PRESERVING INTRA-PATIENT VARIANCE IMPROVES PHYLOGENETIC INFERENCE OF HIV TRANSMISSION

AUGUST GUANG
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
TRANSMISSION NETWORKS AS PHYLOGENIES
Phylogenetic trees have information about the 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, phylogeny 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
but with rare exceptions they are summaries
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

A

B

C
if all the variation was accounted for
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

ACAC
ACAC
ATTTC
TTTC

S

1

A = 0.5
T = 0.5
C = 0

1

A = 0
T = 0
C = 1

E
indels are hidden states

S 1
A = 0.5
T = 0.5
C = 0

TCA
ACA
ACA

- = 1

0.25
0.75

A = 0
T = 0
C = 1

A = 1
T = 0
C = 0

E 1
indels are hidden states

A C A A
A C A A
T - A T
T C A T

insertion

S 1

M 0.75

D

I 0.25

M 1

M 0.75

M

E
we can build read profiles from read alignments
sample genomes from those profiles

A1: ACAATGACAAATGGCAAA
A2: ACATGAAACTGGCA
B1: TATGAAATGGCAAA
B2: TCATGAAACTGGCA
C1: TATGACAAATGGCAAA
C2: TCATGAAACTGGCA
and build alignments and trees with those samples

A1  ACAATGACCAATGGGCAAA
B1  TATAATGAAATGGGCAAA
C1  TATAATGACCAATGGGCAAA

A2  ACATGGAACTGGGCA
B2  TCATGAAACTGGGCA
C2  TCATGAAACTGGGCA
our "synthetic" approach

Read Profile

simulate n sequences

G

M

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

T

Diagram:

- Read Profile
- Simulate n sequences
- Structure labeled G
- Structure labeled M
- Structure labeled T

Diagram shows a flow of data and processes leading to different structures or entities.
each step in this workflow is a high-dimensional inference problem...

- D->G: HMMer
- G->M: mafft
- M->T: RAxML or MrBayes
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
consensus tree set with HMM consensus is unresolved
TRANSIMISSIM (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

- D: ART
- G: pyvolve
- V: SimPhy
- T: binary mapping
- N: outbreaker
bootstrap results on simulations

red = consensus
blue = synthetic
bootstraps for synthetic were better than consensus for all splits on true tree
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](https://bitbucket.org/aguang/ms_hiv)
- Transmissim: [https://github.com/aguang/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
a. Whole-genome NGS reads for Patient 1

Read1  ATGGCATATGGAGCATGATGGC
Read2  TGATGCATCGCTGATGCCATAT
Read3  TGGATGCATCGCTGATGGCA

b. HMMER alignment to reference pHMM

Reference1  TGGATGCATCGCTGATGGCATATGGCATATT
Reference1  TG–ATGCATCGCTGATGGCATTTATGGCATT

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–ATGCATCGCTGATGCCATAT-------------------
Synthetic1 T--ATGCATCGCTGATGCCATATT--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
bootstraps differed significantly on certain splits
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...
K-means clustering of ML trees from synthetic sequences and ML tree from consensus alignment

HMM Consensus Alignment

synthetic approach