Parallelization of MCMC based Phylogenetic analysis to greatly reduce run time

BioNet November Seminar

Presented by: David Yang
MSc. Student
University of Calgary

November 15th 2023

Table of Content

Introduction

Methods

Results

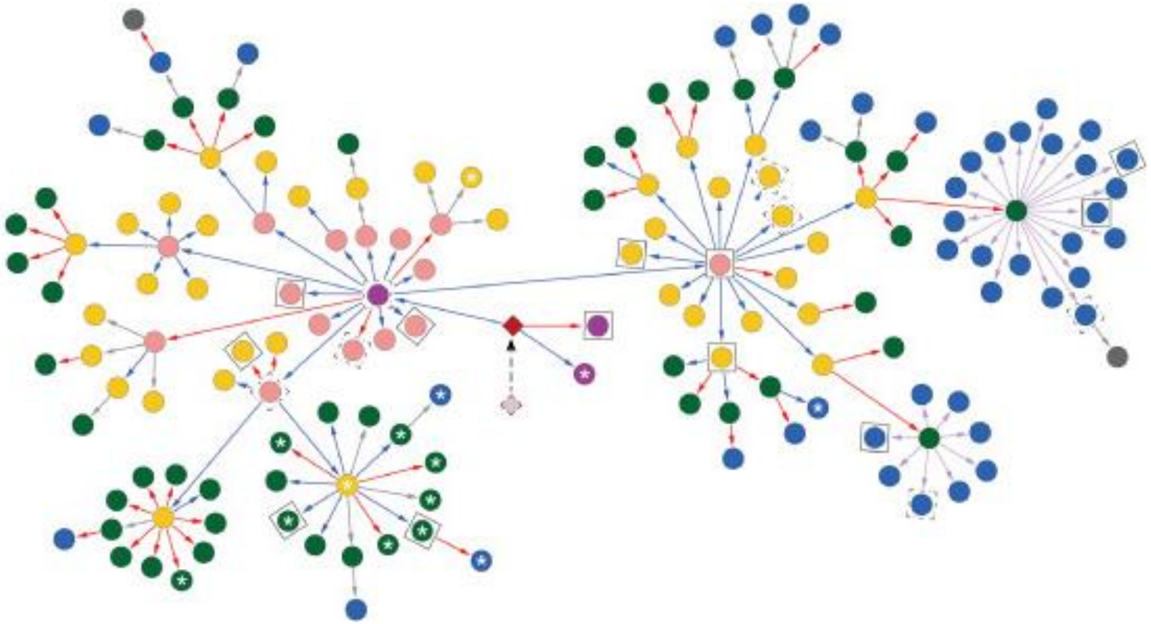
Conclusion

UNIVERSITY OF CALGARY

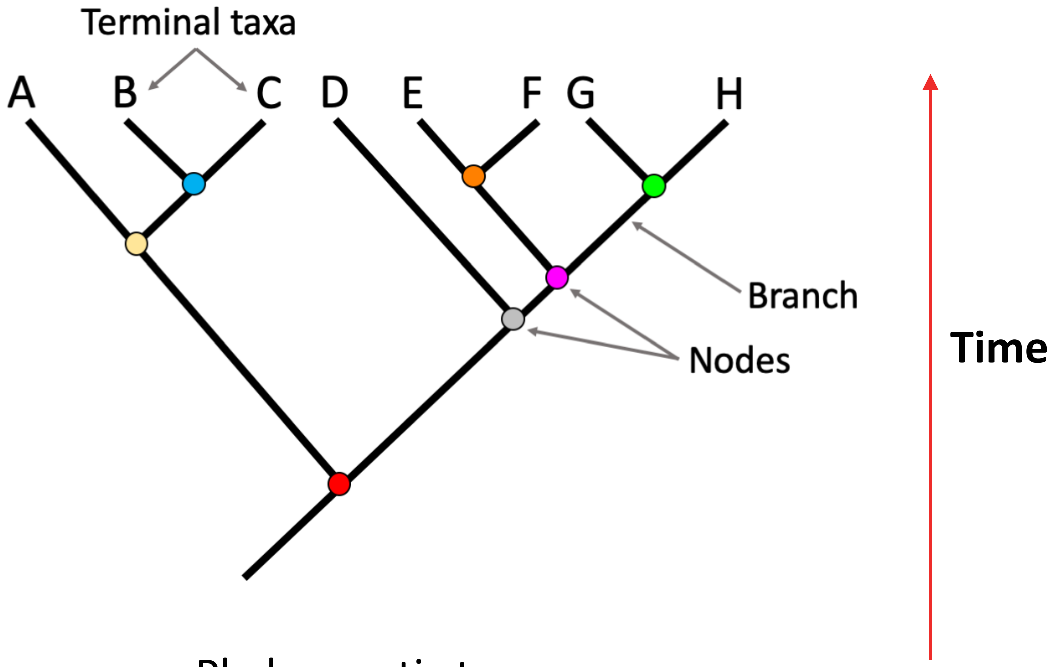
Introduction

Phylogenetic analysis

Phylogenetics is the study of **evolutionary history** and relationships through the analysis of heritable traits such as **genomic sequences**



Transmission network



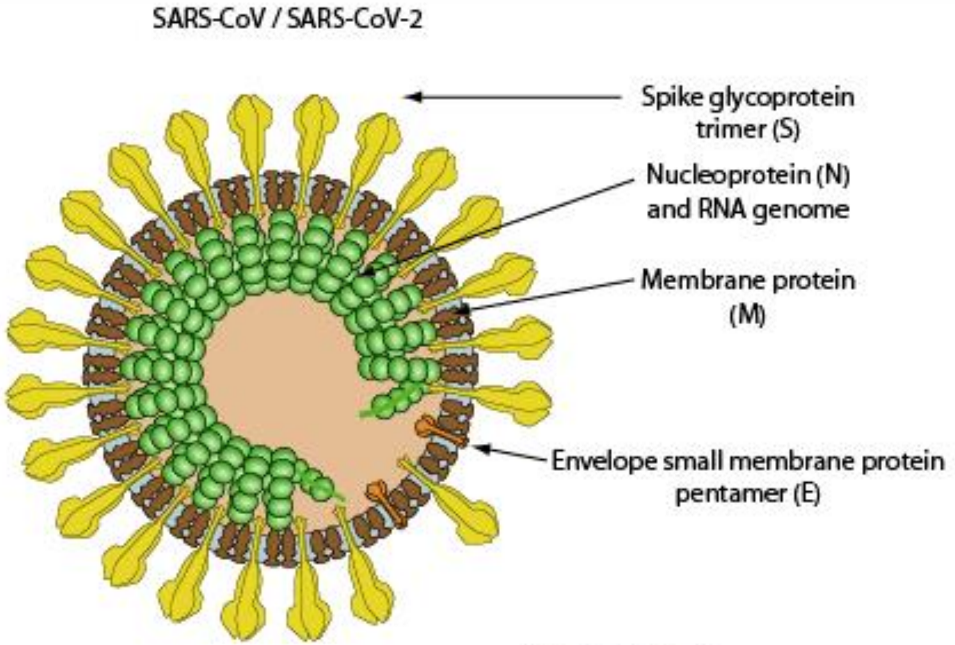
Phylogenetic tree

Relevance to COVID-19 Pandemic

Viral agent of COVID-19: **SARS-CoV-2**

Outcomes from phylogenetic analyses (genomic epidemiology):

