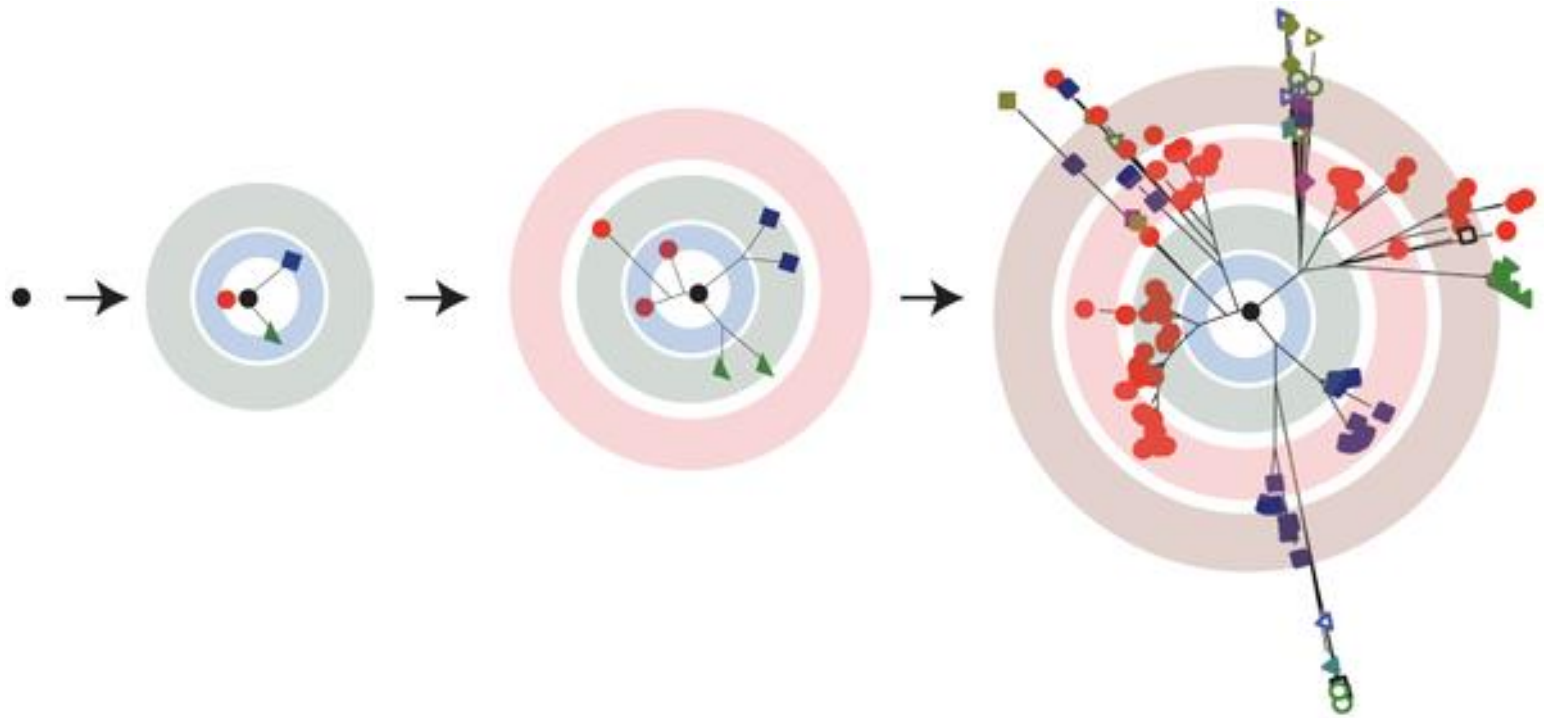
Algorithms for Analysis and Applications of High-Throughput Sequencing of Intra-Host Viral Populations

Alex Zelikovsky, Georgia State University

10th International Conference on Bioinformatics
Genomic and Evolution of Pathogens
11/21/2015 at Georgia Tech

RNA Virus: Intra-Host Population

High mutation rate (~10^{-4})



Lauring & Andino, PLoS Pathogens 2011

Intra-Host Viral Population: Curse or Blessing?

Curse – for sequencing:

- very low variability vs error rate
- analogue to very low signal-to-noise ratio
  - Mutation = signal rate 0.05%
  - Error rate = noise rate 0.1% / 2%

Blessing – for transmission inference

- Just a single sequence/no variation ➔ no information
  - limited information for relatedness
  - for inferring direction of transmission

This talk: deal first with **Curse** and then with **Blessing**

Intra-Host Viral Population Reconstruction from Single Amplicon NGS Reads

Introduction

- Viral spectrum reconstruction for RNA virus
- Technology: SMRT sequencing technologies (PacBio)
  - Long (up to 10 000bp)
  - High error rate (~15%→ 3%)
  - Low coverage (30k→100k reads)

Existing Algorithms

- **PredictHaplo** (Francesca Di Giallonardo et al.)
  - Probabilistic (Bayesian mixture) model with Dirichlet process to estimate number of haplotypes
  - Markov chain via Monte Carlo sampling for inference

- Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing (Dario A. Dilernia et al.)

ML Problem Formulation

- **Given:** set of reads *R* from unknown haplotype set *H'*
- **Find:** set of haplotypes $H=\{H_1,\ldots,H_k\}$ with corresponding frequencies $F=\{f_1,\ldots,f_k\}$ maximizing $\Pr(R|H)$

NOTE: Given haplotypes, the frequencies can be reliably estimated via Expectation-Maximization

  - similarly to transcriptome quantification

Alignment

- Ideal: Multiple Sequence Alignment of all reads
- **Challenge:**
  - too many indels (10% of 2300bp sequences)
  - in too many reads (10K-30K)
- **Solution:**
  - Pairwise alignment to reference BWA (Li H. and Durbin R. (2009))
  - B2W (Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N (2009))
- **Error rate:**
  - After alignment the error rate reduces significantly
    - Majority of errors are random lengthy insertions
    - Alignment removes random insertions

Extract signal from noise

Assumption: Noise is **random** / signal is **not!**

For 2 positions I and J:

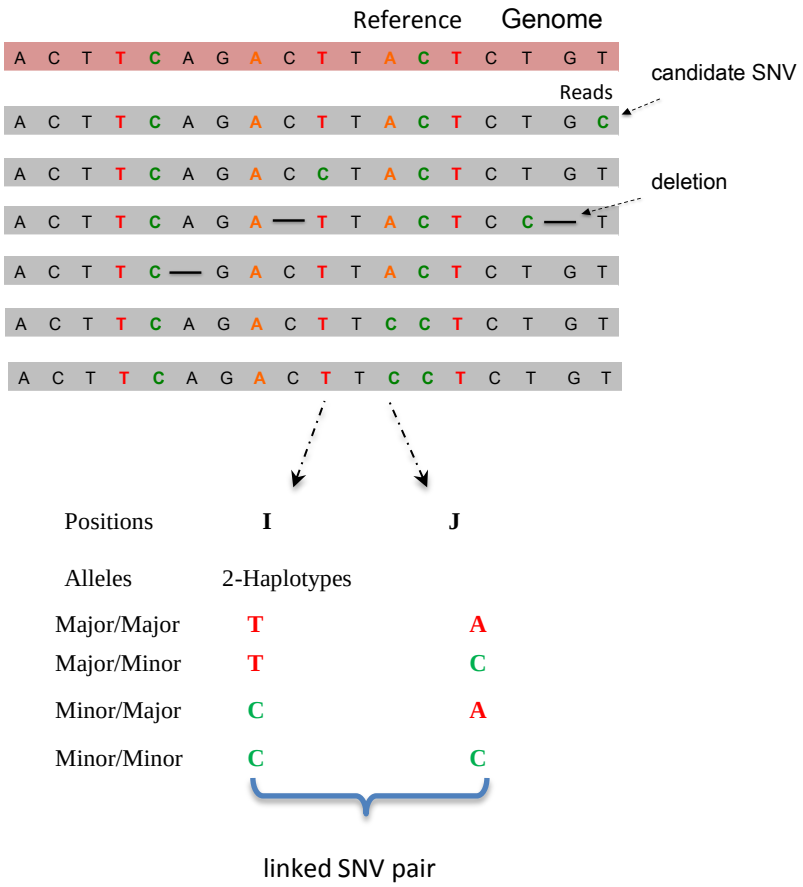
- Major/Major haplotype 11
- Major/Minor haplotype 12
- Minor/Major haplotype 21
- Minor/Minor haplotype 22

Theorem: Minor/Minor does not exist, then for expected number of reads Ekl (k,l=1,2)

E11*E22<= E12*E21

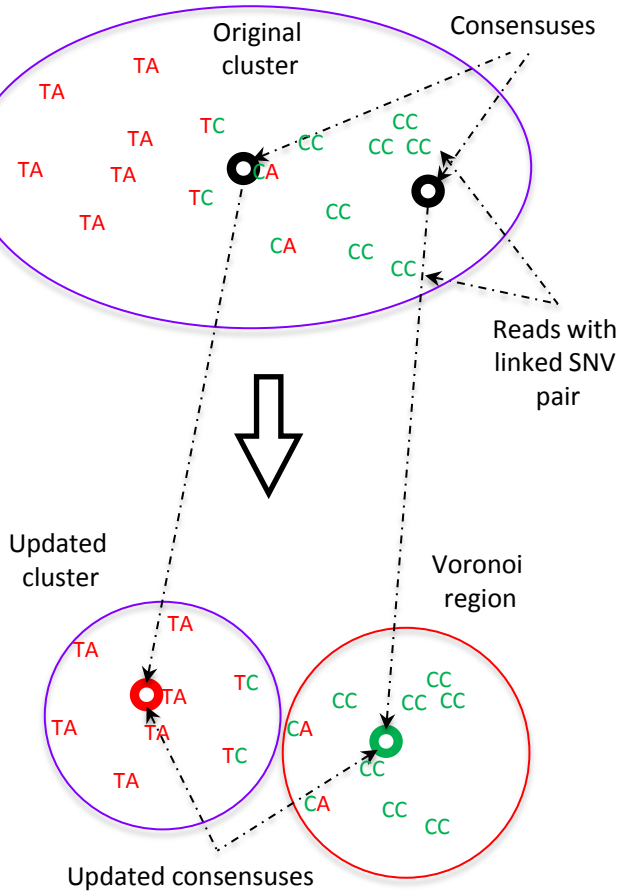
Definition: let X be binomial distribution with p = (A12*A21/A11*n),
Akl (k,l=1,2) and n = observed number of reads

