PRESERVING INTRA-PATIENT VARIANCE IMPROVES PHYLOGENETIC INFERENCE OF HIV TRANSMISSION

AUGUST GUANG

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

BROWN

acknowledgments

My CFAR Research Group:

Casey Dunn, Associate Professor of Evolutionary Biology

Rami Kantor, Associate Professor of Medicine

Mia Coetzer, Assistant Professor of Medicine

Mark Howison, Director of Data Science

My DunnLab Research Group:

Casey Dunn (again)

Zachary Lewis, Postdoc

Catriona Munro, PhD Candidate

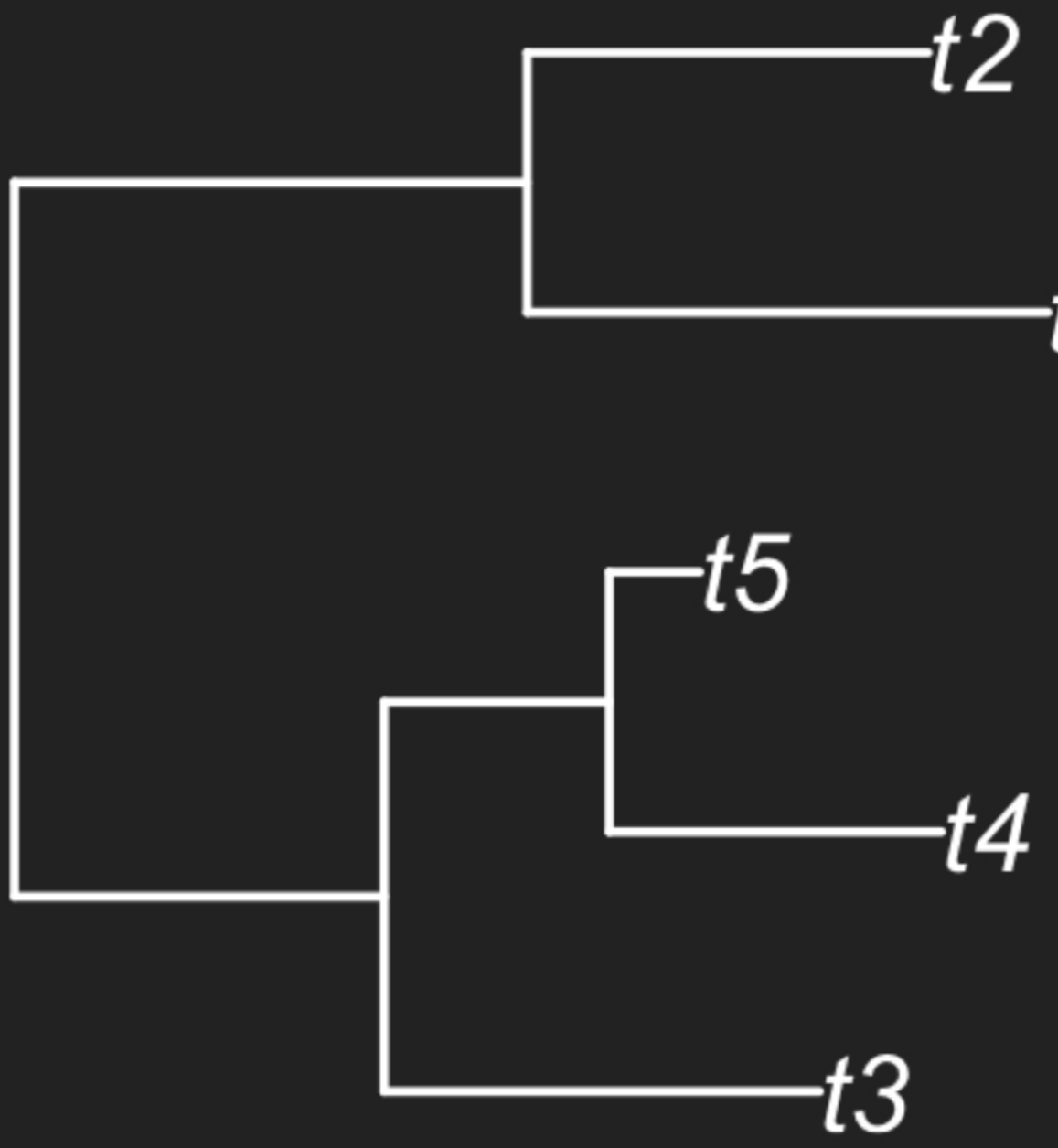
Alex Damian Serrano, PhD Candidate

Colin MacLean, Research Programmer & Charles Lawrence, Professor of Applied Mathematics

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

BROWN

TRANSMISSION NETWORKS AS PHYLOGENIES

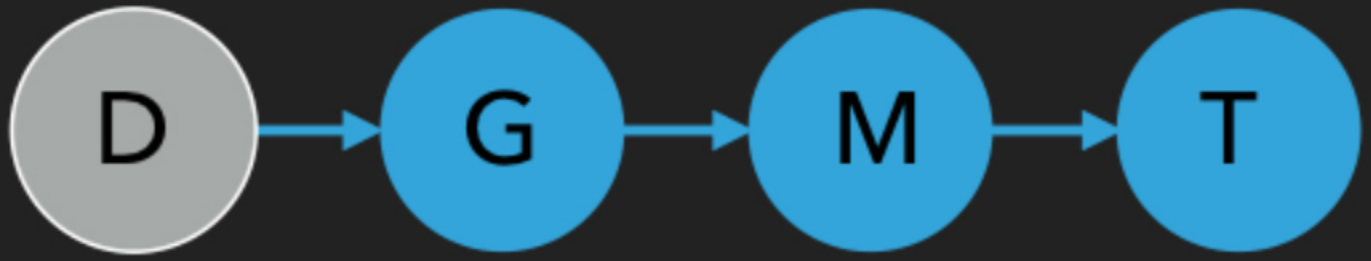


Phylogenetic trees have information about relatedness of organisms at tips

in the absence of reliable patient contact histories, phylogenies can be proxies for transmission networks

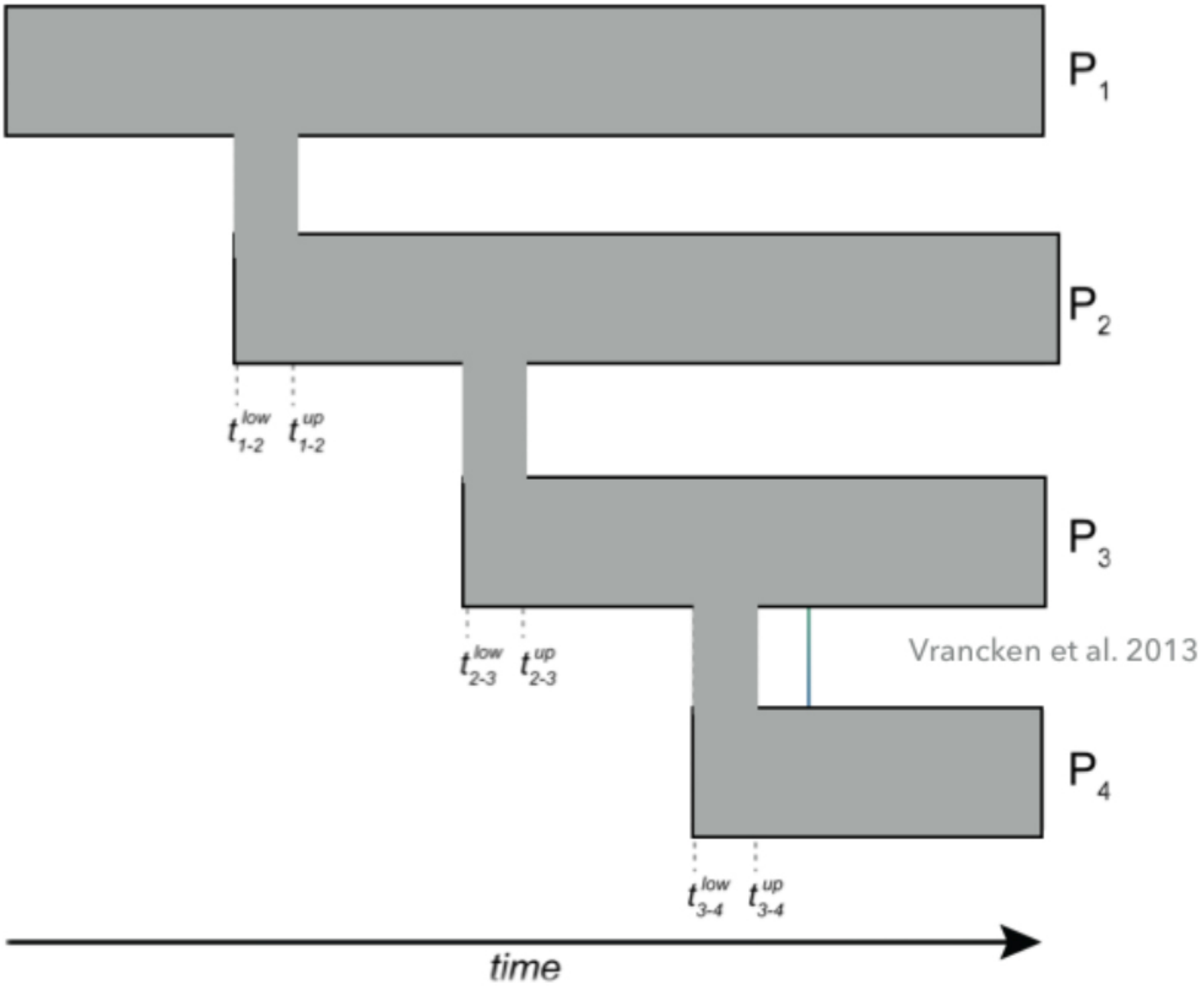
in the absence of reliable patient contact histories, phylogenies can be proxies for transmission networks