- Probable zoonotic origin was found
- Elucidation of multiple episodes of the Founder effect during the early pandemic
- Identification of Infection sources, “super-spreaders” and asymptomatic individuals


SARS-CoV / SARS-CoV-2
Spike glycoprotein trimer (S)
Nucleoprotein (N) and RNA genome
Membrane protein (M)
Envelope small membrane protein pentamer (E)

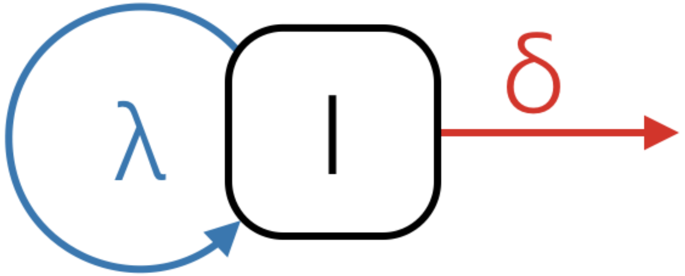
© ViralZone 2020
SIB Swiss Instute of Bioinformatics

Bayesian Phylogenetics

Advantageous over conventional methods especially for **viral outbreaks**

Incorporates existing knowledge through various **prior models**:

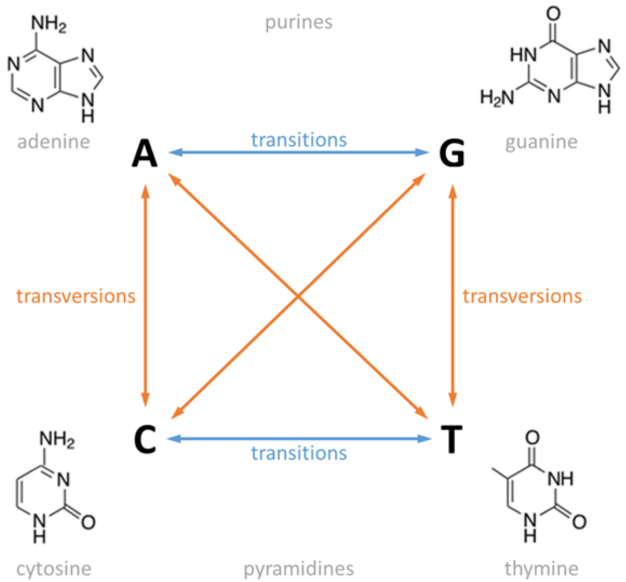
Birth-death skyline model



Assumption: sampled individual does not remain infectious

λ – transmission rate
δ – rate of becoming non-infectious

Generalised time reversible model



Assumptions:

Independent transversions

Independent transitions

Unequal base frequencies

Limitations with Phylogenetic Analysis

Inference through phylogenetic analyses is a complex and exhaustive process:

# rooted trees = $\frac{(2n-3)!}{2^{n-2}(n-2)!}$

# unrooted trees = $\frac{(2n-5)!}{2^{n-3}(n-3)!}$

# sequences (n)	# rooted trees	# unrooted trees
2	1	1
5	105	15
10	3.44e7	2.03e6
15	2.13e14	7.91e12
20	8.20e21	2.21e20

The goal to obtain the **maximum clade credibility** (MCC) tree - most reasonable evolutionary relation given the estimated Bayesian parameters

Often accompanies long run-times and intensive computational demand

Current methods

Many studies have proposed optimization methods:

- **RAxML** provides features for parallelizing maximum likelihood calculations to speed up computations
- **matOptimize**, which was inspired by the overwhelmingly number of SARS-CoV-2 sequences available, also optimizes maximum parsimony based phylogenetic analyses through parallelization and memory-efficient data structures

Lack of significant breakthroughs in terms reducing run-times for Bayesian phylogenetic analyses

In this paper, we propose a parallelization method

for Bayesian phylogenetic analysis using BEAST2

Methodology

Datasets

Simulation data: Two HIV1 simulated sequence datasets each with 10 subsamples of 1000 sequences were generated with the simulation software FAVITES using established parameters for HIV from literature:

i. 1st dataset – perfect sequence sampling

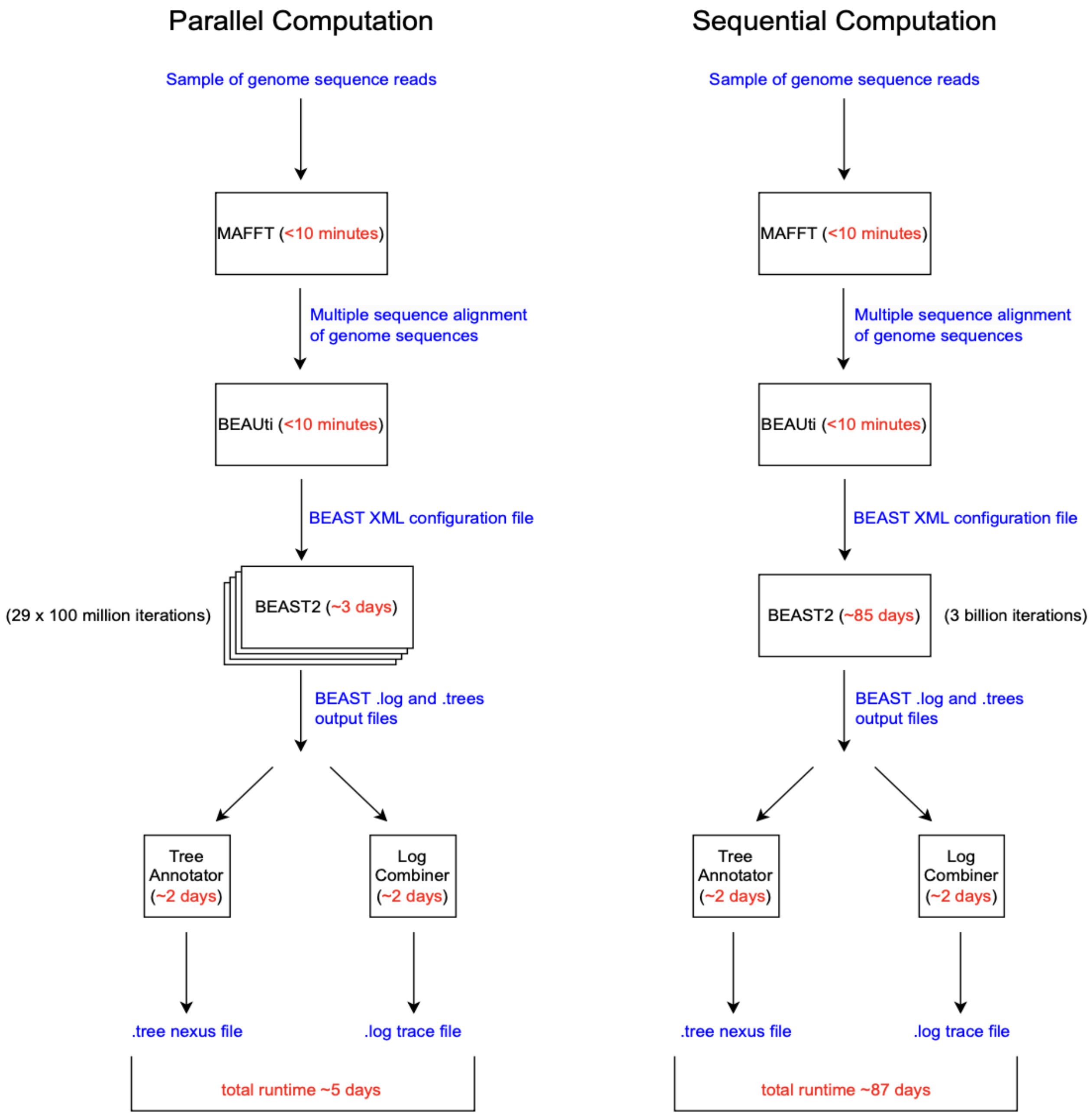
ii. 2nd dataset – 10% sequence sampling rate

SARS-CoV-2 data: Six total subsamples of 1000 sequences of SARS-CoV-2 were obtained from GISAID database using random (Augur) and weighted (Nybbler) sampling from February 1st, 2020 to October 31st, 2020.

