**Working report 1: stress sub-optimal sequence coverage**

Run simulation with 4000 individuals and 8000 individuals, per each run consider different cases as defined below. **Scenario\*i\_A** is for 4000, and **Scenario\*i\_B** is for 8000.

# **Preparing different strains of HIV-1**

Five variants will be considered in the experiments: HIV-1 subtype A, B, C, D, and G; and only the ***pol gene*** will se used in the simulations.

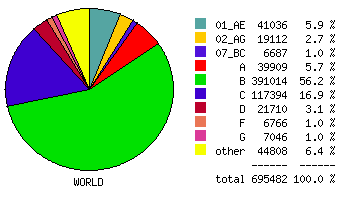
## **Subtype B**

Use the reference sequence in the data base at <https://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html>

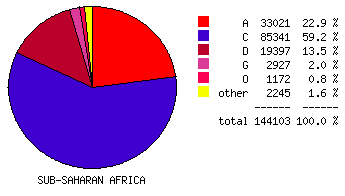
(one sequence for subtype B)

## **Subtypes A, C, D, and G**

Current HIV-1 subtypes variants found in the World and in Sub-Saharan Africa: <https://www.hiv.lanl.gov/components/sequence/HIV/geo/geo.comp>



*Figure 1: Pie slice of frequency of subtypes of HIV-1 in the World*



*Figure 2: Pie slice of frequency of subtypes of HIV-1 in Sub-Saharan Africa*

We choose: subtype **A**, **C**, **D**, and **G**; and **9 sequences were retained per strain subtype.**

* **A**: full length 4-11 & 15 ([https://www.hiv.lanl.gov/components/sequence/HIV/search/d\_search.comp?ssam\_subtype=A%20OR%20A1%20OR%20A2&ssam\_organism=HIV-1&ssam\_sample\_georegion=ssa&ssam\_sample\_country=[A-Z](https://www.hiv.lanl.gov/components/sequence/HIV/search/d_search.comp?ssam_subtype=A OR A1 OR A2&ssam_organism=HIV-1&ssam_sample_georegion=ssa&ssam_sample_country=[A-Z)])
* **C**: full length 279-294 (<https://www.hiv.lanl.gov/components/sequence/HIV/search/search.comp>) 27/10/2017 [remove 282 & 283 double 281 / remove 285 double 284 / remove 288-281, double 292] >>
* **D**: not full length (+8k), 671-677 & 6,9 (<https://www.hiv.lanl.gov/components/sequence/HIV/search/search.comp>)
* **G**: full length (+8k), 2, 6, 13, 583, 780, 894-897 (<https://www.hiv.lanl.gov/components/sequence/HIV/search/search.comp>)

Within these choosen virus retain one with less gaps and retrieve the pol gene (done with MEGA alignment). File with the four different strains of pol gene: “***HIV\_1\_A\_C\_D\_G\_pol.fas***”, file with a single strain are ***HIV\_1\_A\_single\_pol.fas, HIV\_1\_C\_single\_pol.fas, HIV\_1\_D\_single\_pol.fas,*** and ***HIV\_1\_G\_single\_pol.fas***.

Remane:

* **hiv.seq.A.pol.i.fasta**
* **hiv.seq.B.pol.i.fasta** [without any gaps]
* **hiv.seq.C.pol.i.fasta**
* **hiv.seq.D.pol.i.fasta**
* **hiv.seq.G.pol.i.fasta**

## For all subtypes, call in R the single sequence (pol gene) per subtype, deal with gaps (e.g.: delete gaps > Ref: Steve Evans and Tandy Warnow, Phylogenetic analyses of alignments with gaps) and simulate evolution of each on a coalescent tree of 30 tips. Before, rename taxon labels in the input sequence files (avoid error with seq-gen ”Tree is missing from end of sequence file”).

Taxons were:

1 3012

>A1.UG.-.UG031.AB098330

1 3012

>B.Ref

1 3012

>C.ZM.2002.02ZM108.AB254141

1 3012

>D.SN.1990.SE365.AB485648

1 3012

>G.GH.2003.GHNJ175.AB231893

After deleting gaps, strains A,B,D, and G decrease in lengths. I renamed the files

* **hiv.seq.A.pol.j.fasta > 3006**
* **hiv.seq.B.pol.j.fasta > 3012**
* **hiv.seq.C.pol.j.fasta > 2949**
* **hiv.seq.D.pol.j.fasta > 2985**
* **hiv.seq.G.pol.j.fasta > 2988**

and the taxons names: Seq.A, Seq.B, Seq.C, Seq.D, and Seq.G.