If Prob(X>A22) <0.01/(N choose 2), then
These two minor alleles are **linked**

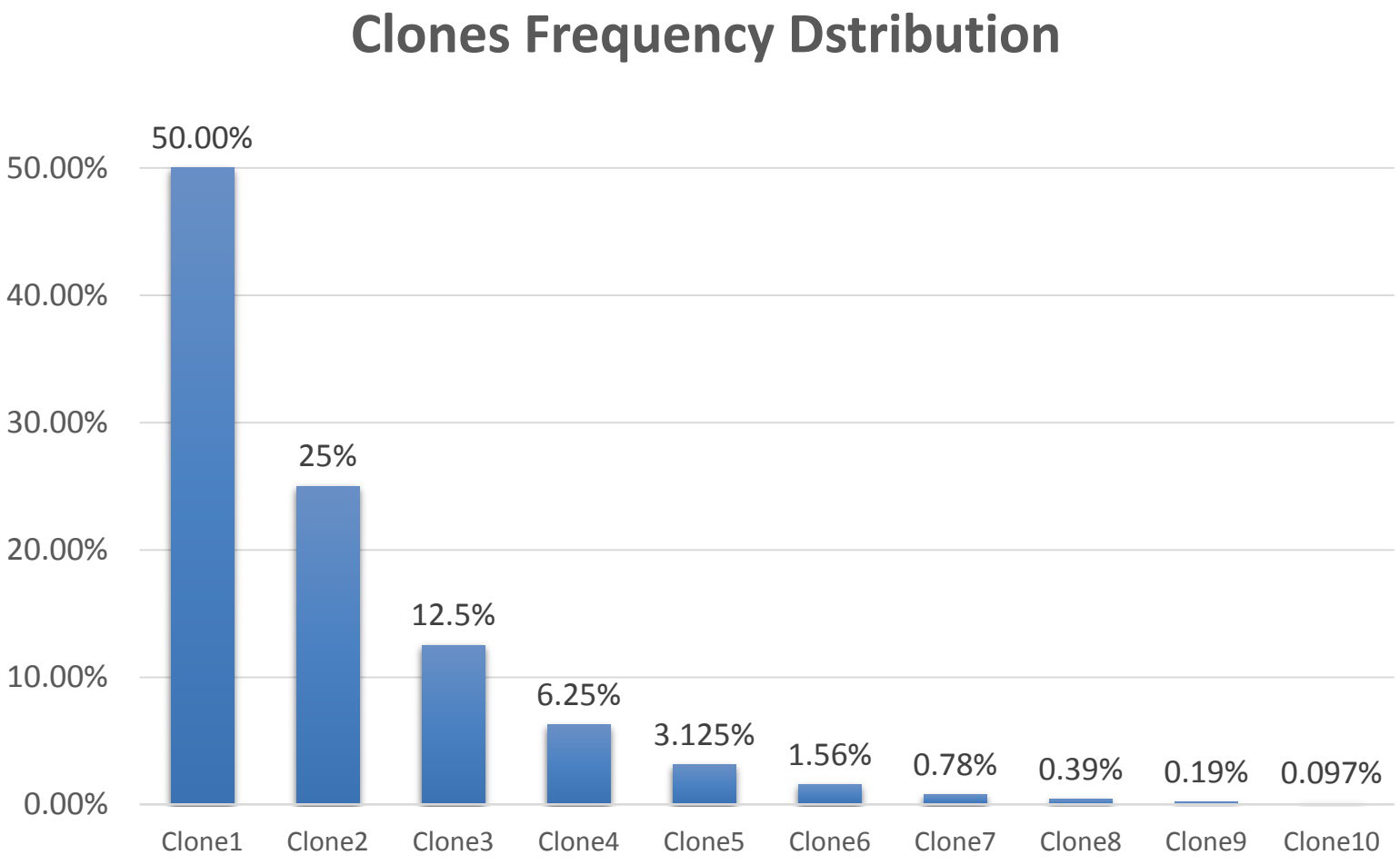


Haplotyping

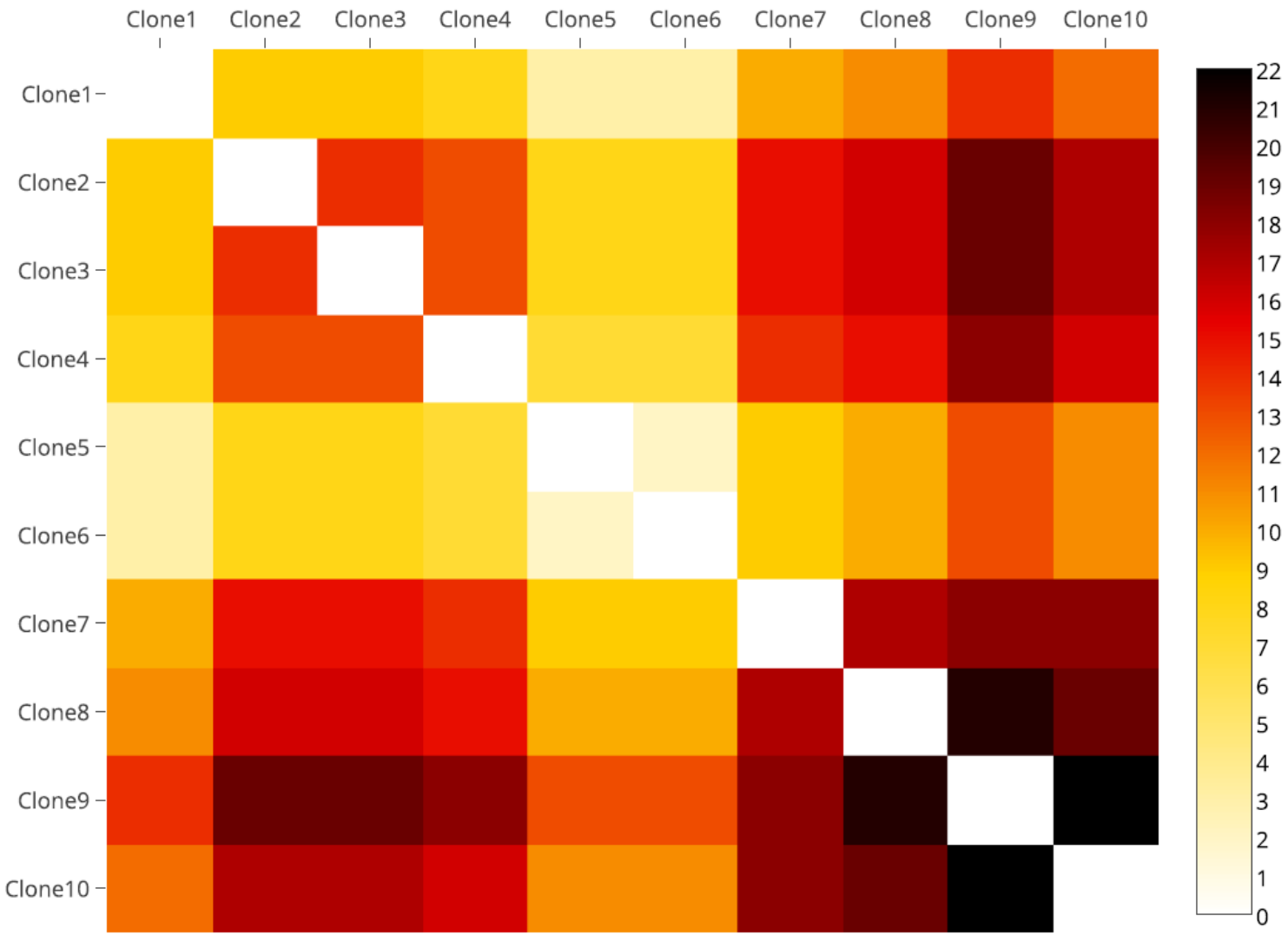
- **Initial cluster C** contains all reads
- Label C complex
- **Repeat**
  - If there is a complex cluster
    - Find pair of linked SNVs
    - **If** it exists split that cluster on 2 parts **else** label current cluster as simple.
- **Until** *all clusters are simple*
- **Calculate frequencies** with ***k*GEM**



