Dating HIV-1 mutation dispersion using Bayesian Inference

National and Kapodistrian UNIVERSITY OF ATHENS

Author: Belias Michael

Student in MSc in Biostatistics

Athens, 20 April 2016

Contents

1 Introduction

1.1 Epidemiology

Epidemiology is the study of the distribution and determinants of health-related states or events (including disease), and the application of this study to the control of diseases and other health problems. Various methods can be used to carry out epidemiological investigations: surveillance and descriptive studies can be used to study distribution; analytical studies are used to study determinants(World Health Organization, 2016).

1.2 Molecular epidemiology

A definition of molecular epidemiology (Schulte, 1993) defines it as "the use of biological markers or biological measurements in epidemiological research." This definition limits molecular epidemiology only to biomarker study. A straightforward classification of a biomarker classifies its use into four types, those that: (1) improve assessment of exposure, (2) identify the underlying mechanisms of disease and disease transmission, (3) identify subgroups of the population that are more susceptible to the effects of pathogens or pathogenic substances, and (4) identify subgroups of cases with more homogenous disease to better clarify the role of various aetiologic agents (Rabkin & Rothman, 2001). Overall, molecular epidemiology studies measure biologic response (such as mutations) to specific exposures (mutagens) and assess the interplay of host characteristics such as genotype and phenotype in gene expression, and development of disease and response to therapy. Molecular epidemiology is also useful in diagnosis, prognosis, and follow-up of therapeutic results. Molecular epidemiology is a technique-based discipline. (Chattopadhyay, 2011)

1.3 HIV

The human immunodeficiency virus (HIV) is a lentivirus (a subgroup of retroviruses) that causes HIV infection and acquired immunodeficiency syndrome (AIDS) (Weiss, 1993). Human immunodeficiency viruses are separated in two major groups (HIV-1 and HIV-2) and are the result of multiple cross-species transmissions of simian immunodeficiency viruses (SIVs) naturally infecting African primates. Most of these transfers resulted in viruses that spread in humans to only a limited extent. However, one transmission event, involving SIV-cpz from chimpanzees in south-eastern Cameroon, gave rise to HIV-1 group M-the principal cause of the AIDS pandemic (Greene, 2007).

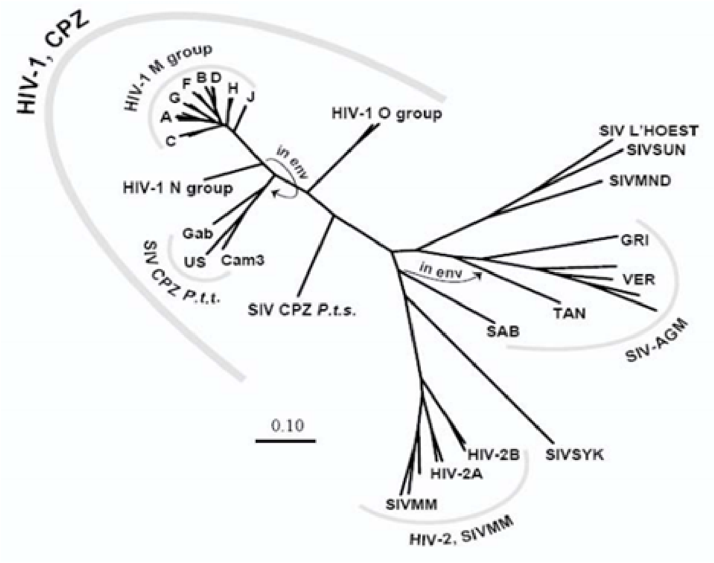
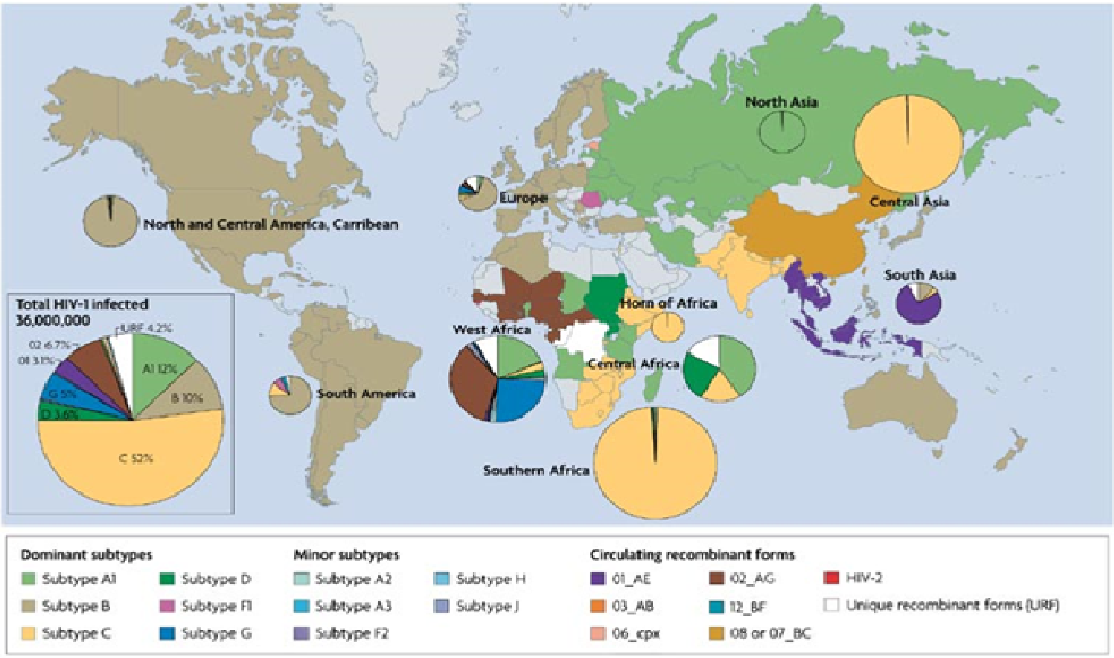


Fig 0.1 Phylogenetic relationship of primate lentivirus. The pol gene of primate lentivirus was used to generate the unrooted tree. Two of the viruses presented in the tree (HIV-1 group N and SIVagm SAB have mosaic genomes. The small arrows in the tree indicate where the sequences would branch in an env gene tree ((Bette Korber and Watkins, n.d.) (accessed on November 2004).

1.4 Group M

Group M, which is responsible for the majority of infections in the worldwide HIV-1 epidemic, can be further subdivided into 10 recognized phylogenetic subtypes, or clades (A to K), which are approximately equidistant from one another. Within group M, the average inter-subtype genetic variability is 15% for the gag gene and 20-30% for the env gene while the pol gene shows only a 10% genetic divergence (Camacho R., 2006). Clades B and D are more closely related to each other than to other subtypes, and clade D is considered the early clade B African variant, but their original designation as subtypes is retained by authors for consistency with earlier published works ((Gao, 1996), (Louwagie, 1993)). Classification of HIV-1 subtypes was originally based on the sub genomic regions of individual genes. However, with an increasing number of viral isolates available worldwide and improvements in sequencing methods, HIV-1 phylogenetic classifications are currently based either on nucleotide sequences derived from multiple sub genomic regions (gag, pol, and env) of the same isolates or on full-length genome sequence analysis. This approach has revealed virus isolates in which phylogenetic relations with different subtypes switch along their genomes(Peeters, 2001). According to recent studies, on a global scale the most prevalent HIV-1 genetic forms are subtypes A, B, and C, with subtype C accounting for more than 50% of all HIV-1 infections worldwide.



Nature Reviews | Microbiology

Fig 0.2 The frequency of each HIV-1 subtype and recombinant form was estimated in each country based on published findings. A complete breakdown of subtype prevalence per country and the countries present in each region are listed in the Supplementary information S1 (table). The countries are colour-coded based on the dominant HIV-1 group main (M) subtype. The countries coloured grey have a low level of HIV-1 prevalence or were not represented in the scientific literature related to HIV-1 subtype prevalence. The pie charts depict the proportion of each subtype or recombinant form in each geographical region. The size of the pies is proportional to the number of HIV-1 infected individuals in that particular region. (Ariën et al., 2007).

1.5 Subtype B

Subtype B is the main genetic form in western and central Europe, the Americas, and Australia and is also common in several countries of South-east Asia, northern Africa, and the Middle East and among South African and Russian homosexual men(Buonaguro et al., 2007).

1.6 AIDS

Acquired Immune Deficiency Syndrome (AIDS) was first recognized as a new disease in 1981 when increasing numbers of young homosexual men succumbed to unusual opportunistic infections and rare malignancies (Greene, 2007). A retrovirus, now termed human immunodeficiency virus type 1 (HIV-1), was subsequently identified as the causative agent of what has since become one of the most devastating infectious diseases to have emerged in recent history ((Barre-Sinoussi et al., 1983); (Gallo et al., 1984); (Popovic et al., 1984)). HIV-1 spreads by sexual, percutaneous, and perinatal routes; however, 80% of adults acquire HIV-1 following exposure at mucosal surfaces, and AIDS is thus primarily a sexually transmitted disease ((Hladik and McElrath, 2008); (Cohen et al., 2011)). Since its first identification almost three decades ago, the pandemic form of HIV-1, also called the main (M) group, has infected at least 60 million people and caused more than 25 million deaths(Merson et al., 2008). Developing countries have experienced the greatest HIV/AIDS morbidity and mortality, with the highest prevalence rates recorded in young adults in sub-Saharan Africa (http://www.unaids.org/). Although antiretroviral treatment has reduced the toll of AIDS-related deaths, access to therapy is not universal, and the prospects of curative treatments and an effective vaccine are uncertain ((Barouch, 2008); (Richman et al., 2009)). Thus, AIDS will continue to pose a significant public health threat for decades to come. One of the major characteristics of lentiviruses is their extensive genetic variability, which is the result of the high error rate, the recombinogenic properties of the reverse transcriptase enzyme and the fast turnover of virions in HIV infected individuals. (Drosopoulos et al., 1998)

1.7 Phylogenetic Analysis

Phylogenetics is the study of the evolutionary history and relationships among individuals or groups of organisms (e.g. species, or populations).These relationships are discovered through phylogenetic inference methods that evaluate observed heritable traits, such as DNA sequences or morphology under a model of evolution of these traits. The result of these analyses is a phylogeny (also known as a phylogenetic tree) - a hypothesis about the history of evolutionary relationships. The tips of a phylogenetic tree can be living organisms or fossils. Phylogenetic analyses have become central to understanding biodiversity, evolution, ecology, and genomes. (Wikipedia, 2016)

1.8 Viral Phylodynamics

Viral phylodynamics is defined as the study of how epidemiological, immunological, and evolutionary processes act and potentially interact to shape viral phylogenies. A great research on viral phylodynamics has focused on transmission dynamics in an effort to shed light on how these dynamics impact viral genetic variation (Grenfell, 2004). Viral Phylodynamics can help us in three problems :

- Viral origins: The rapid rate of evolution in viruses allows molecular clock models to be estimated from genetic sequences, thus providing a per-year rate of evolution of the virus. With the rate of evolution measured in real units of time, it is possible to infer the date of the most recent common ancestor (MRCA) for a set of viral sequences.

- Viral Spread : Phylodynamic models may provide insight into epidemiological parameters that are difficult to assess through traditional surveillance means.

- Viral control efforts :

Phylodynamic approaches can also be useful in ascertaining the effectiveness of viral control efforts, particularly for diseases with low reporting rates. For example, the genetic diversity of the DNA-based hepatitis B virus declined in the Netherlands in the late 1990s, following the initiation of a vaccination program (Ballegooijen et al., 2009).

In our case we will mostly study the first issue.

1.9 Molecular Clock Assumption

A hypothesis that predicts a constant rate of molecular evolution among species. It is also a method of genetic analysis that can be used to estimate evolutionary rates and timescales using data from DNA or proteins. This is a very strict assumption not applicable to viruses like HIV, so the use of ** Relaxed Molecular Clocks** is necessary. The relaxed molecular clock is a statistical model of molecular evolution that allows the evolutionary rate to vary amongst organisms (in our case HIV)(Ho, 2013).

1.9.1 Relaxed Molecular Clocks

Relaxed molecular clock models take into account rate variation across lineages and have been proposed in order to obtain better estimates of divergence times (Drummond et al., 2006). They represent an intermediate position between the ‘strict’ molecular clock hypothesis and Joseph Felsenstein’s many-rates model(Felsenstein, 2001) and are made possible through MCMC techniques that explore a weighted range of tree topologies and simultaneously estimate parameters of the chosen substitution model. It must be remembered that divergence dates inferred using a molecular clock are based on statistical inference and not on direct evidence.

1.10 Molecular Epidemiology in HIV

Molecular epidemiology of (HIV-1) subtype B has been studied extensively since it’s discovery, as it quickly became the dominant subtype of the virus in the economically developed world, especially among men who have sex with other men (Felsenstein, 2001). With international transportation becoming more affordable than in the past, interaction between persons who live with the virus has increased, and that appears to have been a reason for the spread of the epidemic within the European region, as migration pathways of the virus have been observed between the countries of Europe (Paraskevis et al., 2009).

2 Data Collection

2.1 Initial Dataset

This study is based on the 2009 study of Magiorkinis et al (Magiorkinis, 2009). The original Dataset had 8,370 sequences of HIV-1 Subtype - B which were collected from 79 countries across the world. The collection of the sequences was divided into two strategies : * Systematic research through the PubMed Database * The addition of two highly acknowledged and sophisticated European researches CATCH and SPREAD.