To simulate the sequence under the coalescent tree I used **frequencies** from the inputs sequences, and the rates c(3.37,14.50,1.44,1.21,14.50,1.00) from <http://www.math.mcgill.ca/ivrbik/vignette.html> for all. I simulate the sequence under GTR+Gamma (category 4, and shape 0.9) for all.

I got pools of different virus strains: **A.pool.gene.pol.fasta**, **B.pool.gene.pol.fasta**, **C.pool.gene.pol.fasta**, **D.pool.gene.pol.fasta**, and **G.pool.gene.pol.fasta**, each with 30 sequences.

Construct a phylogenetic tree of **A.pool.gene.pol.fasta**, **B.pool.gene.pol.fasta**, and **C.pool.gene.pol.fasta**

**Monday 6 November 2017**

After testing if the sequence in pool remain the same subtypes (using subtyping tool http://www.bioafrica.net/rega-genotype/html/subtypinghiv.html), I found not, and rethiking on having same gene length for all subtypes (A,B,C,D, and G) and making sure that the virus subtypes remain the same, I did the following:

1. Make same length the different strains

* **hiv.seq.A.pol.j.fasta > 3012**
* **hiv.seq.B.pol.j.fasta > 3012**
* **hiv.seq.C.pol.j.fasta > 3012**
* **hiv.seq.D.pol.j.fasta > 3012**
* **hiv.seq.G.pol.j.fasta > 3012**

2. Make sure all didn’t loose their subtyping by using <http://www.bioafrica.net/rega-genotype/html/subtypinghiv.html>. Each strain was subdivided in four subsequences due to the fact the subtyping tool can handle sequence less than 1000bp. Part.1.800, Part.2.800, Part.3.800, and Part.4.612.

**RESULTS**:<http://bioafrica.mrc.ac.za/rega-genotype/genotype.php?cmd=list&job=457289831>

We choose subtypes A, B, and G for further work because all their subsequences were identified by the subtyping tool.

3. Donwload per each part of the subtypes chosen the sequences (when a sequence is tested for subtyping, it is compared to other sequences), and use jModelTest2.1.3 to get parameter values (frequencies, transmission rates, proportion of invariant sites, and gamma shape). Explanation: for e.g. for A.pol.part.1.800 we download its alignment and proceed by jModelTest, the same for A.pol.part.2.800, A.pol.part.3.800, and A.pol.part.4.612 and we will have an average for each parameter of interest. We will do the same for subtype A, and G.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Part | Freq.A | Freq.C | Freq.G | Freq.T | Ra | Rb | Rc | Rd | Re | Rf | p-inv | Gamma-shape |
| A.part.1 | 0.3968 | 0.1597 | 0.2395 | 0.2041 | 2.5456 | 8.1515 | 1.0124 | 1.1789 | **12.9379** | **1.0000** | 0.4160 | **1.0700**  **~~0.9~~** |
| A.part.2 | **0.3935** | **0.1708** | **0.2060** | **0.2297** | **2.9114** | **12.5112** | **1.2569** | **0.8559** | 15.7725 | 1.0000 | **0.5230** | 1.9700 |
| A.part.3 | 0.3903 | 0.1584 | 0.2287 | 0.2227 | 1.4780 | 8.3455 | 0.7475 | 0.5091 | 10.2495 | 1.0000 | 0.3760 | 0.8170 |
| A.part.4 | 0.3952 | 0.1695 | 0.2289 | 0.2064 | 4.4667 | 11.9476 | 2.0865 | 2.8635 | 26.7825 | 1.0000 | 0.1510 | 0.3150 |
| B.part.1 | 0.3913 | 0.1640 | 0.2303 | 0.2144 | 2.5049 | 8.5117 | 1.0556 | **1.2966** | **12.6824** | **1.0000** | 0.4610  **~~0.5~~** | 1.2760 |
| B.part.2 | 0.3941 | 0.1698 | 0.2116 | 0.2245 | 2.3657 | **10.7771** | **1.0675** | 0.7766 | 12.4312 | 1.0000 | 0.4680 | 1.4780 |
| B.part.3 | **0.3857** | **0.1609** | **0.2234** | **0.2300** | **2.2228** | 11.3781 | 0.8362 | 0.5263 | 13.6281 | 1.0000 | 0.4170 | **0.9410** |
| B.part.4 | 0.3973 | 0.1596 | 0.2406 | 0.2025 | 4.4488 | 11.5945 | 2.3689 | 2.7909 | 29.1640 | 1.0000 | 0.2000 | 0.3180 |
| G.part.1 | 0.3910 | 0.1671 | 0.2409 | 0.2010 | 3.1392 | **9.9166** | **1.3332** | **1.2652** | 14.6654 | 1.0000 | 0.3740 | 0.9520 |
| G.part.2 | 0.3927 | 0.1686 | 0.2085 | 0.2302 | **2.6304** | 11.2629 | 1.1085 | 0.7065 | **14.9356** | **1.0000** | **0.5120** | 1.6980 |
| G.part.3 | **0.3987** | **0.1563** | **0.2202** | **0.2249** | **1.4520** | 8.5121 | 0.7428 | 0.5943 | 9.9156 | 1.0000 | 0.3990 | **0.9460** |
| G.part.4 | 0.3909 | 0.1766 | 0.2256 | 0.2069 | 4.7363 | 13.9468 | 2.2234 | 2.5832 | 29.1633 | 1.0000 | 0.2070 | 0.3360 |

