Validation of integration date estimation methods using a simulation of HIV latent genomes

Bradley R. Jones and Jeffrey B. Joy

University of British Columbia; BC Centre for Excellence in HIV/AIDS

I have no conflicts of interest

CAHR 2020, BSP6.01

Correspondence: bjones@cfenet.ubc.ca

UBC

NIH National Institutes of Health

BRITISH COLUMBIA CENTRE *for* EXCELLENCE *in* HIV/AIDS

Providence HEALTH CARE How you want to be treated.

BRITISH COLUMBIA Ministry of Health

Background

- **The HIV persistent reservoir is the main barrier to an effective HIV cure**
- However, many features of the persistent reservoir are uncertain such as the timing of proviral integration into the reservoir
- Sophisticated phylodynamic tools are necessary to reveal the secrets of the persistent reservoir, but these tools require validation
- A common method to validate phylodynamic tools is with *in silico* computer simulation
- However, there does not exist an *in silico* tool for simulating HIV genomes within host incorporating the persistent reservoir
- **We present software that simulates HIV within host genome evolution including the persistent reservoir**
- **We use this software to compare date estimation methods for proviral integration dates in the persistent reservoir**

Simulation model

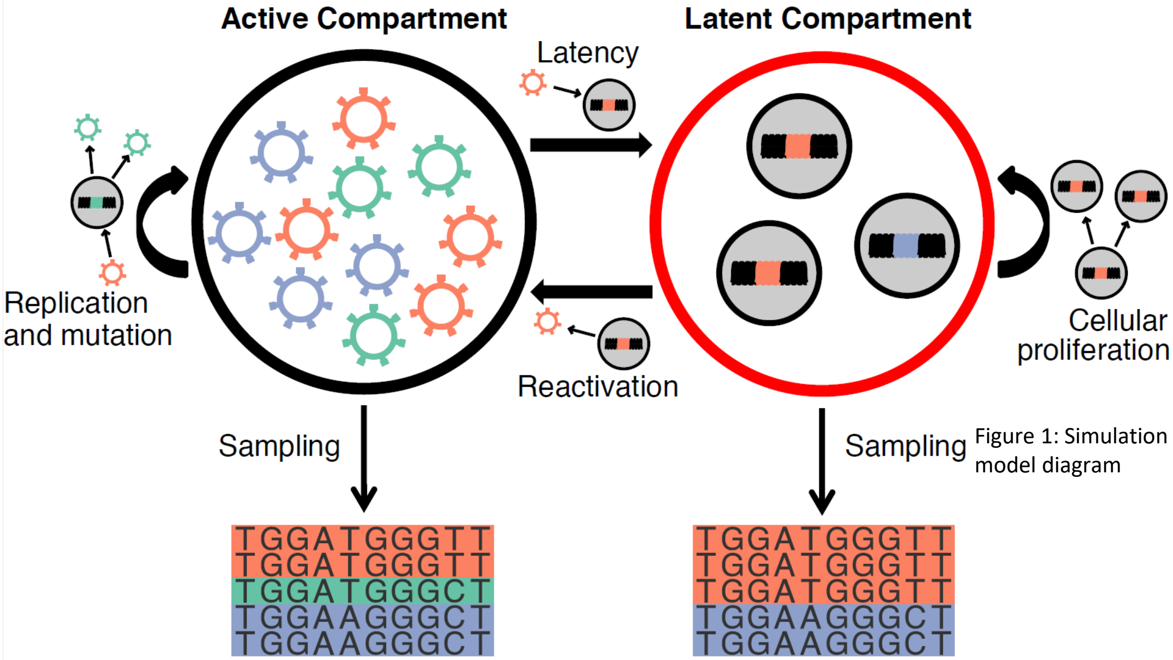


Figure 1: Simulation model diagram

- We extended the Java software, SANTA-SIM, to include multiple compartments
- **In our modified SANTA-SIM, we created a model of HIV within host evolution incorporating an active compartment where genomes can mutate and replicate and a latent compartment where genomes can proliferate, but cannot mutate (Figure 1)**
- Fitness was modelled using associated mutation to human leukocyte antigens
- Viral genomes from the active compartment can become latent moving to the latent compartment and viral genomes in the latent compartment can reactivate moving to the active compartment