2.1.1 The European Data-Set

The European part of the Data-Set included sequences from two European studies : the Combined Analysis of Resistance Transmission over Time of Chronically and Acute Infected HIV Patients (CATCH) and the SPREAD (Strategy to Control SPREAD of HIV Drug Resistance) collaboration. The SPREAD study included 4480 newly diagnosed patients sampled during 9/2002-12/2007 from 20 European countries and Israel. In the prospective setting a standardised sampling The subtype B global dispersal strategy was designed to include representative sampling from all countries ((Vercauteren et al., 2009); (Wensing, 2008)). For the purpose of this study we included only those classified as subtype B from both the CATCH and the SPREAD studies. The CATCH study 2208 antiretroviral naïve individuals from 18 European countries and Israel during 1996-2002 (Wensing et al., 2005).

2.1.2 The Non-European Data-Set

The Non-European Data-Set was collected through systematic bibliographic search in PubMed using the following keywords “HIV-1”, “molecular epidemiology”, “resistance”, “subtype B” and “pol” in various combinations, cleaned of different Subtypes only keeping the subtype B . The later sampling criteria were: i) in cases where more than one study was available for one country, we included only those sequences isolated from different areas of that country, ii) in cases where the sampling areas were not described in the studies from the same country, to avoid redundant sequences, we only included sequences from the largest study, iii) similarly for studies performed at the same centres or cities, iv) from longitudinal studies concerning mainly resistance to antiretroviral therapy, we included only the oldest available sequence per patient, and v) we excluded studies concerning mother to child transmission.(Magiorkinis, 2009)

2.2 Monophyletic clusters extraction

In subsequent projects, several monophyletic clusters were extracted from Maximum Likelihood Phylogenetic trees. Those were identified as subtrees with a common ancestral node, whose strains had been submitted in a specific area, accounting for more than 75% of the total strains within the subtree. Each monophyleticncluster had to consist of more than 10 sequences, in order to be eligible for inclusio n.This process was carried out manually, by visual inspection using Dendroscope (Huson and Scornavacca, 2012). This concluded into 25 clusters with 3511 total sequences allocated as shown in Figure

Table 1: 24 Monophyletic Clustered Data-Set sizes as extracted in previews studies

	Cluster ID	Total
1	Cluster 0	4
2	Cluster 1	207
3	Cluster 2	118
4	Cluster 3	185
5	Cluster 4	144
6	Cluster 5	61
7	Cluster 6	53
8	Cluster 7	51
9	Cluster 8	58
10	Cluster 9	37
11	Cluster 10	46
12	Cluster 11	32
13	Cluster 12	30
14	Cluster 13	33
15	Cluster 14	33
16	Cluster 15	33
17	Cluster 16	36
18	Cluster 17	37
19	Cluster 18	39
20	Cluster 19	44
21	Cluster 20	46
22	Cluster 21	143
23	Cluster 22	241
24	Cluster 23	1790
25	Cluster 24	5
26	Cluster 25	5
	Overall	3511

3 Methods

3.1 Operating Systems - Statistical Packages and R-Packages Used.

For the completion of my thesis 2 main Operating systems were used:

- Windows 8.1 - Windows 10
- Ubuntu Linux 15.04

The Statistical Packages were :

1.R (R Core Team, 2015) R version 3.2.3 (2015-12-10)

2.R-Studio(RStudio Team, 2015) `RStudio Desktop 0.99.892`

3.Python 2.7.11

The packages used :

- for phylogenetic analysis and Beast Input Manipulation

a. phangorn (Schliep, 2011)

b. ape (Paradis et al., 2004)

c. Rphylip (Scott A. Chamberlain, 2013)

d. phytools (Revell, 2012)

e. ips (Heibl, 2008 onwards)

f. XML (Lang and CRAN Team, 2015)

g. ggtree (Yu et al., submitted)

- Other Packages:

a. ggplot2 (Wickham, 2009)

b. knitr (Leisch and Peng, submitted)

c. rworldmap (South, 2011)

d. png (Urbanek, 2013)

e. RMySQL (Ooms et al., 2015)

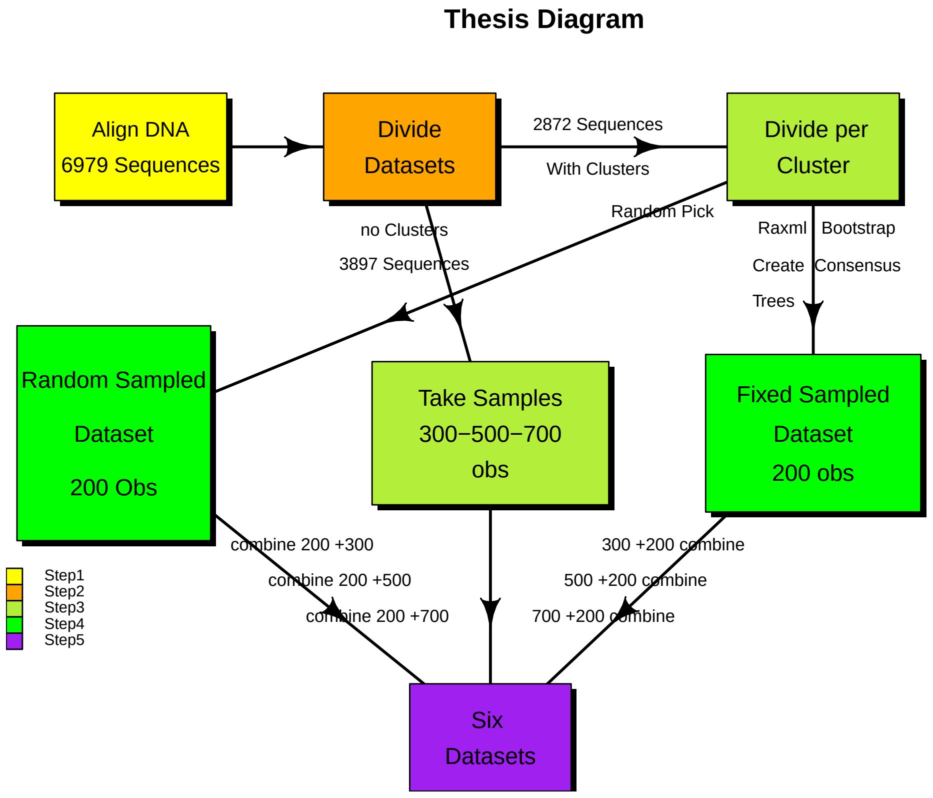
f.

For the writing I used the `Rmarkdown` (Allaire et al., 2016) package with **LaTeX**.

3.2 Analysis

3.2.1 The Study outline

The diagram below shows the work that was done.



The original Data-Set had 8348 but 6979 sequences left after we deleted all sequences without sampling years. The most vital information for my study part was kept:

- The ID
- The Submission Country
- The region
- The monophyletic Cluster it belongs
- and the sequence.

The Monophyletic Clusters as presented in the 2.2 chapter (Monophyletic clusters extraction) had many observations of NULL sampling year, so they were deleted. As shown in the following table cluster 24 was erased completely.

Table 2: 24 Monophyletic Clustered Data-Sets (with and wihout Year Information)

	Cluster ID	With Sampling Years Frequency	Total	Percentage
1	Cluster 0	4	4	100
2	Cluster 1	100	207	48.31
3	Cluster 2	16	118	13.56
4	Cluster 3	132	185	71.35
5	Cluster 4	92	144	63.89
6	Cluster 5	40	61	65.57
7	Cluster 6	15	53	28.3
8	Cluster 7	25	51	49.02
9	Cluster 8	27	58	46.55
10	Cluster 9	17	37	45.95
11	Cluster 10	39	46	84.78
12	Cluster 11	20	32	62.5
13	Cluster 12	19	30	63.33
14	Cluster 13	22	33	66.67
15	Cluster 14	26	33	78.79
16	Cluster 15	17	33	51.52
17	Cluster 16	27	36	75
18	Cluster 17	28	37	75.68
19	Cluster 18	34	39	87.18
20	Cluster 19	33	44	75
21	Cluster 20	40	46	86.96
22	Cluster 21	136	143	95.1
23	Cluster 22	230	241	95.44
24	Cluster 23	1728	1790	96.54
25	Cluster 24	0	5	0
26	Cluster 25	4	5	80
	Overall	2871	3511	81.77

as we may observe there is a vast difference in the percentage of the sequences with the Year information ranging from 0 % to 100%.

3.2.2 Alignment of the Sequences.

The sequences were then aligned using mafft version 7 a multiple sequence alignment program for Unix-like operating systems(Katoh, 2002). I preferred to keep the Speed-oriented method because accuracy-oriented methods were time consuming. A Maximum Likelihood tree for the verification of the Clusters above was extracted with RaXmL (Stamatakis, 2014) version 8.2.4 in www.phylo.org (Miller et al., 2010) server.

3.2.3 Division of the Dataset (Pseudo-Pgylogenetic Analysis)

The 6979 Sequences were divided into two Data-Sets, regarding the Cluster information.

- 2871 had be classified into Clusters, while
- 3879 had no Cluster information.

The Clustered Data-Set was then divided into the 24 monophyletic Data-Sets, the size of each is shown below in Table 3 .

Table 3: 24 Monophyletic Clustered Data-Set Size

	Cluster ID	With Sampling Years Frequency
1	Cluster 0	4
2	Cluster 1	100
3	Cluster 2	16
4	Cluster 3	132
5	Cluster 4	92
6	Cluster 5	40
7	Cluster 6	15
8	Cluster 7	25
9	Cluster 8	27
10	Cluster 9	17
11	Cluster 10	39
12	Cluster 11	20
13	Cluster 12	19
14	Cluster 13	22
15	Cluster 14	26
16	Cluster 15	17
17	Cluster 16	27
18	Cluster 17	28
19	Cluster 18	34
20	Cluster 19	33
21	Cluster 20	40
22	Cluster 21	136
23	Cluster 22	230
24	Cluster 23	1728
25	Cluster 24	0
26	Cluster 25	4
	Overall	2871

Two different approaches for sub-sampling were followed. The first was to sub-sample at random the Clusters making sure that at least were represented from each, and the second was to pick the most genetically diverse of each cluster.

For the first part for very small Clusters (below 10) such as Cluster 0 and 25 were not Sub-sampled but used as they were and all clusters were randomly sub-sampled to the size of 10.

For the second two parameters were taken care of:

- The size of the monophyletic and
- The Genetic Diversity the Cluster had.

The small Clusters (below 10 size) were represented as whole, medium clusters were sampled depending the genetic diversity they had, while the Cluster 23 was randomly sub-sampled with 30 sequences. The reason was that many Monophyletics were severely decreased and lost their homogeneity while Cluster 23 had lost only 3.5% of it's sequences.

In order to find the Genetic Diversity we run a Phylogenetic Analysis in each one and picked the Best Tree exported by RaXmL v8.2.4 (Stamatakis, 2014) with the following options:

a. raxmlHPC-PTHREADS-SSE3 (multicore)
b. -T 4 (threads)
c. -f a (+ ML search)
d. -N autoMRE - (a stopping Criterion) (Pattengale et al., 2010)
e. -m GTRGAMMA (GTRGAMMA model GTR with Gamma Distribution of rates across sites)
f. -p 123 -x 123 (Bootstrap seeds)

In the (Stamatakis, 2014) and (Pattengale et al., 2010) it is suggested to choose the autoMRE option to define the end of the Phylogenetic bootstrapping and the auto_MRE is the appropriate for small sequences (< 200). In our case we analysed all of them using the autoMRE bootstop method (even Cluster 23, which had 1700 sequences).

The appropriate model was GTR with Gamma heterogeneity model of rate.

The output of the RaXmL runs with these options includes , a bootstrap file , two bipartition files with and without the branchlengths and a Maximum Likelihood Best tree. So our next step will be to gather the inofrmation from the Bootstrap file and create a consensus Tree.

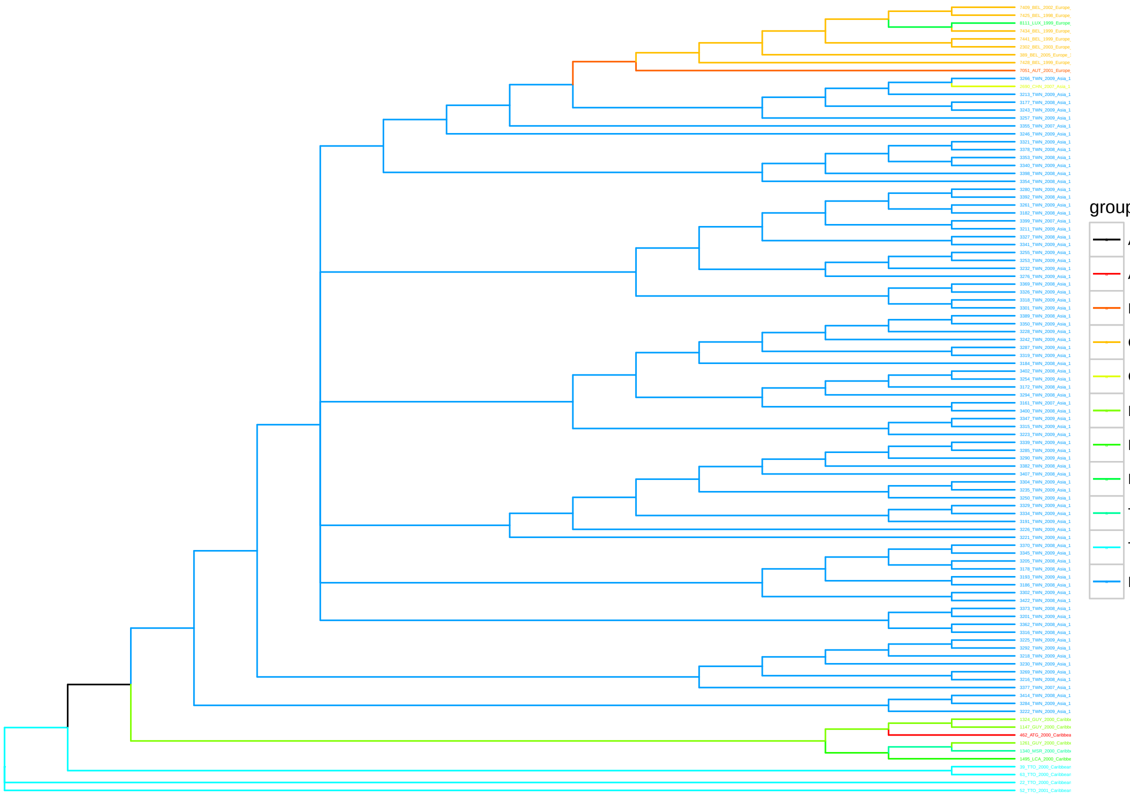
This procedure is appropriate in order to get a phylogenetic signal of the viruses.

