HIV-phyloTSI: IAVI workshop

Andrea Brizzi

2023-11-12

Code and slides

Intrahost diversity and recency of infections

Code and slides

Why do we care about recency?

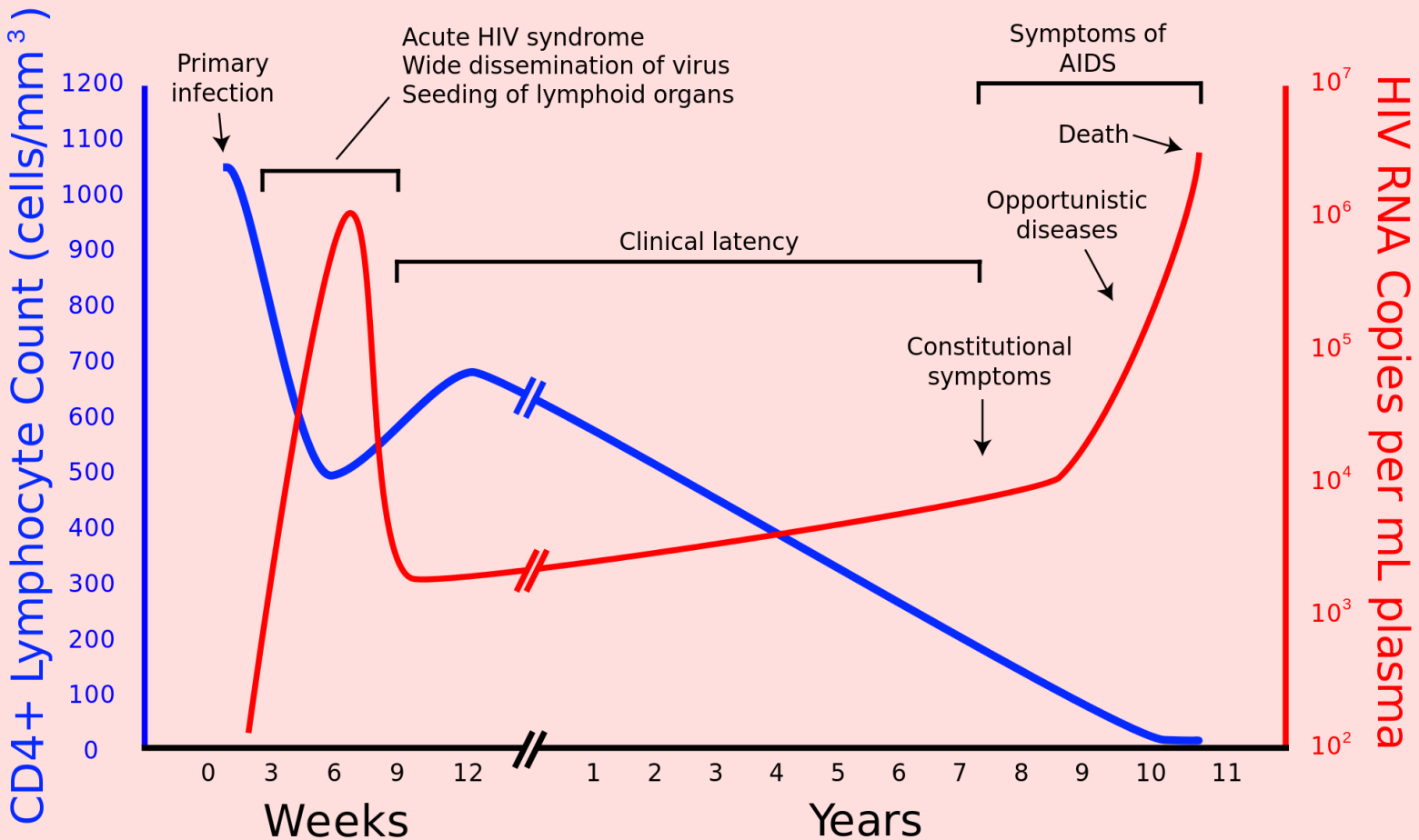
- incidence estimation
- identifying sub-populations suffering from recent infections are happening.
- determining generation intervals.

but getting data is hard:

- longitudinal cohorts as gold standard
- alternatives which are discussed in this section

Code and slides

What is recency? Natural history of HIV



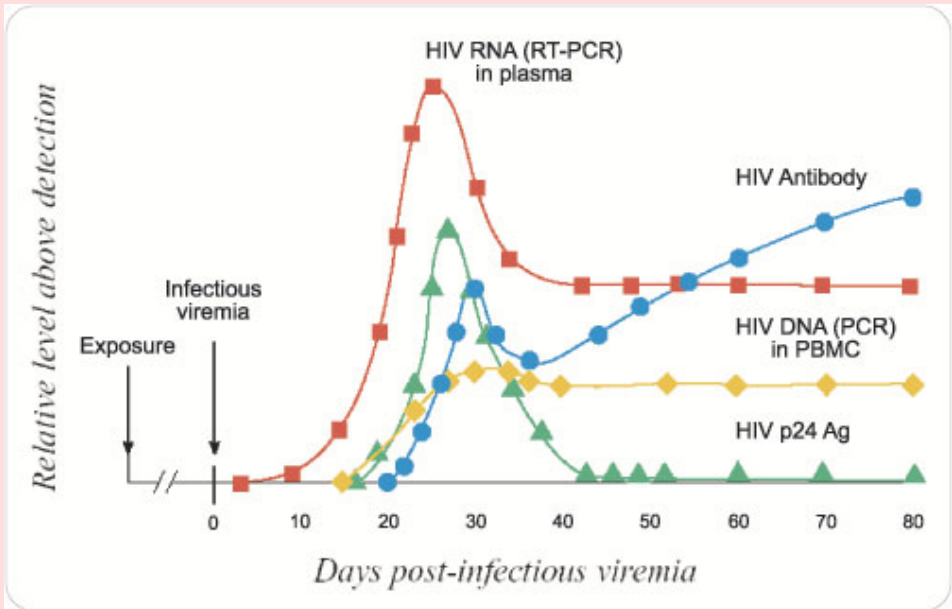
dynamics define infection stages: (CDC: 2014) ; (Fiebig et al. 2003)

Code and slides

Lab assays to identify infection stages

Based on the quality and type of immune response.

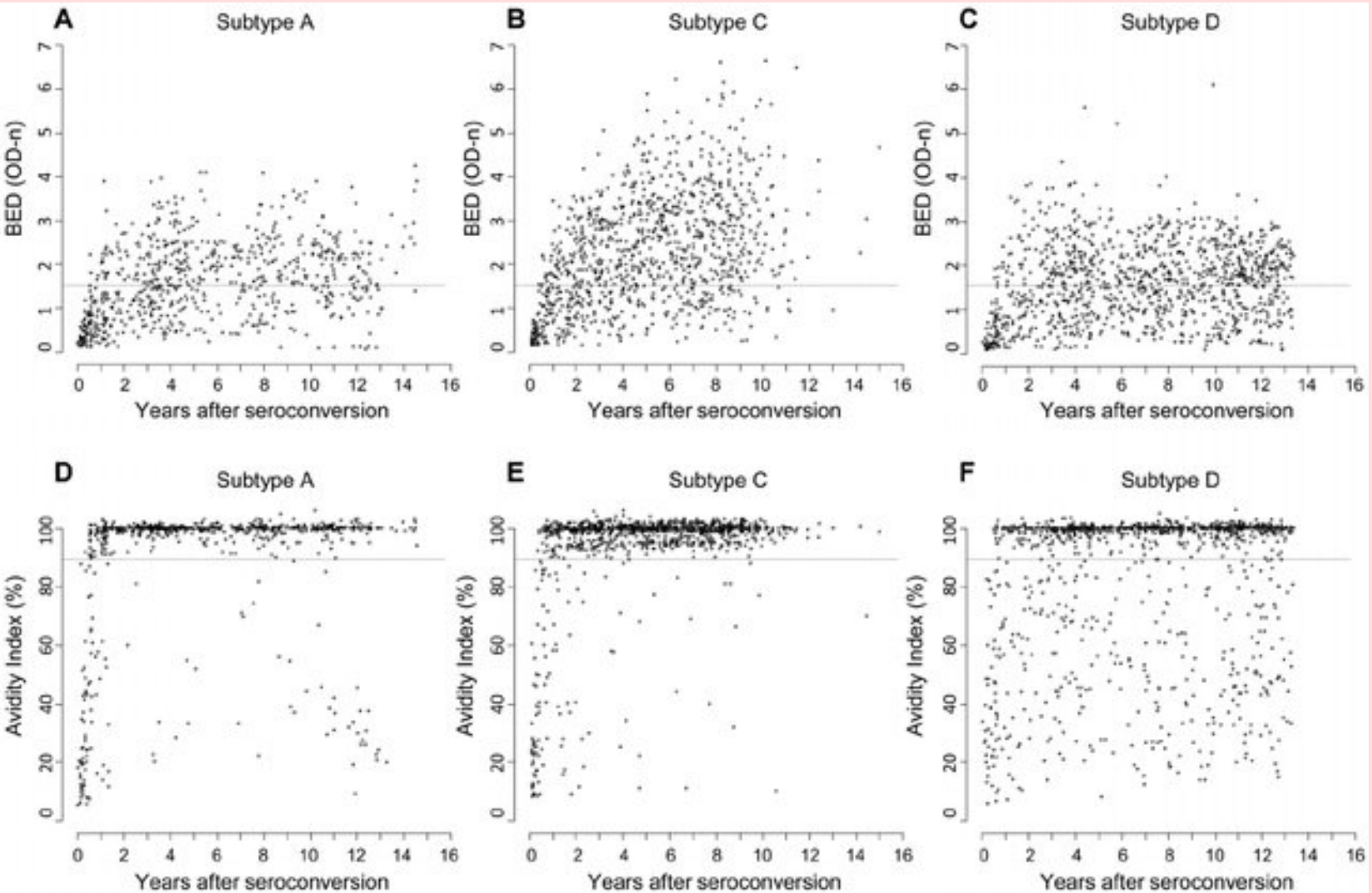
- **Less sensitive Enzyme Immunoassays**, antigen test (1998)
- **Lag-Avidity** (2012)
- **Bed-CEIA** (2011)



Code and slides

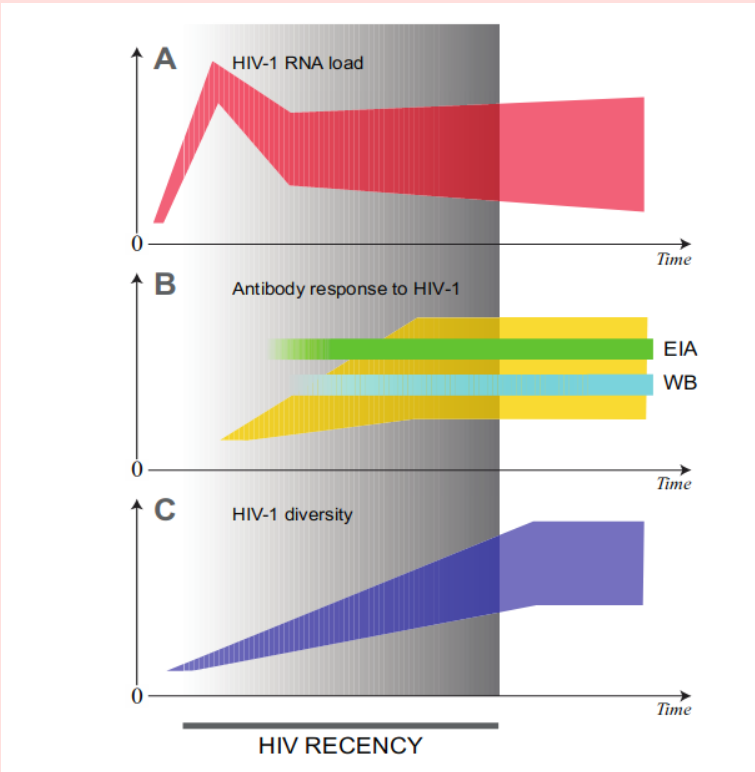
This is great, but specificity and sensitivity are hard to control:

- individual level variation, confounding subtypes, ART status…



Code and slides

An alternative unit of time: genetic diversity



- HIV infection is a multi-faceted phenomenon: immune response is just one facet.
- HIV-1 diversity grows as a function on time since infection.
- NGS is more informative than Sanger (Carlisle et al. (2019))

Code and slides

Case study: recency in Botswana

Ragonnet-Cronin et al. (2022) developed recency classifier

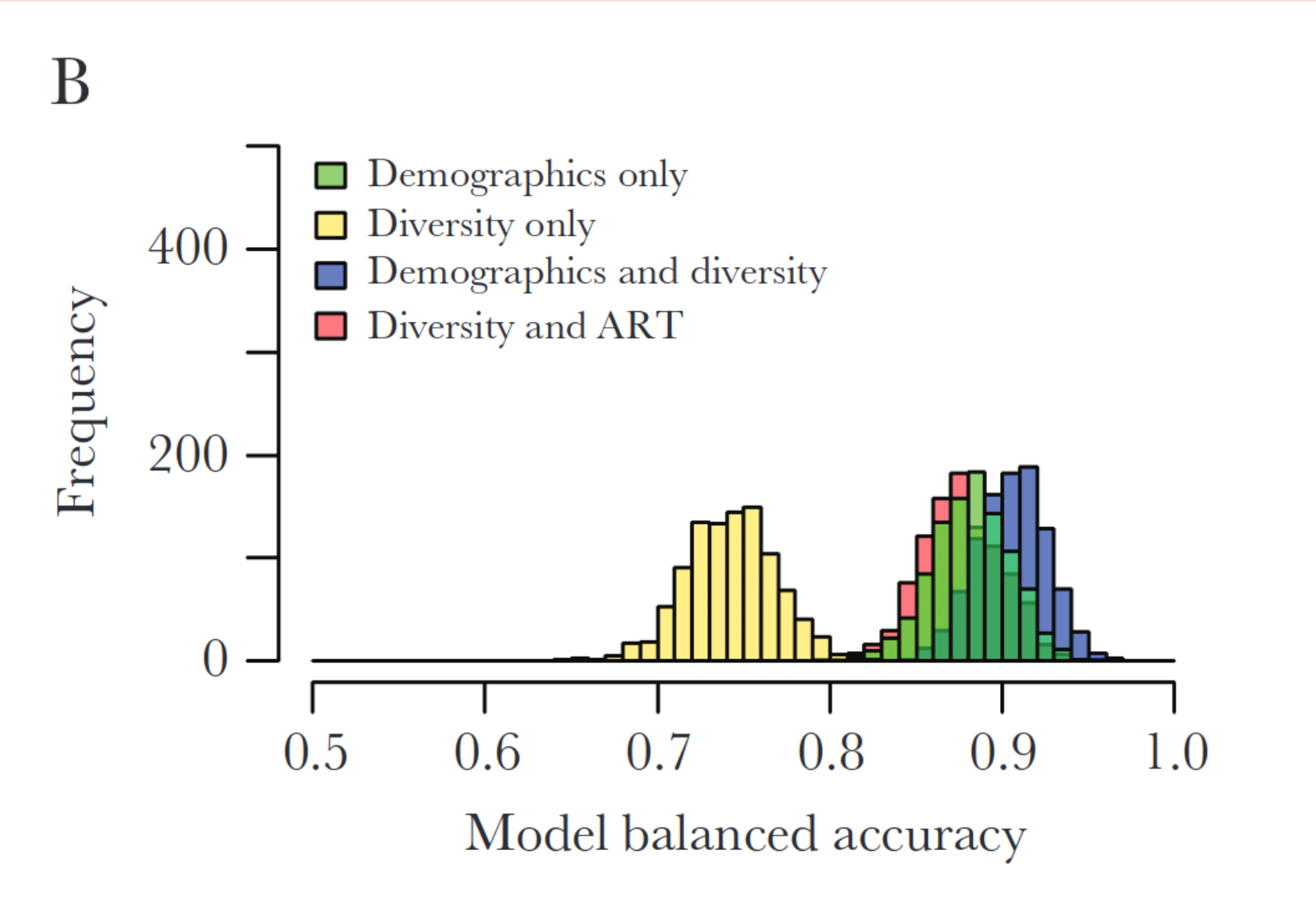
- **end goal**: characterizing undiagnosed cases in Botswana (Bhebhe et al. (2022))
- **data**: demographic data, some testing data and individual level NGS
- **how**: train ML model on known recent and known chronic infections, and use it to classify unknown.

comments on model:

1. still required cohort study, and *not directly generalisable* to other contexts.
2. *simple measures of diversity*: entropy and mean pairwise distance
3. these alone are *not enough* to provide good classification

Code and slides

Case study: recency in Botswana



Code and slides

HIV-phylo TSI

Golubchik et al. (2022) tackles weaknesses of previous approach:

1. **Pre-trained**: no need for expensive training data.
2. Captures more complex **measures of divergence**.
3. Comparatively **low false recency rate** using exclusively NGS data.

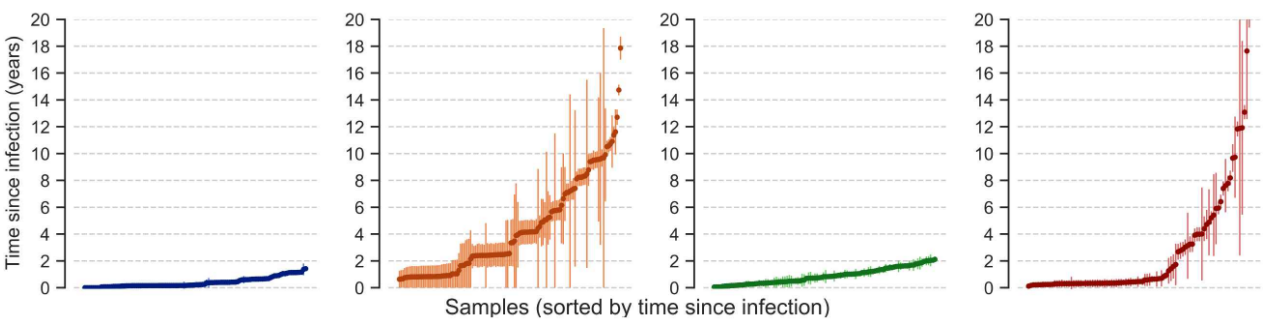
Code and slides

HIV-phylo TSI: training data

sequences from PANGEA and BEEHIVE with known infection range.

	UW Partners in Prevention (UWP)	Rakai CCS (RAK)	BEEHIVE (BEE)	MRC Uganda (MRC)
Participants	108	132	104	94
Samples	150	132	104	94
Age, median (range)	33.3 (19.0 - 55.1)	32.0 (17.0 - 50.0)	32.9 (20.4 - 62.8)	29.0 (17.0 - 75.4)
Proportion female	0.43	0.49	0.03	0.45
Years since infection, mean (range)	0.4 (0.0 - 1.4)	2.9 (0.6 - 17.9)	0.8 (0.1 - 2.1)	0.5 (0.1 - 21.0)
Predominant subtypes	A1, D, C	D, A1, A1D	B	D, A1, B
log10 VL, mean (range)	4.9 (3.4 - 6.8)	4.8 (3.2 - 6.0)	4.6 (3.1 - 5.6)	4.4 (2.9 - 5.2)
CD4 count, mean (range)	n/a	407 (15 - 1626)	501 (140 - 1062)	639 (197 - 1597)
Proportion of infections: <=12 months	0.83	0.23	0.59	0.63
<=18 months	1.0	0.27	0.79	0.65

Duration of infection (seroconversion interval)