Estimating sequences ages

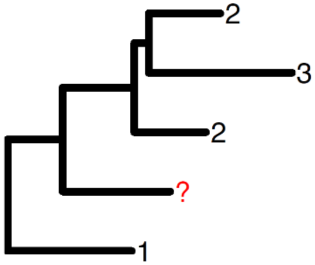
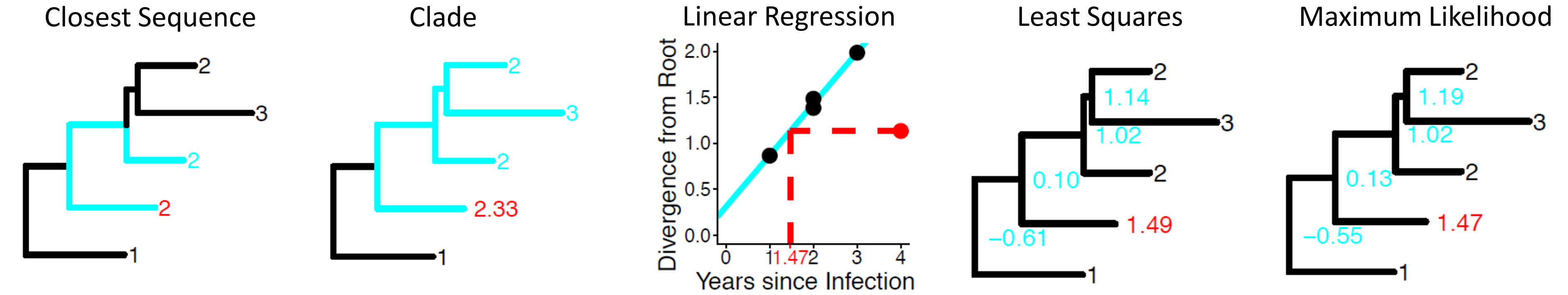


Figure 2: Diagram of input data. We start with a phylogeny and the dates of active genomes (black) and attempt to estimate the integration dates of the unknown active genomes (red)

As an application of our simulation, we compared the ability of different phylodynamic date estimation methods to recover the integration dates of proviral genomes in the persistent reservoir. Each method starts with a phylogeny of the genomes and the collection dates of active genomes to estimate the dates of the latent genomes (Figures 2 and 3).

Figure 3: Date estimation methods


Closest Sequence

Clade

Linear Regression

Least Squares

Maximum Likelihood

In the Closest Sequence (CS) method, we find the closest active sequence via patristic distance (teal) to the query sequence (red) and assign its date

In the Clade (CD) method, we find the smallest subtree (teal) that contains the query sequence and at least one active sequence and assign the mean date of the sequences in the clade

In the Linear Regression (LR) method, we infer a linear regression (teal) between the collection date and divergence of the active sequences and using this regression we estimate the query date

In the Least Squares (LS) method, we select dates for the internal nodes (teal) and the query sequences (red) to minimize the square difference between the dates and the branch lengths

In the Maximum Likelihood (ML) method, we select dates for the internal nodes (teal) and the query sequences (red) to maximize their likelihood

Comparing the methods

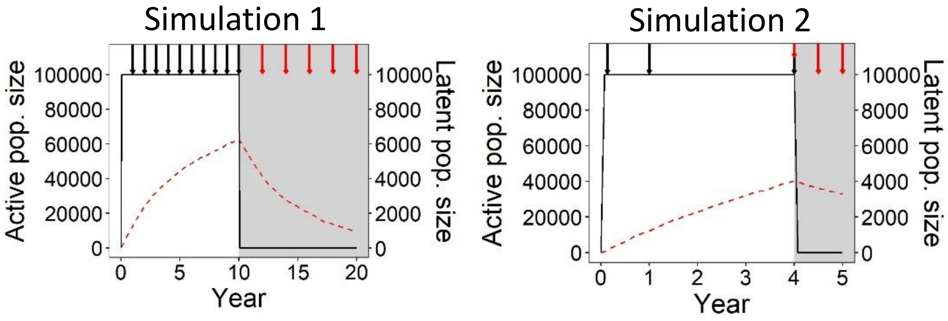


Figure 4: Simulation genome population sizes (solid black line (active), dotted red line (latent)), sampling time points (black arrows (active), red arrows (latent)) and cART (grey shading)