the phylogenetic workflow



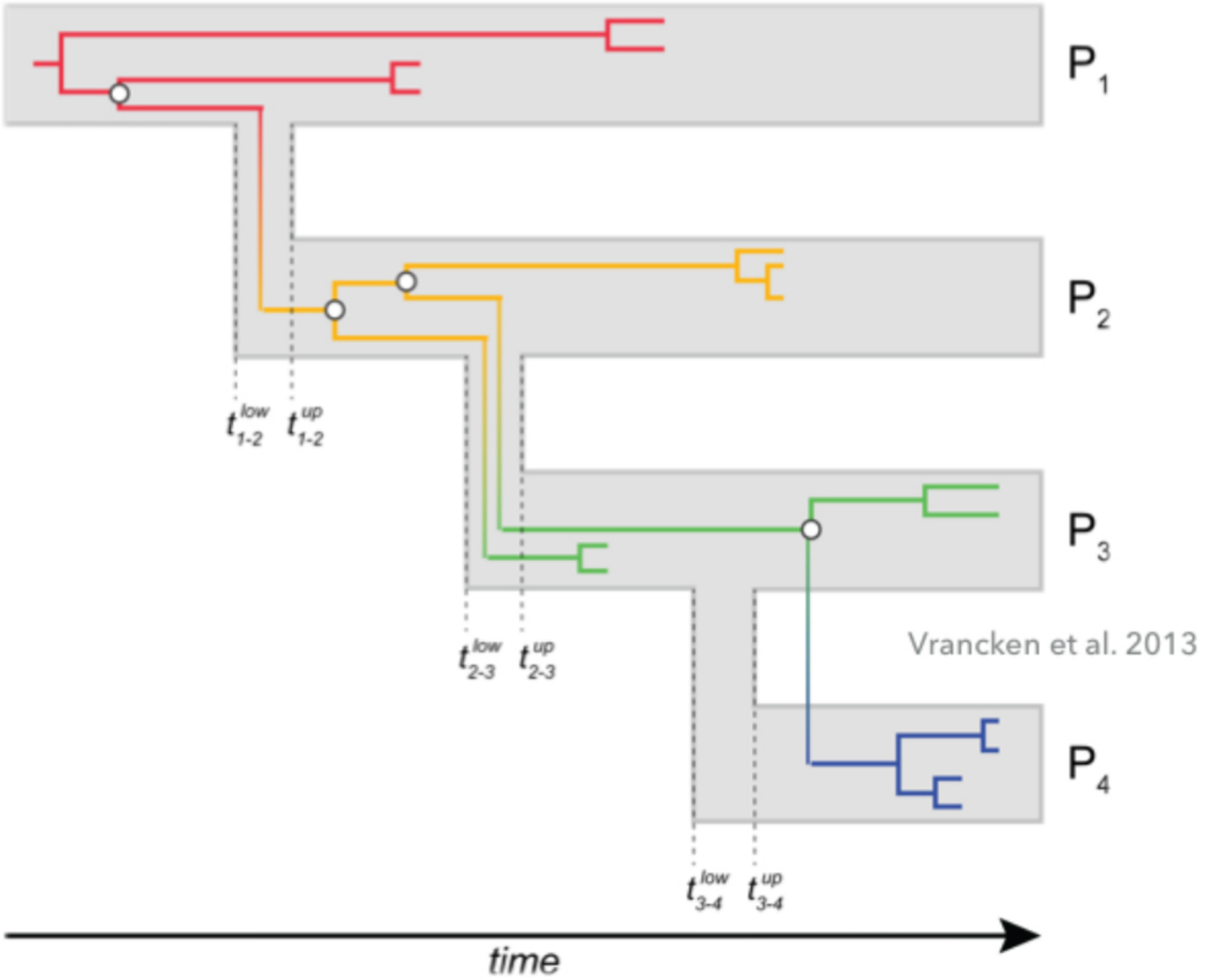
- D: Data: reads
- G: Genomes
- M: Multiple sequence alignment
- T: Transmission tree

we assume the tips of these trees are single entities



Vrancken et al. 2013

but with rare exceptions they are summaries



Vrancken et al. 2013

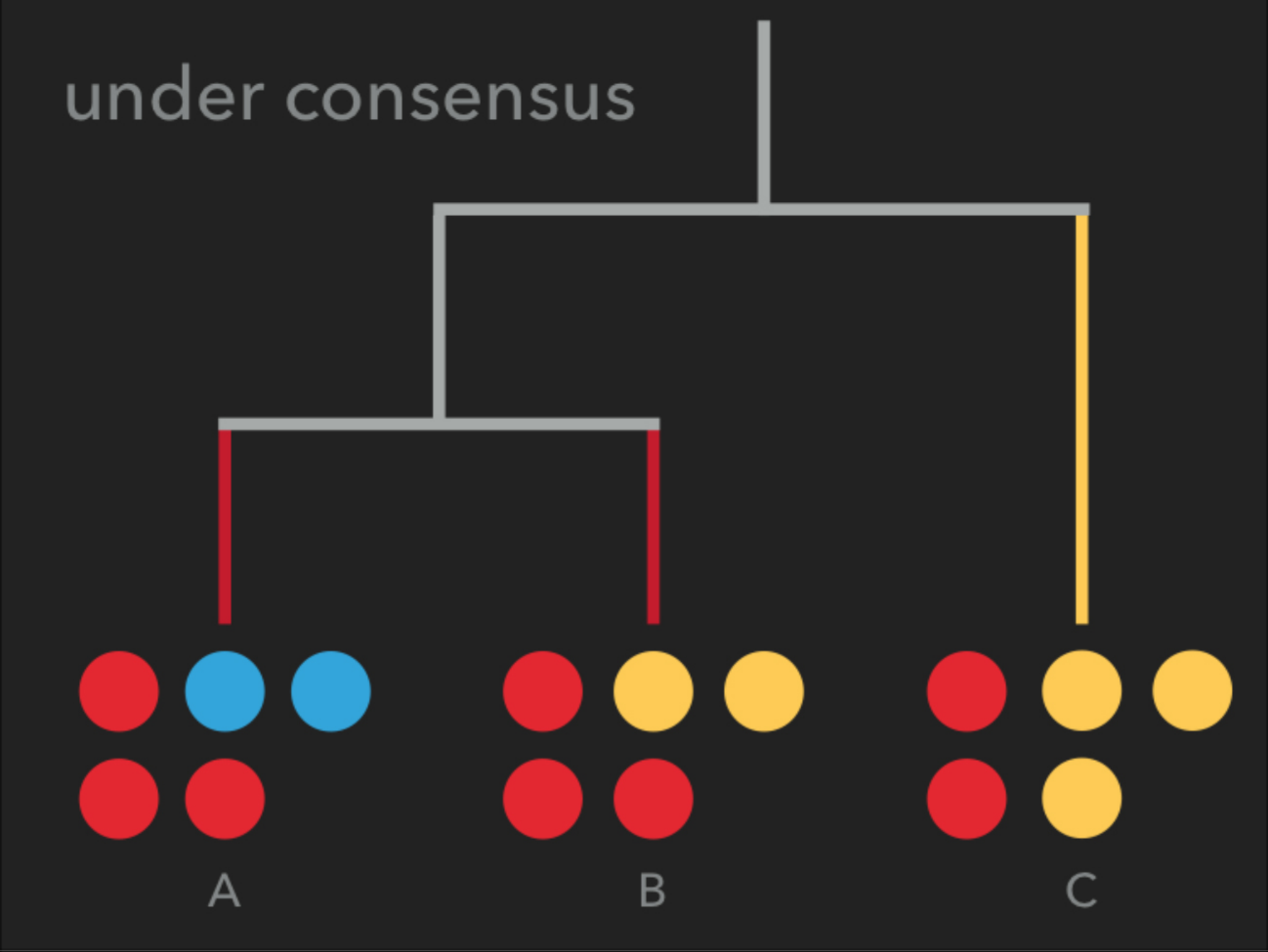
this is not an issue if we suspect that the variation can be summarized by the mode, i.e. consensus genome

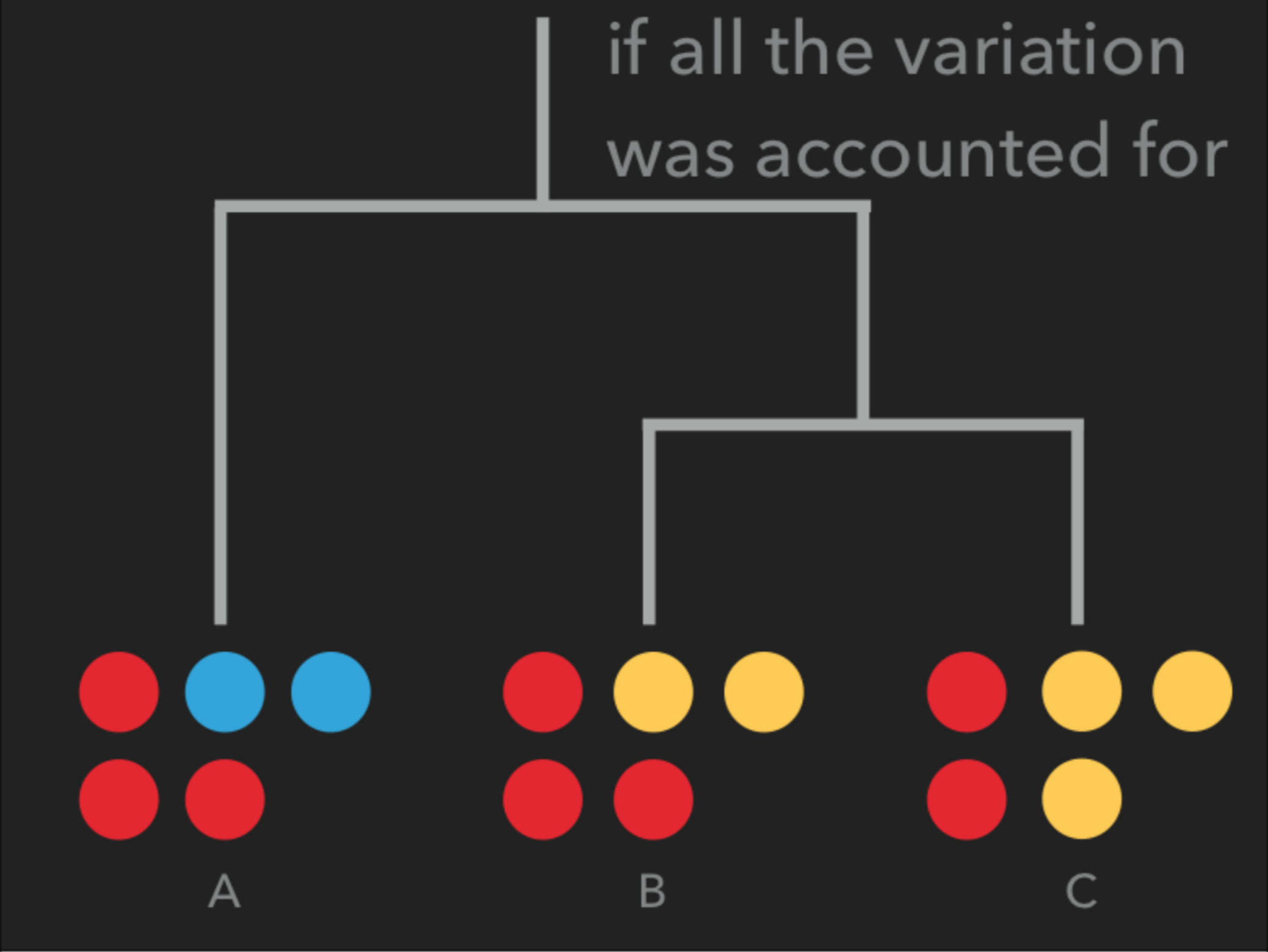
but is it an issue if it cant?