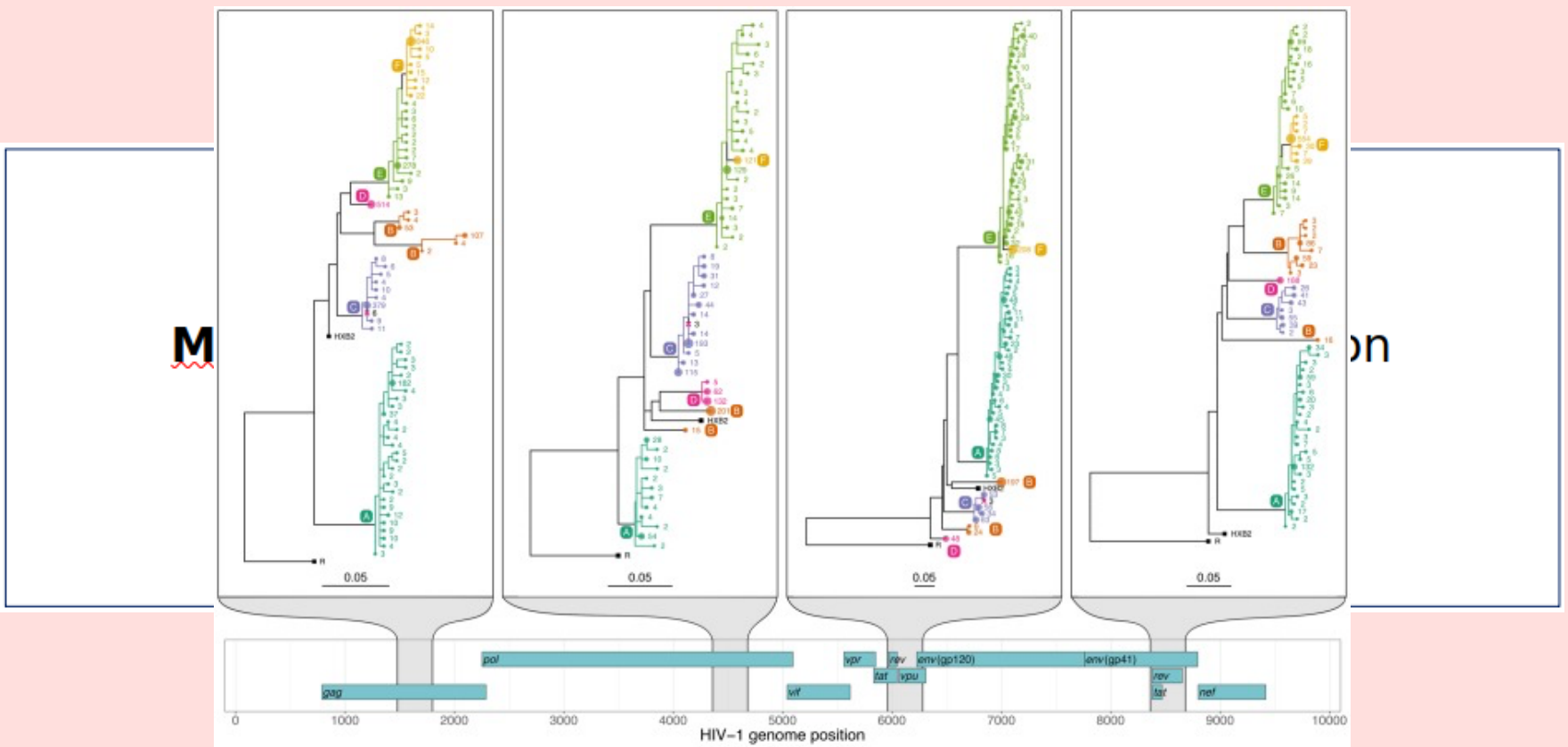


Code and slides

HIV-phylo TSI: intrahost diversity

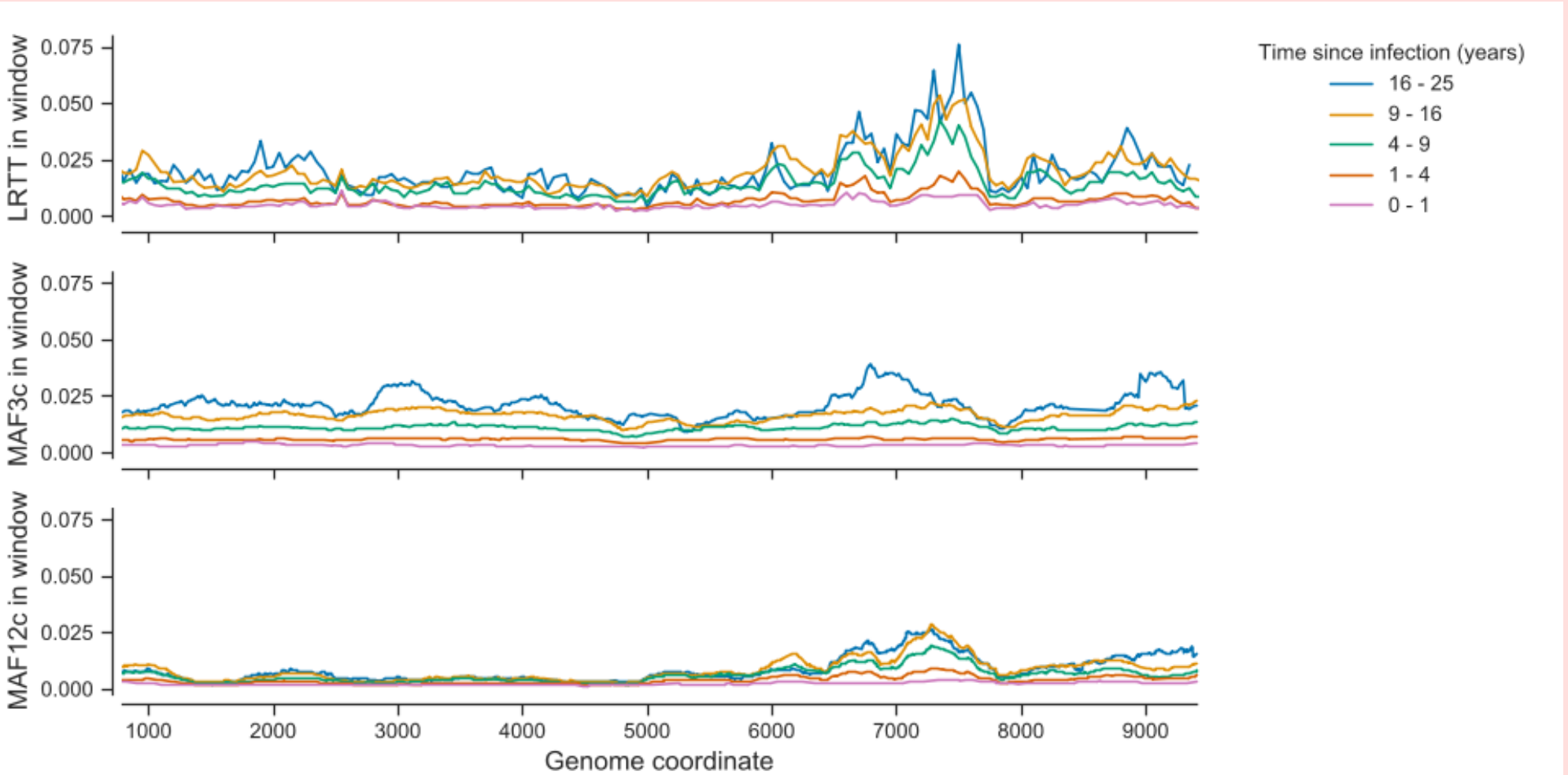
choice of predictors based on viral diversification following infection.

1. “Accumulated” mutations on codons.
2. “Depth” of host-specific subtree in phylogeny. (LRTT)



Code and slides

But why these measures of intrahost diversity?



Code and slides

HIV-phylo TSI model

different signal in different part of the genome. How to combine predictors?

- take average over gene
- let the model pick up the informative covariates.

- Best performing set of features chosen through LOOCV
- Regressor performs well as a classifier too, but FRR (~10%)

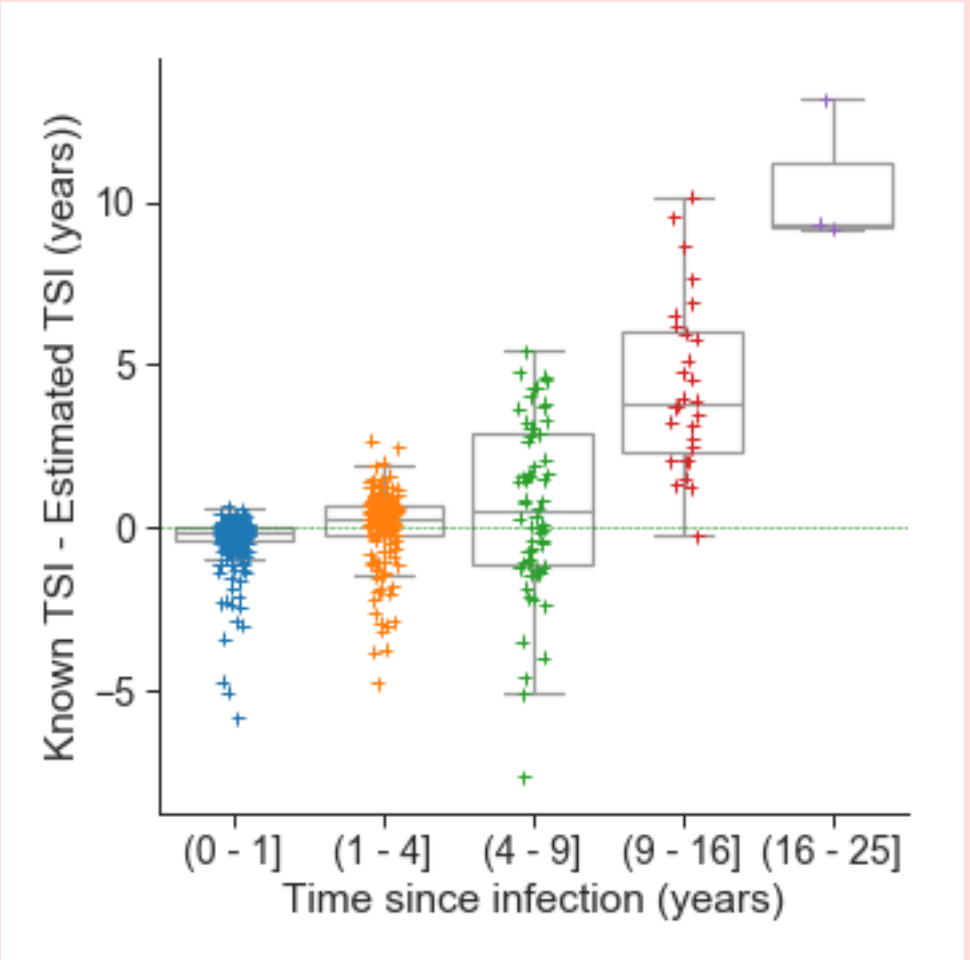
Code and slides

HIV-phylo TSI performances: bias

- bias is low in infections < 9 years.
- range of individual-level errors can be quite wide.

For applications:

- estimates are not precise at the individuals level BUT
- there is signal in medians taken among subpopulation.

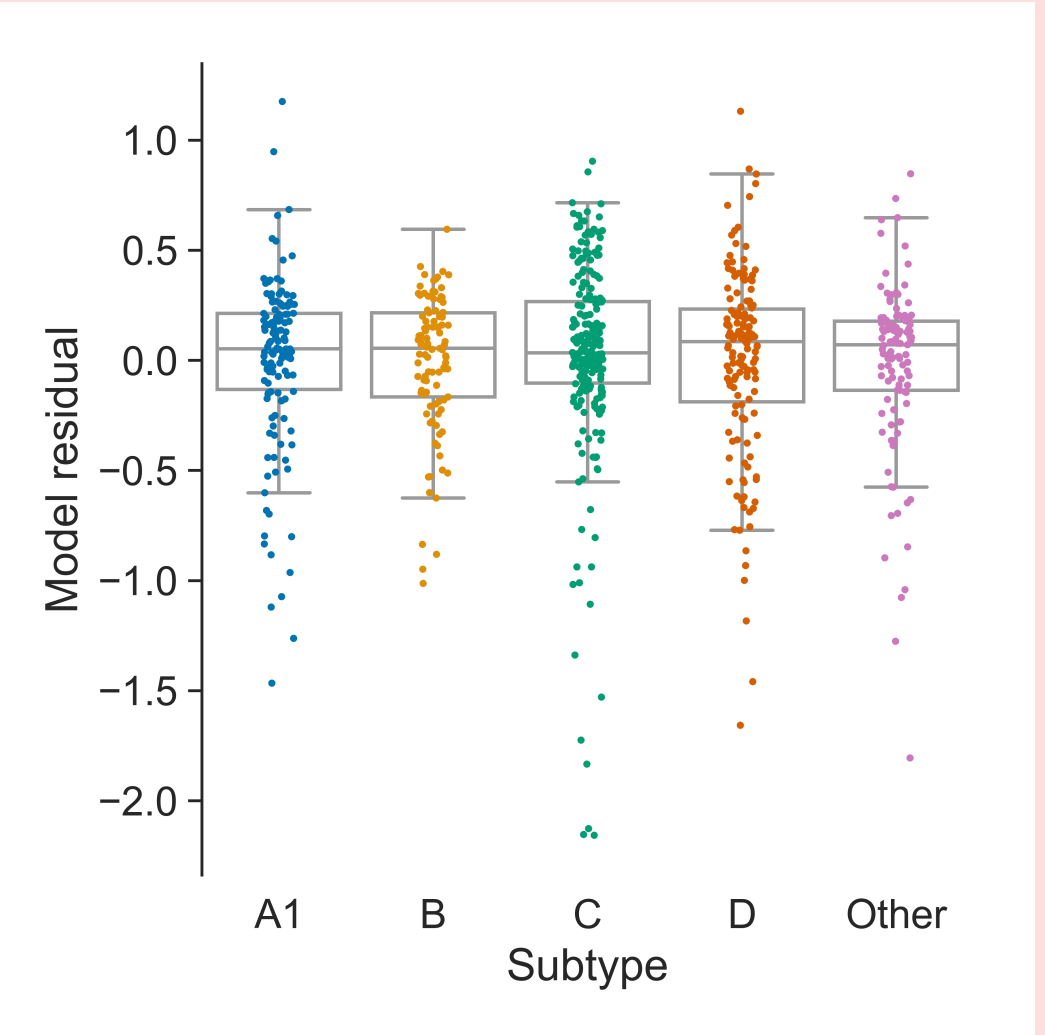


Code and slides

HIV-phyloTSI performances: robustness to subtype

Code and slides

model bias was close to 0 for all subtypes included in the dataset.