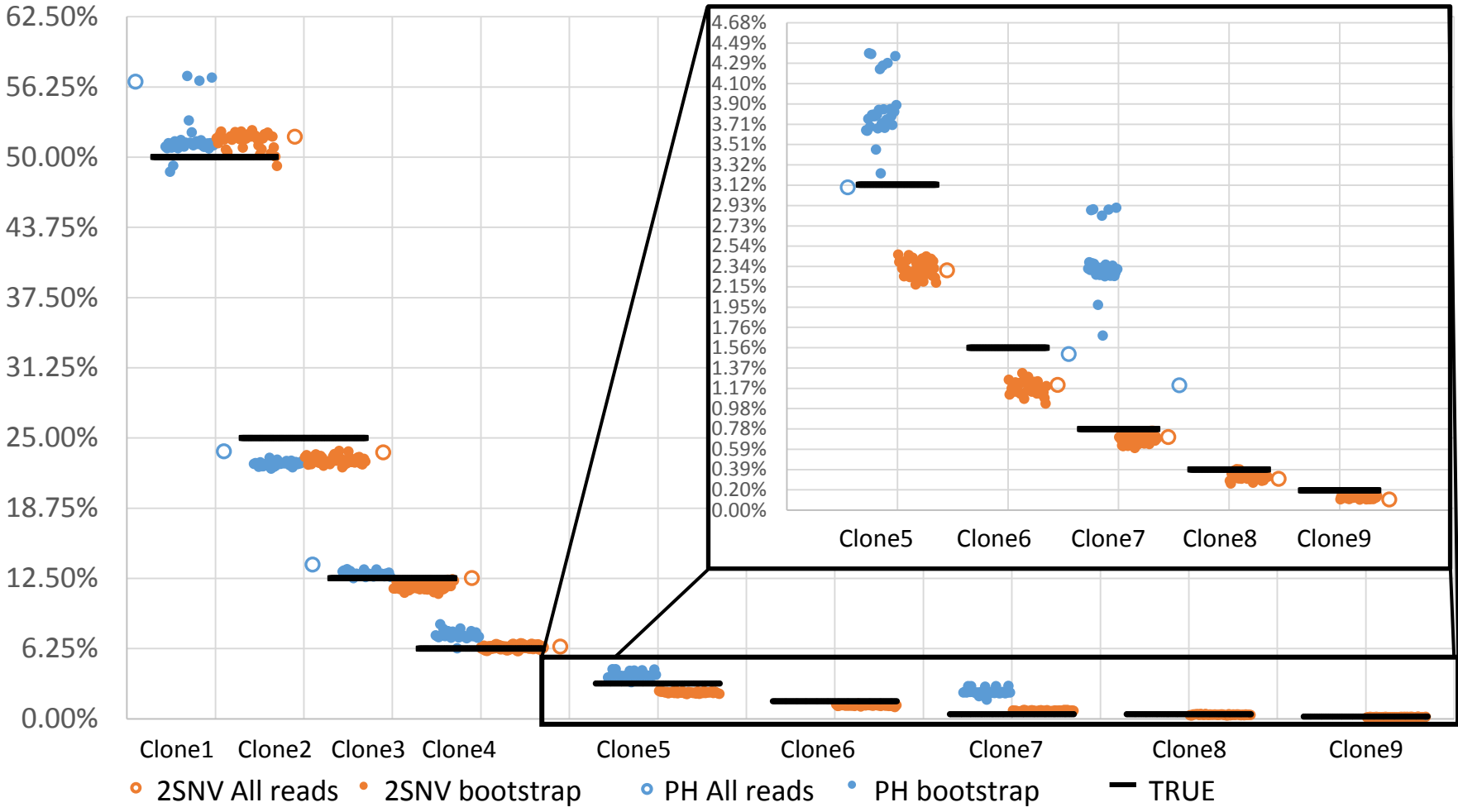
Experimental Setup

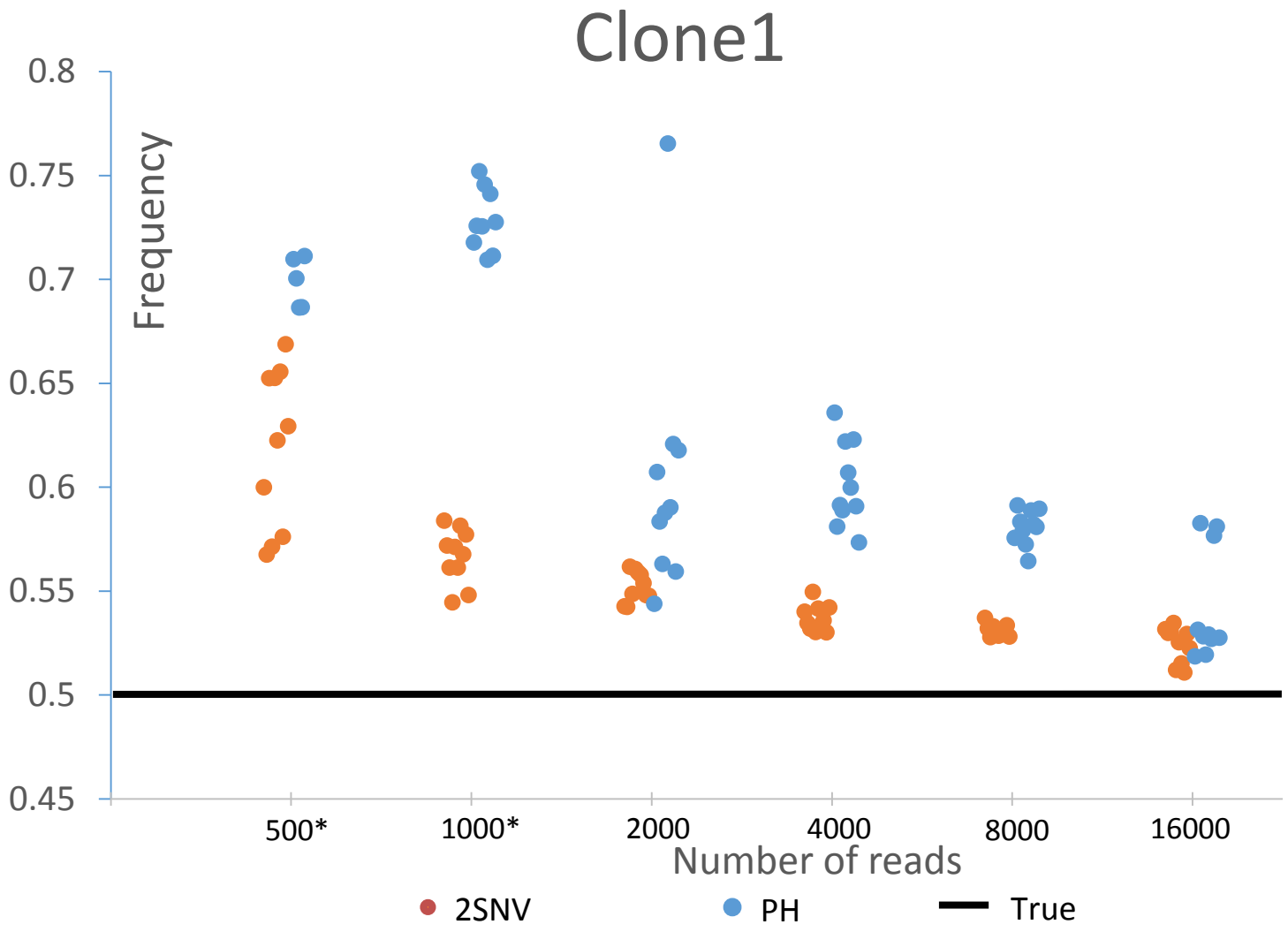


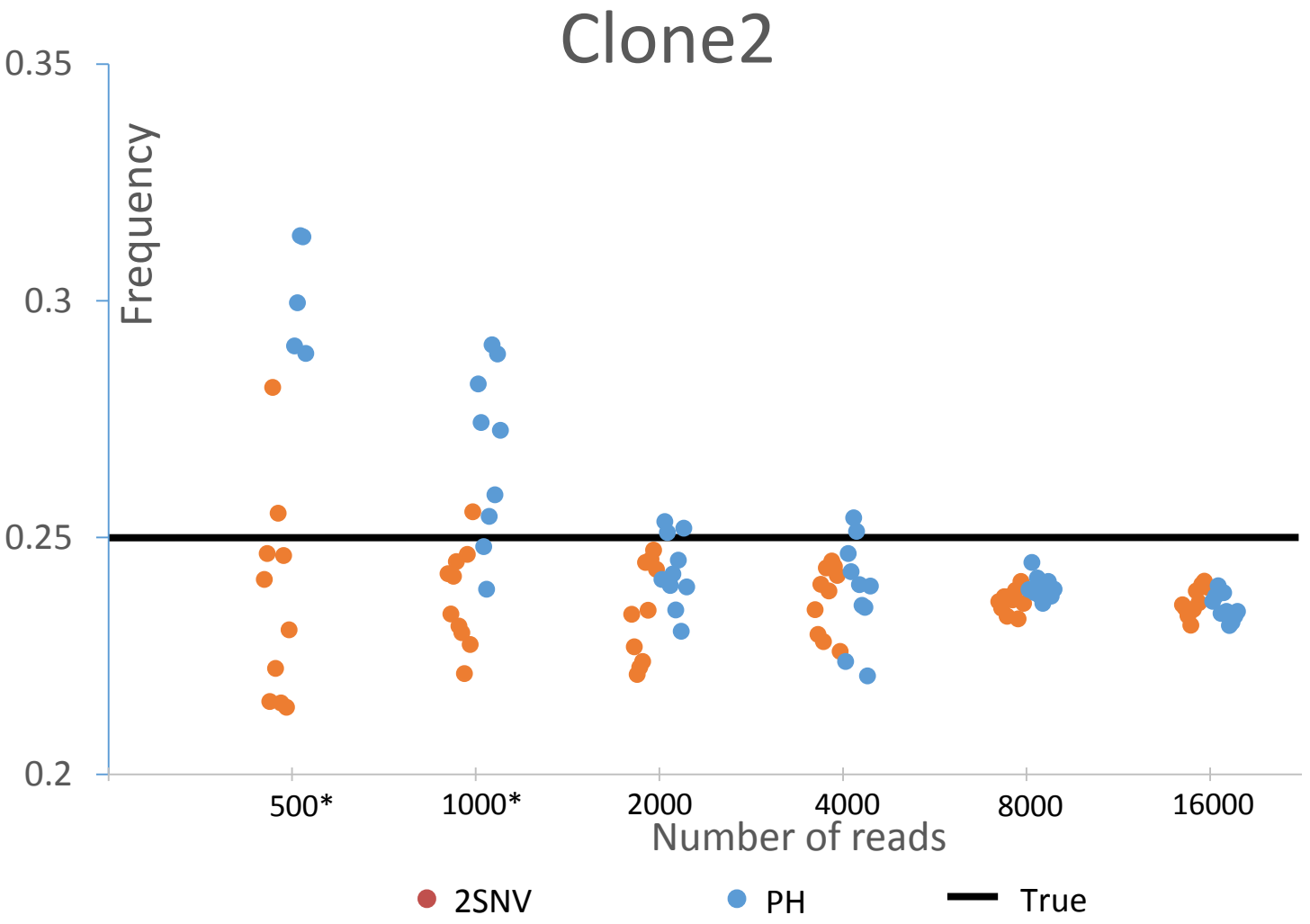
Edit Distance Heatmap

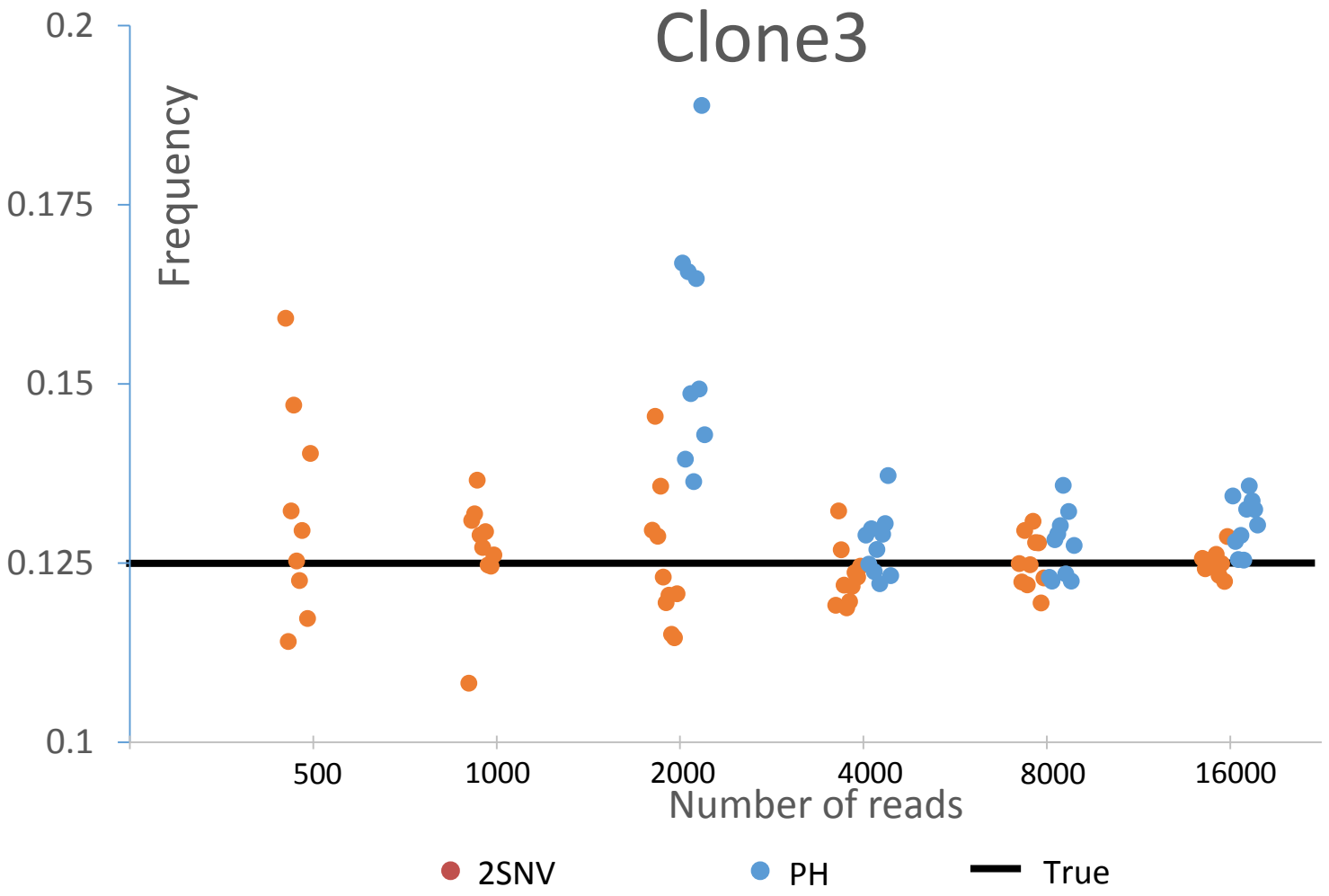