after all, we ultimately care about the transmission tree…

PRESERVING INTRA-PATIENT VARIATION IS IMPORTANT

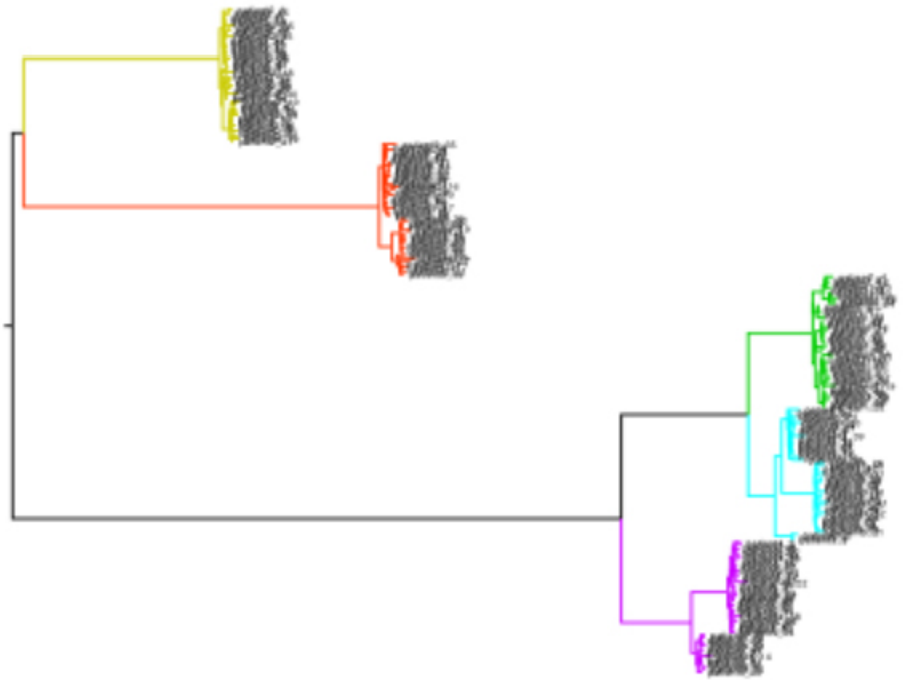
a simple thought experiment


under consensus
A
B
C


if all the variation
was accounted for
A
B
C

some simple simulations

RAxML tree from consensus genome



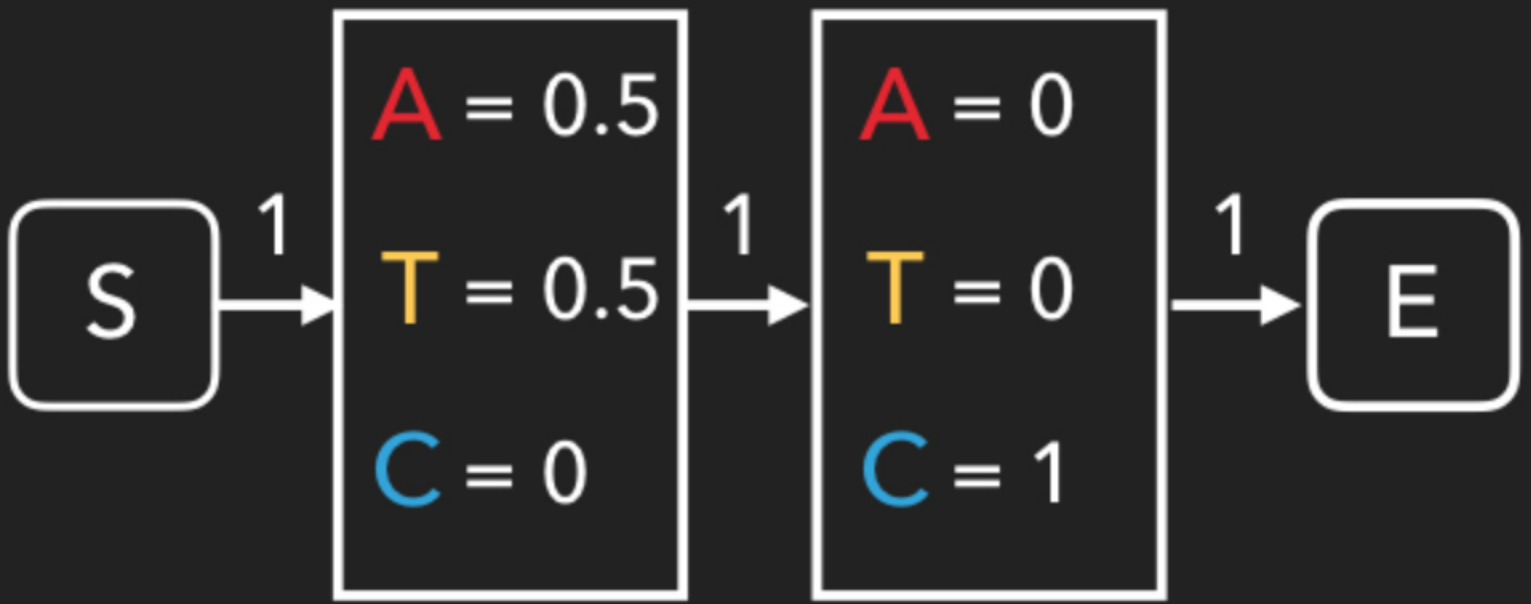
true tree

so we don't recover the same tree...

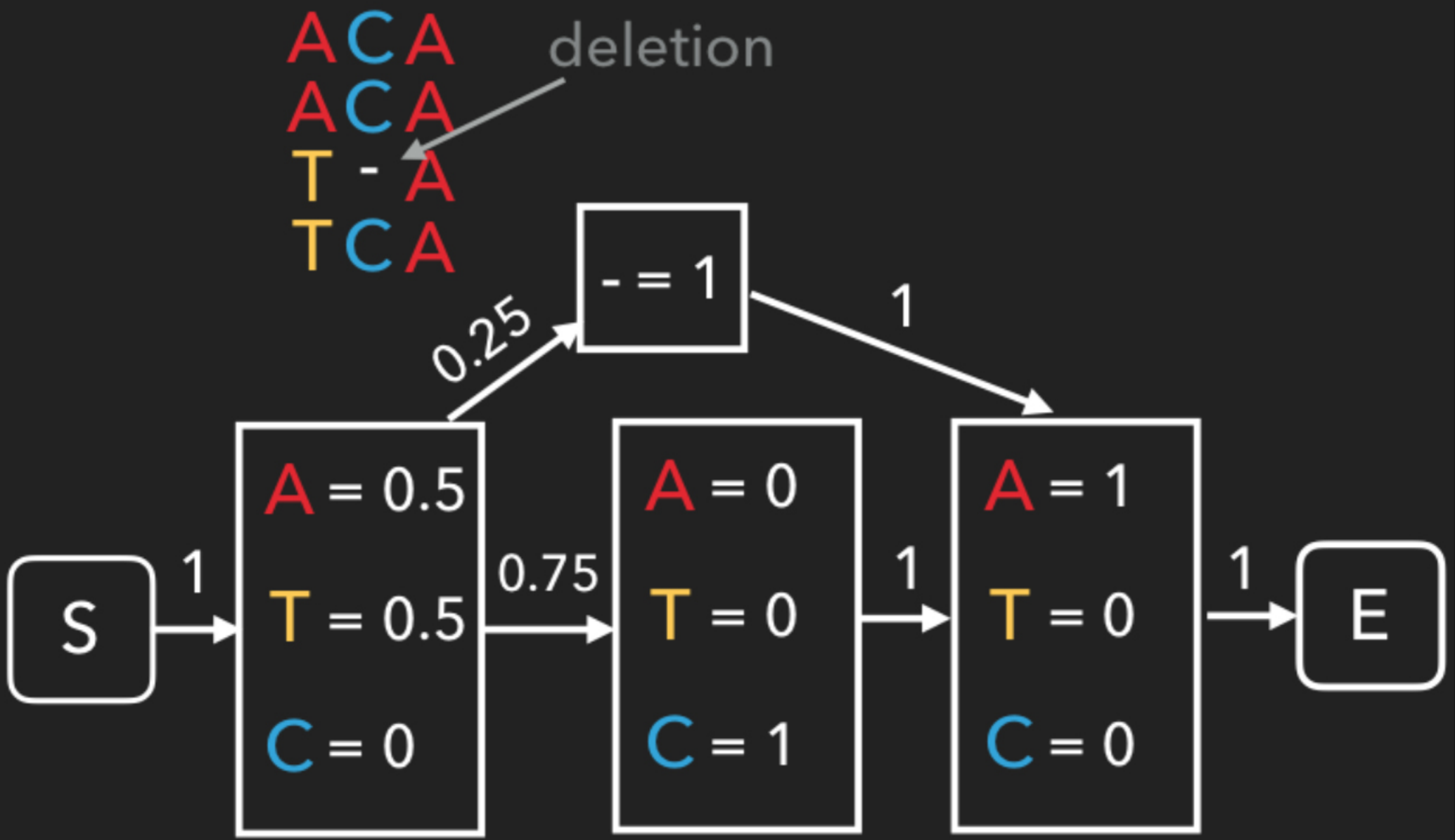
let's try to account for that variation!