The results and the subsampling strategy are shown in Chapter 3.3.

3.3 SubSampling the Monophyletics

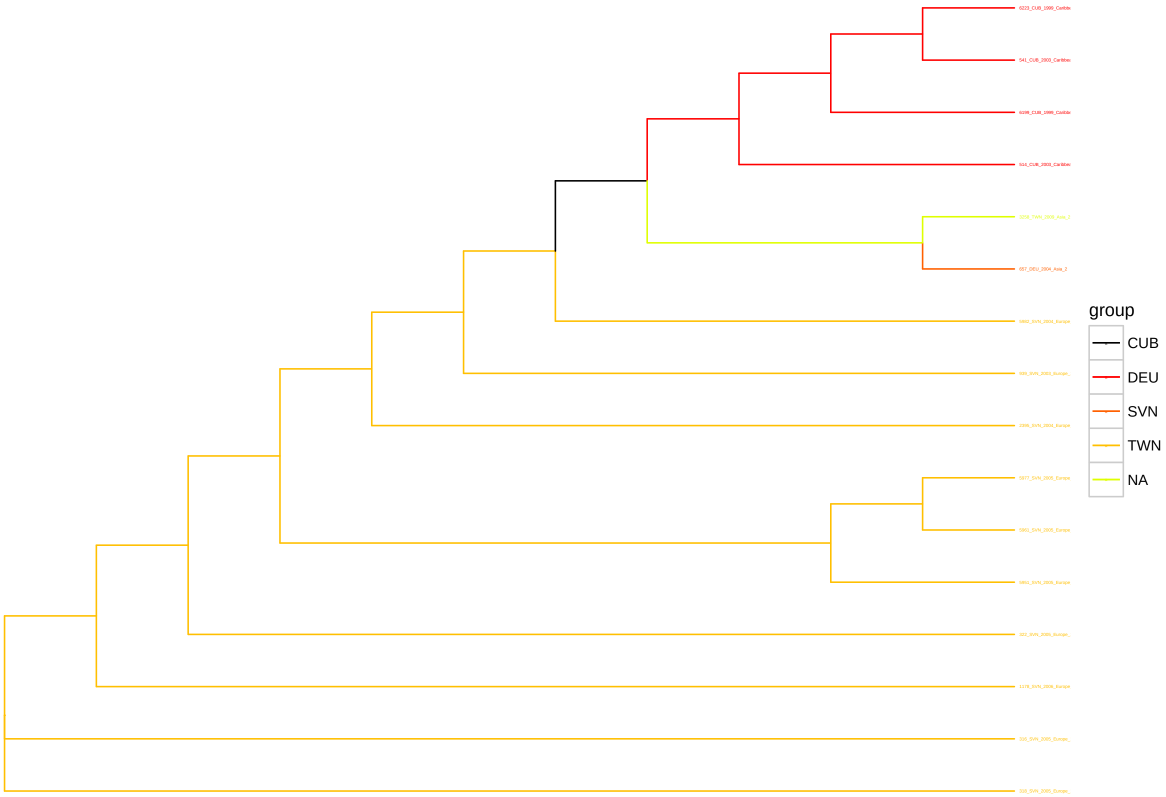
3.3.1 Cluster 1

The 1st Monophyletic Cluster is Cluster 1, it has 100 sequences 48.31% of the original size 207 , as we may observe the the tree is divided into 3 subclusters the Taiwan - Belgium (purple - yellow), the blue (Trinidad and Tobacco) and the green so we pick 2 randomly out of each sub-cluster.



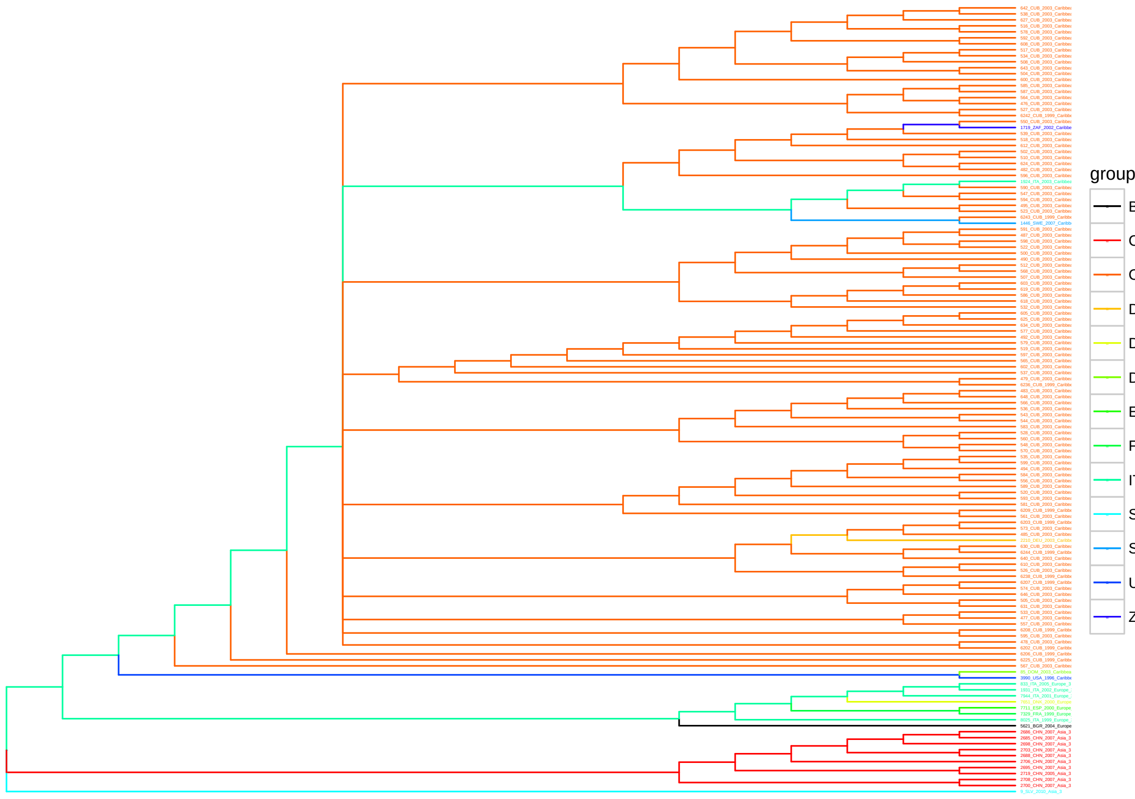
3.3.2 Cluster 2

The 2nd Monophyletic Cluster is Cluster 2, it has 16 sequences 13.56% of the original size which was 118, as we may observe the tree is not divided in any subcluster. This is a subcluster of the original Cluster 2 , which was mostly populated by Asian DNA samples, so we Randomly pick 4.



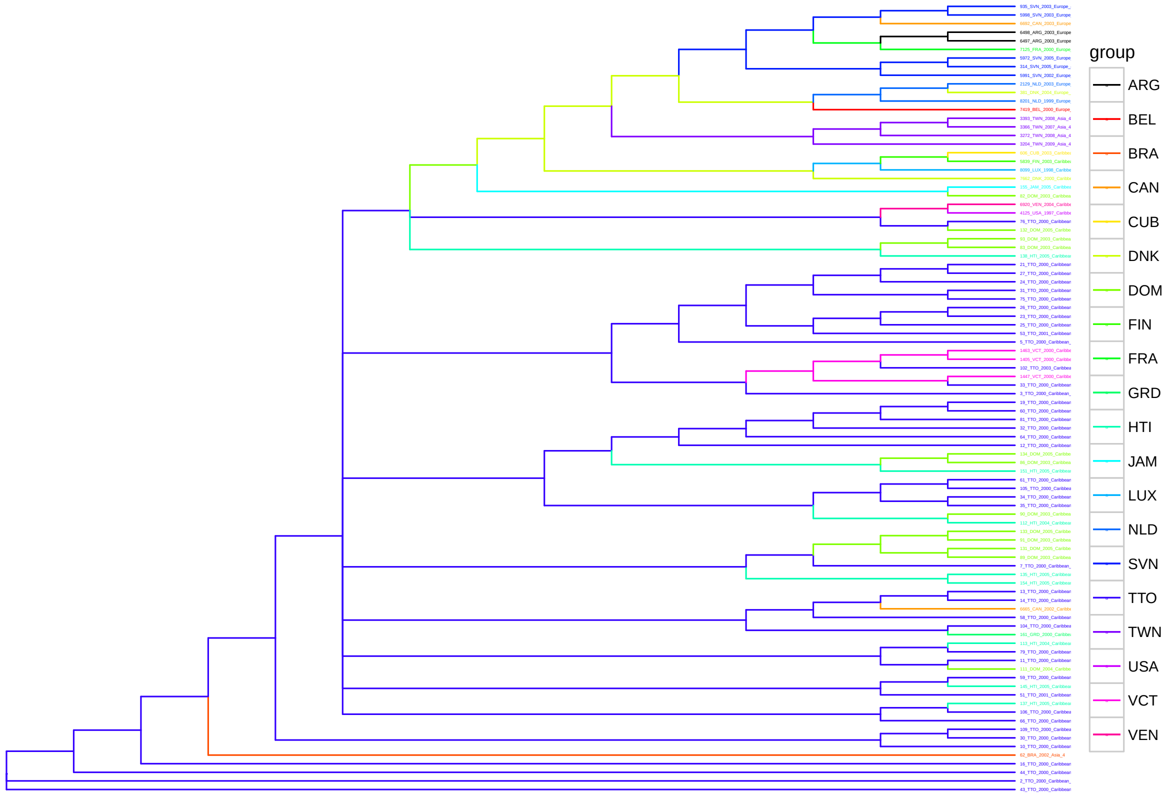
3.3.3 Cluster 3

The 3rd Monophyletic Cluster is Cluster 3, it has 132 sequences 71.35% of the original size which was 185, as we may observe the tree is divided into 2 subclusters the Cuban and a Chinese - European we pick 3 randomly out of Cuban and 4 of the others.



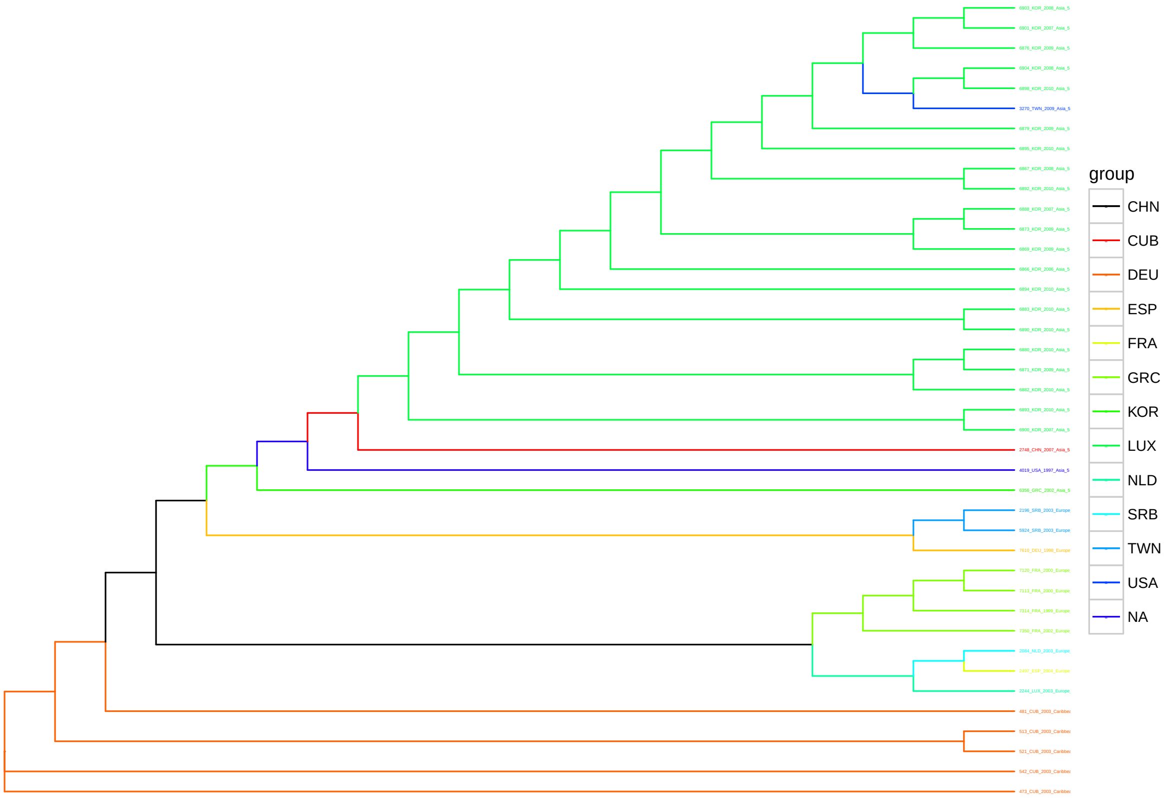
3.3.4 Cluster 4

The 4th Monophyletic Cluster is Cluster 4, it has 93 sequences 64.14% of the original size which was 144, as we may observe the tree is not divided into any particular subset, so we pick randomly 8 sequences, because of close phylogenetic distance.