# **Chosen parameters for subtypes A, B, and G**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subtype | Freq.A | Freq.C | Freq.G | Freq.T | Ra | Rb | Rc | Rd | Re | Rf | p-inv | Gamma-shape |
| A | **0.3935** | **0.1708** | **0.2060** | **0.2297** | **2.9114** | **12.5112** | **1.2569** | **0.8559** | **12.9379** | **1.0000** | **0.5230** | **0.9** |
| B | **0.3857** | **0.1609** | **0.2234** | **0.2300** | **2.2228** | **10.7771** | **1.0675** | **1.2966** | **12.6824** | **1.0000** | **0.5** | **0.9410** |
| G | **0.3987** | **0.1563** | **0.2202** | **0.2249** | **1.4520** | **9.9166** | **1.3332** | **1.2652** | **14.9356** | **1.0000** | **0.5120** | **0.9460** |

We simulate a pool (30 sequences) of pol gene for eah subtype using a coalescent tree and parameters from the table above. We select randomly a subsequence for each subtype to text if really we maintain the pure strain for each subtype, and we found we have same strains (<http://bioafrica.mrc.ac.za/rega-genotype/genotype.php?cmd=list&job=2130285380>).

Even with the original sequence, better to fix the proportion of invariant sites at 80% to still getting the same strain.

# **Running Simpact**

# A. 4000 individuals(1800M&2200W), SEEDS: 40 @10

# Simpact version is: 0.21.0

# Current simulation time is 40.0002

# Number of events executed is 80998

# Started with 4000 people, ending with 8289 **FOR SCENARIOS 3 & 4**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Seed | 10 | 18 | 19 | 21 | **22** | **24** | **27** | 32 | 37 |
| Size | 6 >B | 20 >B | 3 >B | 3 >B | **1403 >A** | **716 >G** | **370 >A** | 20 >B | 5 >B |
|  |  |  |  |  |  |  |  |  |  |

# B. 8000 individuals(3800M&4200W), SEEDS: 40 @10

# Current simulation time is 40.0001

# Number of events executed is 171362

# Started with 8000 people, ending with 17017 (difference is 9017) **FOR SCENARIOS 3 & 4**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Seed | 4 | **12** | 19 | 21 | **22** | **24** | 28 | 34 | 39 |
| Size | 3 >B | **1079 >G** | 4 >B | 3 >B | **3032 >A** | **809 >G** | 3 >B | 3 >B | 4 >B |
|  |  |  |  |  |  |  |  |  |  |

**Important parameters when simulating sequences:** proportion invariant sites (to still have the same strain which is evolving, e.g.: @70% you still getting a HIV sequence but not recognizable which type, but a@80% it is the best), and scale of branch lengths (to have a realistic tree which when calibrated internal node ages are reasonable, e.g.: @1 internal nodes will be of age more than 400 years, at @0.0045 which is the value of substitution rate per year we have good results).