profile Hidden Markov Models

A C
A C
T C
T C



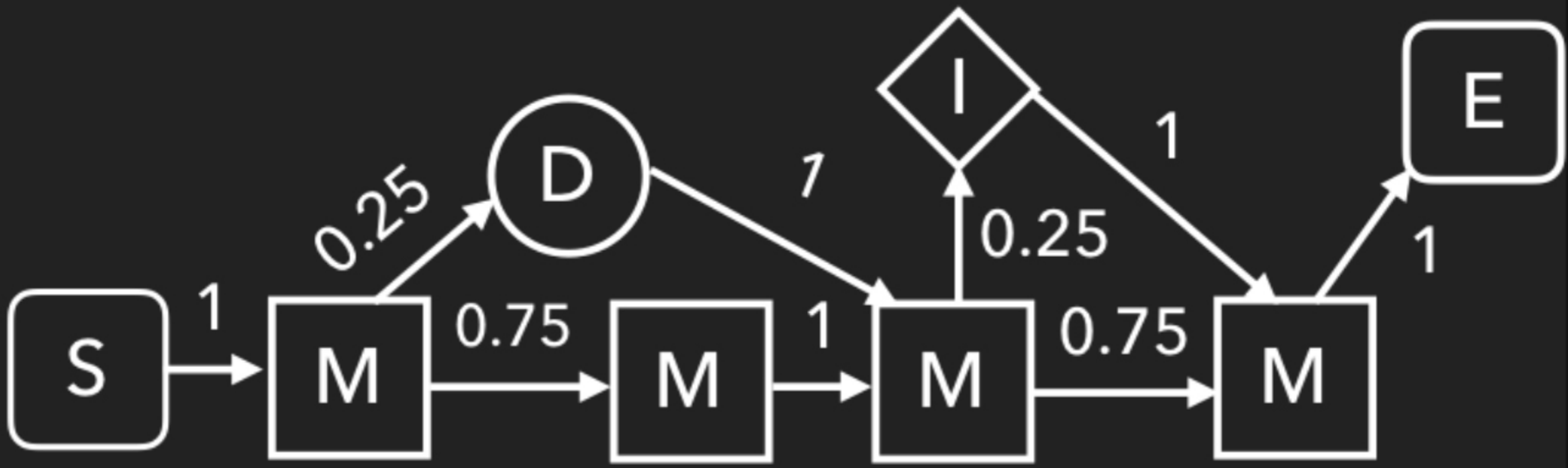
indels are hidden states



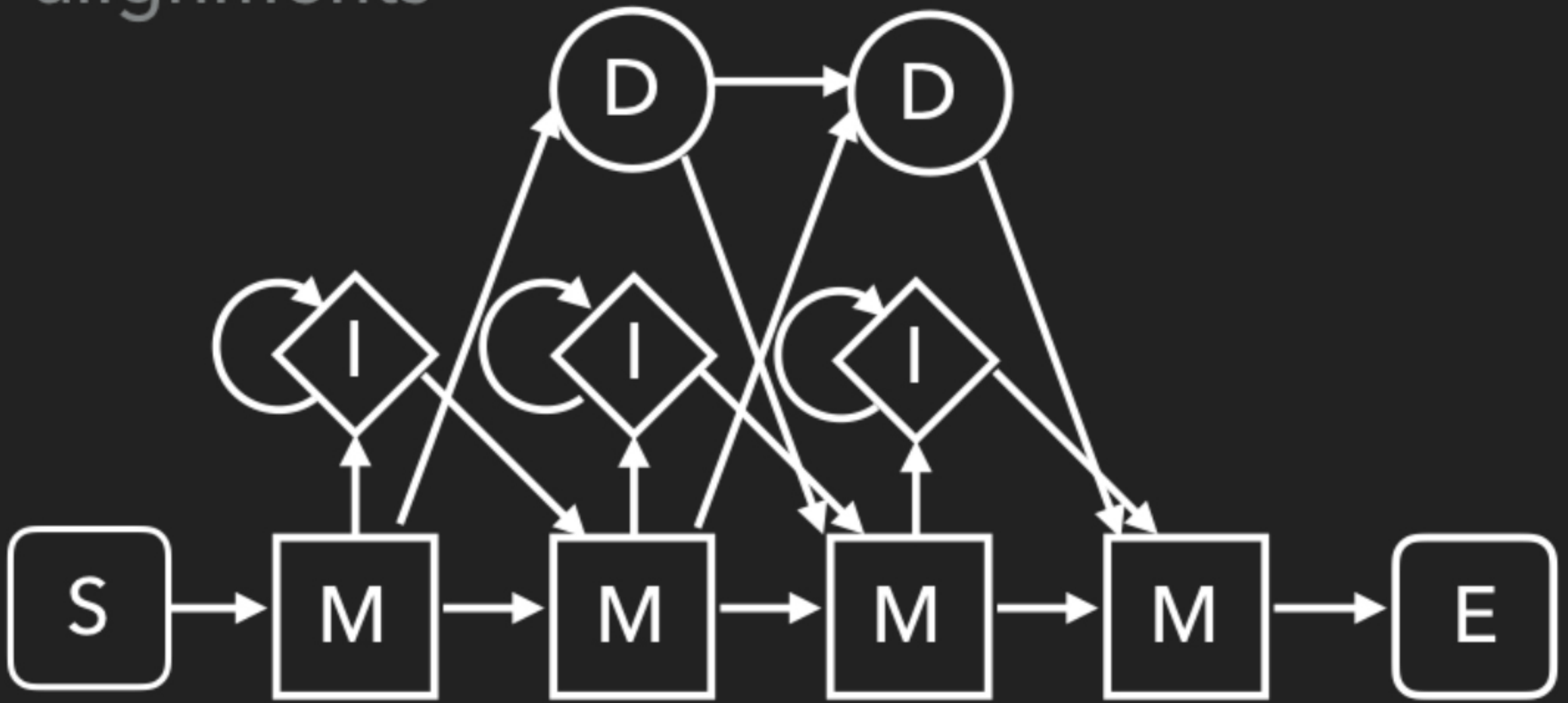
indels are hidden states

```
ACAAT
ACA  T
T-A  T
TCA  T
```

insertion



we can build read profiles from read alignments



sample genomes from those profiles

A1 ACAATGACAATGGCAA
A2 ACATGAACTGGCA
B1 TATGAAATGGCAA
B2 TCATGAACTGGCA
C1 TATGACAATGGCAA
C2 TCATGAACTGGCA

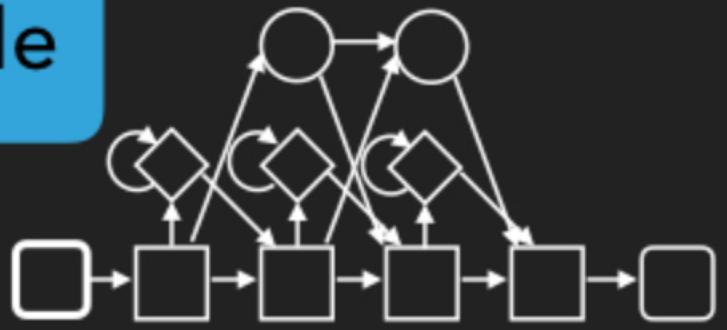
and build alignments and trees
with those samples

```
A1 ACAATGACAATGGCAA
B1 T  A  TGA  AATGGCAA
C1 T  A  TGACAATGGCAA
```

```
A2 ACATGAACTGGCA
B2 TCATGAACTGGCA
C2 TCATGAACTGGCA
```