Software used

- **MAFFT**: create multiple sequence alignments for nucleotide sequences
- **Beauti**: Configuration for BEAST2 analysis
- **BEAST2**: Bayesian analysis of molecular sequences
- **Logcombiner**: Combining the output from multiple BEAST2 runs
- **Treeannotator**: Finding maximum clade credibility tree
- **TreeCMP**: Calculating similarity between phylogenetic trees

Figure 1: Overview of the Bayesian phylogenetic analyses performed in the study. Differences in methodology between MCMC ran in parallel and sequentially are shown along with the corresponding run-time for each step in the procedure.



Parallel Computation

Sample of genome sequence reads

MAFFT (<10 minutes)

Multiple sequence alignment of genome sequences

BEAUti (<10 minutes)

BEAST XML configuration file

(29 x 100 million iterations)

BEAST2 (~3 days)

BEAST .log and .trees output files

Tree Annotator (~2 days)

Log Combiner (~2 days)

.tree nexus file

.log trace file

total runtime ~5 days

Sequential Computation

Sample of genome sequence reads

MAFFT (<10 minutes)

Multiple sequence alignment of genome sequences

BEAUti (<10 minutes)

BEAST XML configuration file

BEAST2 (~85 days)

(3 billion iterations)

BEAST .log and .trees output files

Tree Annotator (~2 days)

Log Combiner (~2 days)

.tree nexus file

.log trace file

total runtime ~87 days

UNIVERSITY OF CALGARY

Analysis of results (difference between parallel and sequential)

Two main components:

1. Comparison of parameter estimates (i.e., substitution rates & gamma rate parameter) from parallel computations with sequential computations and/or ground-truth
2. Comparison of phylogenetic trees (i.e., distance metrics) from parallel computations with sequential computations and/or ground-truth

Statistical tests such as U-test and T-test were performed to check significance

Results

Simulation dataset #1 – perfect sampling

GTR estimates

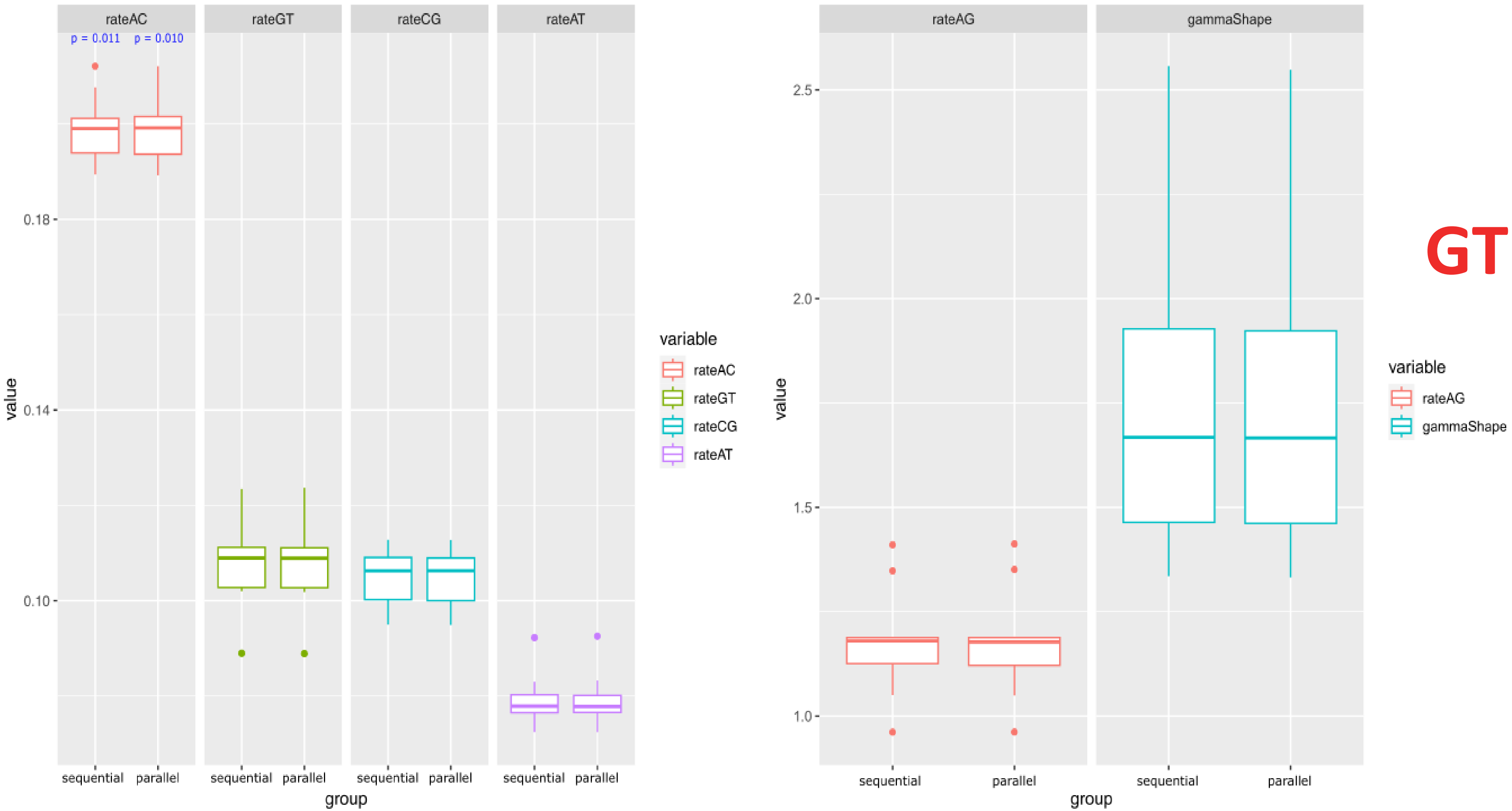
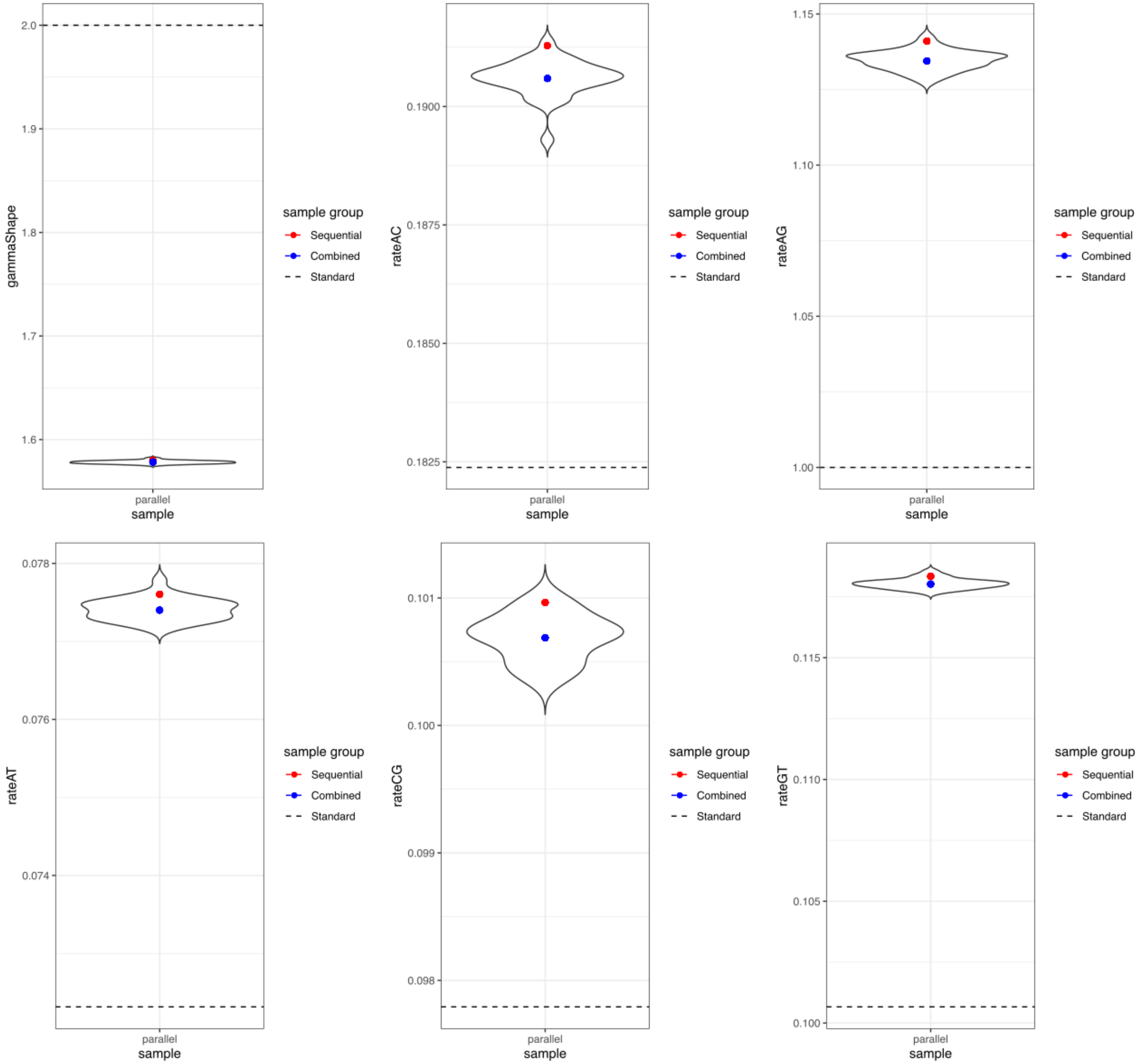