# **Define Transmission Network Parameters Estimated From HIV Sequences**

**References:**

* Andrew J. Leigh Brown et al., Transmission Network Parameters Estimated From HIV Sequences for a Nationwide Epidemic, 2011
* Gareth J. Hughes, et al., Molecular Phylodynamics of the Heterosexual HIV Epidemic in the United Kingdom, 2009
* Paraskevis D et al., Phylogenetic reconstruction of a known HIV-1 CRF04\_cpx transmission network using maximum likelihood and Bayesian methods, 2004

# **Big steps:**

* **Network creation:** the tree data shoule be processed using a custom R script utilizing the mrca function in the R package ape; a network with a time depth of 31 years, representing the maximum depth and maximum number of connections between the sequences, will be also created for comparison. Clusters will be identified using the clusters function within the R package igraph.
* **Network Shape Analysis:** the degree of a node (number of links per node) approximates the number of possible infected contacts an individual has had within the period of the network. We will use the statnet package within R to fit several models to the degree distribution of the nodes within our networks.
* **Network Structure:** The network structure was reconstructed using the time to the most recent common ancestors estimated for all individuals at a given time depth

# **Scenario 1**

* one subtype of the virus (HIV-1-A) for all seeds
* complete sampling for a transmission network of one seed
* same sampling time interval (e.g.: five or three years) for a transmission network for all seeds

## **Analysis 1.A:**

* tract SEED 22 (with 1403 individuals > subdivide in **subdivide in 20%=281, 30%=421, 40%=561, 50%=701, 60%=842, 70%=982, 80%=1122, 90%=1263, 100%=1403**) for a “random” complete sampling for a transmission network of one seed and look on effect of incomplete sequence *coverage* on the transmission network by comparing the true transmission network with one constructed from phylogenetic tree: randomize the 3032 sequences, subdivide them according %, construct phylogenetic trees, reconstruct transmission networks, compare with true transmission networks of same individuals > get a table of results.

## **Analysis 1.B:**

* tract SEED 22 (with **3032 individuals >**  **20%=606**, **30%=910**, **40%=1213**, **50%=1516**, **60%=1819**, **70%=2122**, **80%=2426**, **90%=2729**, **100%=3032**) for a “random” complete sampling for a transmission network of one seed and look on effect of incomplete sequence coverage on the transmission network by comparing the true transmission network with one constructed from phylogenetic tree: randomize the 1403 sequences, subdivide them according %, construct phylogenetic trees, reconstruct transmission networks, compare with true transmission networks of same individuals > get a table of results.

# **Scenario 2**

* one subtype of the virus (HIV-1-A) for all seeds
* complete sampling for a transmission network of all seeds
* same sampling time interval (e.g.: five or three years) for a transmission network of all seeds

## **Analysis 2.A:**

* cross-sectional sampling for all seeds (10, 18, 19, 21, 22, 24, 27, 32, 37) in a period of past 7 years from the end of simulation backward (we got sequences for seeds 22, 24, and 27), rename Ids to avoid duplication due to the fact each transmission network has Ids named in ordinal order; hence, put prefix of the seed on these Ids. Main files: combined sequences, and their sampling times. Overall we got 1098 sequences (randomize),  **subdivide in 20%=220, 30%=330, 40%=440, 50%=549, 60%=659, 70%=769, 80%=878, 90%=988, 100%=1098**

## **Analysis 2.B:**

* cross-sectional sampling for all seeds (10, 18, 19, 21, 22, 24, 27, 32, 37) in a period of past 7 years from the end of simulation backward (we got sequences for seeds 22, 24, and 27), rename Ids to avoid duplication due to the fact each transmission network has Ids named in ordinal order; hence, put prefix of the seed on these Ids. Main files: combined sequences, and their sampling times. Overall we got 1098 sequences (randomize),  **subdivide in 20%=381, 30%=571, 40%=762, 50%=952, 60%=1142, 70%=1333, 80%=1523, 90%=1714, 100%=1904**

# **Scenario 3**

* different subtypes of the virus (HIV-1-A-B-G) for all seeds
* complete sampling for a transmission network of one seed
* same sampling time interval (e.g.: five or three years) for a transmission network of one seed

## **Analysis 3.A:**

* cross-sectional sampling for all seeds (10, 18, 19, 21, 22, 24, 27, 32, 37) in a period of past 7 years from the end of simulation backward (we got sequences for seeds 22-**A**, 24-**G**, and 27-**A**), rename Ids to avoid duplication due to the fact each transmission network has Ids named in ordinal order; hence, put prefix of the seed and strain subtype on these Ids. Main files: combined sequences, and their sampling times. Overall we got 1098 sequences (randomize since they are not same clade subtype),  **subdivide in 20%=220, 30%=330, 40%=440, 50%=549, 60%=659, 70%=769, 80%=878, 90%=988, 100%=1098**