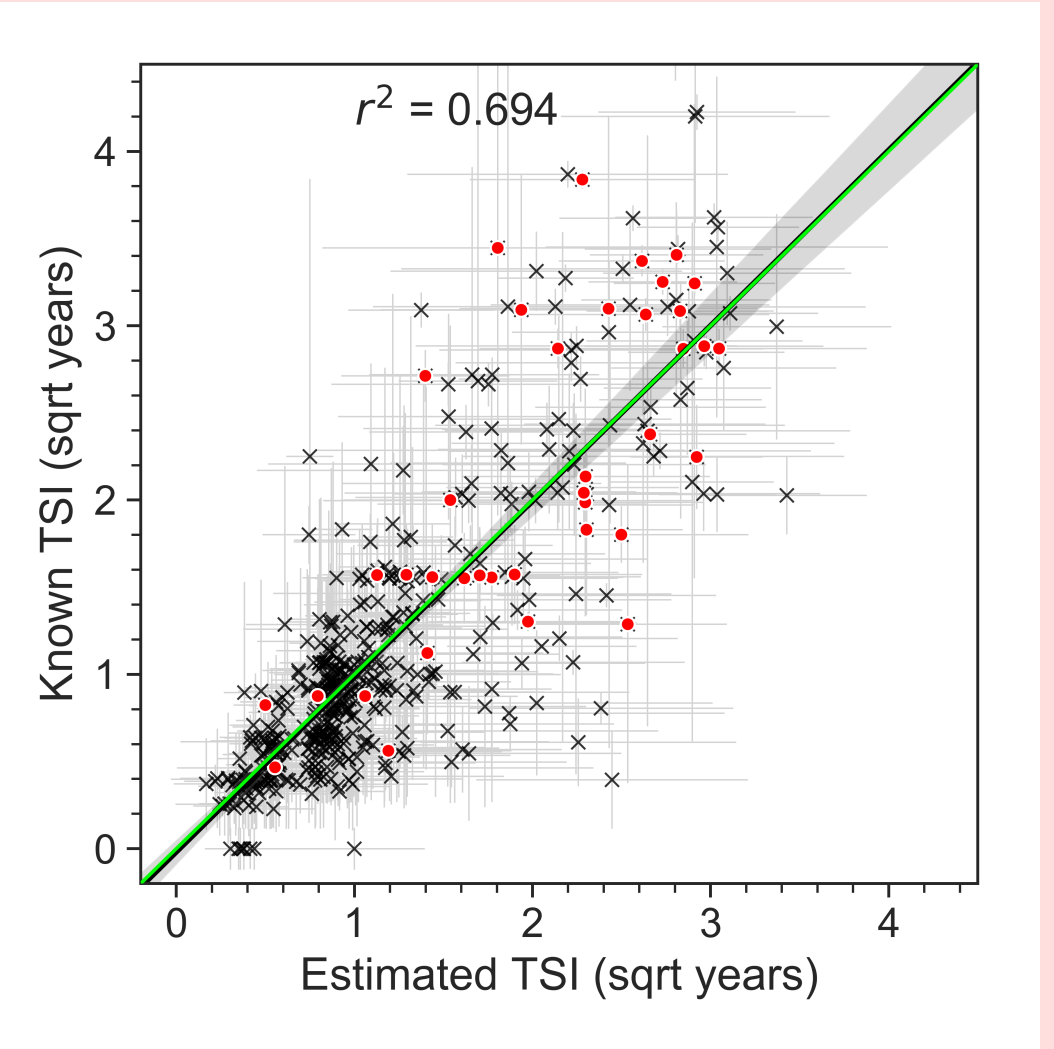


Code and slides

HIV-phyloTSI performances: robustness to ART (?)

Some individuals in the training data reported prior ART use.

- followed pattern
- BUT: small sample size
- AND: no predictions for suppressed individuals.



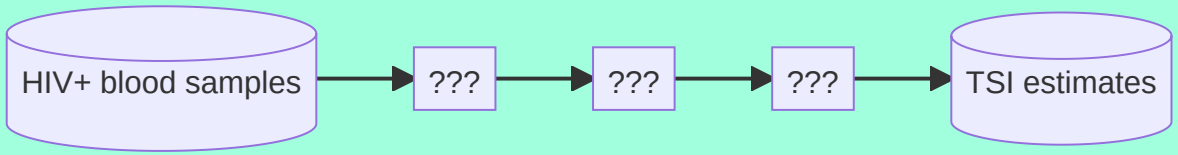
Code and slides

Code and slides

HIV-PhyloTSI bioinformatics pipeline

Code and slides

Demistifying the process



- In this section, want to **describe** processes and **data manipulation** carried out by the pipeline.
- **Not** the actual **implementation** of the pipeline.

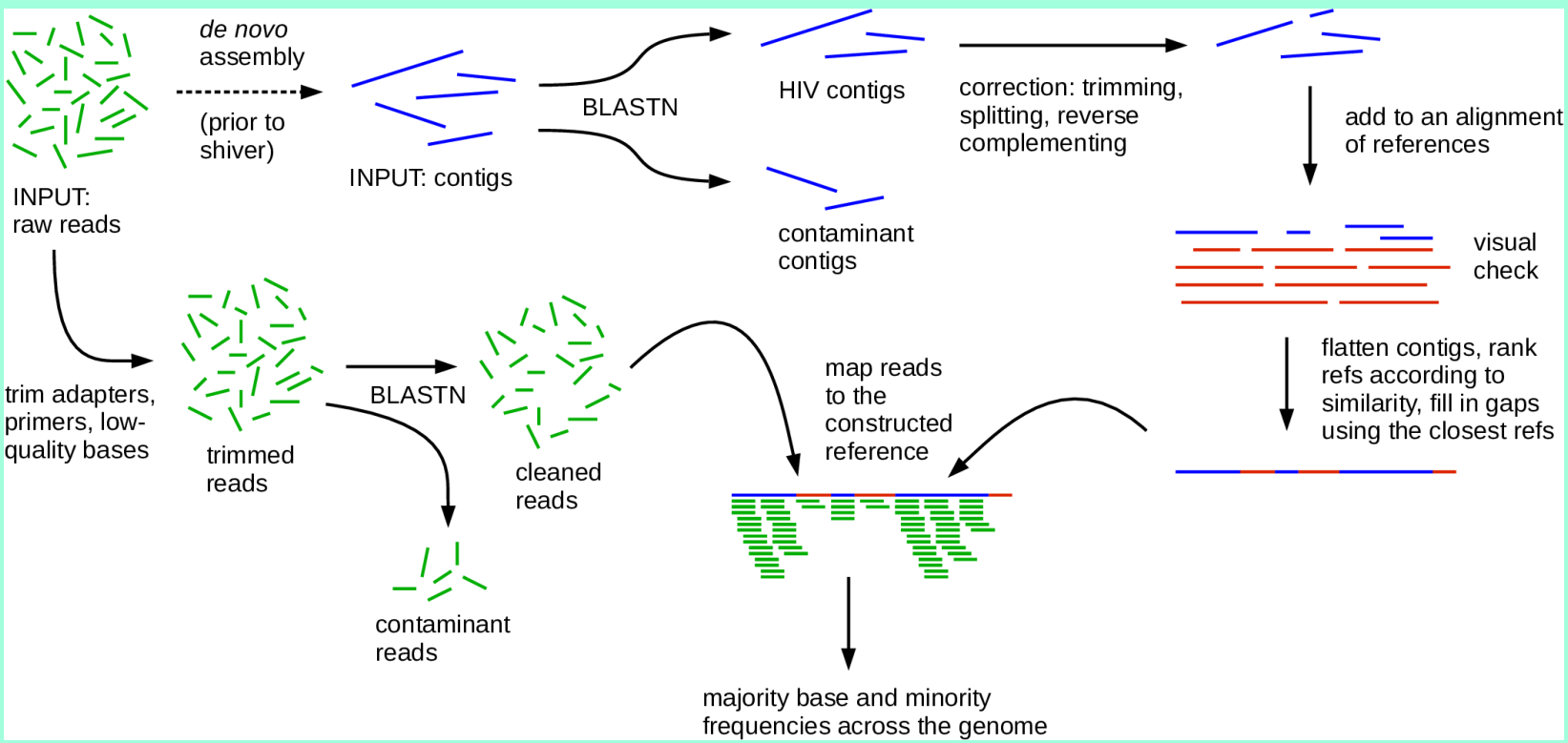
Main steps:

- Next Generation Sequencing (NGS)
- Constructing phylogenies **MANY!**
- Analysing phylogenies

Code and slides

Next Generation Sequencing

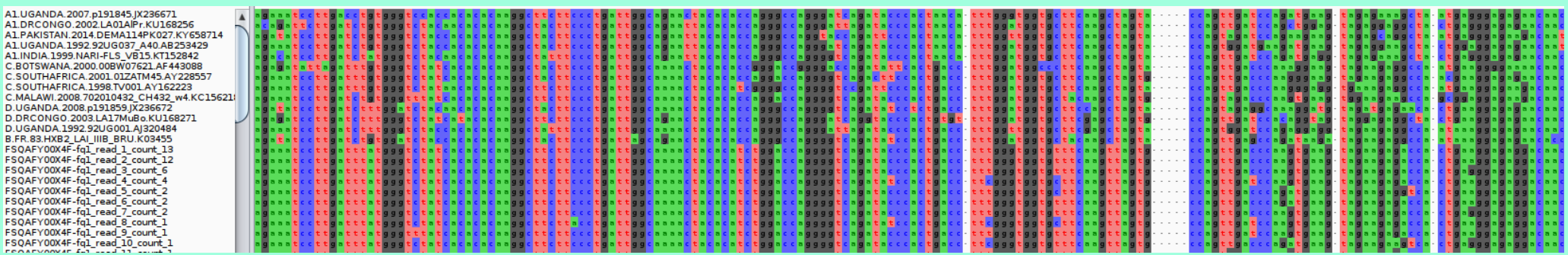
- simultaneous sequencing of multiple viral particles to capture within-host diversity
- PANGEA protocols: I (2012) and II (2018)
- TSI pipeline starts with the `*.bam`, `*_ref.fasta` produced by `shiver`



Code and slides

Grouping & multiple sequence alignment

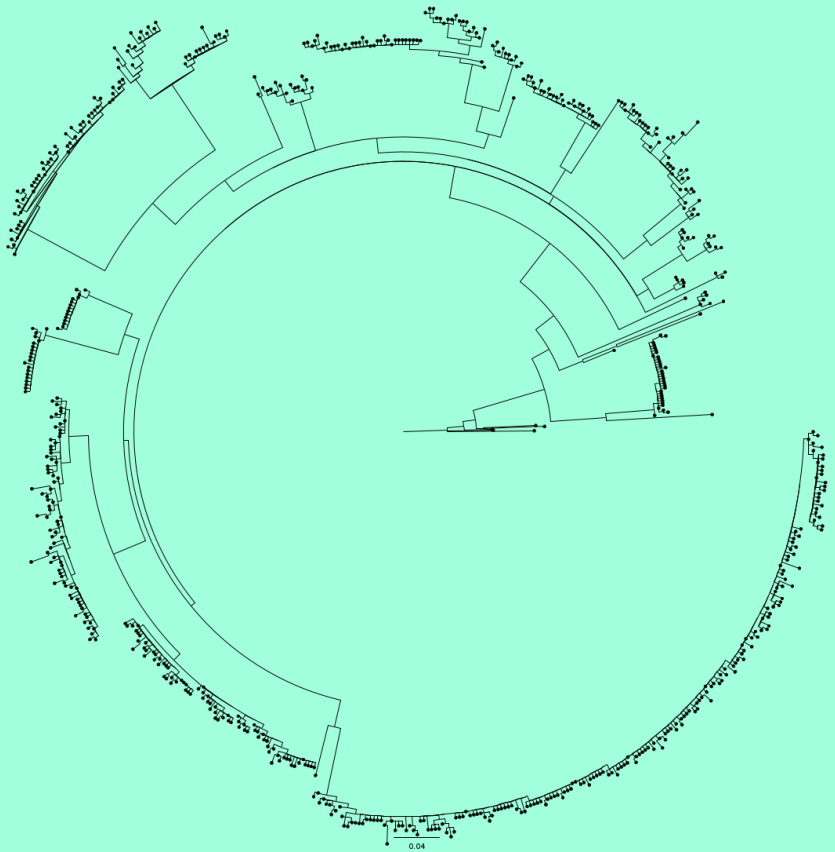
- Many sequences, many hosts, many genome windows -> **group by host!**
- aligning to compare “apples to apples”: MAFFT (Katoh 2002)