Figure 2: Boxplots from ten parameter estimates from MCMC phylogenetic analyses ran in parallel and sequentially on simulated HIV data with perfect sampling rate. Significant p-values from the U-test (as a line) and t-test (directly above) are labelled. Figures with different vertical axis scaling were used due to differences in the ranges of values.

GTR estimates

Figure 3: Violin plots from parameter estimates from twenty-nine independent MCMC chains (100-million iterations) from the phylogenetic analysis performed on the first replicate of sequences in the HIV dataset perfect sampling rate. The parameter estimates for all 29 MCMC chains combined ("combined") and parameter estimates from the MCMC run sequentially ("sequential") on the same dataset are also shown. Horizontal dashed lines represent the true value (standard) for each parameter.



Tree similarity

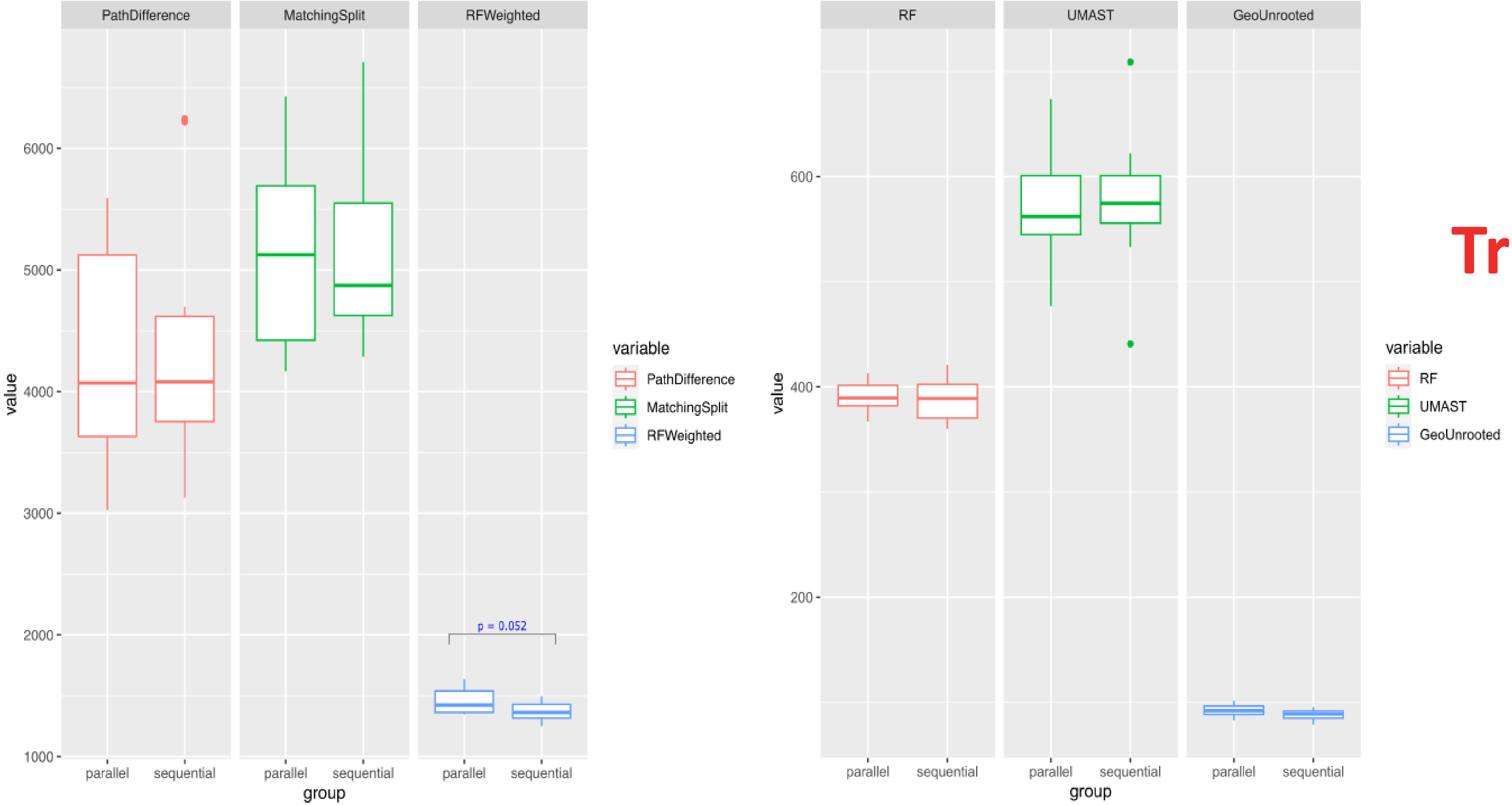


Figure 4 : Boxplots from ten distance metrics calculated by comparing MCC phylogenetic trees obtained from MCMC with “true” trees defined in the simulated HIV data with perfect sampling rate. Significant p-values from the U-test comparing the distances from the sequential and parallel samples are labelled.