our "synthetic" approach

Read Profile



simulate n sequences

G

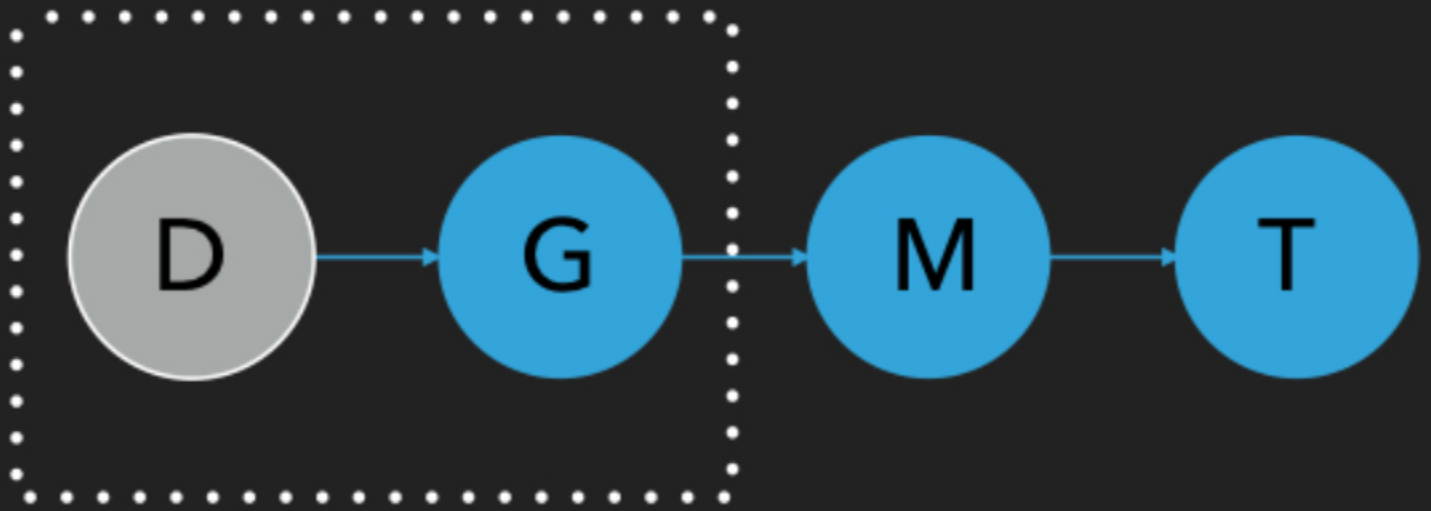
M

T

BLUE WATERS

SUSTAINED PETASCALE COMPUTING

each step in this workflow is a high-dimensional inference problem…

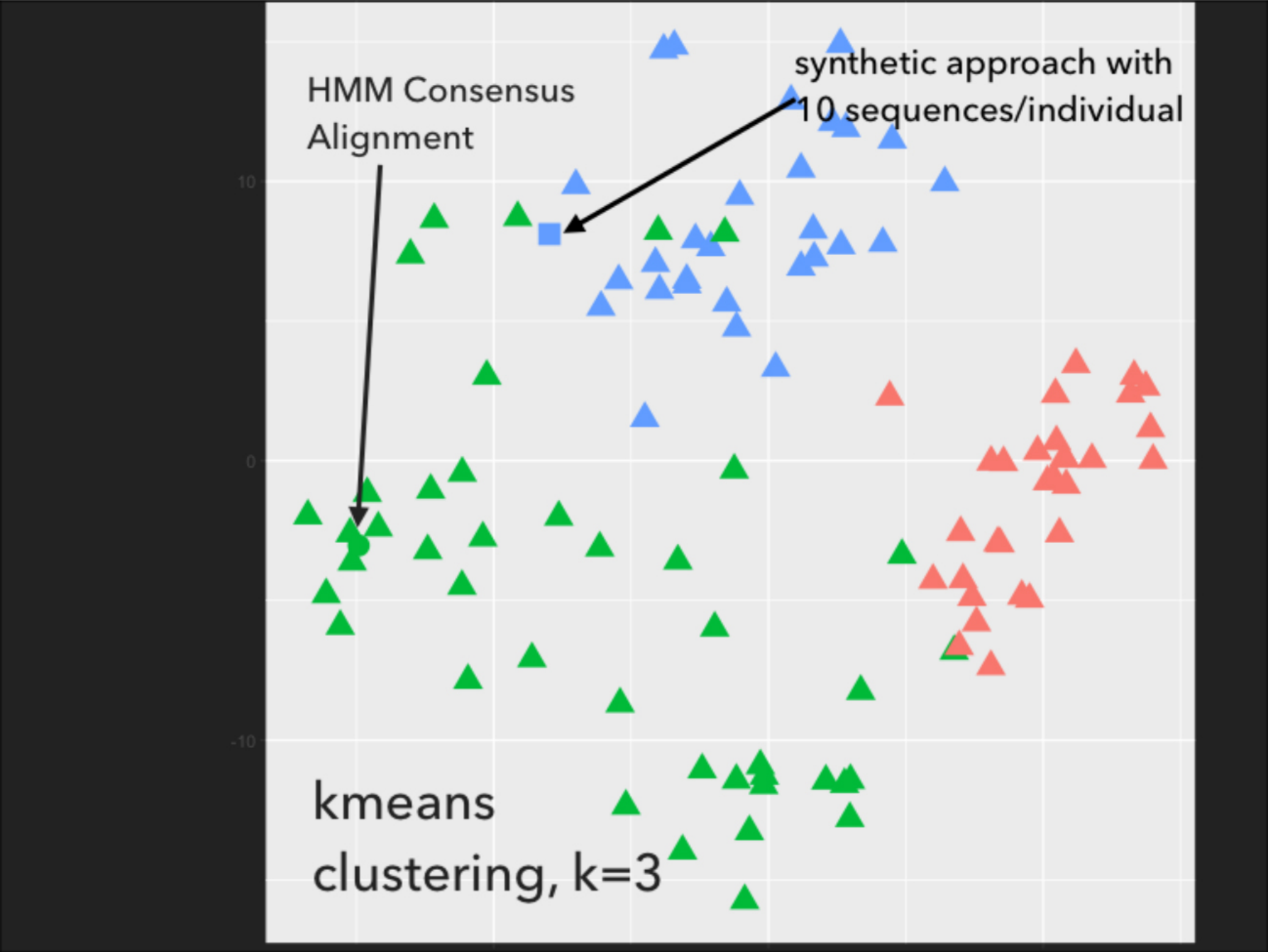


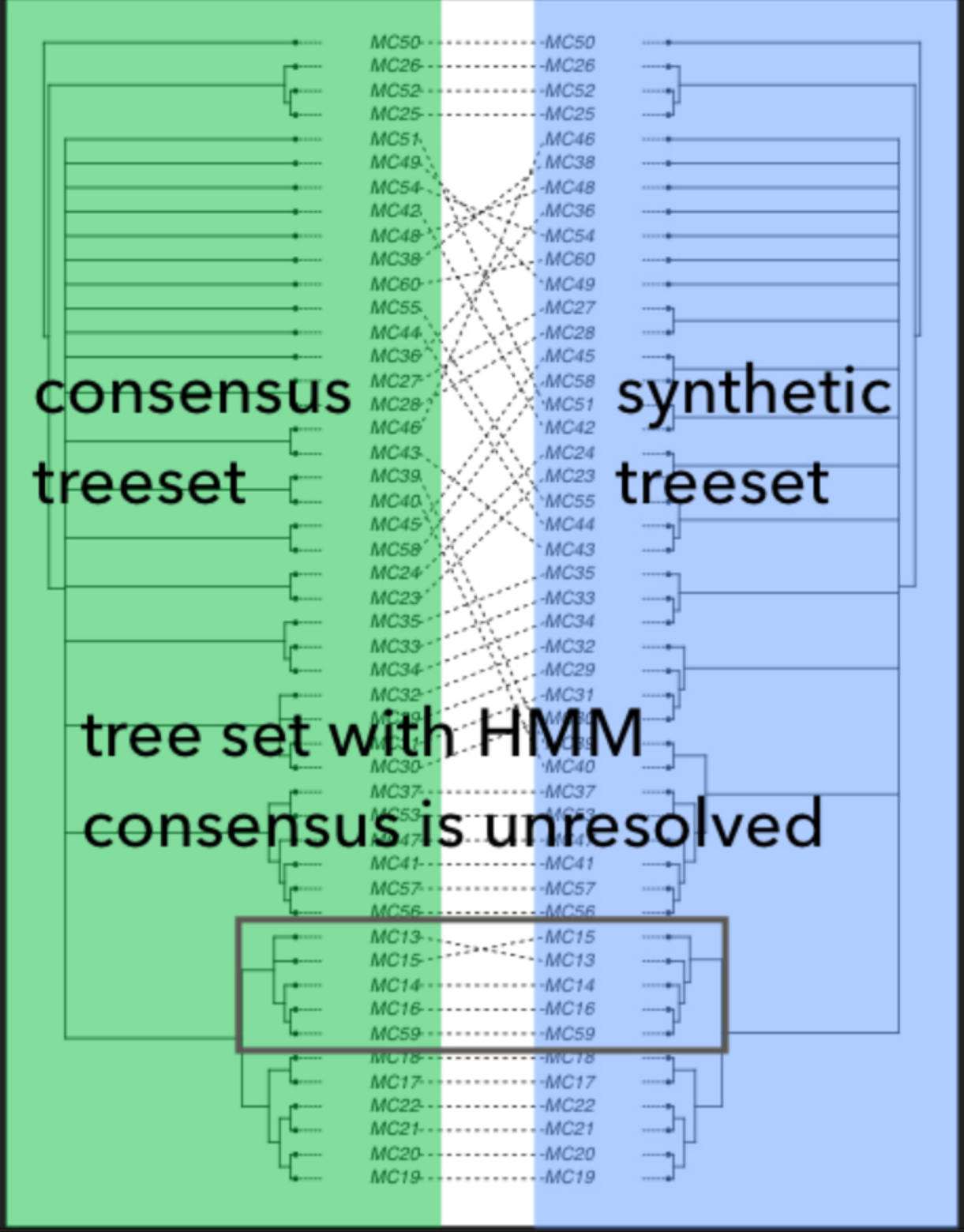
- D->G: HMMer
- G->M: mafft
- M->T: RAxML or MrBayes

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

HIV DATASET

- Individuals newly-diagnosed with HIV in 2013
- Knew transmission history for 5 individuals
- Ran consensus approach; synthetic approach with 10 sequences/individual (collapsed tree); 100 runs of synthetic approach with 1 sequence/individual
- Computed Robinson-Foulds distance between trees from all approaches and performed Multidimensional Scaling


HMM Consensus
Alignment
synthetic approach with
10 sequences/individual
10
0
-10
kmeans
clustering, k=3


consensus
treeset
synthetic
treeset
tree set with HMM
consensus is unresolved

TRANSMISSIM (SIMULATED TRANSMISSION NETWORKS, PHYLOGENIES, GENOMES, READS)

a generative model of patient reads from transmission events

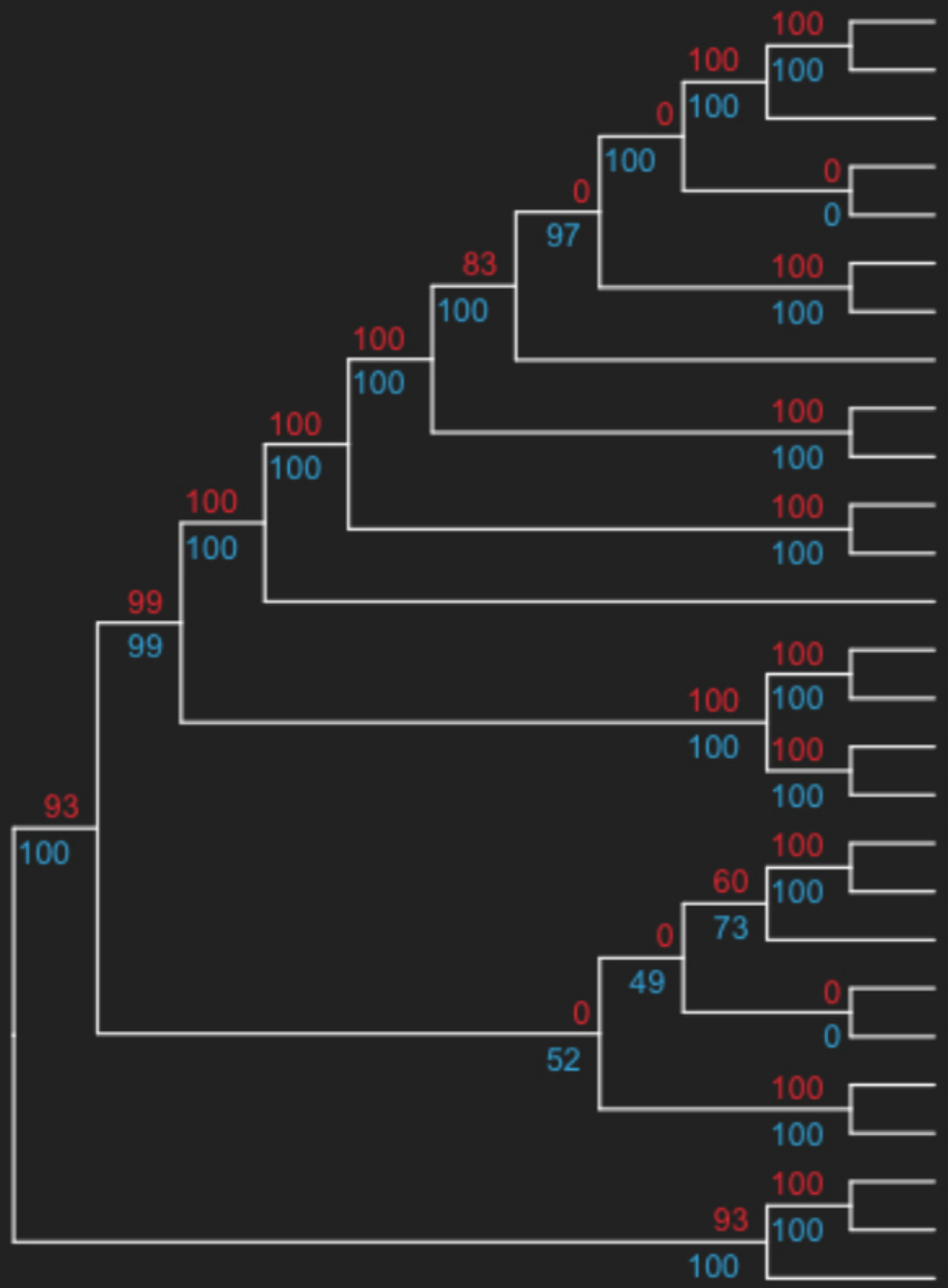
- N: Transmission network
- T: Transmission tree
- V: Viral phylogeny
- G: Genomes
- D: Data: reads