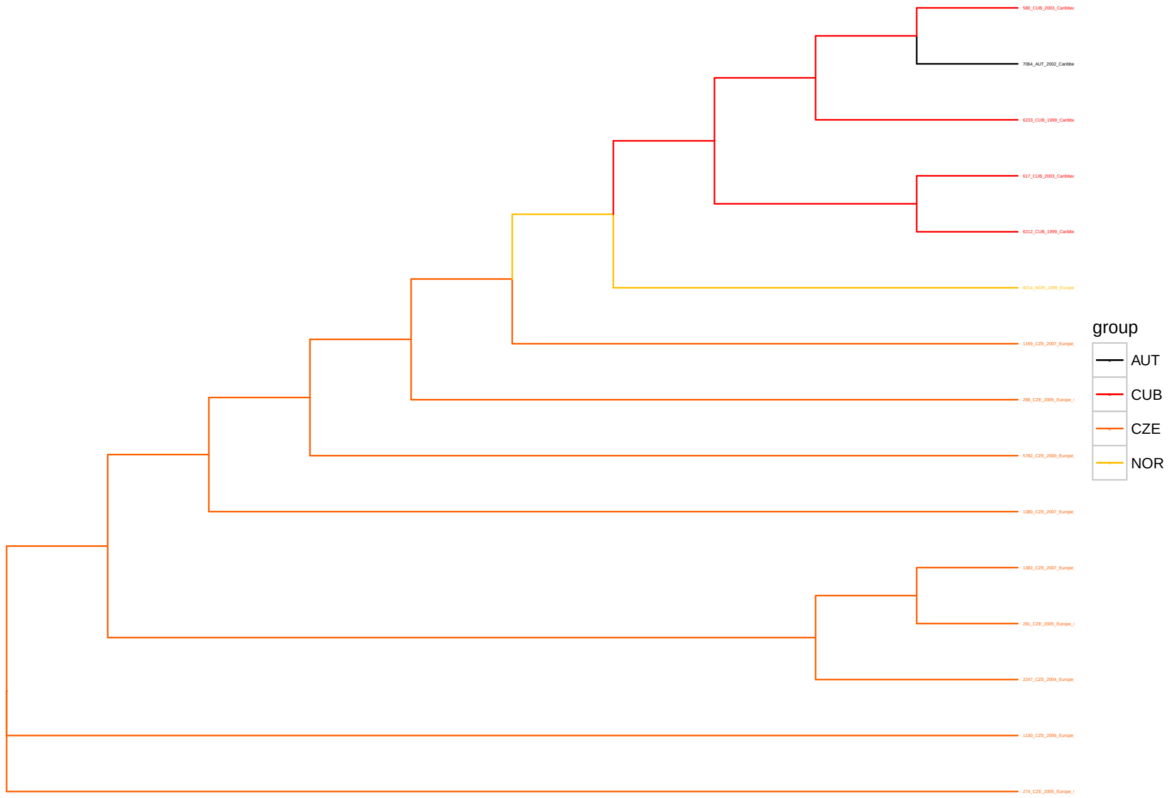
3.3.5 Cluster 5

The 5th Monophyletic Cluster is Cluster 5, it has 40 sequences 65.57% of the original size which was 61, as we may observe the tree is divided into 2 subclusters a Cuban , a European and a Korean, so we pick 2 randomly out of each sub-cluster.



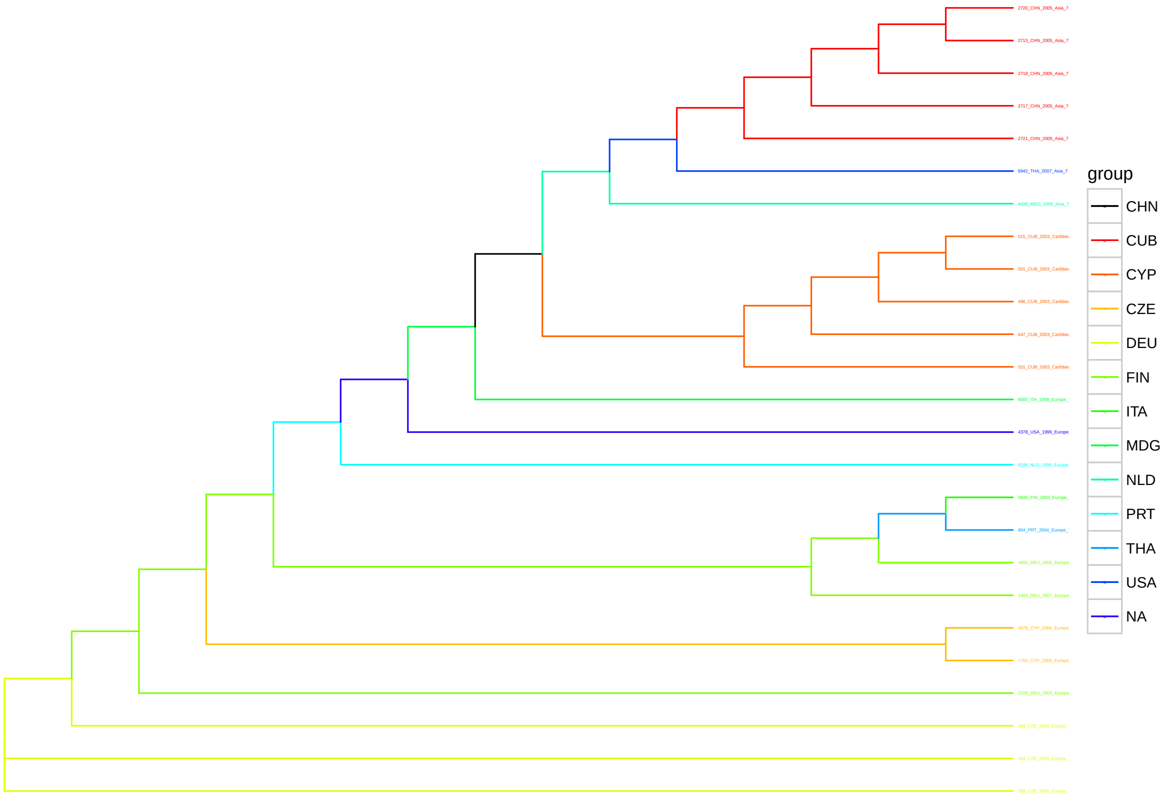
3.3.6 Cluster 6

The 6th Monophyletic Cluster is Cluster 6, it has 15 sequences 28.3% of the original size which was 53, as we may observe the tree is divided into 2 subclusters a Cuban , a European and we pick 2 randomly out of each sub-cluster.



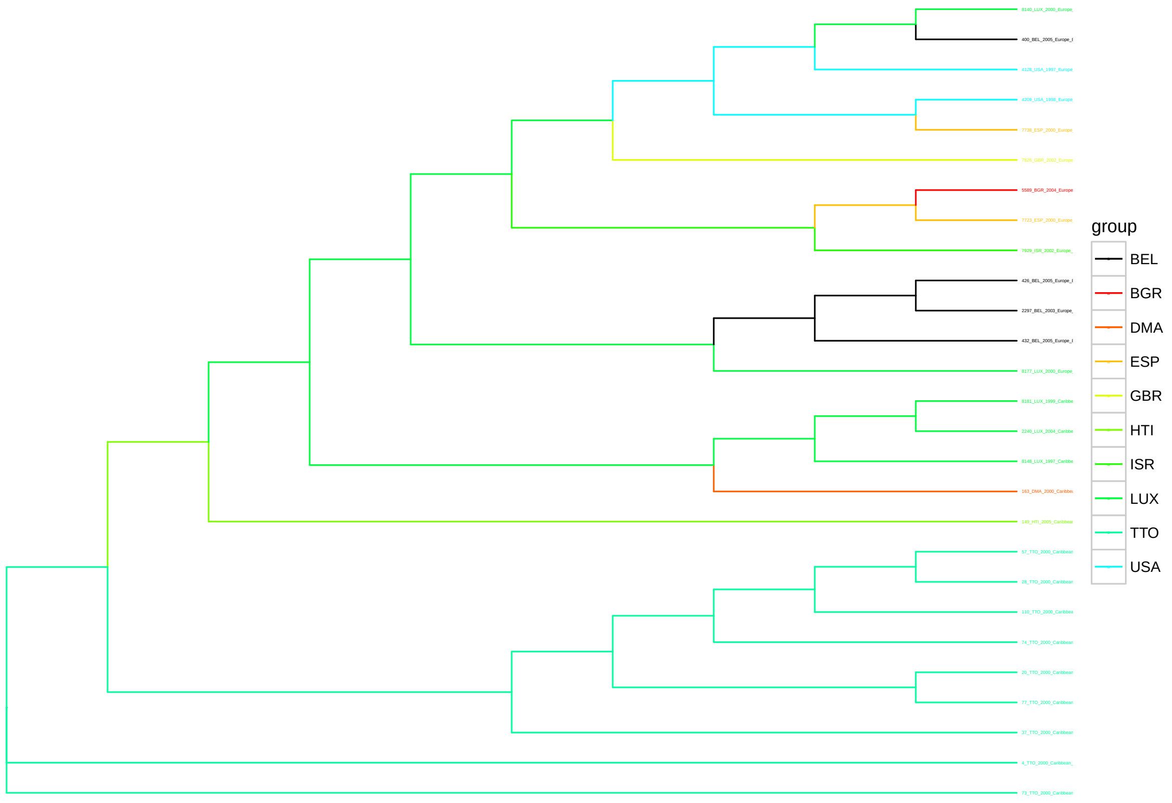
3.3.7 Cluster 7

The 7th Monophyletic Cluster is Cluster 7, it has 25 sequences 49.02% of the original size which was 51, as we may observe the tree is divided into 3 subclusters a Cuban, a Chinese - European, and a only European so we pick 4 randomly out of each sub-cluster, because they appear to have great distance and we need to representation.



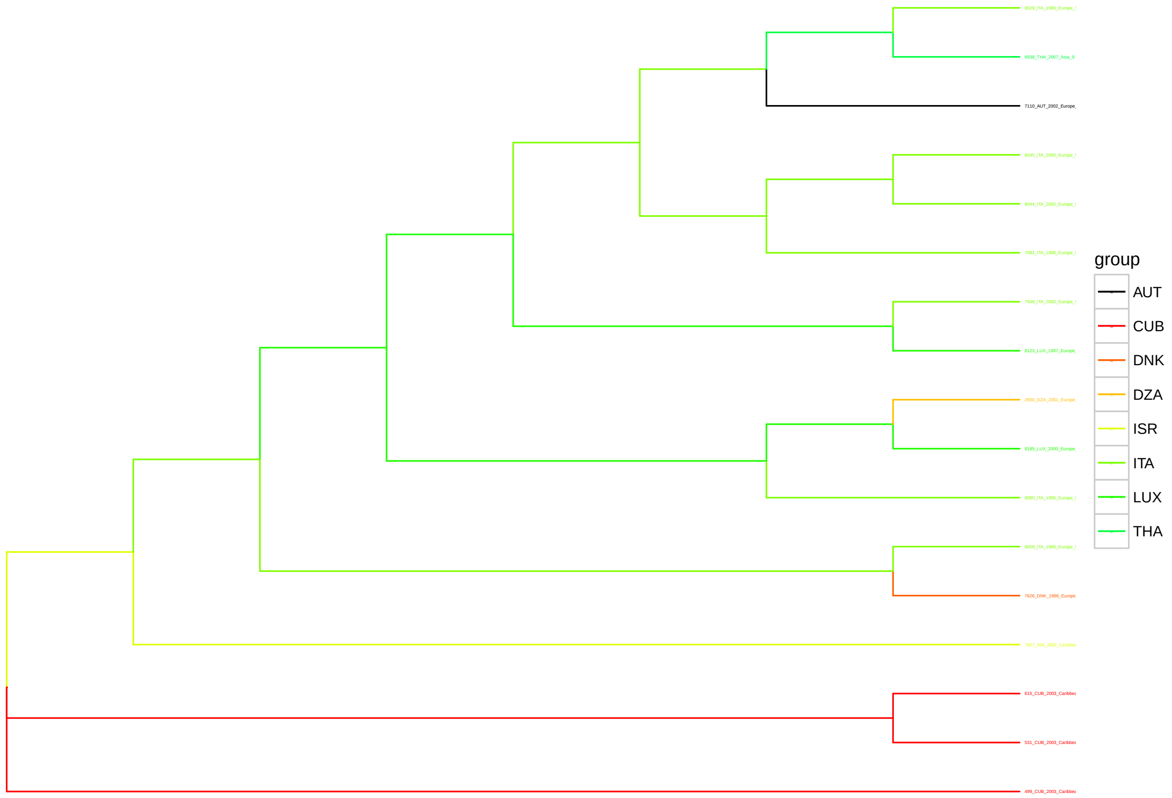
3.3.8 Cluster 8

The 8th Monophyletic Cluster is Cluster 8, it has 27 sequences 46.55% of the original size which was 58, as we may observe the tree is divided into 2 subclusters a Caribbean with mostly Trinidad and Tobacco sequences and a mixed one with American and European sequences. We pick 2 from the solid first and 6 from the second because of the great distance between them.