Simulation dataset #2 – 10% sampling

GTR estimates

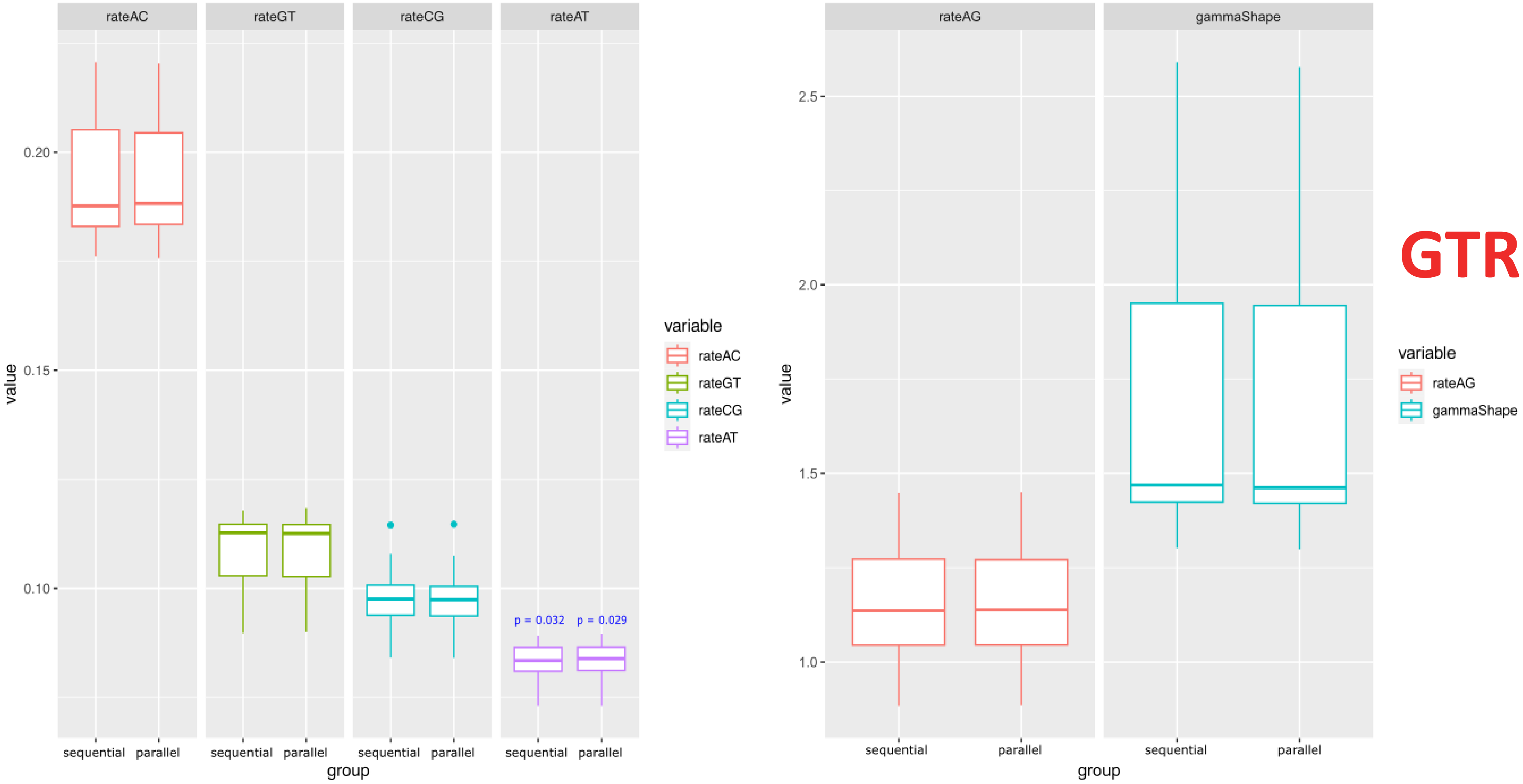
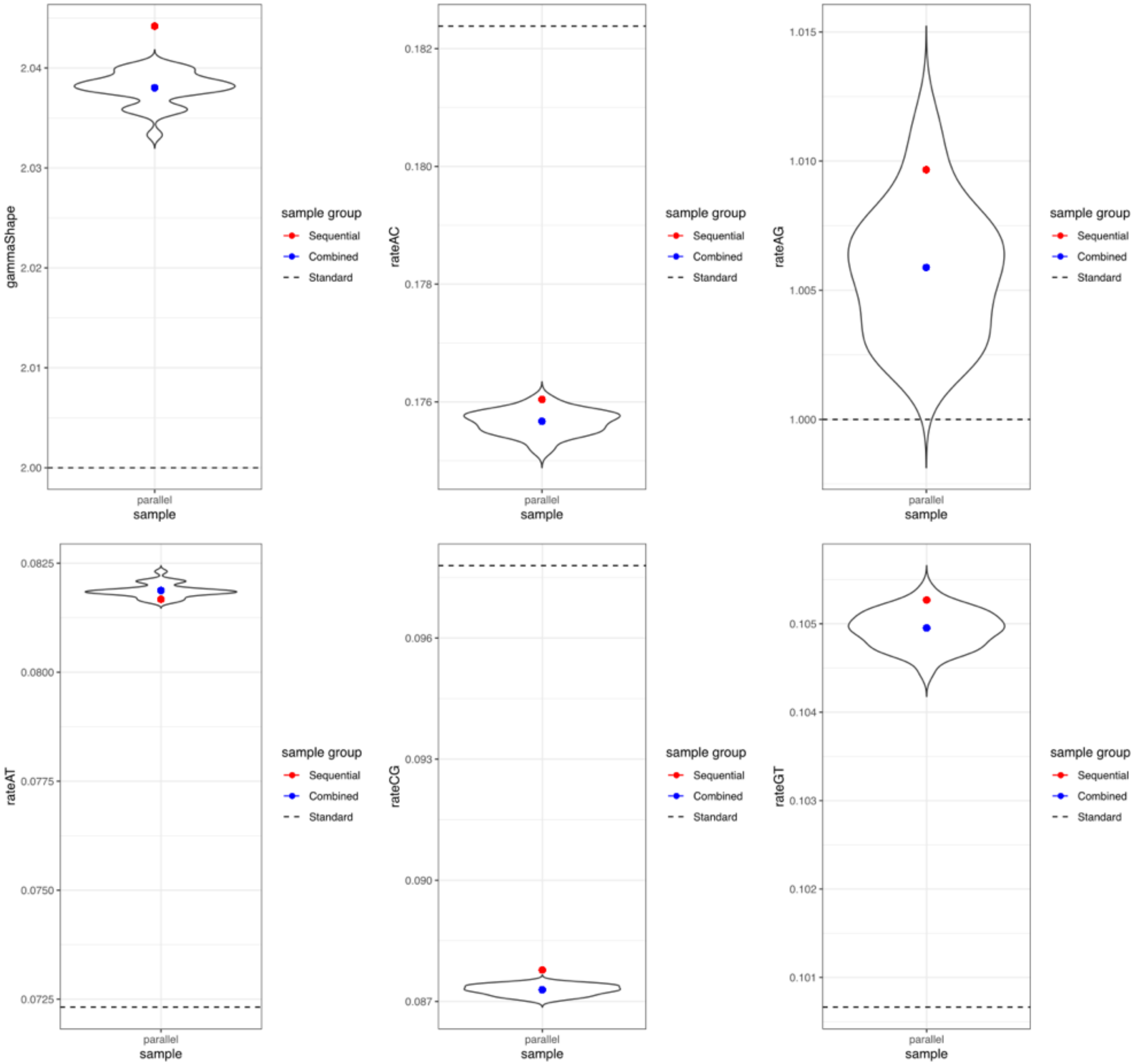