Code and slides

Inferring phylogenies

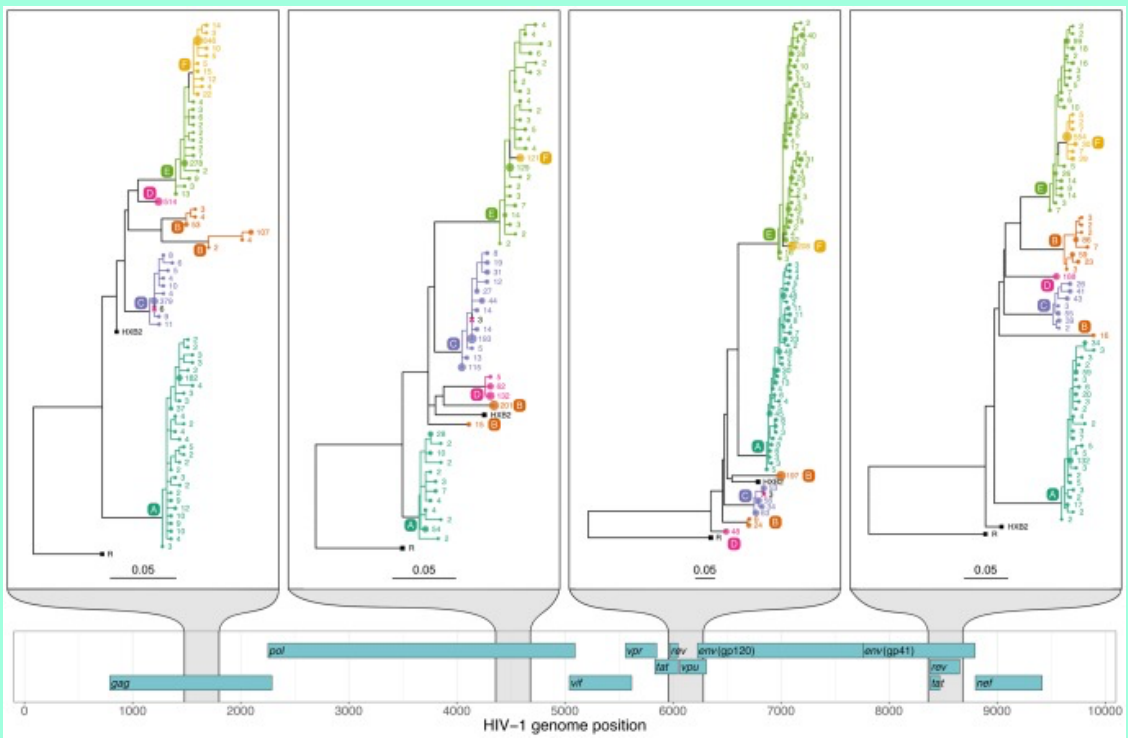
- Phylogenies are made through IQTREE by group and window through IQTREE (Nguyen et al. 2015)



Code and slides

Analysing phylogenies

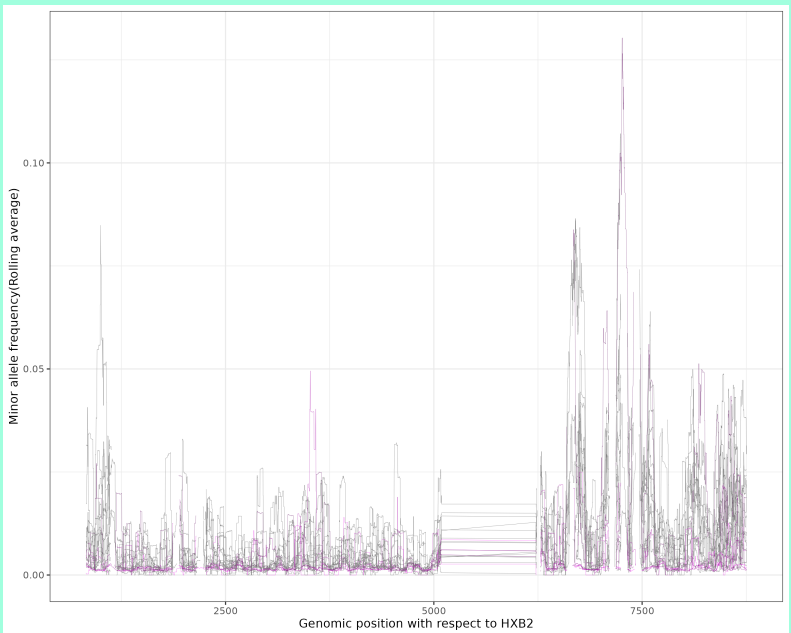
- `phyloscanner` Wymant et al. (2017) summarizes each tree through summary statistics:
- `patStats.csv`: contains LRTT, number of tips, etc...



Code and slides

Minor Allele Frequencies

- $\mathrm{MAF} = (1 - \text{proportion of majority bases})/\text{depth}$
- Evaluated at first 2 codos positions (`MAF12c`) and/or third codon position (`MAF3c`)



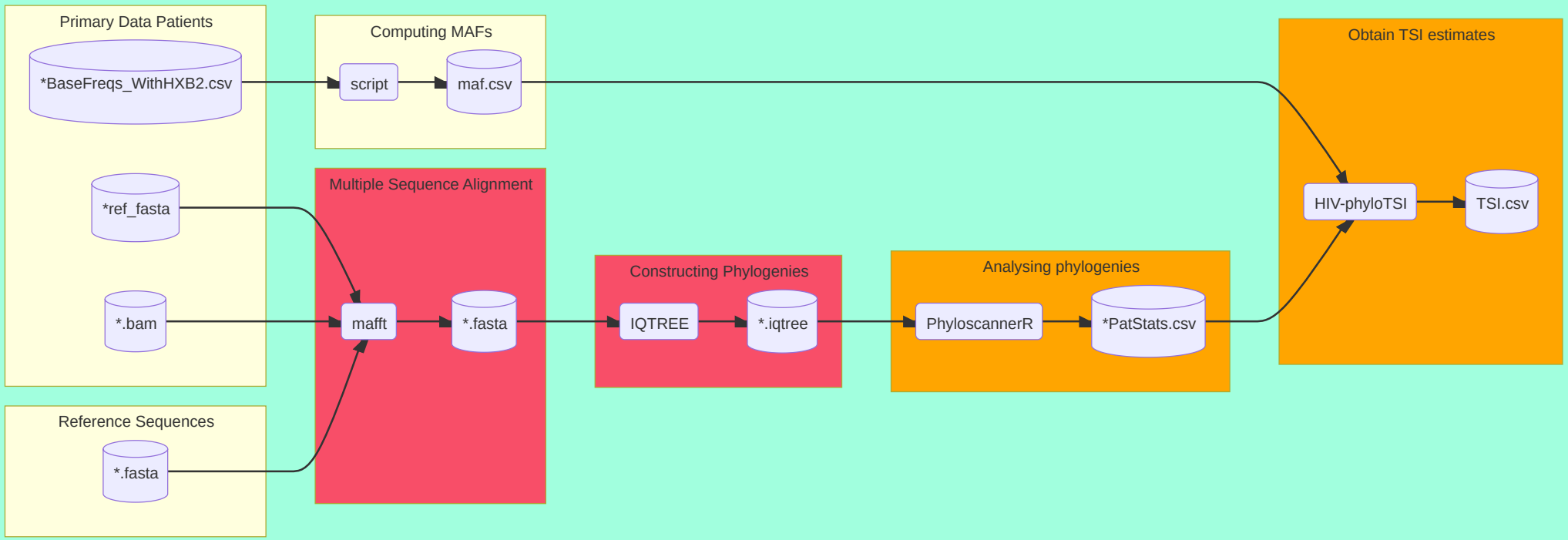
Code and slides

HIV-phyloTSI

- described in Golubchik et al. (2022), code
- Focus of the practical session,
- takes as input features of phylogenies and MAFs.
- returns a `csv` file with TSI estimates and some uncertainty range.

Code and slides

Summary



> **Note**
>
> Steps run for each group and window are shown in red, while those running by group in orange

Code and slides

But what is HIV-phyloTSI doing?

- **Input**: divergence measures which are differently informative depending on position on the genome.
- **ML algorithm** to capture complexity:
  - Random Forest: averages over many simple decision trees
  - is interpretable: we can understand which variables are important.

Code and slides

HIV-PhyloTSI obtaining TSI estimates

Code and slides

Plan

- divide in group of 4, each goup with a machine running Linux or MacOS
- installing dependencies
- running HIV-phyloTSI on 20 sequences .
- if there is time, visualising outputs in R.

Code and slides

Preliminary Tools

There are two tools that will make our lives easier:

- **Terminal access** allows to run programs and give instructions to computer. <> (`Ctrl+Alt+T` on Linux, `Terminal` on MacOS, see WSL for Windows).
- **Conda** allows to download the exact requirements needed for HIV-phyloTSI. Installation instructions can be found here

Installing dependencies

The 'ingredients' to run the analyses are: the **ML algorithm**; the **data** and the **code dependencies**. The below code chunk allows you to download everything that is needed for the analyses.

```
# change directory to where you want to install HIV-phyloTSI repo
cd $HOME/git            # this is where I install git packages
# cd $HOME && mkdir git && cd git

# clone directories necessary to run the analysis
# BDIs code and workshop materials
git clone git@github.com/BDI-pathogens/HIV-phyloTSI.git
git clone git@github.com/abriz97/HIV-phyloTSI-workshops.git
# store paths to 2 above directories
DIR_WORKSHOP="$(pwd)/HIV-phyloTSI-workshops"
DIR_PROGRAM= "$(pwd)/HIV-phyloTSI"

# install python dependencies for HIV-phyloTSI and load the environment
conda env create -f HIV-phyloTSI-workshops/hivphylotsi.yml
conda activate hivphylotsi
```

Note

When interacting with a terminal, learning the power of the `$` operator is key. The operator allows to evaluate variables (eg. `$HOME`) or to evaluate commands surrounded by brackets (eg. `$(pwd)`).

Running the algorithm

Once all the ingredients are there, we can start cooking. It is relatively simple to run the analyses, even though we need to be precise in the way we specify the paths to the input data.

```
# Run HIV-phyloTSI on input data.
python $DIR_PROGRAM/HIV-phyloTSI.py \
    -d $DIR_PROGRAM/Model \
    -p $DIR_WORKSHOP/input/ptyr1_patStats.csv \
    -m $DIR_WORKSHOP/input/phsc_input_samples_maf.csv \
    -o $DIR_WORKSHOP/output/ptyr1_tsi_workshop.csv

# print header of output to make sure it exists:
head $DIR_WORKSHOP/HIV-phyloTSI-workshops/output/ptyr1_tsi_workshop.csv
```

Note

The first 2 lines point to `$DIR_PROGRAM` because they refer to the code we want to use. On the other hand, the bottom 3 lines refer to the input data and output paths, and this is why they point to `$DIR_WORKSHOP`.

Visualising results

I provide some **R functions and scripts** to visualise results, which **can be found** in the github **repository**:

- script: `$DIR_WORKSHOP/src/workshop_analyses.R`
- functions: `$DIR_WORKSHOP/src/R/workshop_R_helpers.R`