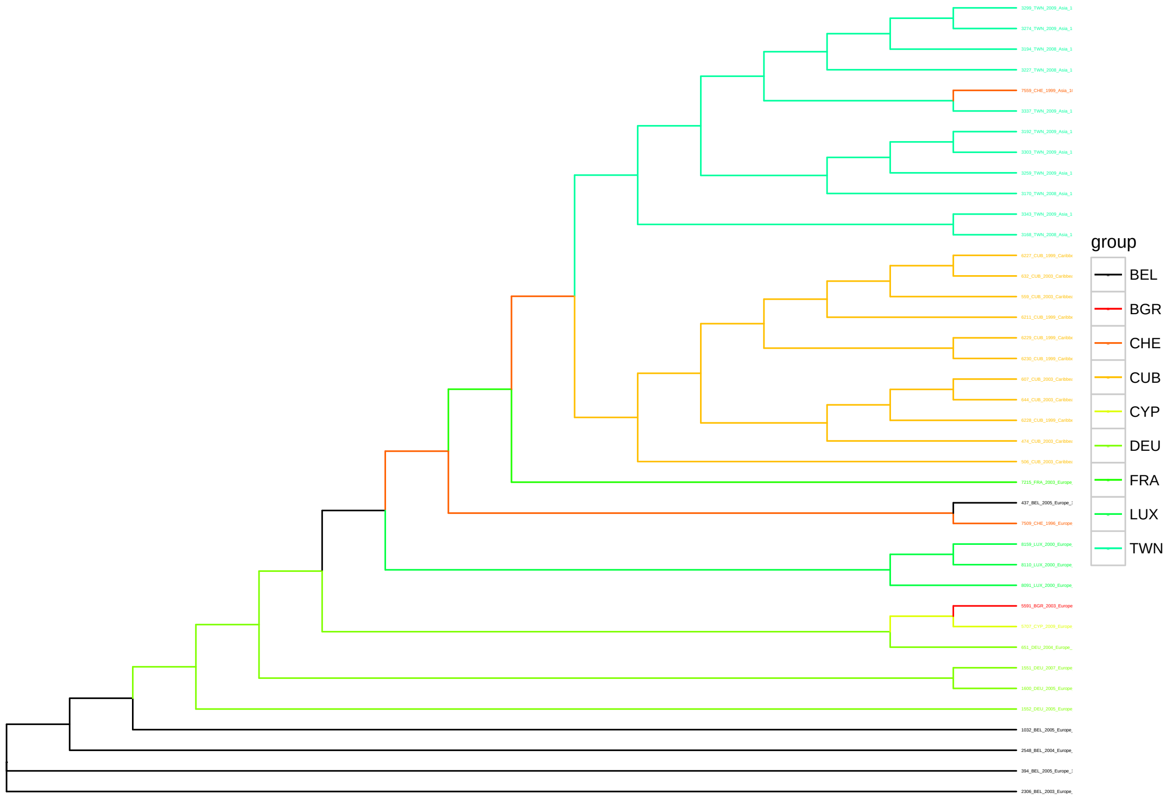
3.3.9 Cluster 9

The 9th Monophyletic Cluster is Cluster 9, it has 17 sequences 45.95% of the original size which was 37 , as we may observe the tree is divided into 2 subclusters a solid Cuban and a European so we pick 2 randomly out of each sub-cluster.



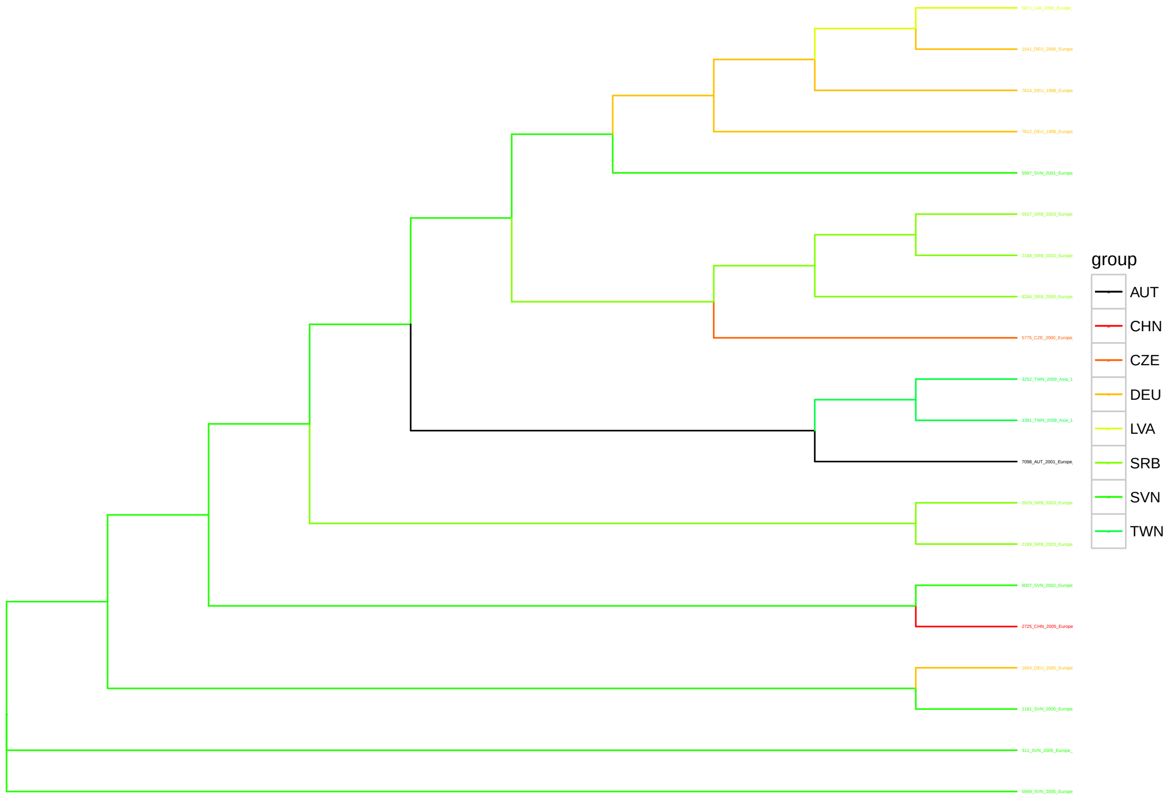
3.3.10 Cluster 10

The 10th Monophyletic Cluster is Cluster 10, it has 39 sequences 84.78% of the original size which was 46, as we may observe the tree is divided into 3 subclusters a solid Cuban a solid Taiwanesse and a European so we pick 2 randomly out of each sub-cluster.



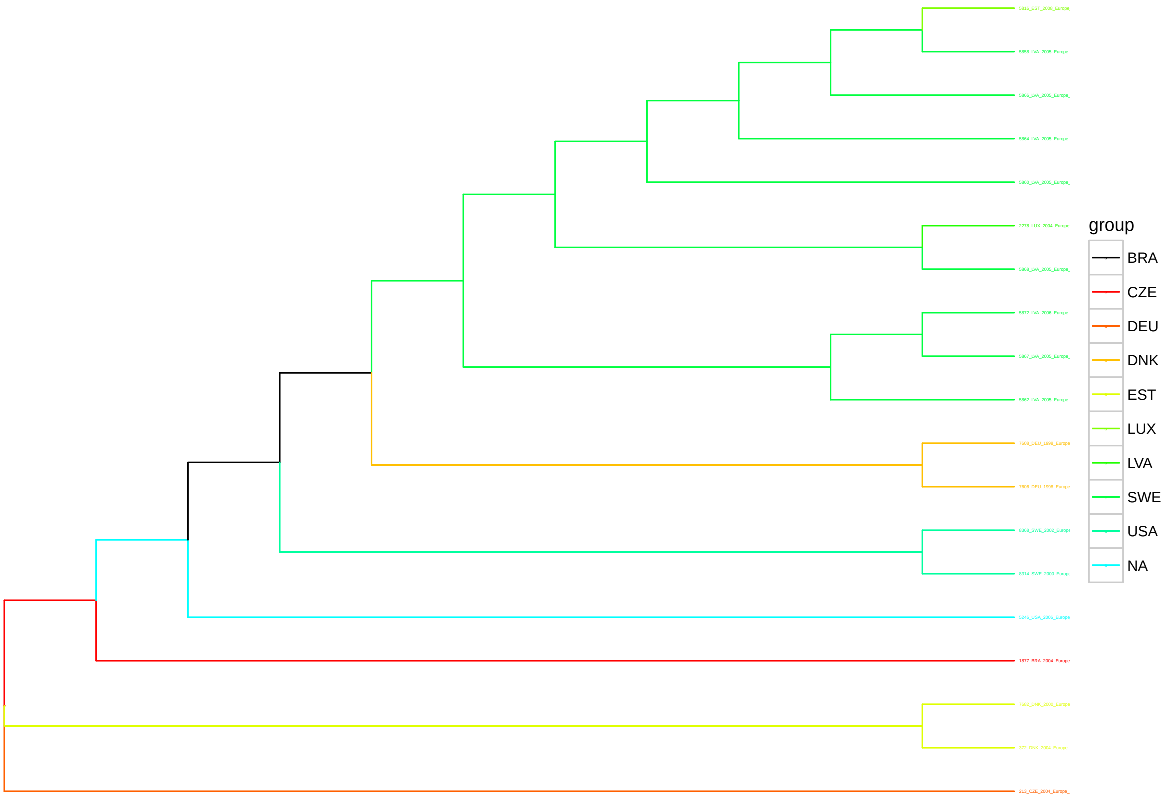
3.3.11 Cluster 11

The 11th Monophyletic Cluster is Cluster 11, it has 20 sequences 62.5% of the original size 32, as we may observe the tree is divided into many subclusters 8 countries in a 20 sequences dataset with almost equal representation so we pick 8 randomly.



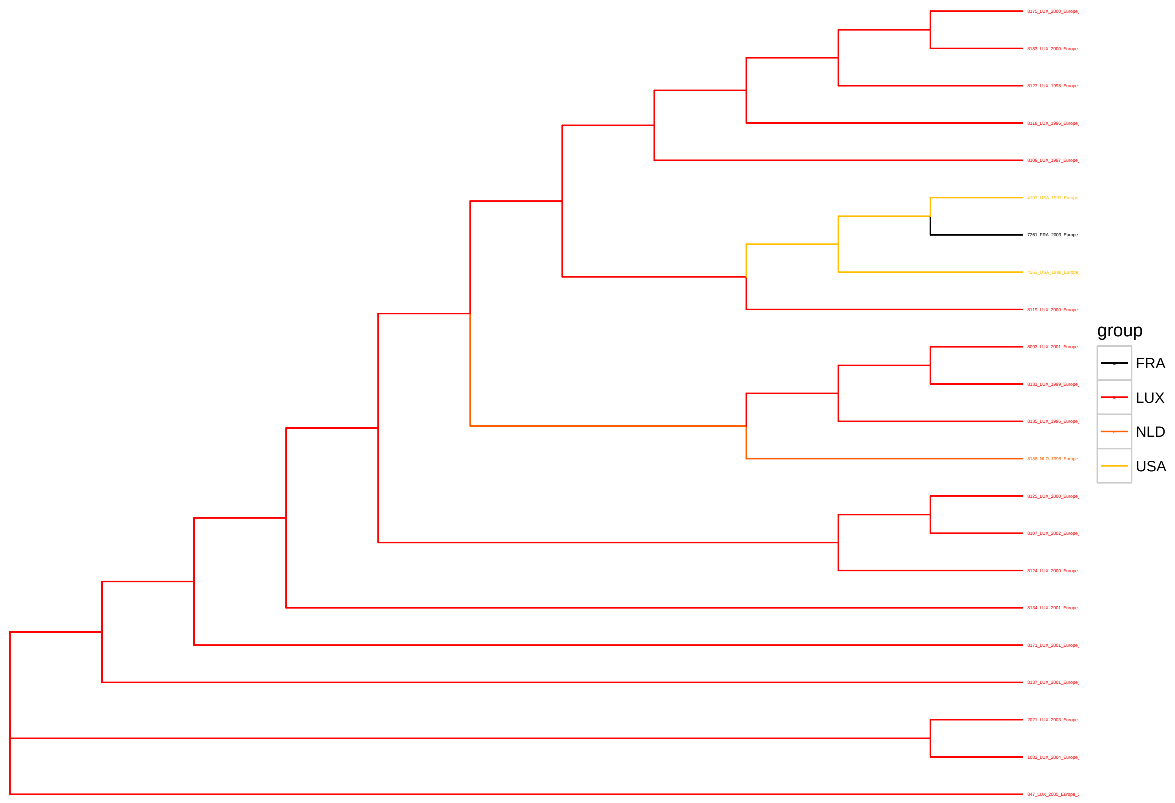
3.3.12 Cluster 12

The 12th Monophyletic Cluster is Cluster 12, it has 19 sequences 63.33% of the original size 30, as we may observe the the tree is divided into 2 subclusters a solid latvian we pick 2 randomly and aloose European we pick 6