a generative model of patient reads from transmission events

- N: outbreaker
- T: binary mapping
- V: SimPhy
- G: pyvolve
- D: ART

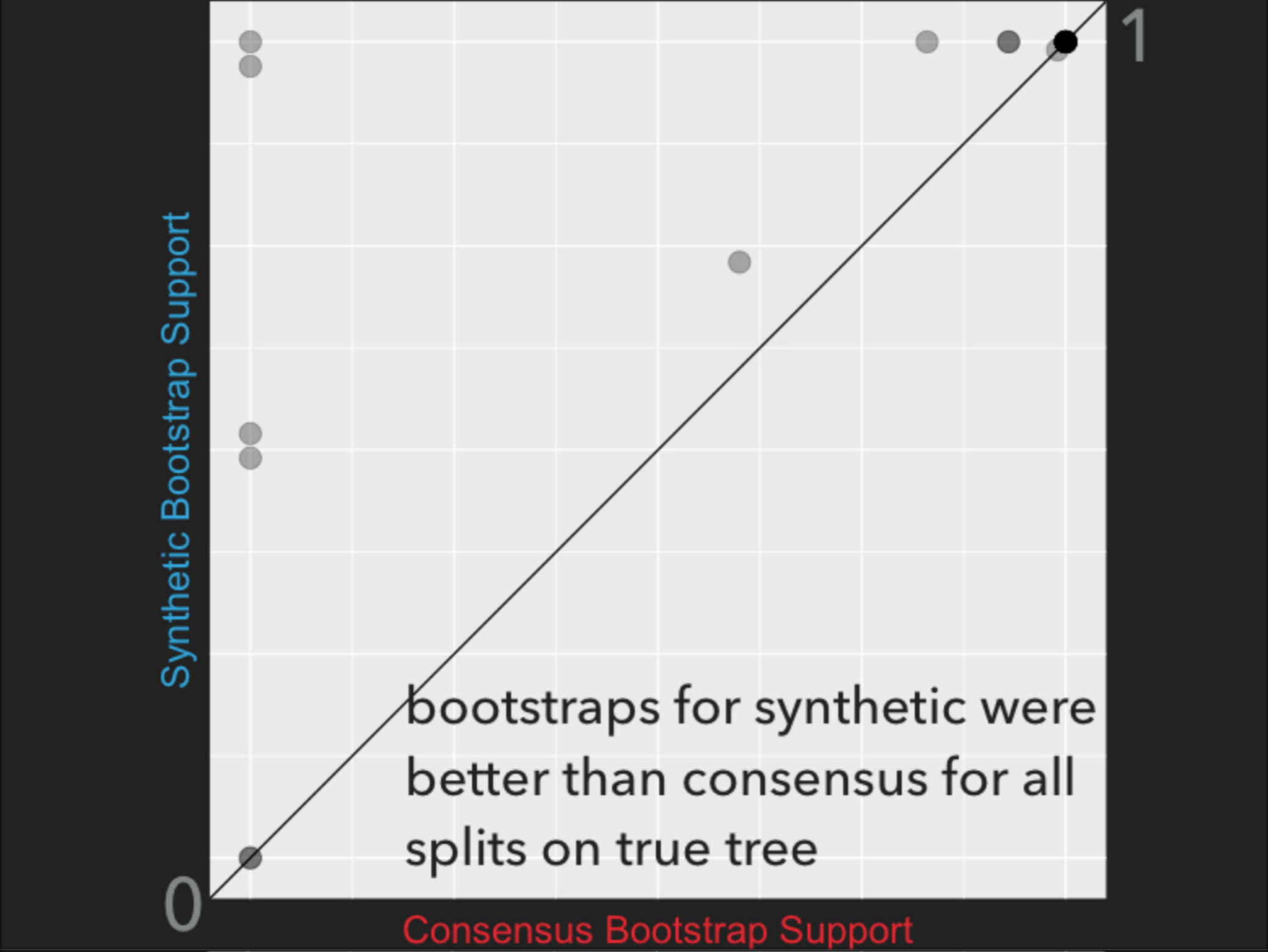
BLUE WATERS
SUSTAINED PETASCALE COMPUTING

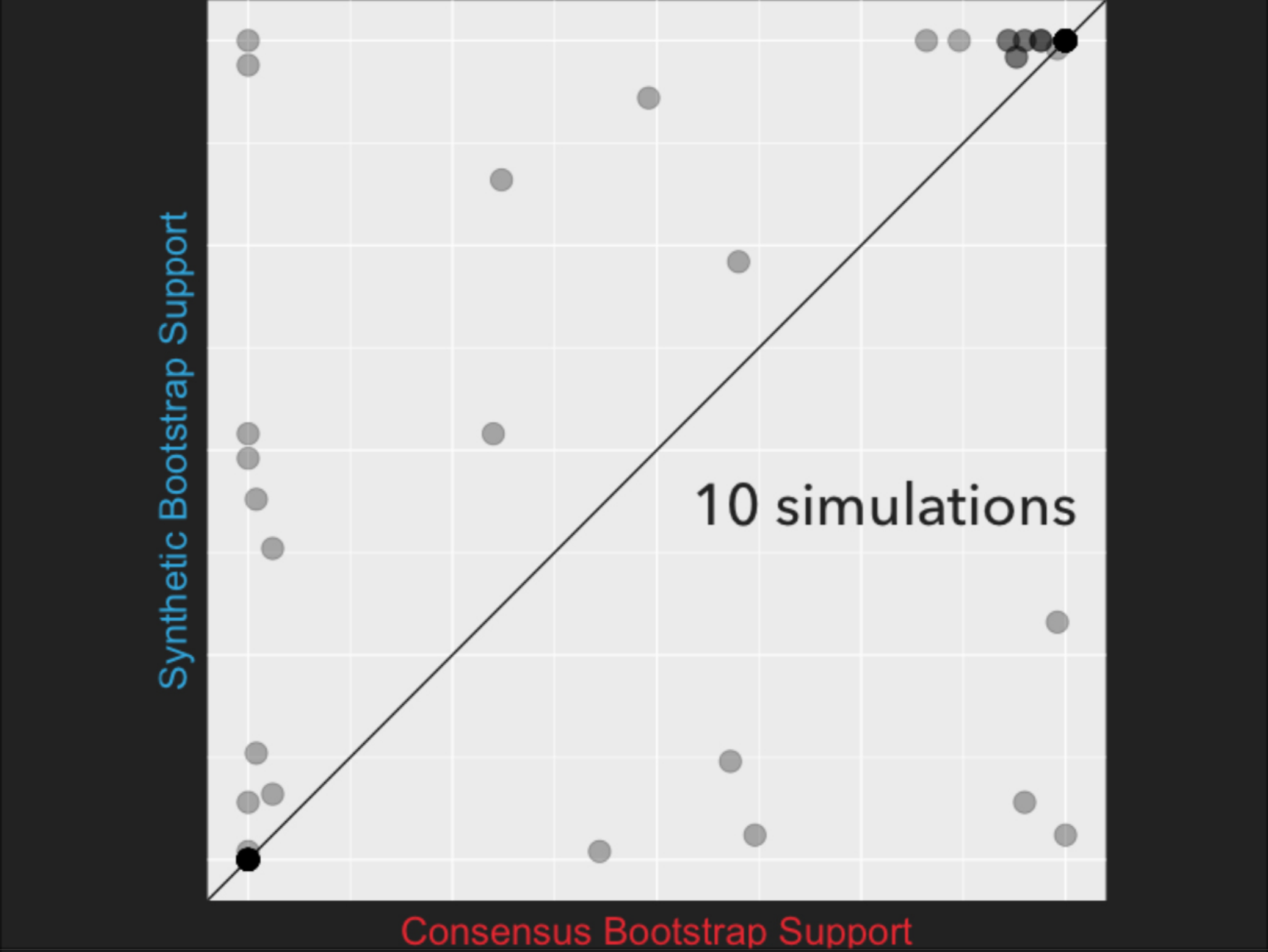
bootstrap results on simulations



red = consensus

blue = synthetic


Synthetic Bootstrap Support
0
1
Consensus Bootstrap Support
bootstraps for synthetic were better than consensus for all splits on true tree


Synthetic Bootstrap Support
Consensus Bootstrap Support
10 simulations

WHY BLUE WATERS?

1. 10,000 node hours: Create MDS density plot with 10,000 trees to look for conclusive regions of variation
2. remaining node hours: additional simulations for validation with parameter sweeps

BLUE WATERS PRODUCTS

- Reproducible manuscript and figures: https://bitbucket.org/aguang/ms_hiv
- Transmissim: https://github.com/aguang/transmissim

acknowledgments

My CFAR Research Group:

Casey Dunn, Associate Professor of Evolutionary Biology

Rami Kantor, Associate Professor of Medicine

Mia Coetzer, Assistant Professor of Medicine

Mark Howison, Director of Data Science

My DunnLab Research Group:

Casey Dunn (again)

Zachary Lewis, Postdoc

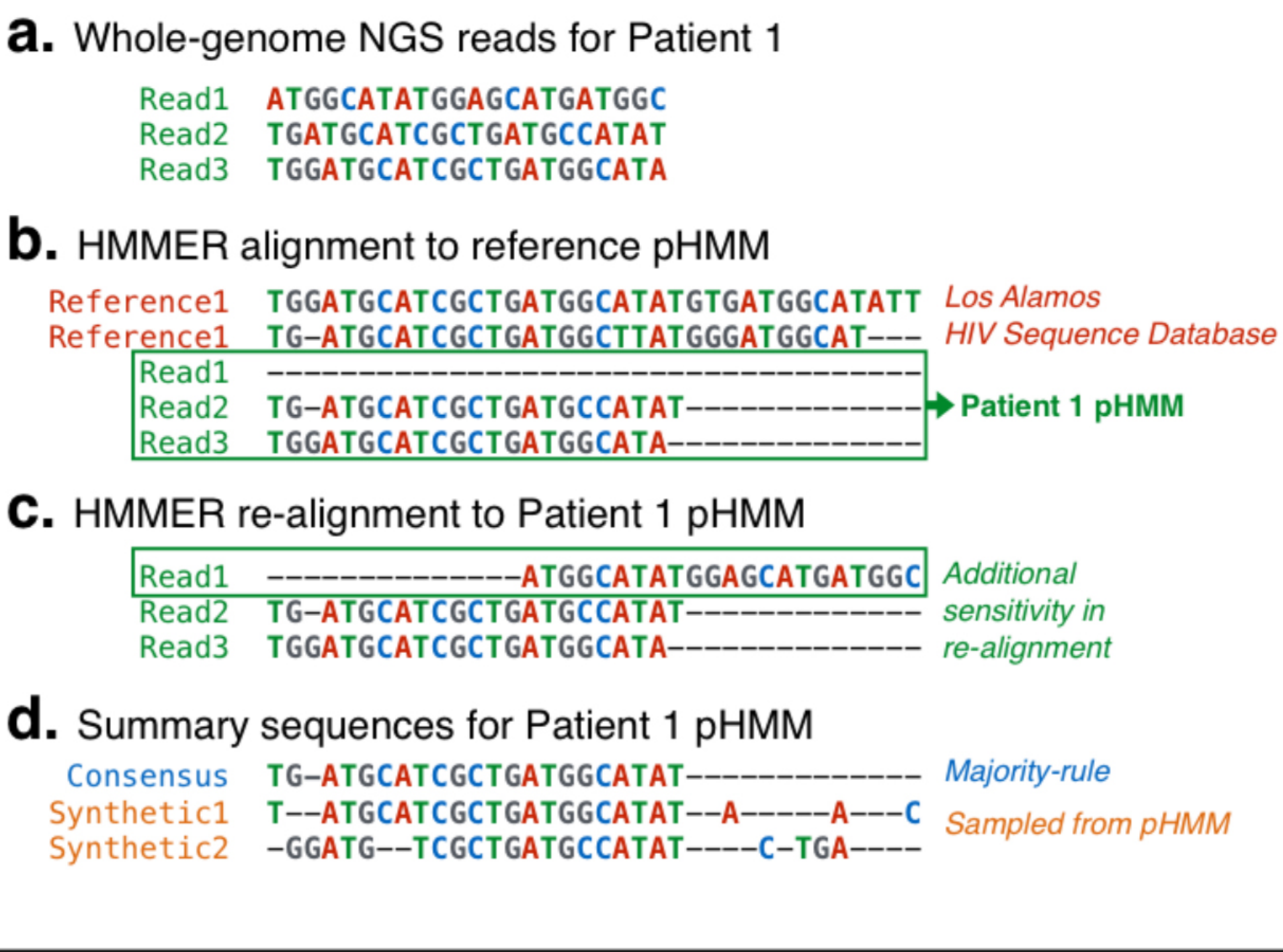
Catriona Munro, PhD Candidate

Alex Damian Serrano, PhD Candidate

Colin MacLean, Research Programmer & Charles Lawrence, Professor of Applied Mathematics