- **Two types of simulations were performed (Simulation 1 and Simulation 2) with different sampling schemes and infection periods (Figure 4)**
- 100 replicates of each type of simulation were performed for 200 total simulated data sets
- We inferred maximum likelihood phylogenies using RAxML from each data set and applied each date estimation method
- **Least Squares was the most accurate in Simulation 1: lower root mean squared error (RMSE) and higher concordance** (all Friedman and all paired t-tests: p<0.01; Figure 5A-B)
- **Linear Regression was the most accurate in Simulation 2: lower RMSE and higher concordance** (all Friedman and all paired t-tests (except LR versus LS): p < 0.01; Figure 5C-D)
- **RSME and concordances of the estimates of Least Squares and Linear Regression were not significantly different in Simulation 2** (paired t-test: p = 0.74 and p = 0.26 respectively; Figure 5C-D)

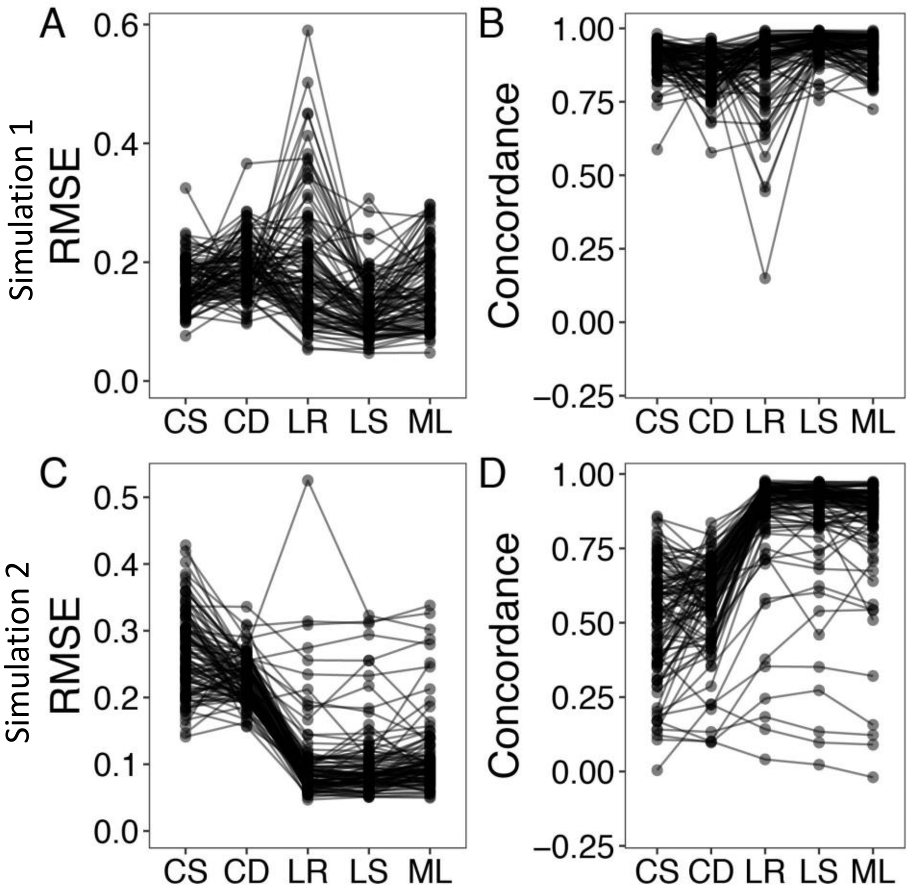


Figure 5: Root mean squared error (RMSE) and concordance between the estimated date and actual integration date of the sampled genomes for each date estimation method. Each dot represents one simulation (n=100) and lines connect the same simulation across methods. CS: Closest Sequence, CD: Clade, LR: Linear Regression, LS: Least Squares, ML: Maximum Likelihood