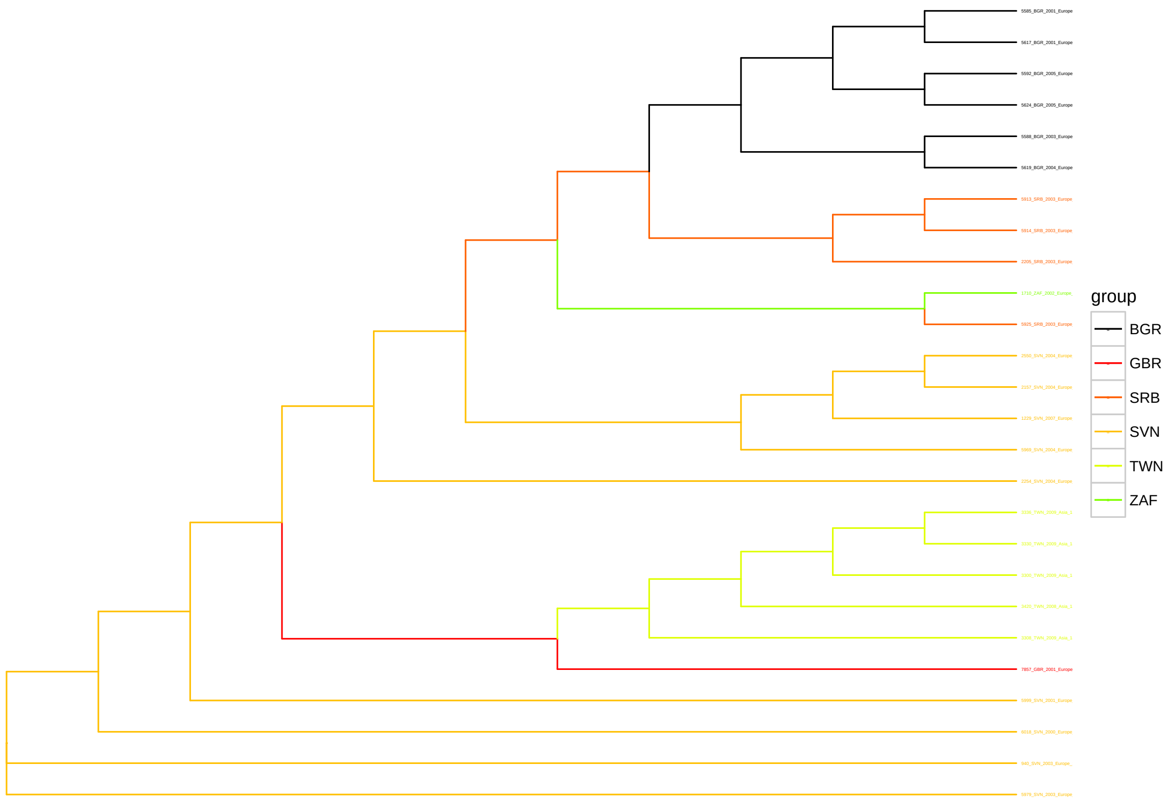
3.3.13 Cluster 13

The 13th Monophyletic Cluster is Cluster 13, it has 22 sequences 66.67% of the original size 33, as we may observe the tree is not divided to any subcluster so we pick randomly 5 sequences.



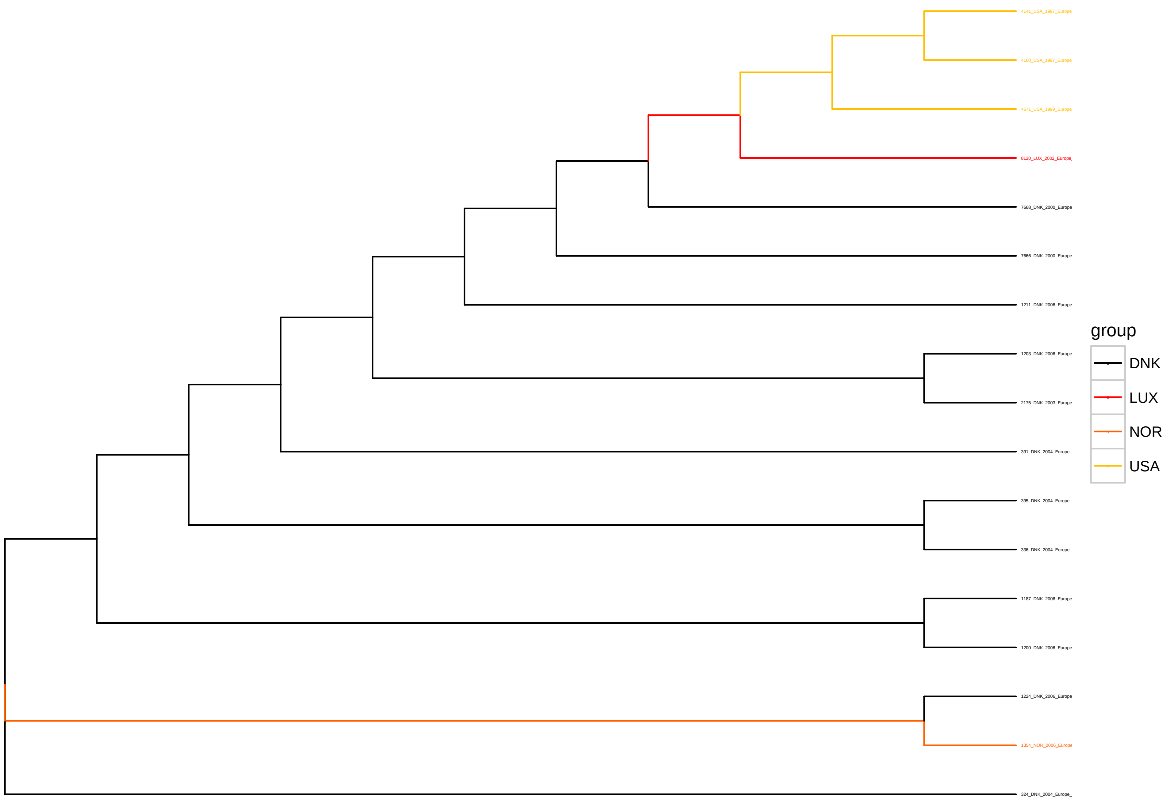
3.3.14 Cluster 14

The 14th Monophyletic Cluster is Cluster 14, it has 26 sequences 78.79% of the original size 33, as we may observe the tree is divided into 3 subclusters a Serbo-Bulgarian, a Slovenian and a Taiwanianand we pick 2 randomly out of each sub-cluster.



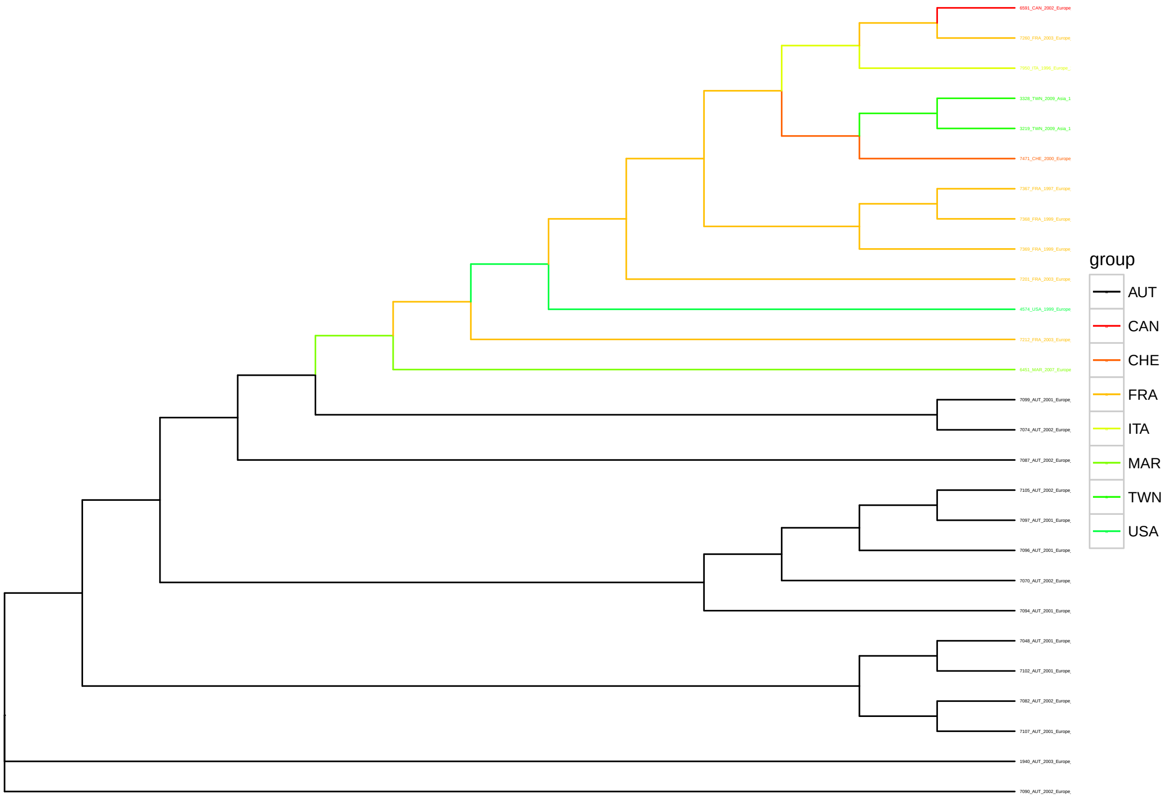
3.3.15 Cluster 15

The 15th Monophyletic Cluster is Cluster 15, it has 17 sequences 51.52% of the original size 33, as we may observe the tree is divided into 2 subclusters a Danish and an Euro-American we pick 2 randomly out of each sub-cluster.



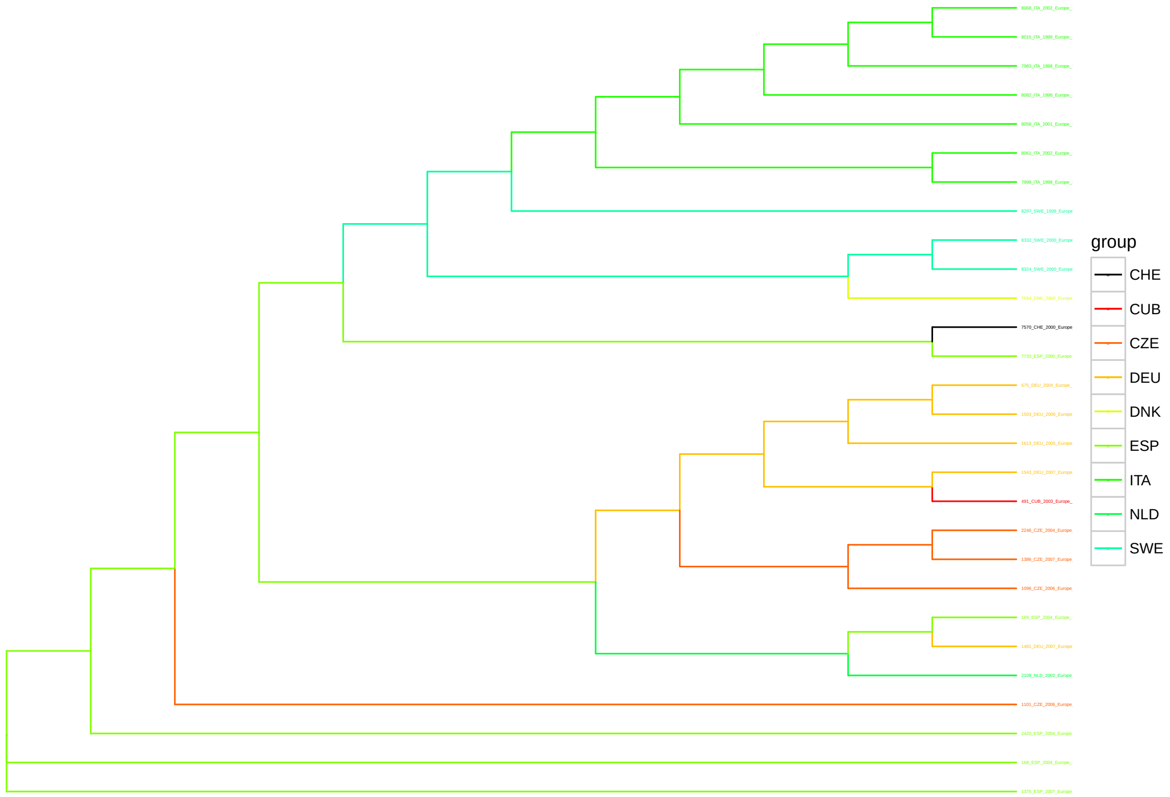
3.3.16 Cluster 16

The 16th Monophyletic Cluster is Cluster 16, it has 27 sequences 75% of the original size 36, as we may observe the the tree is divided into 2 sub clusters an Austrian and a French-Asian , and we pick 3 randomly out of each sub-cluster.