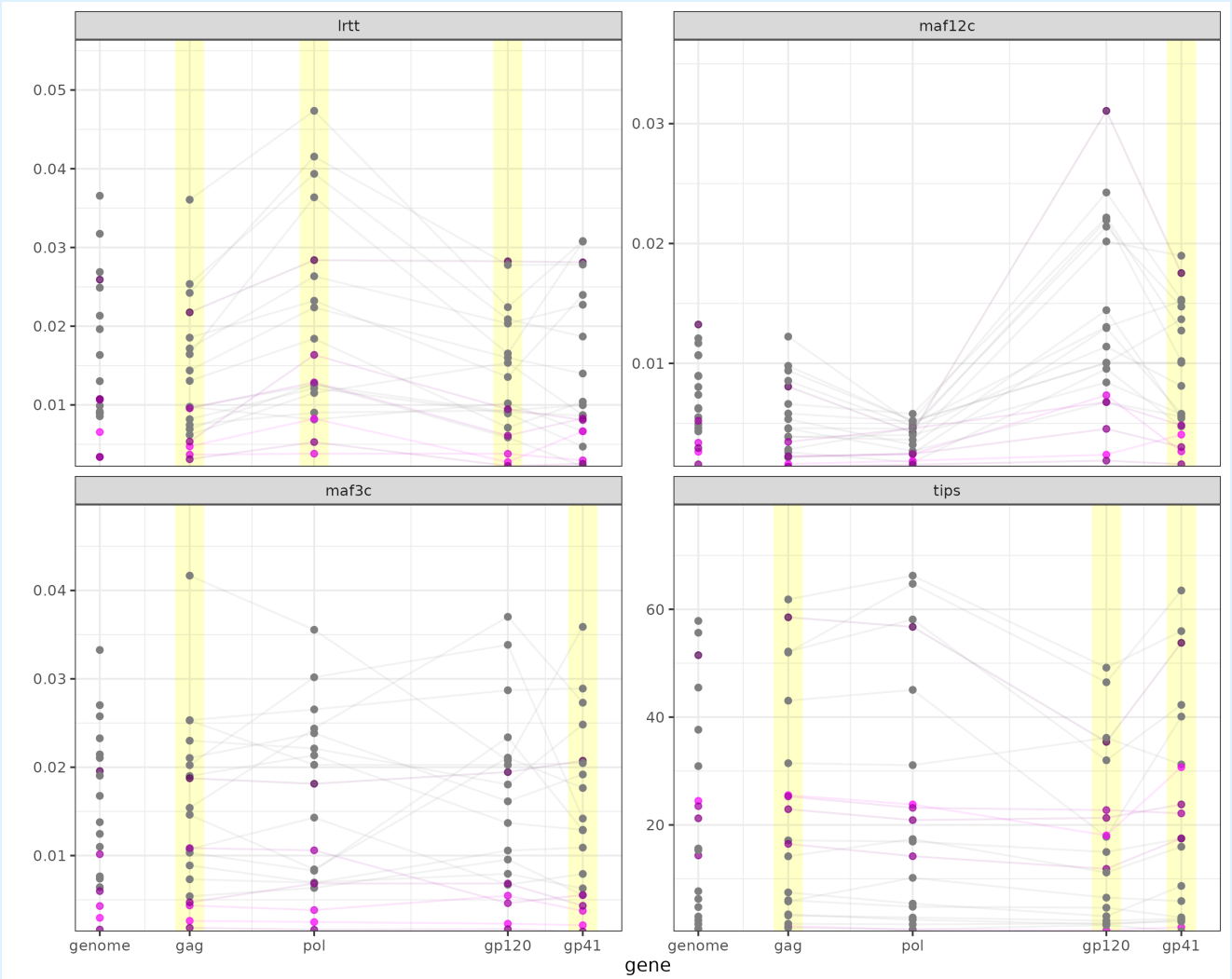
Again, we can use conda to **install the necessary packages**:

```
# install few R dependencies for visualisation and load the environment
cd $DIR_WORKSHOP
conda env create -f workshopR.yml
conda activate workshopR
```

I will be showing snippets of the above code together with the plots they produce.
You can reproduce the steps by opening up the script in RStudio.

Predictors used

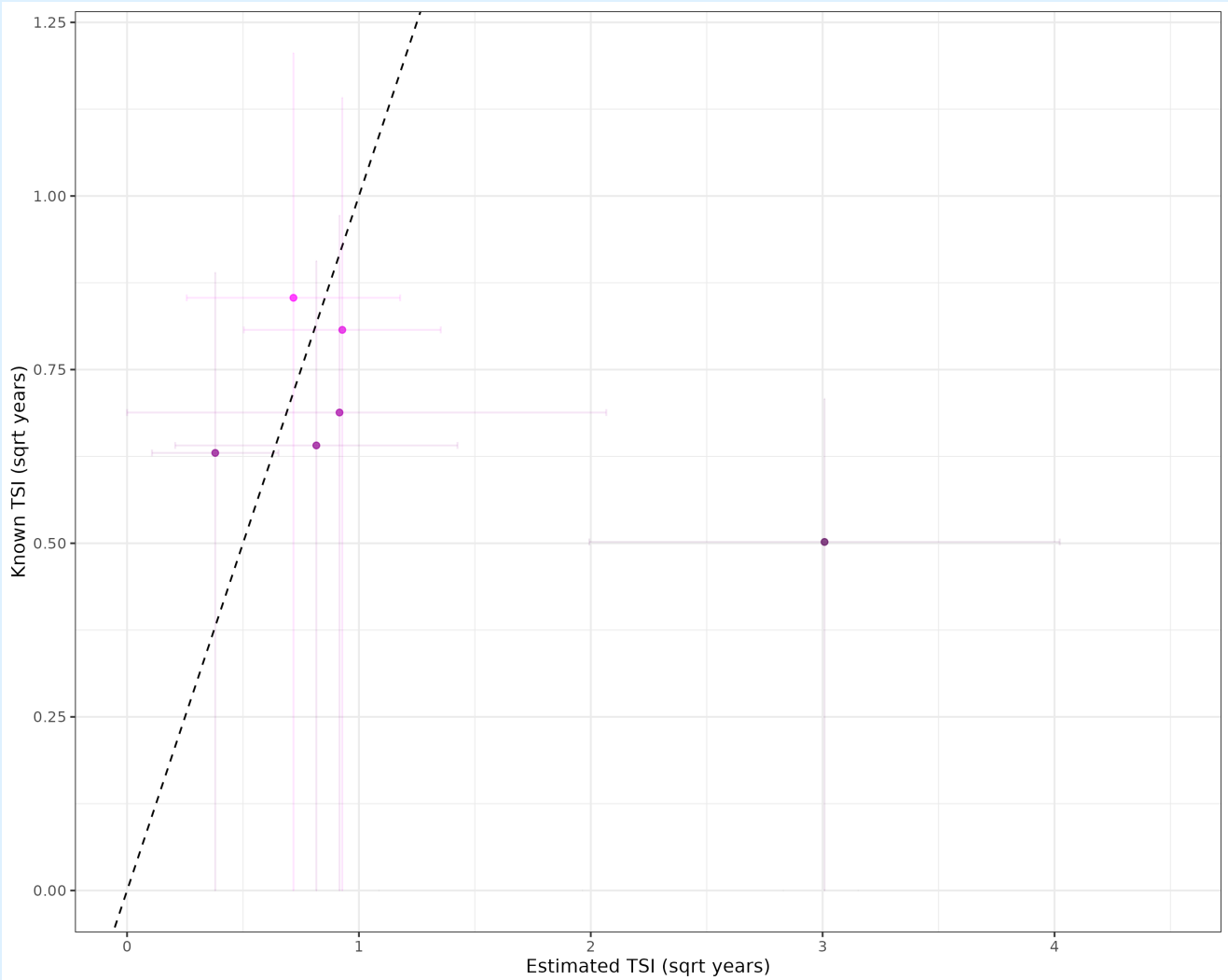
```
dtsi  <- fread(file.path(git_root, "data/ptyr1_tsi.csv"))
plot_predictors_from_tsi_output(dtsi, exclude="dual")
```



Code and slides

Evaluation on seroconverters

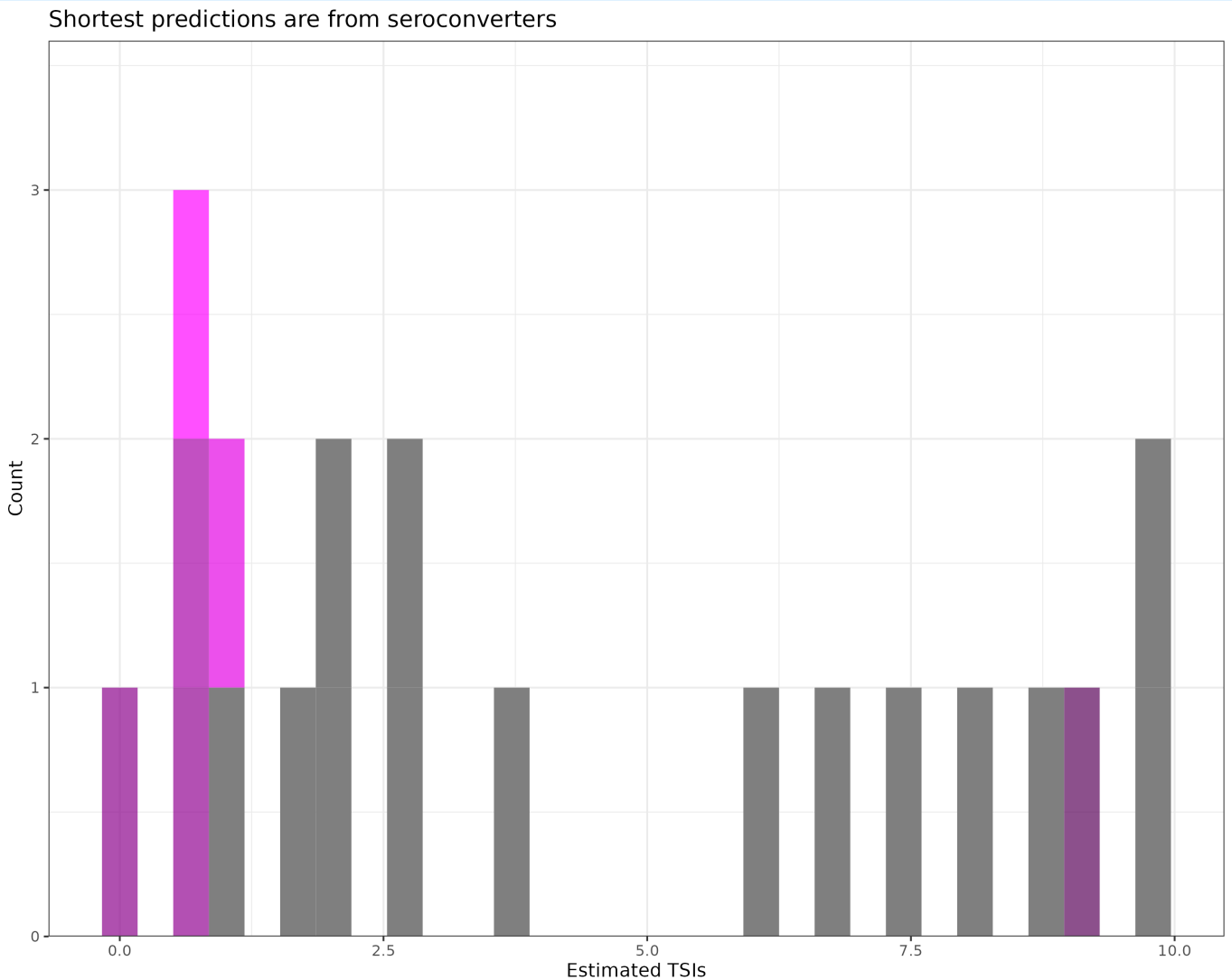
```
dall  <- merge(dtsi, ddates)
plot_cross_interval_tsisero(dall, sqroot=TRUE)
```