Figure 5: Boxplots from parameter estimates from MCMC phylogenetic analyses ran in parallel and sequentially on simulated HIV data with 10% sampling rate. Significant p-values from the U-test (as a line) and t-test (directly above) are labelled. Figures with different vertical axis scaling were used due to differences in the ranges of values

GTR estimates

Figure 6: Violin plots of parameter estimates from twenty-nine independent MCMC chains (100-million iterations) from the phylogenetic analysis performed the first replicate of sequences in the HIV dataset 10% sampling rate. The parameter estimates for all 29 MCMC chains combined (“combined”) and parameter estimate from the MCMC ran sequentially (“sequential”) on the same data are also shown. Horizontal dashed lines represent the true value (standard) for each parameter.



Tree similarity

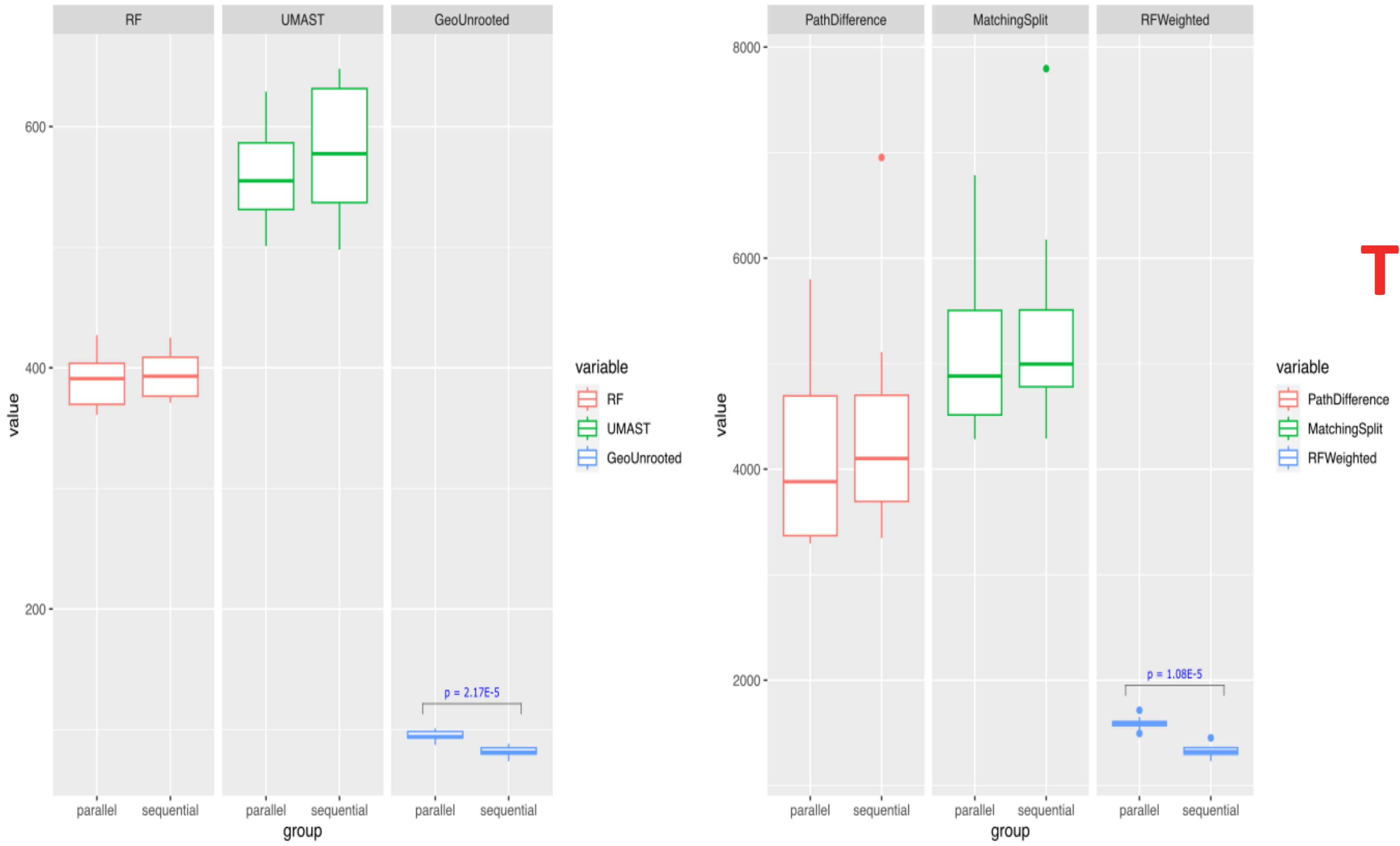


Figure 7: Boxplots from ten distance metrics calculated by comparing MCC phylogenetic trees obtained from MCMC with “true” trees defined in the simulated HIV data with 10% sampling rate. Significant p-values from the U-test comparing the distances from the sequential and parallel samples are labelled.