Comparison with empirical data

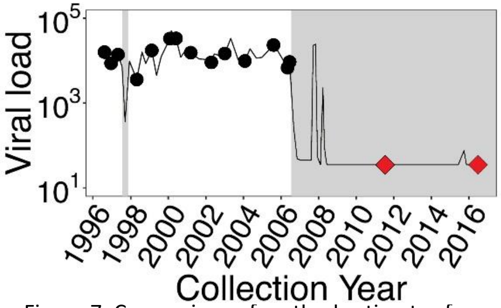
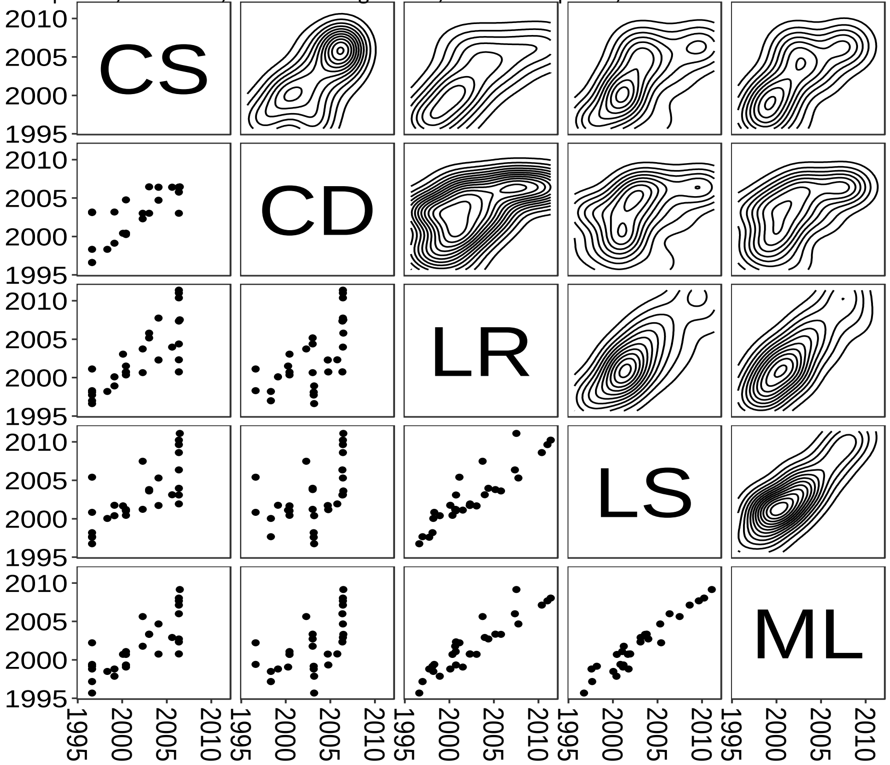


Figure 6: Plasma viral load (black line), sampling history (black dots (plasma) and red diamonds (provirus)) and cART (grey shading) of participant

Figure 7: Comparison of method estimates for each proviral sequence. CS: Closest Sequence, CD: Clade, LR: Linear Regression, LS: Least Squares, ML: Maximum Likelihood



- We collected HIV RNA *nef* sequences from plasma and HIV DNA *nef* sequences from an HIV infected individual (Figure 6)
- We inferred a maximum likelihood phylogeny from the sequences using RAxML and applied each date estimation method
- **Linear Regression, Least Squares and Maximum Likelihood had the most similar estimates with concordances ranging from 0.89-0.92** (Figure 7)
- The least concordant methods were Maximum Likelihood and Clade (concordance: 0.48), but the estimates of these methods were significantly correlated (Pearson correlation coefficient: 0.48, $p < 0.01$; Figure 7)

Conclusions and next steps

- **Least Squares was the best at recovering the integration dates with the lowest RMSEs and highest concordances overall**
- Our software is an valuable tool to validate phylodynamic software used in the context of the HIV persistent reservoir
- Next we plan to use simulation model to estimate parameters of the persistent reservoir by comparing to empirical data