMY-tree\_v2.0.jar also the Perl scripts AMY-tree\_v2.0.pl and getStatusReference.pl for those who want to run AMY-tree without the graphical user interface. Also the latest phylogenetic Y chromosomal tree (UpdatedTree\_v2.1.txt), its corresponding mutation conversion file (MutationConversion\_v2.1.txt), the file used for the quality control (qualityControl\_v2.0.txt) are included.

### **How to cite?**

Please cite this script as:

Van Geystelen, A., R. Decorte, and Larmuseau, M.H.D. (2013). "Updating the Y-chromosomal phylogenetic tree for forensic applications based on whole genome SNPs." FSI: Genetics 101.

<http://dx.doi.org/10.1016/j.fsigen.2013.03.010>

### **Reporting bugs and comments**

Bugs and comments on the scripts are welcome via

[anneleen.vangeystelen@bio.kuleuven.be](mailto:anneleen.vangeystelen@bio.kuleuven.be) or [maarten.larmuseau@bio.kuleuven.be](mailto:maarten.larmuseau@bio.kuleuven.be).