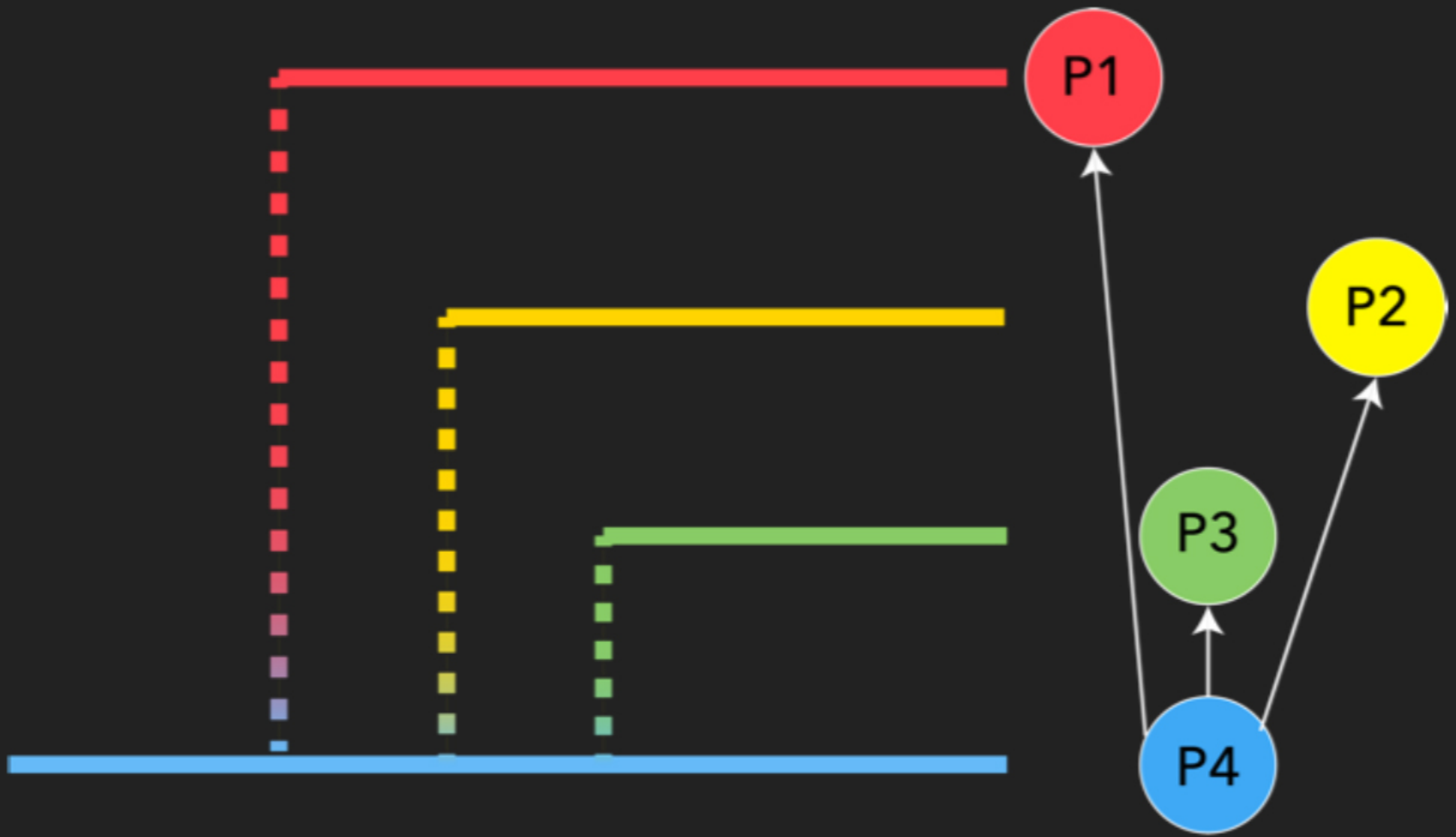
BLUE WATERS
SUSTAINED PETASCALE COMPUTING

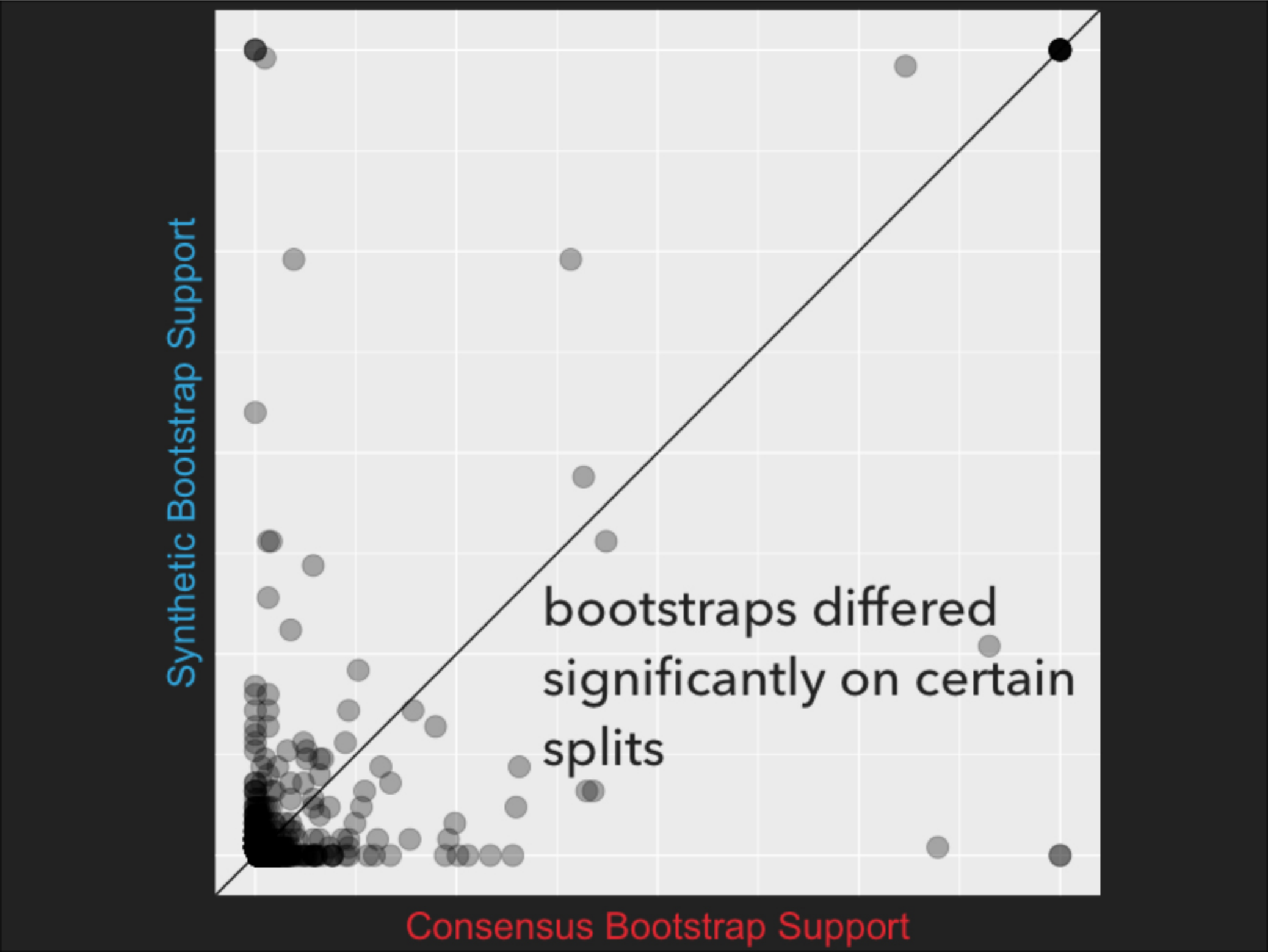
BROWN

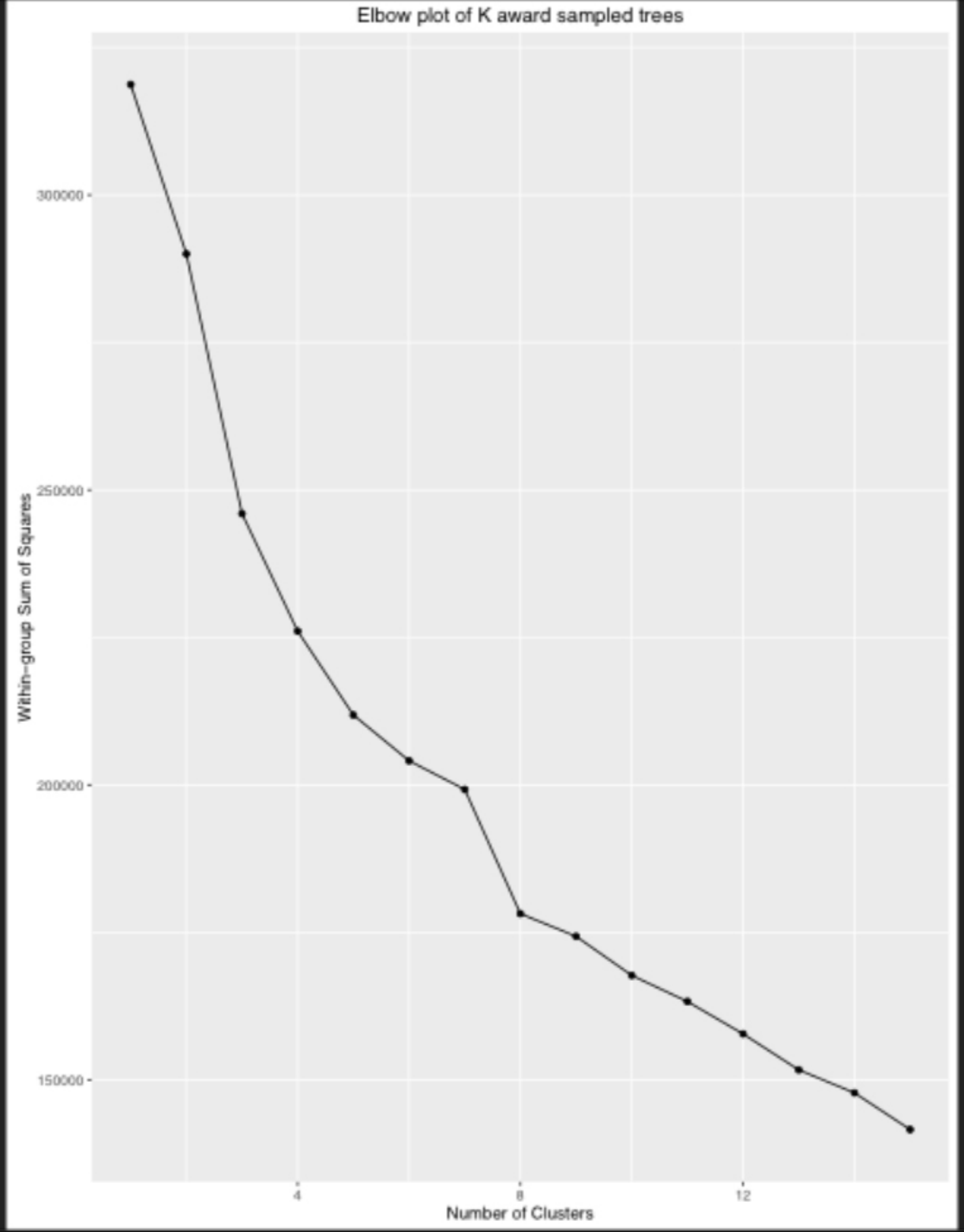

a. Whole-genome NGS reads for Patient 1
Read1 ATGGCATATGGAGCATGATGGC
Read2 TGATGCATCGCTGATGCCATAT
Read3 TGGATGCATCGCTGATGGCATA
b. HMMER alignment to reference pHMM
Reference1 TGGATGCATCGCTGATGGCATATGTGATGGCATATT
Reference1 TG-ATGCATCGCTGATGGCTTATGGGATGGCAT---
Read1 ----------
Read2 TG-ATGCATCGCTGATGCCATAT----------
Read3 TGGATGCATCGCTGATGGCATA----------
Los Alamos HIV Sequence Database
Patient 1 pHMM
c. HMMER re-alignment to Patient 1 pHMM
Read1 ----------ATGGCATATGGAGCATGATGGC
Read2 TG-ATGCATCGCTGATGCCATAT----------
Read3 TGGATGCATCGCTGATGGCATA----------
Additional sensitivity in re-alignment
d. Summary sequences for Patient 1 pHMM
Consensus TG-ATGCATCGCTGATGGCATAT----------
Synthetic1 T--ATGCATCGCTGATGGCATAT--A-----A---C
Synthetic2 -GGATG--TCGCTGATGCCATAT----C-TGA----
Majority-rule
Sampled from pHMM

under certain assumptions, transmission networks have a surjective mapping to a phylogeny (transmission tree)

however, the mapping is not injective and thus not one-to-one

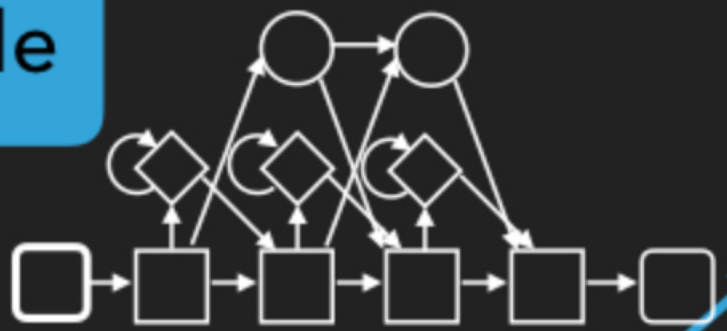



Synthetic Bootstrap Support
Consensus Bootstrap Support
bootstraps differed significantly on certain splits