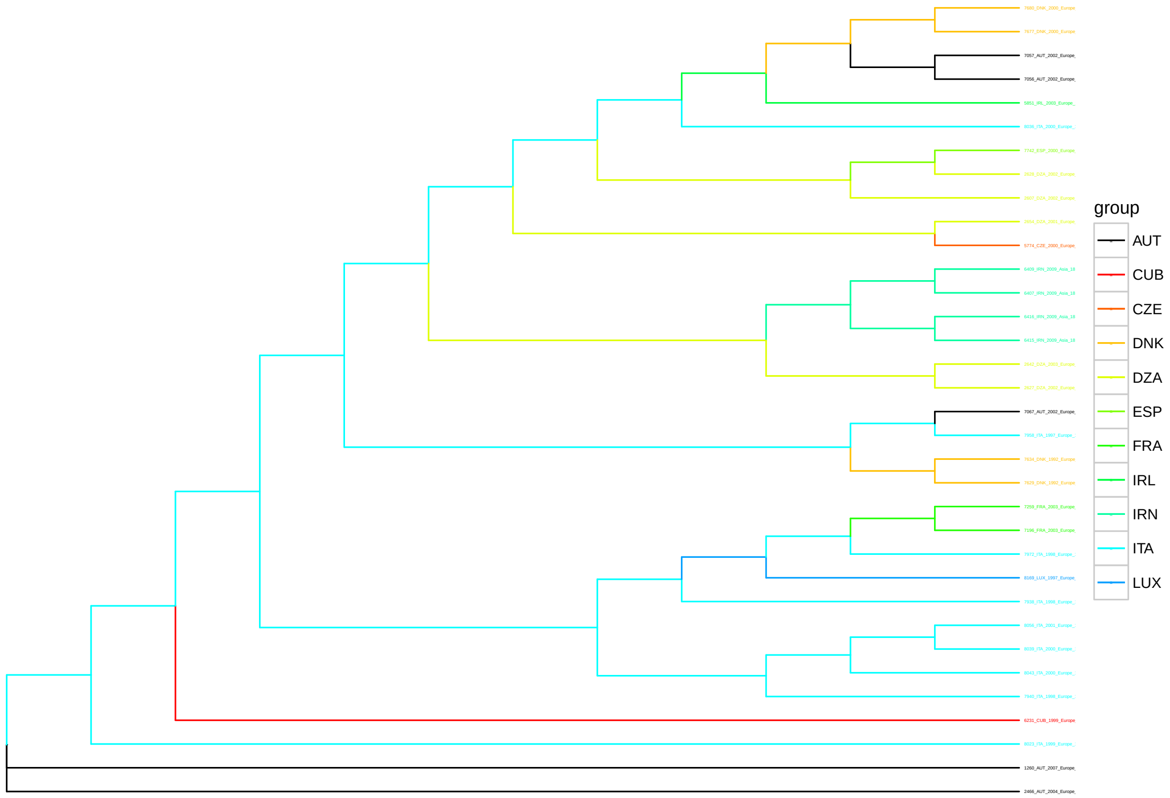
3.3.17 Cluster 17

The 17th Monophyletic Cluster is Cluster 17, it has 28 sequences 75.68% of the original size 37, as we may observe the tree is divided into 3 subclusters a solid Sewdish-Italian, a Spanish and a mixed European and we pick 2 randomly out of each.



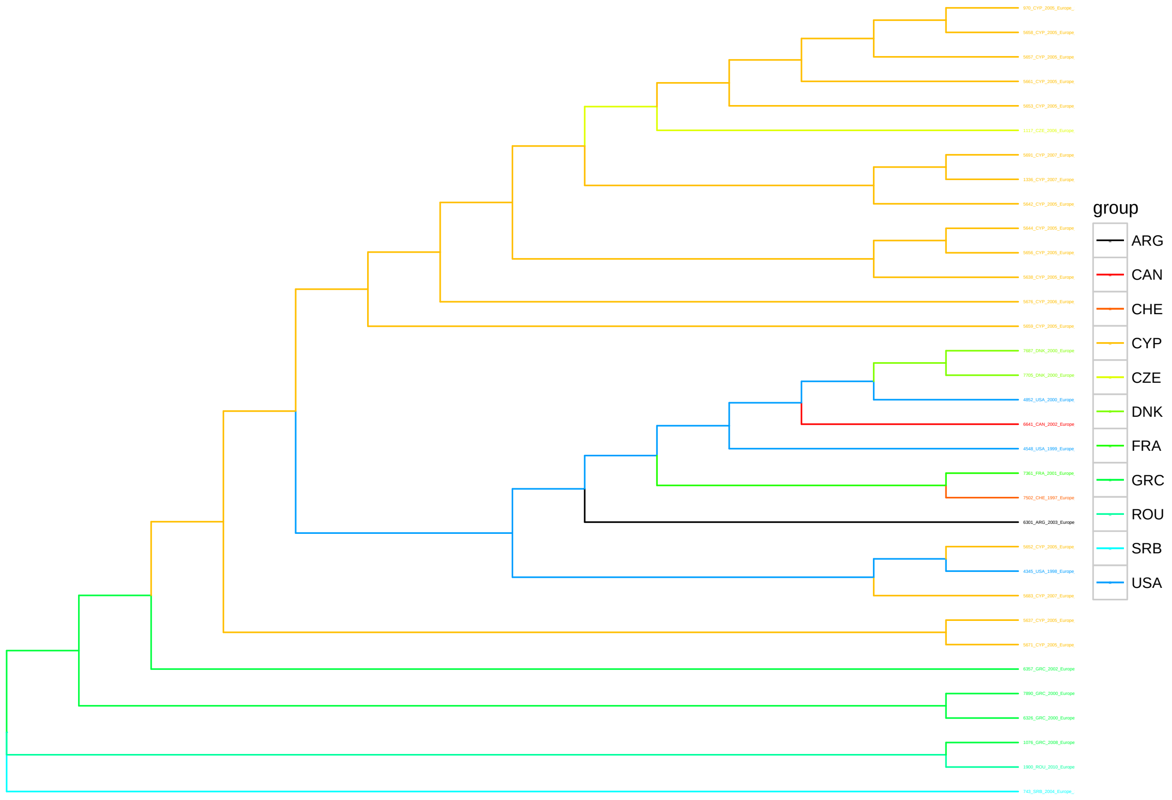
3.3.18 Cluster 18

The 18th Monophyletic Cluster is Cluster 18, it has 34 sequences 87.18% of the original size 39, as we may observe the tree is not divided into any particular sub cluster 11 countries for 34 sequences, almost equally represented. We pick 2 from Italy and random 5 of all others.



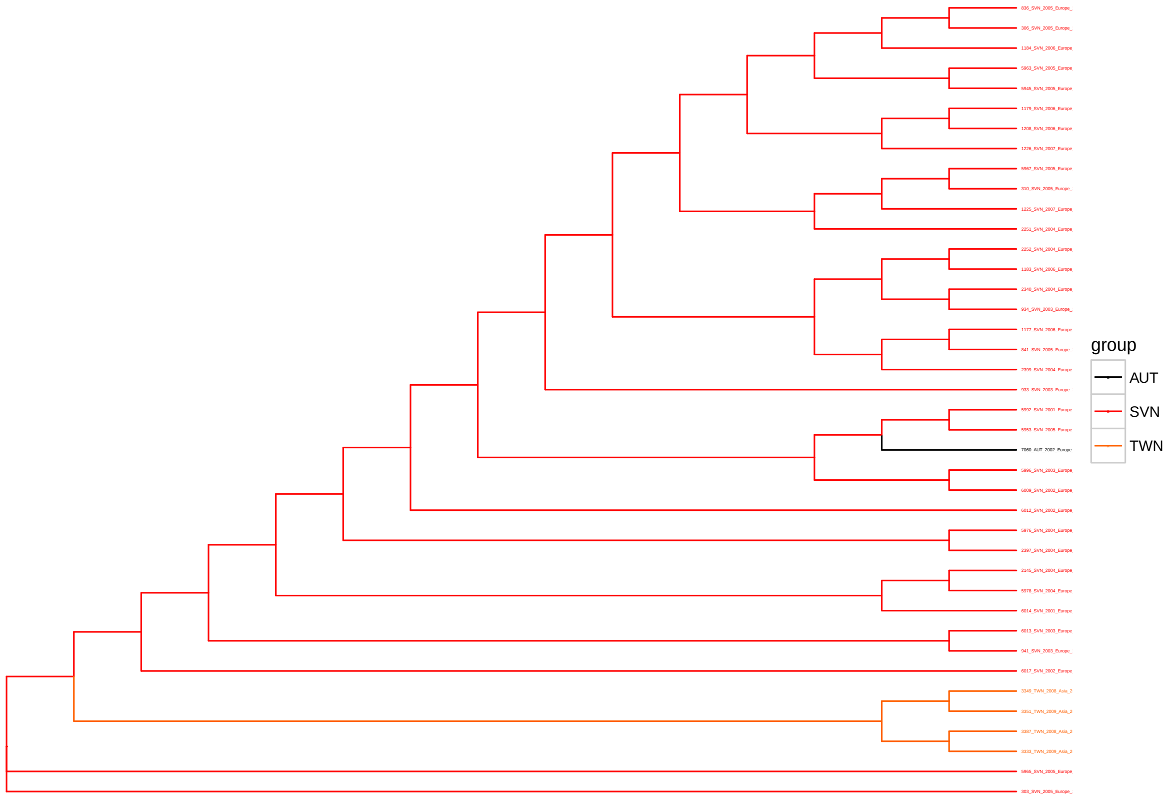
3.3.19 Cluster 19

The 19th Monophyletic Cluster is Cluster 19, it has 33 sequences 75% of the original size 44, as we may observe the tree is divided into 2 subclusters the Greco-Cypriot and all others, so we pick 2 randomly.



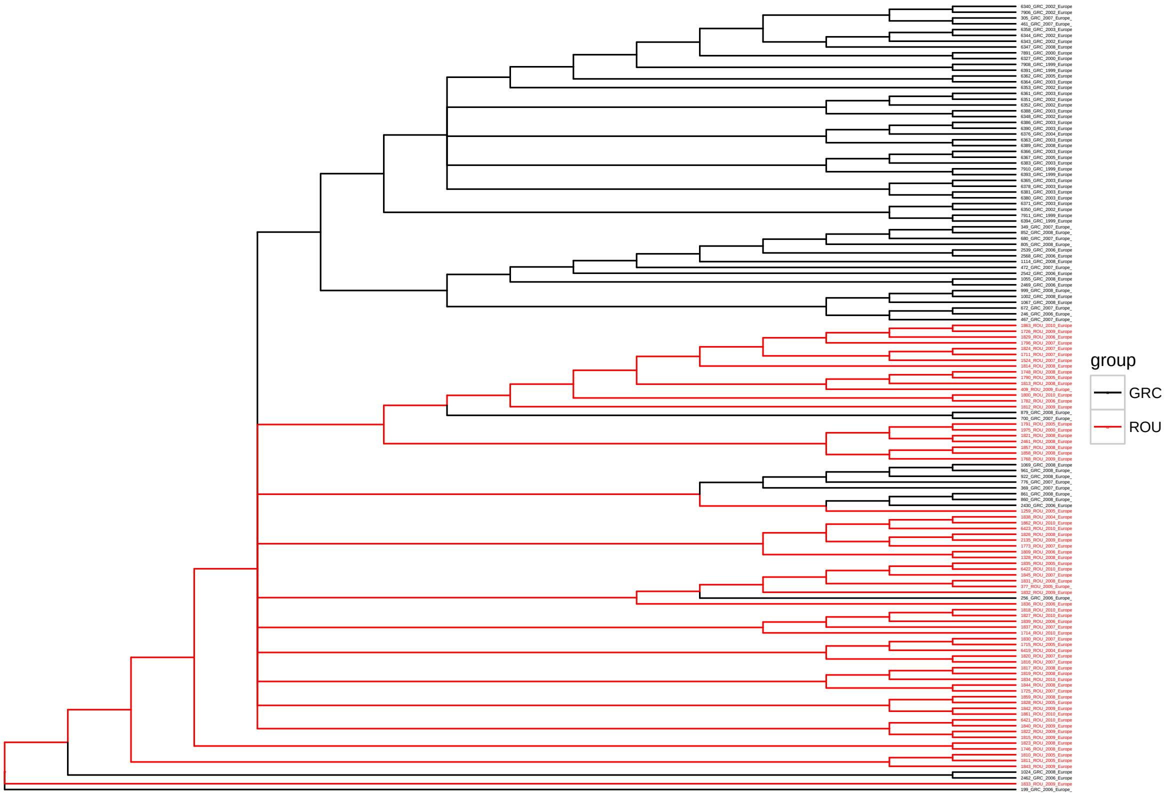
3.3.20 Cluster 20

The 20th Monophyletic Cluster is Cluster 20, it has 40 sequences 86.96% of the original size 46, as we may observe the tree is divided into 2 subclusters a Slovenian and a Taiwanese we pick 2 randomly out of each.