Code and slides

Histogram of estimates TSIs

```
1 p_hist <- plot_histogram_tsi(dtsi)
```



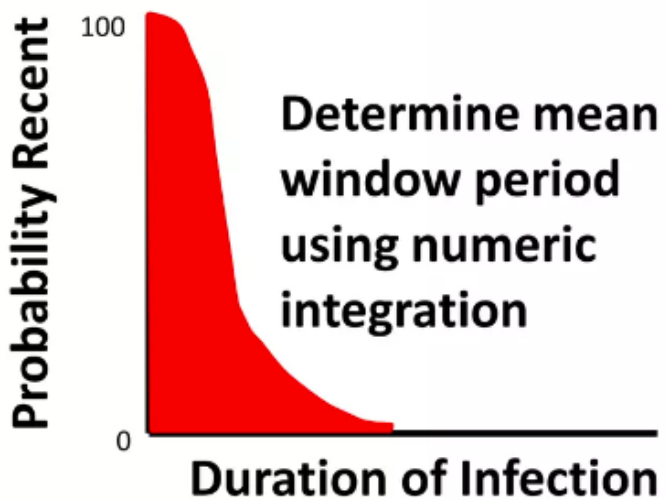
Code and slides

Possible follow-up analyses

Prevalence of recent infections

Following Freeman and Hutchison (1980) and Brookmeyer and Quinn (1995), if incidence is constant, it can be estimated as:

$$I = \frac{\#\text{ recent}}{\#\text{ uninfected} \times MDRI}$$



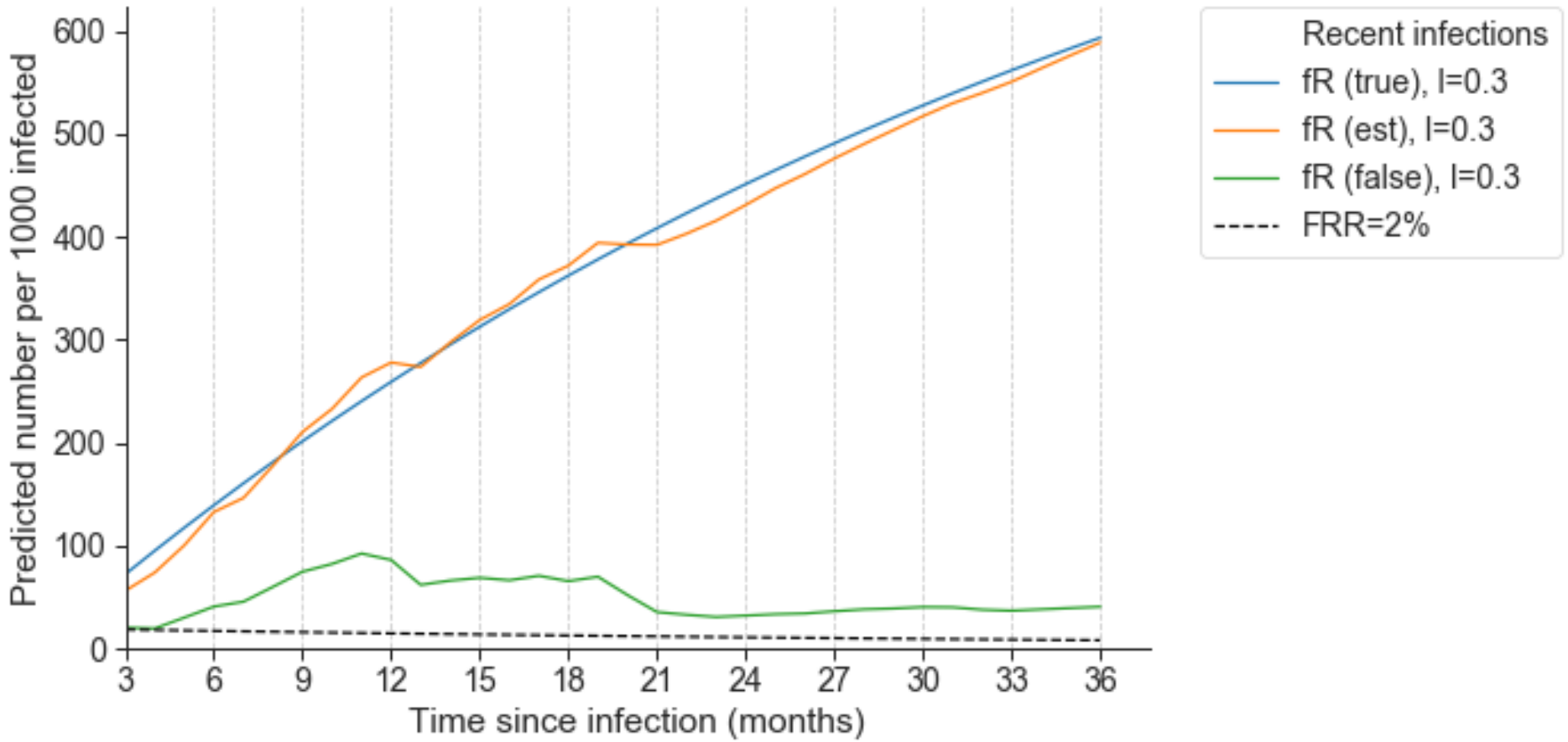
where $MDRI$ = mean duration of recency of infection.

- $MDRI = P[\text{ever being classified as recent}] \times E[TSI|\text{classified as recen}$
- **not a constant**: viral suppression, underlying population.

Here we **focus on the numerator**.

Simulation study:

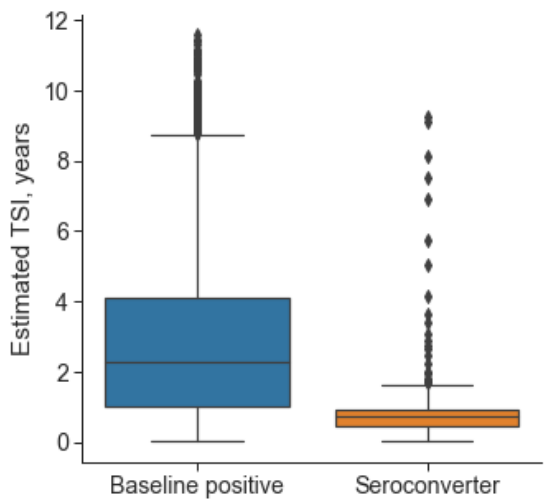
In simulated settings, the prevalence of recent infections is well estimated, despite at least 2% of recent classifications are wrong.

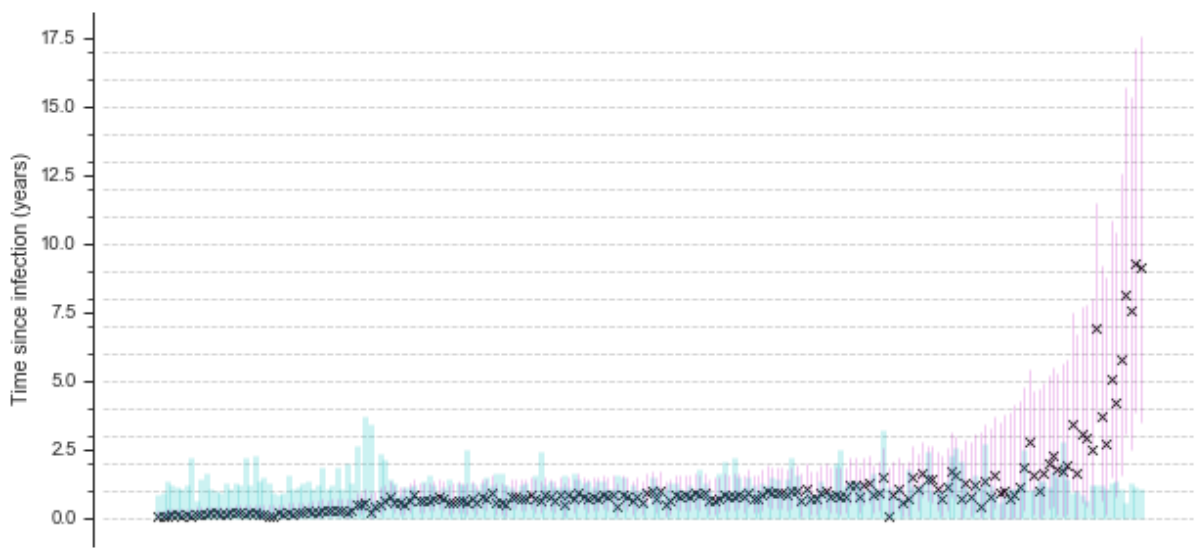


Code and slides

Application in Zambia

When directly applied to real-world data, HIV-phyloTSI generally produces smaller TSI
for known recent infections as compared to people with unknown first positive date:





=> Can be used to compare median TSIs among population subgroups.

Code and slides

Suggestions on TSI comparison among groups

Comparing simple summary statistics can be misleading:

- need to make sure statistic is *reliable*
- need to *chose* summary statistic: **median**

For reliability:

1. make sure groups are large enough (*minimum of 40 per group*)
2. Perform **bootstrap** to estimate uncertainty around

Code and slides

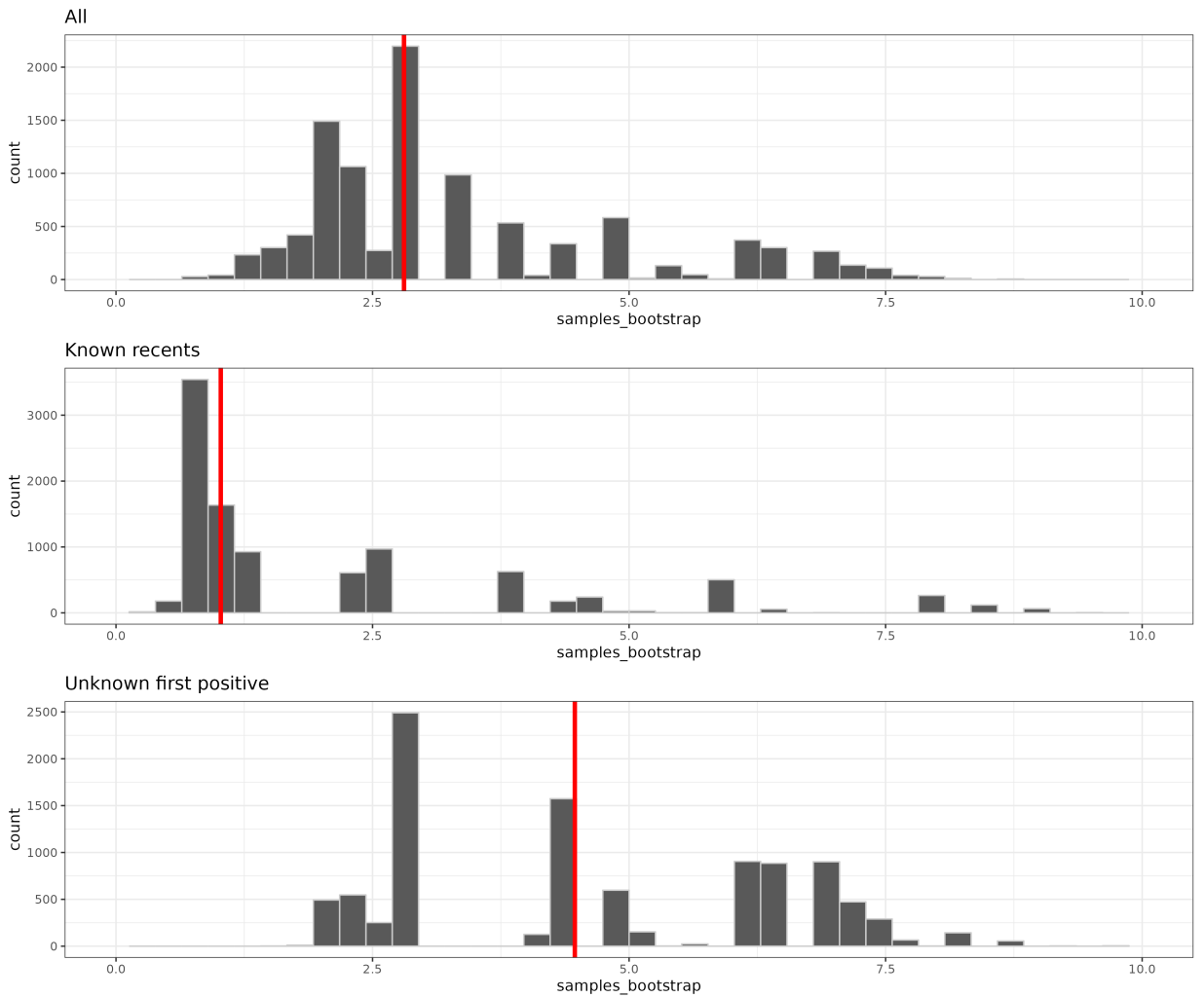
Suggestions on TSI comparison: bootstrap

Bootstrapping: statistical technique which *recycles* analysis data to estimate uncertainty around an estimator (e.g. median).

1. **Synthetic data** sets of the same size are obtained by sampling with replacement from the data.
2. The **estimator is computed** on each synthetic data set.
3. The distribution of the estimators obtained on the syntetic data sets is summarised through 95% quantiles.

Code and slides

Results on sample dataset:



Code and slides

How NOT to perform grouping:

Groups should be made based on covariates different than HIV-phyloTSI inputs or outputs.

Do NOT:

1. Group by estimate of HIV-phyloTSI
2. Group by estimated date of infection
3. Again: chose groups of sizes < 40

Code and slides

Other analyses on population level outputs:

Estimation of generation time distribution

Enriching source-recipient pairs by providing time since infection.