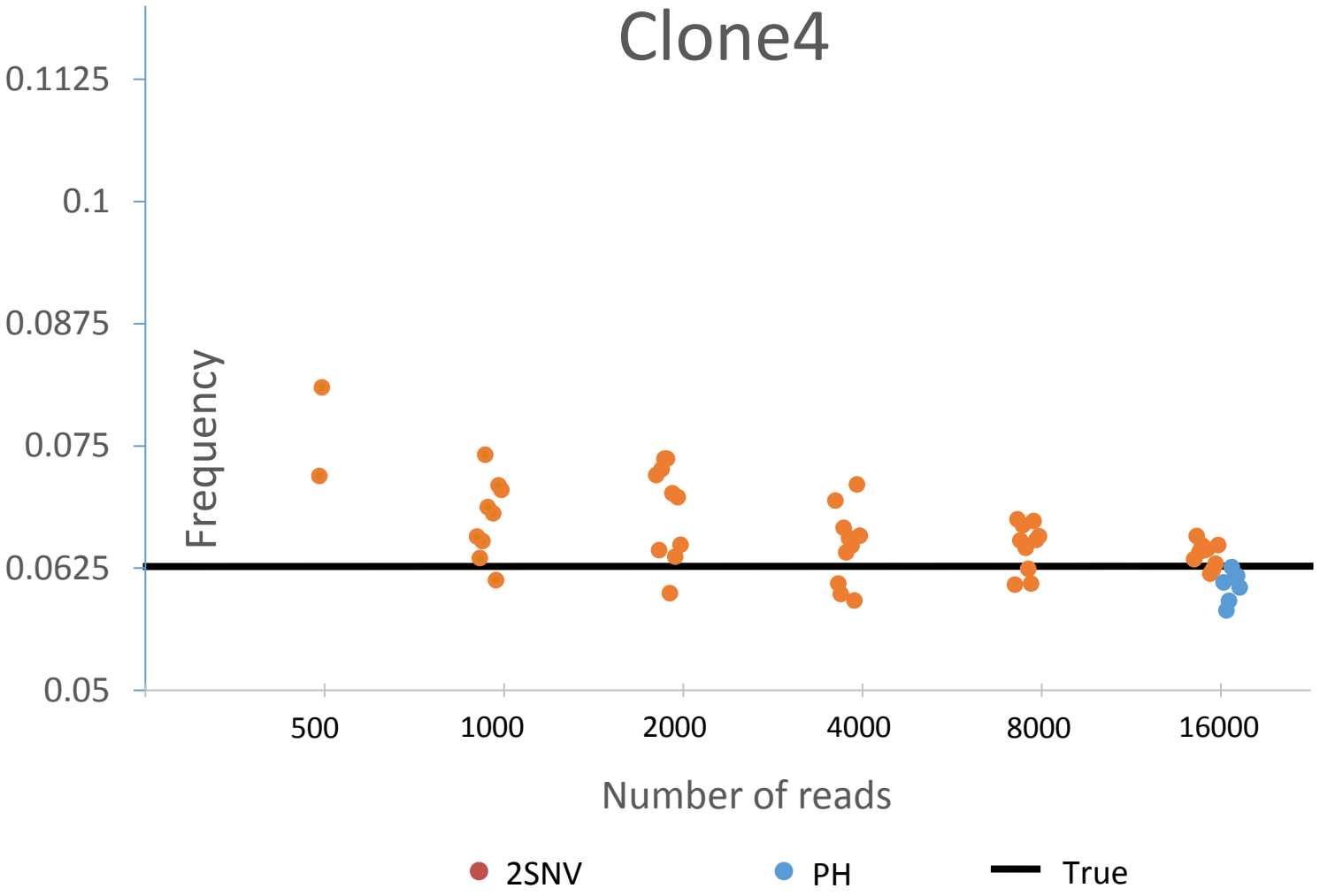
Results

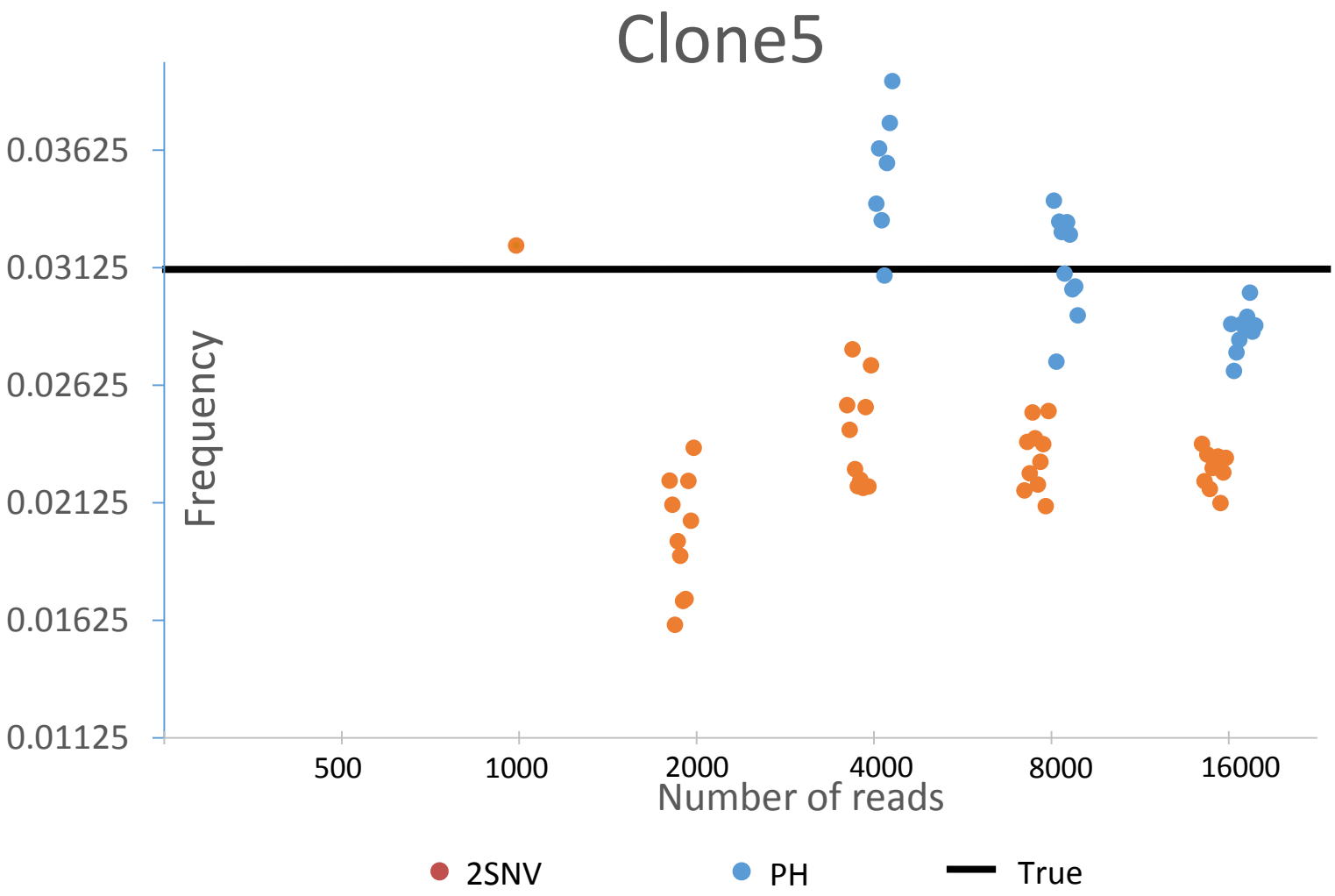

62.50%
56.25%
50.00%
43.75%
37.50%
31.25%
25.00%
18.75%
12.50%
6.25%
0.00%
4.68%
4.49%
4.29%
4.10%
3.90%
3.71%
3.51%
3.32%
3.12%
2.93%
2.73%
2.54%
2.34%
2.15%