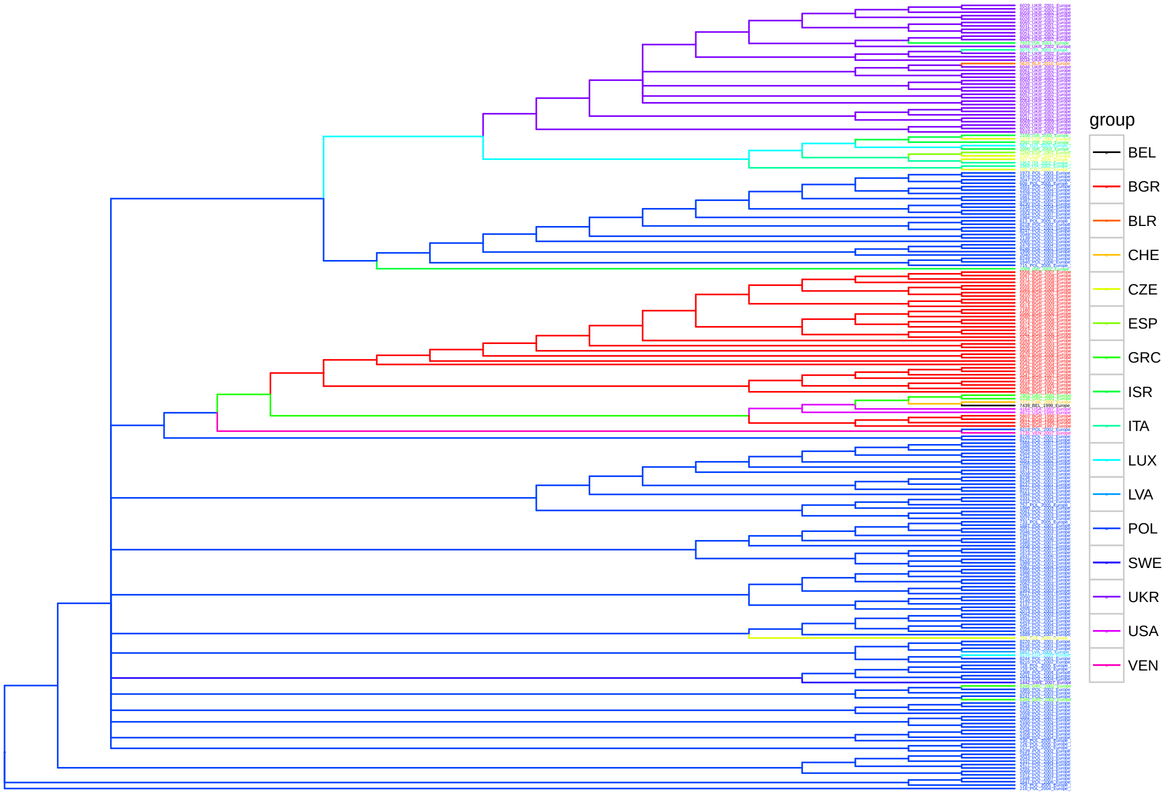
3.3.21 Cluster 21

The 21st Monophyletic Cluster is Cluster 21, it has 136 sequences 95.1% of the original size 143, as we may observe the tree is divided into 2 subclusters Greek and Roumanian and we pick 4 randomly out of each sub-cluster.



3.3.22 Cluster 22

The 22nd Monophyletic Cluster is Cluster 22, it has 230 sequences 95.44% of the original size 241, as we may observe the tree is divided into 3 subclusters a polish, an Ukranian and a Bulgarian we pick 4 randomly out of each sub-cluster.



3.3.23 Cluster 23

The Cluster 23 was randomly picked.

3.3.24 Cluster 25

The 24th Monophyletic Cluster is Cluster 25 we pick them all.

3.3.25 Random Pick from sequences without any Cluster Information

The previous sub sampling method collected 200 sequences and took into consideration :

- The proportional representation of the clusters
- The genetic distance in each cluster
- to keep intact any phylogenetic properties

A randomly choosen Data-Set of 200 sequences was also picked.

The part of the original Dataset with no Cluster information was randomly sampled with 300 sequences, 500 sequences and 700 sequences. Then we combined these DataSets into 3 All Randomly picked DataSets (500 , 700 and 900 sequences) and 3 Semi-Randomly picked Datasets (500 , 700 and 900 sequences)

4 Phylodynamic Analysis

We have 6 datasets in order to run the phylodynamic analysis. The All Random sized 500 , 700 and 900 were we randomly picked 200 sequences out of the Clustered Dataset and 300 - 500 -700 out of the non Clustered Data-set and the Semi Random sized were we picked 200 sequences out of the Consensus Trees out of the clustered and the same random 200 - 500 -700 out of the non Clustered Data-set. In order to speed the phylodynamic Analysis done by the Beast v1.8.x program we will assess each one of the 6 Data-Set's Maximum Likelihood Best Tree as a starting tree. I order to find the ML Best Tree we will use the RaXmL once again, with the same options.

The parameters we inserted into the Beauti program were :

1. The Data-Sets
2. The Guess Date was specified by the information of the tips
3. The Substitution Model was GTR with Gamma Heterogeneity Model and 4 categories
4. The Molecular Clock correction was Uncorrealated Lognormal Relaxed Clock as proposed by (Drummond et al., 2006)
5. The Maximul Likelihood Best Tree reported from RaXmL for each Data-Set
6. Operator Tuning !! <– !!
7. The length of Chain was set to 100,000,000 with Echo and Log at every 10,000

4.1 Results

For all the Data-Sets the origin could not be identified in detail. The Molecular clock analysis revealed that the time of the mostrecent common ancestor (tMRCA) was in 1960 (95%HPD: 1950-1970). The Model suggested a rapid increase in number of infections lasting between 1975 and 1985. After then new infections seem to be detected into Eastern Europe Countries.

References

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., Hyndman, R., 2016. Rmarkdown: Dynamic documents for r.

Ariën, K.K., Vanham, G., Arts, E.J., 2007. Is HIV-1 evolving to a less virulent form in humans? Nature Reviews Microbiology 5, 141–151.

Ballegooijen, W.M. van, Houdt, R. van, Bruisten, S.M., Boot, H.J., Coutinho, R.A., Wallinga, J., 2009. Molecular sequence data of hepatitis b virus and genetic diversity after vaccination. American Journal of Epidemiology 170, 1455–1463.

Barouch, D.H., 2008. Challenges in the development of an HIV-1 vaccine. Nature 455, 613–619.

Barre-Sinoussi, F., Chermann, J., Rey, F., Nugeyre, M., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., Rozenbaum, W., Montagnier, L., 1983. Isolation of a t-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220, 868–871.

Bette Korber, B.F.H., Christian Brander, Watkins, D.I. (Eds.), n.d. HIV molecular immunology database 1999. Theoretical Biology and Biophysics.

Buonaguro, L., Tornesello, M.L., Buonaguro, F.M., 2007. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: Pathogenetic and therapeutic implications. Journal of Virology 81, 10209–10219.

Camacho R., C., 2006. The significance of subtype-related genetic variability: Controversies and unanswered questions. Mediscript.

Chattopadhyay, A., 2011. Oral health epidemiology: Principles and practice. Jones; Bartlett Publishers.

Cohen, M.S., Shaw, G.M., McMichael, A.J., Haynes, B.F., 2011. Acute HIV-1 infection. New England Journal of Medicine 364, 1943–1954.

Drosopoulos, W.C., Rezende, L.F., Wainberg, M.A., Prasad, V.R., 1998. Virtues of being faithful: Can we limit the genetic variation in human immunodeficiency virus? Journal of Molecular Medicine 76, 604–612.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology 4, e88.

Felsenstein, J., 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. Journal of Molecular Evolution 53, 447–455.

Gallo, R., Salahuddin, S., Popovic, M., Shearer, G., Kaplan, M., Haynes, B., Palker, T., Redfield, R., Oleske, J., Safai, B., et, 1984. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. Science 224, 500–503.

Gao, S.G.R., F.; Morrison, 1996. Molecular cloning and analysis of functional envelope genes from human deficiency virus type 1 sequence subtypes a through g 70, 1651–1667.

Greene, W.C., 2007. A history of AIDS: Looking back to see ahead. European Journal of Immunology 37, S94–S102.

Grenfell, B.T., 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303, 327–332.

Heibl, C., 2008 onwards. PHYLOCH: R language tree plotting tools and interfaces to diverse phylogenetic software packages. http://www.christophheibl.de/Rpackages.html.

Hladik, F., McElrath, M.J., 2008. Setting the stage: Host invasion by HIV. Nat Rev Immunol 8, 447–457.

Ho, S.Y.W., 2013. Molecular clocks, relaxed variant. In: Encyclopedia of Scientific Dating Methods. Springer Science Business Media, pp. 1–5.

Huson, D.H., Scornavacca, C., 2012. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Systematic Biology 61, 1061–1067.

Katoh, K., 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Research 30, 3059–3066.

Lang, D.T., CRAN Team, 2015. XML: Tools for parsing and generating xML within r and s-plus.

Leisch, F., Peng, R.D., submitted. Knitr: A comprehensive tool for reproducible research in r. Methods in Ecology and Evolution.

Louwagie, F.E.P., J.; McCutchan, 1993. Molecular cloning and analysis of functional envelope genes from human deficiency virus type 1 sequence subtypes a through g 7, 769–780.

Magiorkinis, K.A., G., 2009. The global spread of hIV-1 subtype b epidemic: A phylogeographic meta-analysis.

Merson, M.H., O'Malley, J., Serwadda, D., Apisuk, C., 2008. The history and challenge of HIV prevention. The Lancet 372, 475–488.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees. In: 2010 Gateway Computing Environments Workshop (GCE). Institute of Electrical & Electronics Engineers (IEEE).

Ooms, J., James, D., DebRoy, S., Wickham, H., Horner, J., 2015. RMySQL: Database interface and 'mySQL' driver for r.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Paraskevis, D., Pybus, O., Magiorkinis, G., Hatzakis, A., Wensing, A.M., Vijver, D.A. van de, Albert, J., Angarano, G., Asjo, B., Balotta, C., Boeri, E., Camacho, R., Chaix, M.-L., Coughlan, S., Costagliola, D., Luca, A.D., Mendoza, C. de, Derdelinckx, I., Grossman, Z., Hamouda, O., Hoepelman, I.M., Horban, A., Korn, K., Kuecherer, C., Leitner, T., Loveday, C., Macrae, E., Maljkovic, I., Meyer, L., Nielsen, C., Coul, E.L.O. de, Ormaasen, V., Perrin, L., Puchhammer-Stockl, E., Ruiz, L., Salminen, M., Schmit, J.-C., Schuurman, R., Soriano, V., Stanczak, J., Stanojevic, M., Struck, D., Laethem, K.V., Violin, M., Yerly, S., Zazzi, M., Boucher, C.A., Vandamme, A.-M., Programme, S., 2009. Tracing the HIV-1 subtype b mobility in europe: A phylogeographic approach. Retrovirology 6, 49.

Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R., Moret, B.M., Stamatakis, A., 2010. How many bootstrap replicates are necessary? Journal of Computational Biology 17, 337–354.

Peeters, M., 2001. Recombinant hIV sequences: Their role in the global epidemic. Theoretical Biology; Biophysics Group.

Popovic, M., Sarngadharan, M., Read, E., Gallo, R., 1984. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. Science 224, 497–500.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Revell, L.J., 2012. Phytools: An r package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3, 217–223.

Richman, D.D., Margolis, D.M., Delaney, M., Greene, W.C., Hazuda, D., Pomerantz, R.J., 2009. The challenge of finding a cure for HIV infection. Science 323, 1304–1307.

RStudio Team, 2015. RStudio: Integrated development environment for r. RStudio, Inc., Boston, MA.

Schliep, K., 2011. Phangorn: Phylogenetic analysis in r. Bioinformatics 27, 592–593.

Scott A. Chamberlain, L.J.R. &, 2013. Rphylip: An r interface for pHYLIP.

South, A., 2011. Rworldmap: A new r package for mapping global data. The R Journal 3, 35–43.

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.

Urbanek, S., 2013. Png: Read and write pNG images.

Vercauteren, J., Wensing, A.M.J., Vijver, D.A.M.C. van de, Albert, J., Balotta, C., Hamouda, O., Kücherer, C., Struck, D., Schmit, J.-C., Åsjö, B., Bruckova, M., Camacho, R.J., Clotet, B., Coughlan, S., Grossman, Z., Horban, A., Korn, K., Kostrikis, L., Nielsen, C., Paraskevis, D., Poljak, M., Puchhammer-Stöckl, E., Riva, C., Ruiz, L., Salminen, M., Schuurman, R., Sonnerborg, A., Stanekova, D., Stanojevic, M., Vandamme, A.-M., Boucher, C.A.B., 2009. Transmission of drug-resistant HIV-1 is stabilizing in europe. The Journal of Infectious Diseases 200, 1503–1508.

Weiss, R., 1993. How does hIV cause aIDS? 260, 1273–1279.

Wensing, A.M.J., Vijver, D.A. van de, Angarano, G., Åsjö, B., Balotta, C., Boeri, E., Camacho, R., Chaix, M.-L., Costagliola, D., Luca, A.D., Derdelinckx, I., Grossman, Z., Hamouda, O., Hatzakis, A., Hemmer, R., Hoepelman, A., Horban, A., Korn, K., Kücherer, C., Leitner, T., Loveday, C., MacRae, E., Maljkovic, I., Mendoza, C. de, Meyer, L., Nielsen, C., Coul, E.L.O. de, Ormaasen, V., Paraskevis, D., Perrin, L., Puchhammer-Stöckl, E., Ruiz, L., Salminen, M., Schmit, J.-C., Schneider, F., Schuurman, R., Soriano, V., Stanczak, G., Stanojevic, M., Vandamme, A.-M., Laethem, K.V., Violin, M., Wilbe, K., Yerly, S., Zazzi, M., Boucher, C.A., 2005. Prevalence of drug-resistant HIV-1 variants in untreated individuals in europe: Implications for clinical management. The Journal of Infectious Diseases 192, 958–966.

Wensing, J., A.M., 2008. Transmission of drug-resistant HIV-1 in europe remains limited to single classes. AIDS 22, 625–635.

Wickham, H., 2009. Ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

Wikipedia, 2016. Wikipedia, the free encyclopedia.

World Health Organization, 2016. Epidemiology- definition.

Yu, G., Smith, D., Zhu, H., Guan, Y., Lam, T.T.-Y., submitted. Ggtree: An r package for visualization and annotation of phylogenetic tree with different types of meta-data. Methods in Ecology and Evolution.