- Zambia (Hall et al. 2021)
- Rakai: (Monod et al. 2023)

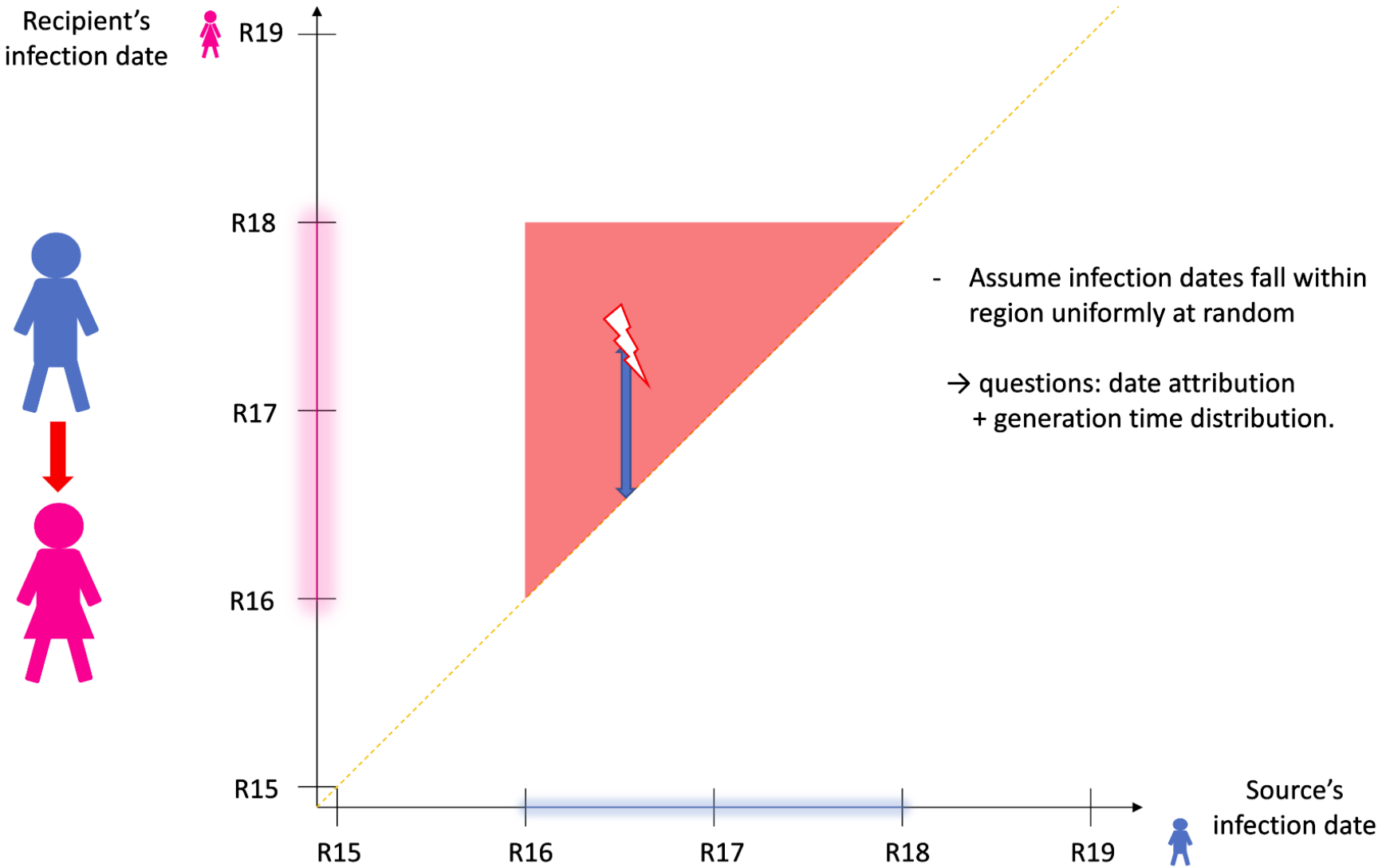
> **Important**
>
> To account for individual level uncertainty, these studies not only make use of central estimates, but also the output prediction/uncertainty range

Dating infection events in Rakai

- Infection pairs data from Rakai Community Cohort Study
- Transmission pairs detected with phyloscanner Ratmann et al. (2020)
- Question: how did transmission pattern change over time?
- Need to date infections

Code and slides

Dating infection events in Rakai



Code and slides

Summing up

- HIV-phyloTSI is a novel algorithm to estimate infection dates.
- alternative to serological assays which allows more control on definition of recency (Robust to subtype and ART usage)
- preliminary analyses and simulation studies demonstrate good performances at population level
- but may be inaccurate at the individual level: very large prediction intervals.
- Can help us explore answers to unanswered questions

Code and slides

References

Bhebhe, Lynnette, Sikhulile Moyo, Simani Gaseitsiwe, Molly Pretorius-Holme, Etienne K. Yankinda, Kutlo Manyake, Coulson Kgathi, et al. 2022. "Epidemiological and Viral Characteristics of Undiagnosed HIV Infections in Botswana." *BMC Infectious Diseases* 22 (1): 710. https://doi.org/10.1186/s12879-022-07698-4.

Bonsall, David, Tanya Golubchik, Mariateresa De Cesare, Mohammed Limbada, Barry Kosloff, George MacIntyre-Cockett, Matthew Hall, et al. 2018. "A Comprehensive Genomics Solution for HIV Surveillance and Clinical Monitoring in a Global Health Setting." Preprint. Genomics. https://doi.org/10.1101/397083.

Brookmeyer, Ron, and Thomas C. Quinn. 1995. "Estimation of Current Human Immunodeficiency Virus Incidence Rates from a Cross-Sectional Survey Using Early Diagnostic Tests." *American Journal of Epidemiology* 141 (2): 166–72. https://doi.org/10.1093/oxfordjournals.aje.a117404.

Carlisle, Louisa A, Teja Turk, Katharina Kusejko, Karin J Metzner, Christine Leemann, Corinne D Schenkel, Nadine Bachmann, et al. 2019. "Viral

Diversity Based on Next-Generation Sequencing of HIV-1 Provides Precise Estimates of Infection Recency and Time Since Infection." *The Journal of Infectious Diseases* 220 (2): 254–65. https://doi.org/10.1093/infdis/jiz094.

Centers for Disease Control and Prevention (CDC). 2014. "Revised Surveillance Case Definition for HIV Infection–United States, 2014." *MMWR. Recommendations and Reports: Morbidity and Mortality Weekly Report. Recommendations and Reports* 63 (RR-03): 1–10.

Duong, Yen T., Maofeng Qiu, Anindya K. De, Keisha Jackson, Trudy Dobbs, Andrea A. Kim, John N. Nkengasong, and Bharat S. Parekh. 2012. "Detection of Recent HIV-1 Infection Using a New Limiting-Antigen Avidity Assay: Potential for HIV-1 Incidence Estimates and Avidity Maturation Studies." Edited by Alan Landay. *PLoS ONE* 7 (3): e33328. https://doi.org/10.1371/journal.pone.0033328.

Fiebig, Eberhard W, David J Wright, Bhupat D Rawal, Patricia E Garrett, Richard T Schumacher, Lorraine Peddada, Charles Heldebrant, et al. 2003. "Dynamics of HIV Viremia and Antibody Seroconversion in Plasma Donors: Implications for Diagnosis and Staging of Primary HIV Infection." *AIDS* 17 (13): 1871–79. https://doi.org/10.1097/00002030-200309050-00005.

Freeman, Jonathan, and George B. Hutchison. 1980. “PREVALENCE, INCIDENCE AND DURATION.” *American Journal of Epidemiology* 112 (5): 707–23. https://doi.org/10.1093/oxfordjournals.aje.a113043.

Gall, Astrid, Bridget Ferns, Clare Morris, Simon Watson, Matthew Cotten, Mark Robinson, Neil Berry, Deenan Pillay, and Paul Kellam. 2012. “Universal Amplification, Next-Generation Sequencing, and Assembly of HIV-1 Genomes.” *Journal of Clinical Microbiology* 50 (12): 3838–44. https://doi.org/10.1128/JCM.01516-12.

Golubchik, Tanya, Lucie Abeler-Dörner, Matthew Hall, Chris Wymant, David Bonsall, George Macintyre-Cockett, Laura Thomson, et al. 2022. “HIV-phyloTSI: Subtype-Independent Estimation of Time Since HIV-1 Infection for Cross-Sectional Measures of Population Incidence Using Deep Sequence Data.” Preprint. HIV/AIDS. https://doi.org/10.1101/2022.05.15.22275117.

Hall, Matthew, Tanya Golubchik, David Bonsall, Lucie Abeler-Dörner, Mohammed Limbada, Barry Kosloff, Ab Schaap, et al. 2021. “Demographics of People Who Transmit HIV-1 in Zambia: A Molecular Epidemiology Analysis in the HPTN-071 PopART Study.” Preprint. HIV/AIDS. https://doi.org/10.1101/2021.10.04.21263560.

Janssen, Robert S. 1998. “New Testing Strategy to Detect Early HIV-1 Infection for Use in Incidence Estimates and for Clinical and

Prevention Purposes." *JAMA* 280 (1): 42. https://doi.org/10.1001/jama.280.1.42.

Katoh, K. 2002. "MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform." *Nucleic Acids Research* 30 (14): 3059–66. https://doi.org/10.1093/nar/gkf436.

Laeyendecker, Oliver, Michal Kulich, Deborah Donnell, Arnošt Komárek, Marek Omelka, Caroline E. Mullis, Greg Szekeres, et al. 2013. "Development of Methods for Cross-Sectional HIV Incidence Estimation in a Large, Community Randomized Trial." Edited by Dimitrios Paraskevis. *PLoS ONE* 8 (11): e78818. https://doi.org/10.1371/journal.pone.0078818.

Monod, Mélodie, Andrea Brizzi, Ronald M Galiwango, Robert Ssekubugu, Yu Chen, Xiaoyue Xi, Edward Nelson Kankaka, et al. 2023. "Growing Gender Disparity in HIV Infection in Africa: Sources and Policy Implications." Preprint. Epidemiology. https://doi.org/10.1101/2023.03.16.23287351.

Moyo, Sikhulile, Eduan Wilkinson, Vladimir Novitsky, Alain Vandormael, Simani Gaseitsiwe, Max Essex, Susan Engelbrecht, and Tulio De Oliveira. 2015. "Identifying Recent HIV Infections: From Serological Assays to Genomics." *Viruses* 7 (10): 5508–24. https://doi.org/10.3390/v7102887.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt Von Haeseler, and Bui Quang Minh. 2015. “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” *Molecular Biology and Evolution* 32 (1): 268–74. https://doi.org/10.1093/molbev/msu300.

Parekh, Bharat S., Debra L. Hanson, John Hargrove, Bernard Branson, Timothy Green, Trudy Dobbs, Niel Constantine, Julie Overbaugh, and J. Steven McDougal. 2011. “Determination of Mean Recency Period for Estimation of HIV Type 1 Incidence with the BED-Capture EIA in Persons Infected with Diverse Subtypes.” *AIDS Research and Human Retroviruses* 27 (3): 265–73. https://doi.org/10.1089/aid.2010.0159.

Ragonnet-Cronin, Manon, Tanya Golubchik, Sikhulile Moyo, Christophe Fraser, Max Essex, Vlad Novitsky, and Erik Volz. 2022. “Human Immunodeficiency Virus (HIV) Genetic Diversity Informs Stage of HIV-1 Infection Among Patients Receiving Antiretroviral Therapy in Botswana.” *The Journal of Infectious Diseases* 225 (8): 1330–38. https://doi.org/10.1093/infdis/jiab293.

Ratmann, Oliver, Joseph Kagaayi, Matthew Hall, Tanya Golubchick, Godfrey Kigozi, Xiaoyue Xi, Chris Wymant, et al. 2020. “Quantifying HIV Transmission Flow Between High-Prevalence Hotspots and

Code and slides

Code and slides