1.95%
1.76%
1.56%
1.37%
1.17%
0.98%
0.78%
0.59%
0.39%
0.20%
0.00%
Clone1
Clone2
Clone3
Clone4
Clone5
Clone6
Clone7
Clone8
Clone9
2SNV All reads
2SNV bootstrap
PH All reads
PH bootstrap
TRUE


Clone1
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
Frequency
500*
1000*
2000
4000
8000
16000
Number of reads
2SNV
PH
True

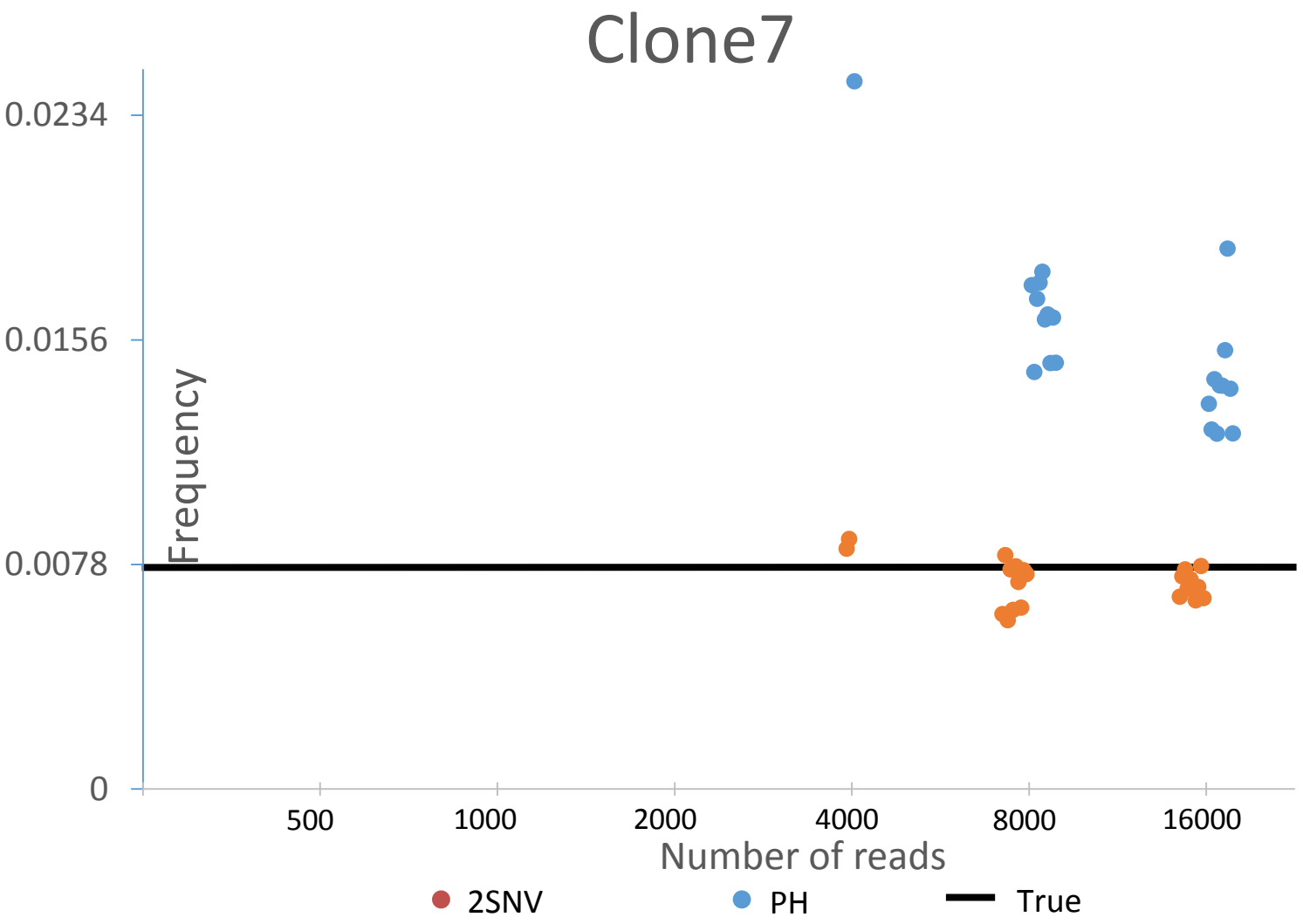

Clone2
0.35
0.3
0.25
0.2
Frequency
500*
1000*
2000
4000
8000
16000
Number of reads
2SNV
PH
True