Real world dataset – SARS-CoV-2

GTR estimates

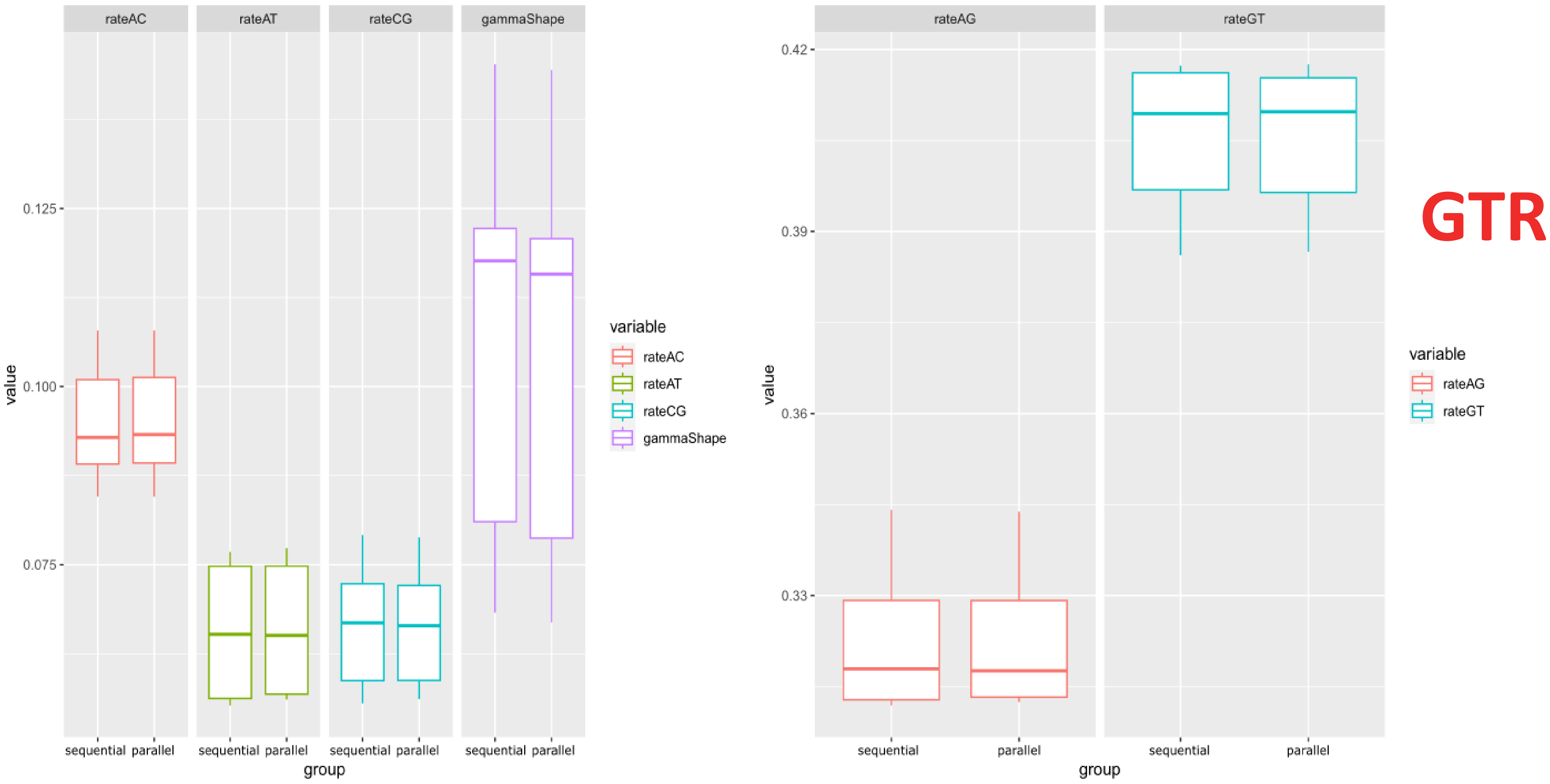


Figure 8: Boxplots from six parameter estimates from MCMC phylogenetic analyses ran in parallel and sequentially on SARS-CoV-2 data. Significant p-values from the U-test (as a line) and t-test (directly above) are labelled.

UNIVERSITY OF CALGARY

Tree similarity

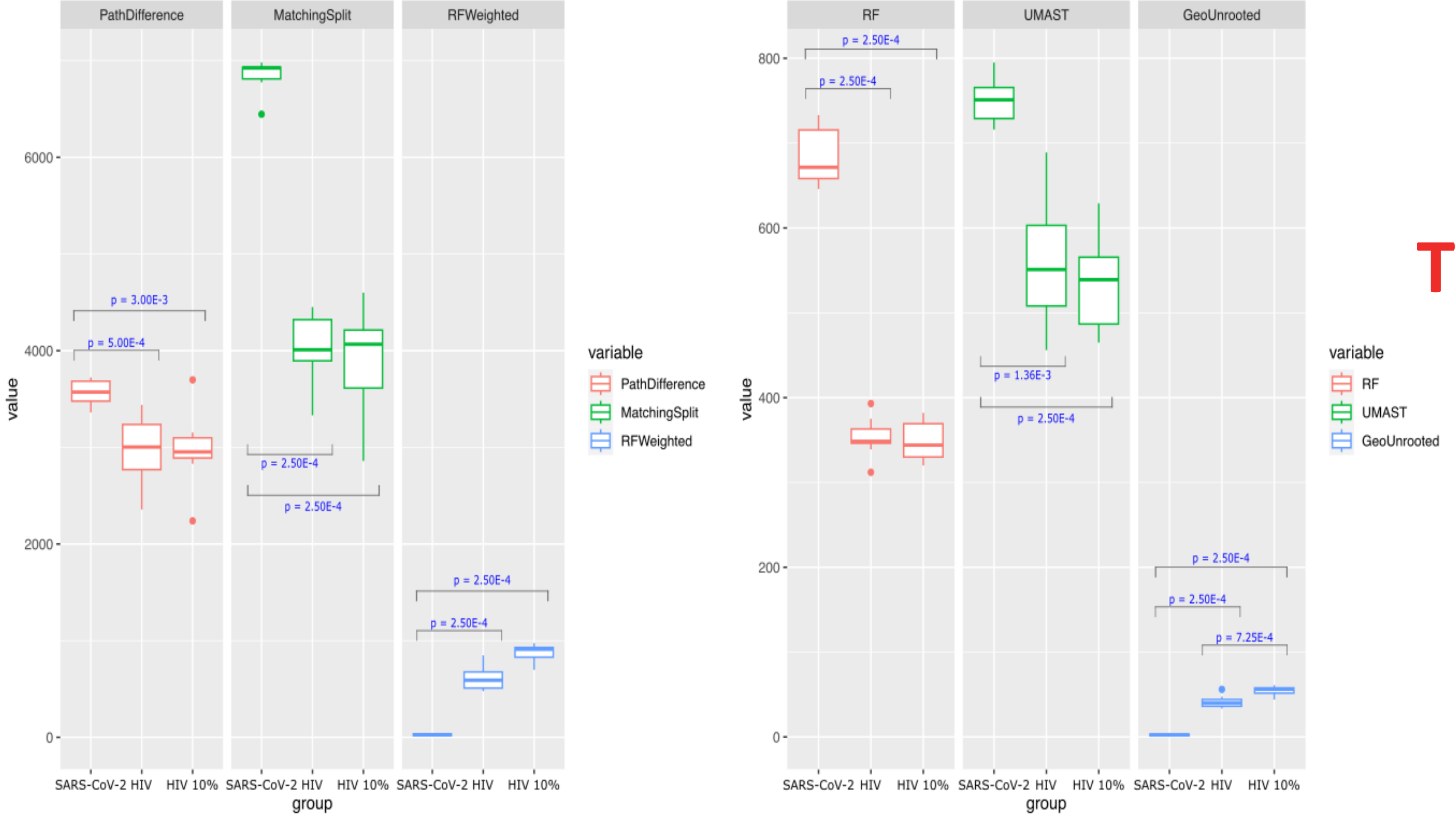


Figure 9: Boxplots from distance metrics calculated by comparing MCC phylogenetic trees obtained from sequential and parallel MCMC runs on simulated HIV data with perfect sampling, simulated HIV data with 10% sampling rate, and SARS-CoV-2 data. Significant p-values from the U-test comparing the distances between the metrics from each dataset are labelled.

Conclusion

Final remarks

Overall experimental run-times were **reduced by almost 84 days** owing to the parallelization

Parameter estimates:

- The sequential MCMC runs provided **no significant advantage** predicting phylogenetic parameters over the parallel MCMC runs in our analyses involving simulated data
- Even in the parameter estimates from SARS-CoV-2 data, **no significant differences** in parameter estimates were observed

Phylogenetic tree predictions:

- Distance metrics calculated in the simulation study also suggested that MCC phylogenetic trees obtained from parallel and sequential MCMC were **mostly similar** except for branch lengths
- In general, our parallelization methodology was less consistent for the real-world SARS-CoV-2 data
- Most of the differences lay in the scaling of the phylogenetic trees and not in the topology of the phylogenetic trees for the simulation dataset whereas in the SARS-CoV-2 dataset, differences were found in the topology as well

Acknowledgements

Supervisor: **Dr. Paul Gordon**

Support from **Dr. Qingrun Zhang** & **Dr. Quan Long** and other lab members

Funding: Alberta Innovates

UNIVERSITY OF CALGARY

Thank you for attending!

