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

Run simulation with 5000 individuals and 8000 individuals, per each run consider different cases as defined below.

# **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.

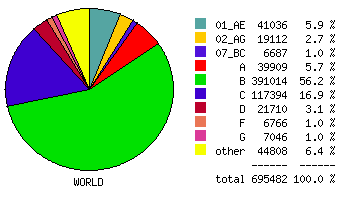
## **Subtype B**

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

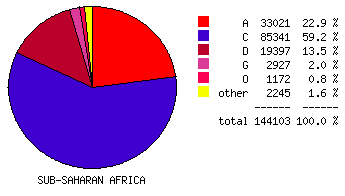
(one sequence for subtype B)

## **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**

# **Scenario 1**

* one subtype of the virus (HIV-1-C) 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

# **Scenario 2**

* one subtype of the virus (HIV-1-C) 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

# **Scenario 3**

* different subtypes of the virus (HIV-1-A-B-C) 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

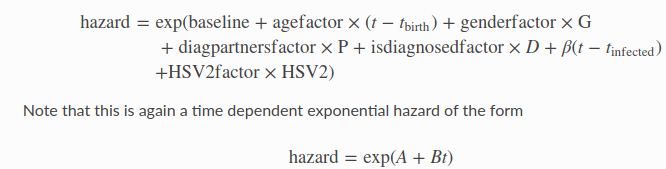
# **Scenario 4**

* different subtypes of the virus (HIV-1-A-B-C) 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.