Elbow plot of K award sampled trees
300000
250000
200000
150000
Within−group Sum of Squares
4
8
12
Number of Clusters

our "synthetic" approach

Read Profile



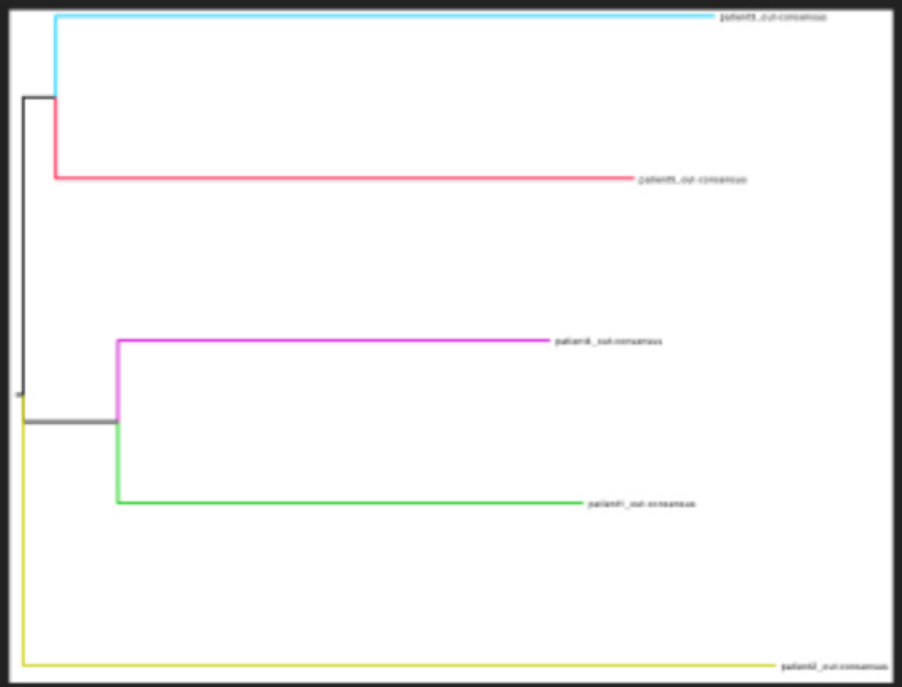
align reads

build profile

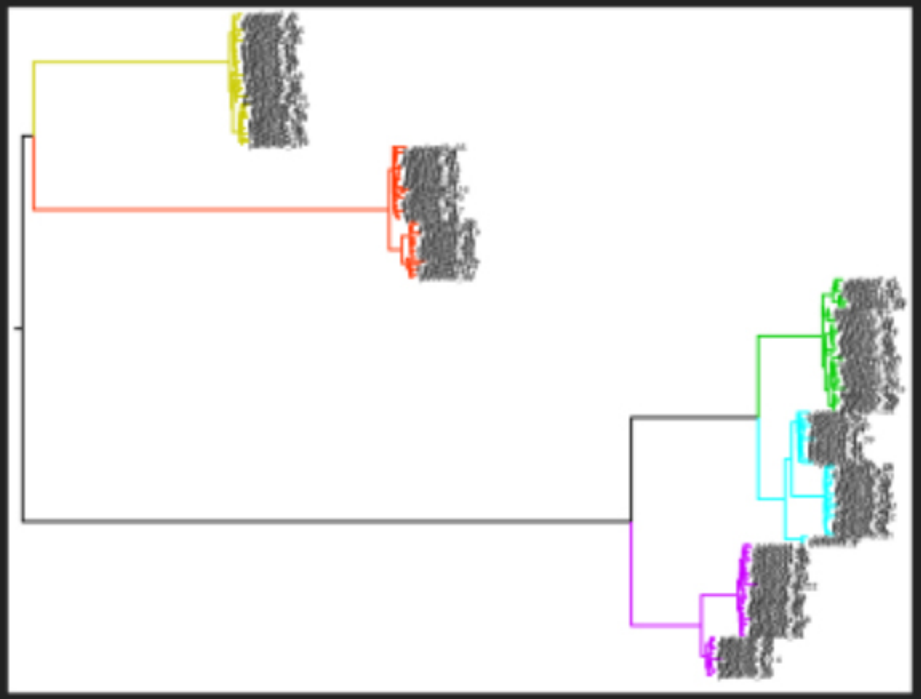
realign reads

Reads

some simple simulations

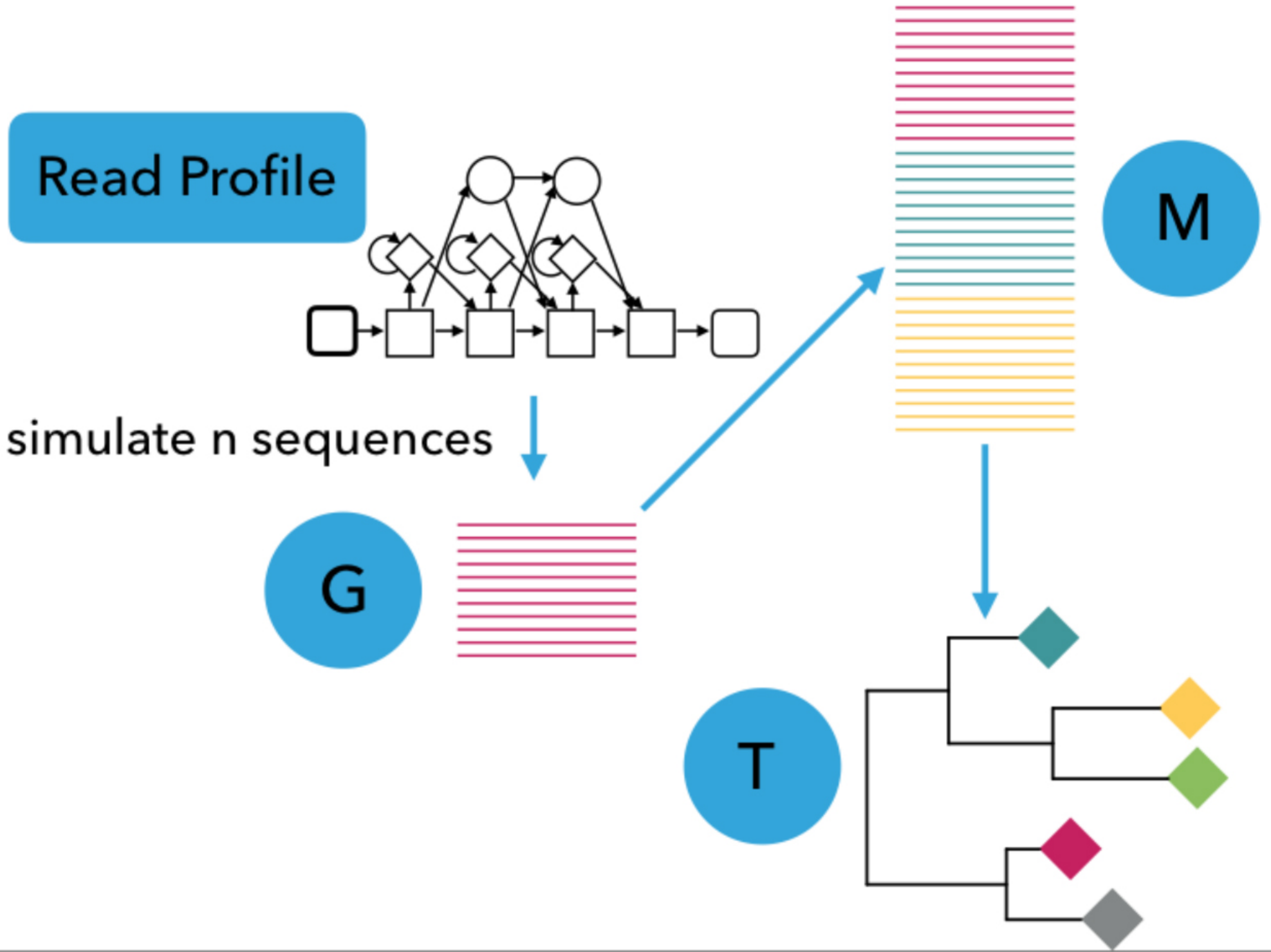


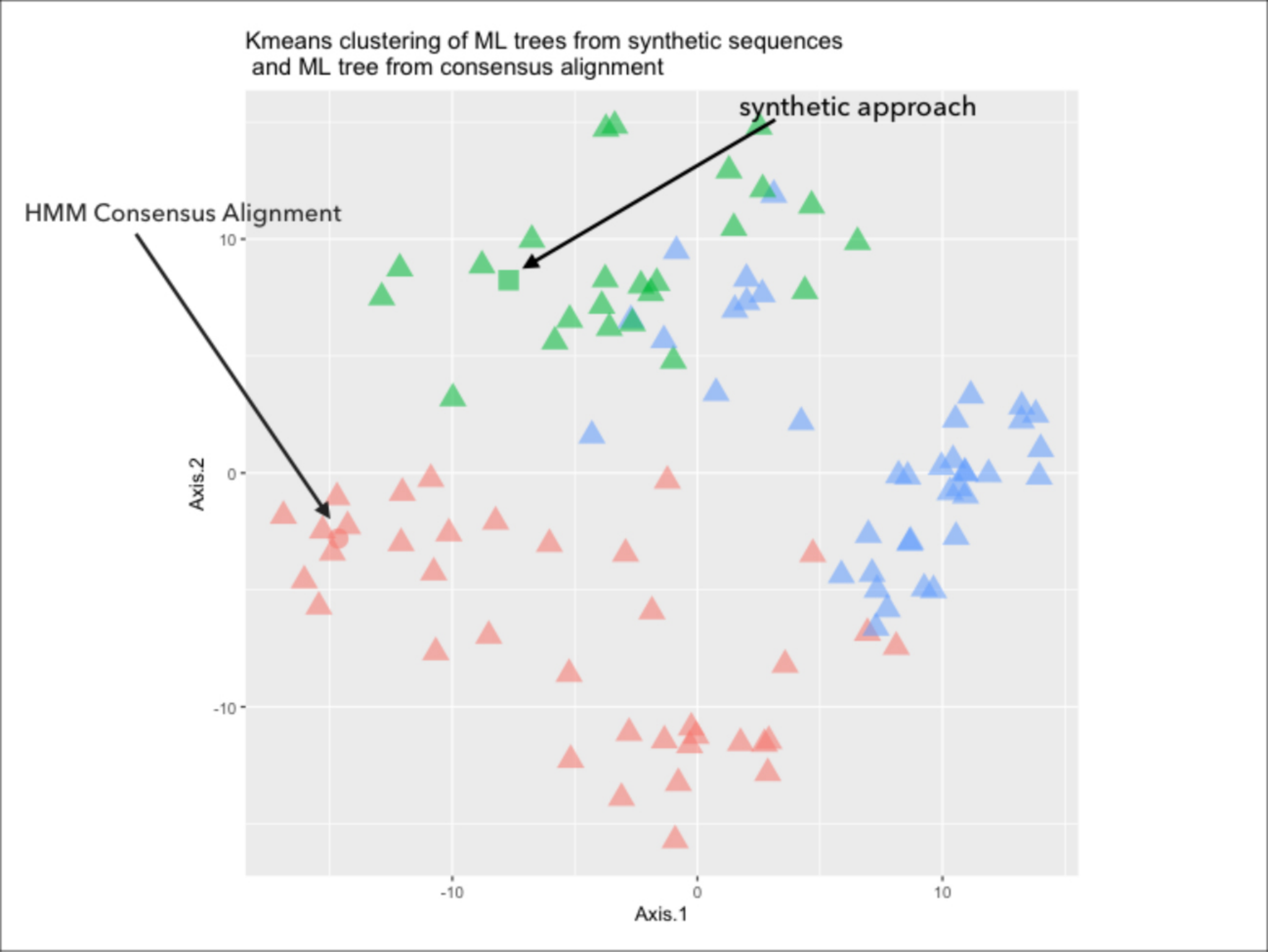
RAxML tree from consensus genome



true tree

so we don't recover the same tree…


Read Profile
simulate n sequences
G
M
T


Kmeans clustering of ML trees from synthetic sequences
and ML tree from consensus alignment
synthetic approach
HMM Consensus Alignment
10
0
-10
Axis.2
-10
0
10
Axis.1