Feel free to ask any questions!

David Yang
For any inquiries I may be reached at david.yang1@ucalgary.ca

Methodology

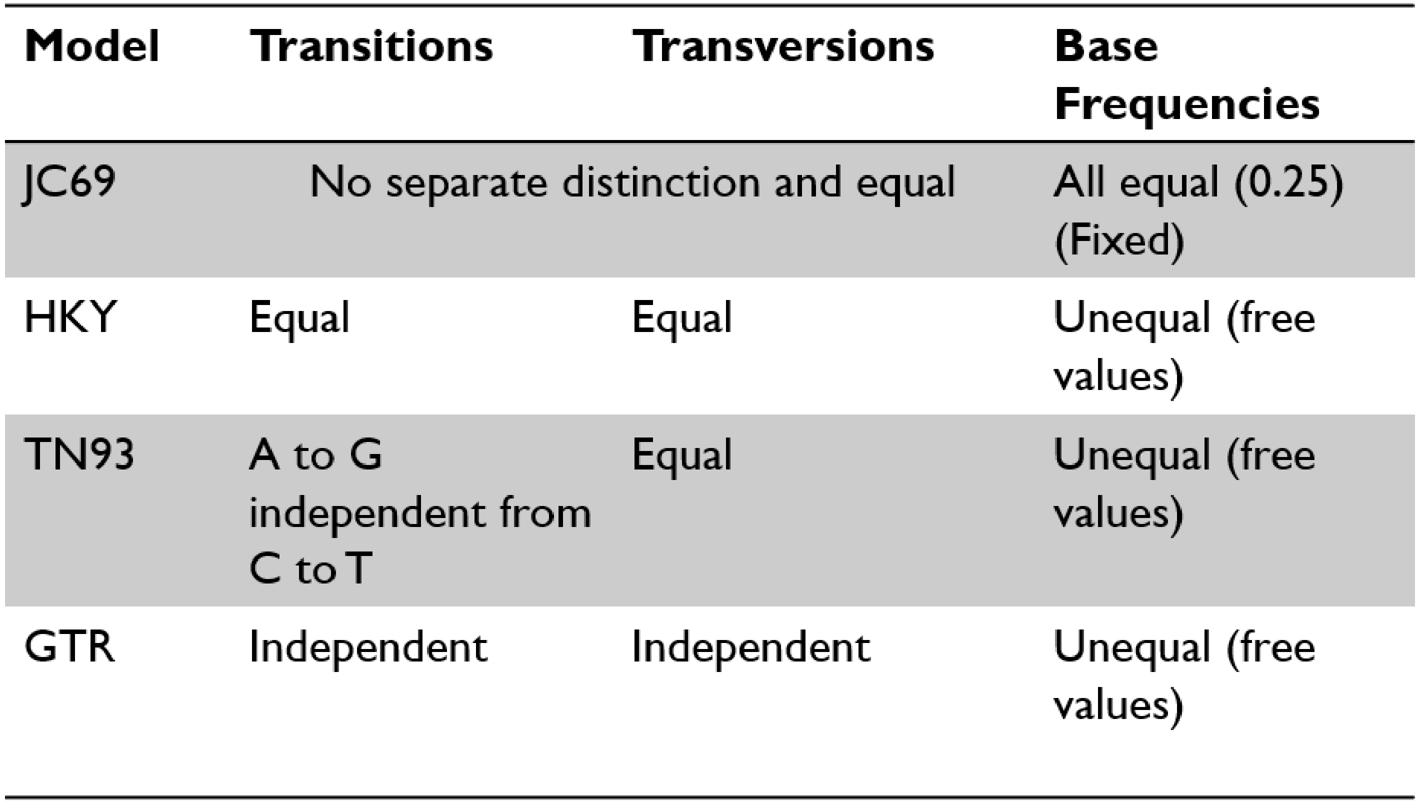
Random sampling

- Randomly select *N* sequences from all sequence data available

Weighted sampling

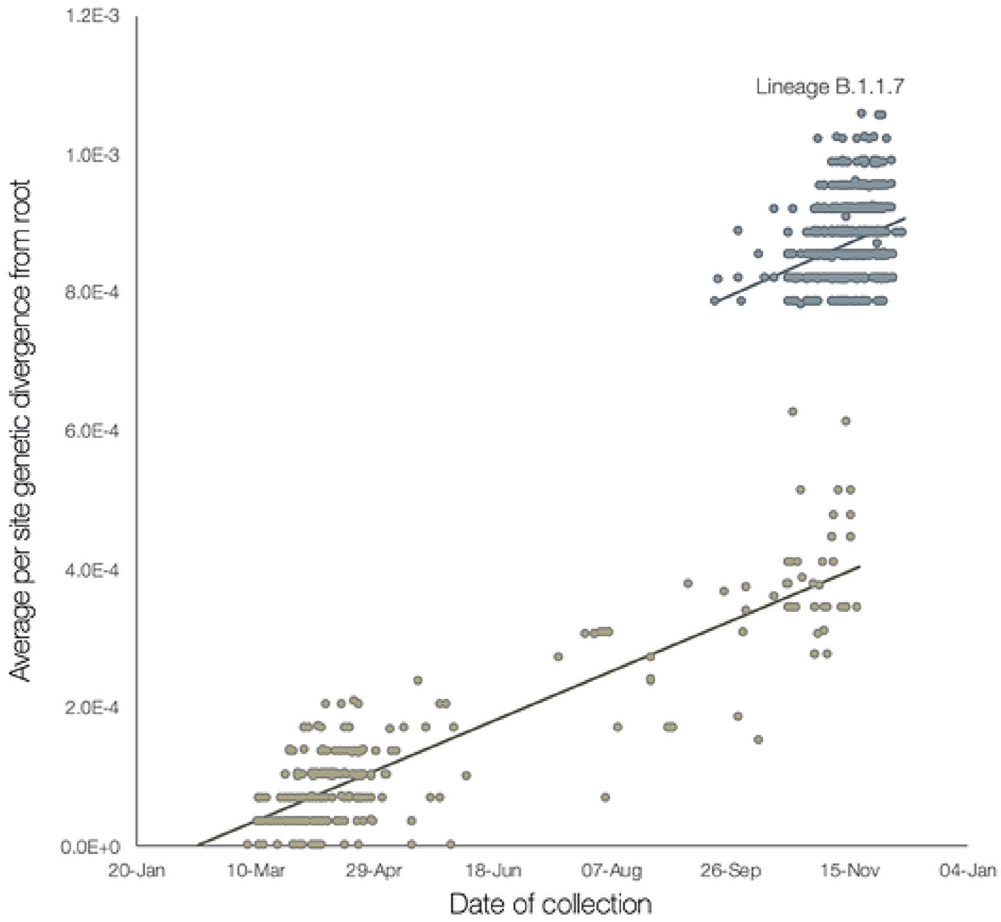
- Select *N* sequences from each country in each month based on SARS-CoV-2 prevalence
- Prevalence estimated with IHME mean estimates

Site models



Model	Transitions	Transversions	Base Frequencies
JC69	No separate distinction and equal		All equal (0.25) (Fixed)
HKY	Equal	Equal	Unequal (free values)
TN93	A to G independent from C to T	Equal	Unequal (free values)
GTR	Independent	Independent	Unequal (free values)

Differences in Clock Rate



(Rambaut et al., 2020)