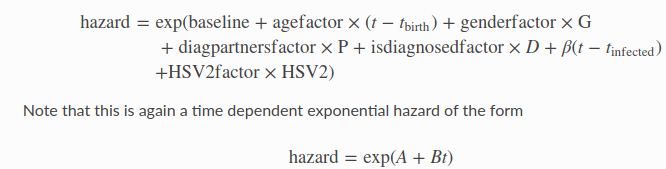
# **Scenario 4**

* different subtypes of the virus (HIV-1-A-B-G) for all seeds
* complete sampling for a transmission network of all seeds
* same sampling time interval (e.g.: five or three years) for a transmission network of all seeds

# **Sequence coverage and missingness mechanisms: Diagnosis event**

When a person gets infected with HIV, either by transmission of the virus or by seeding the population to get the epidemic started, a diagnosis event will get scheduled. When fired, the person is deemed to feel bad enough to go to a doctor and get diagnosed as being HIV-infected. Upon diagnosis, a monitoring event will be scheduled very shortly afterwards, to monitor the progression of the disease and to offer treatment if eligible.

This event is hazard-based, and the hazard is of the following form:



Here is an overview of the relevant configuration options, their defaults (between parentheses), and their meaning:

* **diagnosis.baseline (0)**: Controls the corresponding baselinebaseline value in the expression for the hazard.
* **diagnosis.agefactor (0)**: Controls the corresponding agefactoragefactor value in the expression for the hazard. This allows one to let the age of a person influence the hazard.
* **diagnosis.genderfactor (0)**: Controls the genderfactorgenderfactor parameter in the hazard. This allows you to have a different hazard depending on the gender of the person.
* **diagnosis.diagpartnersfactor (0)**: Corresponds to the value of diagpartnersfactor in the expression for the hazard. The idea is to allow the number of partners that have already been diagnosed to have an effect on a person’s diagnosis time: if a person is not feeling well and knows that some of the partners are infected with HIV, this can be an incentive to go to the doctor sooner.
* **diagnosis.isdiagnosedfactor (0)**: Using this isdiagnosedfactor value in the hazard, it is possible to have a different hazard if the person was diagnosed before. After dropping out of treatment, for example because a person is feeling better and no longer feels the need for treatment, a diagnosis event will be scheduled again. It is reasonable to think that a person may go to the doctor again sooner when he already knows about the HIV infection.
* **diagnosis.beta (0)**: Corresponds to the **β** factor in the hazard expression, allowing one to take the time since infection into account.
* **diagnosis.HSV2factor (0)**: Using the HSV2factorHSV2factor, it is possible to have a different hazard when the person is infected with HSV2.
* **diagnosis.t\_max (200)**: As explained in the section about ‘time limited’ hazards, an exponential function needs some kind of threshold value (after which it stays constant) to be able to perform the necessary calculations. This configuration value is a measure of this threshold.

**For many runs of simulations, just simulate sequences of diagnosed individuals, this means to drop tips of non-diagnosed individuals in the transmission tree. In addition we can also drop tips of individuals who cause negative branch lengths (due to transmission after diagnosis), but we should show that the likelihood of transmitting an infection after being diagnosed is negligeable.**

**Special case: simulate 3 clades, determine sampling interval, depertmine probability of sampling regarding gender, and age; ~~and subdivide the sampling interval in three and the increase sampling from past to present~~.**

Chose a given sequence coverage for a cross-sectional sampling and study the effect of gender and age on transmission network.

Let say we chose 70% of sequence coverage, if we have N individuals for the complete coverage, this means we will have 70%\*N = S in our sample.

1. Gender: Women and men, assign probability of gender to the sampling process, for example within S individuals, women will be 70% and men 30%.
2. Age: subdivide in three groups: young adults, adults, and old; assign probability of sampling for each group, for example gp1-60%, gp2-30%, and gp3-10%
3. Gender and Age: for the women which are 70% of the sample, they are distributed in the three groups, the same for men which are 30% of the sample, for example W.gp1-50%, W.gp2-20%, and W.gp3-30%; and for men we will have M.gp1-60%, M.gp2-10%, and M.gp3-40%.