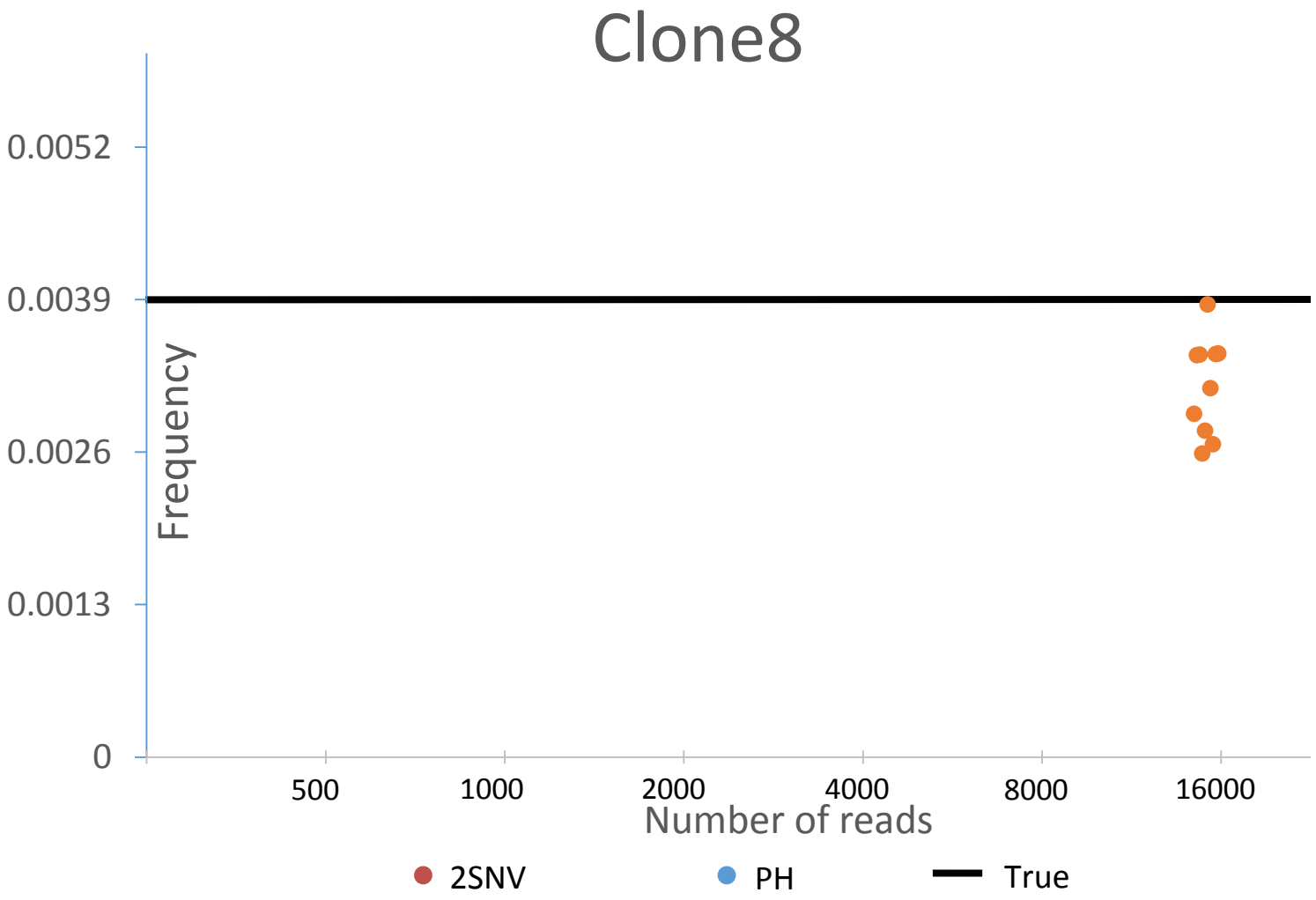

Clone3
Frequency
0.2
0.175
0.15
0.125
0.1
500
1000
2000
4000
8000
16000
Number of reads
2SNV
PH
True

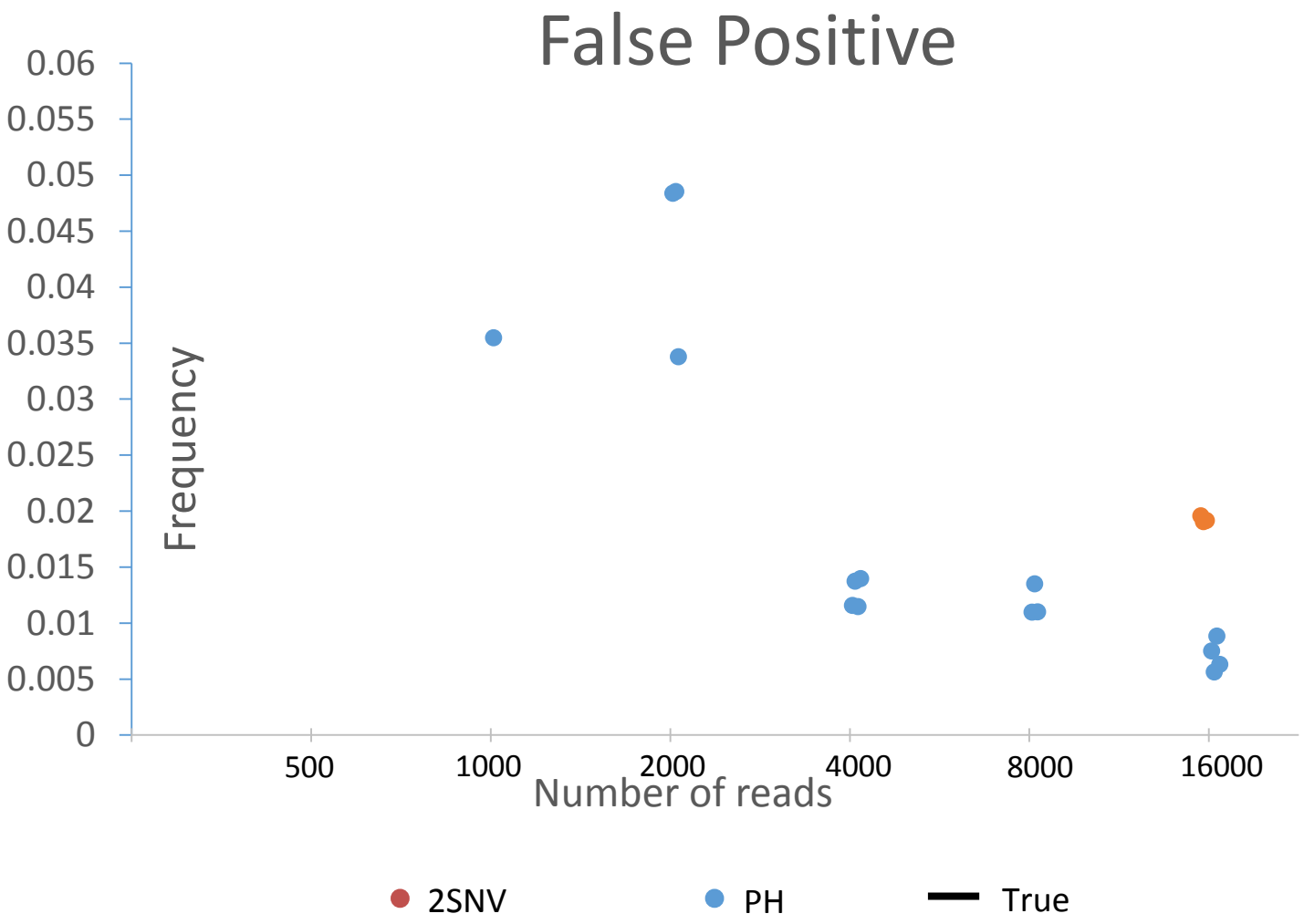

Clone4
0.1125
0.1
0.0875
0.075
0.0625
0.05
Frequency
500
1000
2000
4000
8000
16000
Number of reads
2SNV
PH
True


Clone5
0.03625
0.03125
0.02625
0.02125
0.01625
0.01125
Frequency
500
1000
2000
4000
8000
16000
Number of reads
2SNV
PH
True


Clone6
0.0196
0.0176
0.0156
0.0136
0.0116
0.0096
0.0076
0.0056
Frequency
500
1000
2000
4000
8000
16000
Number of reads
2SNV
PH
True


Clone7
0.0234
0.0156
0.0078
0
Frequency
500
1000
2000
4000
8000
16000
Number of reads
2SNV
PH
True

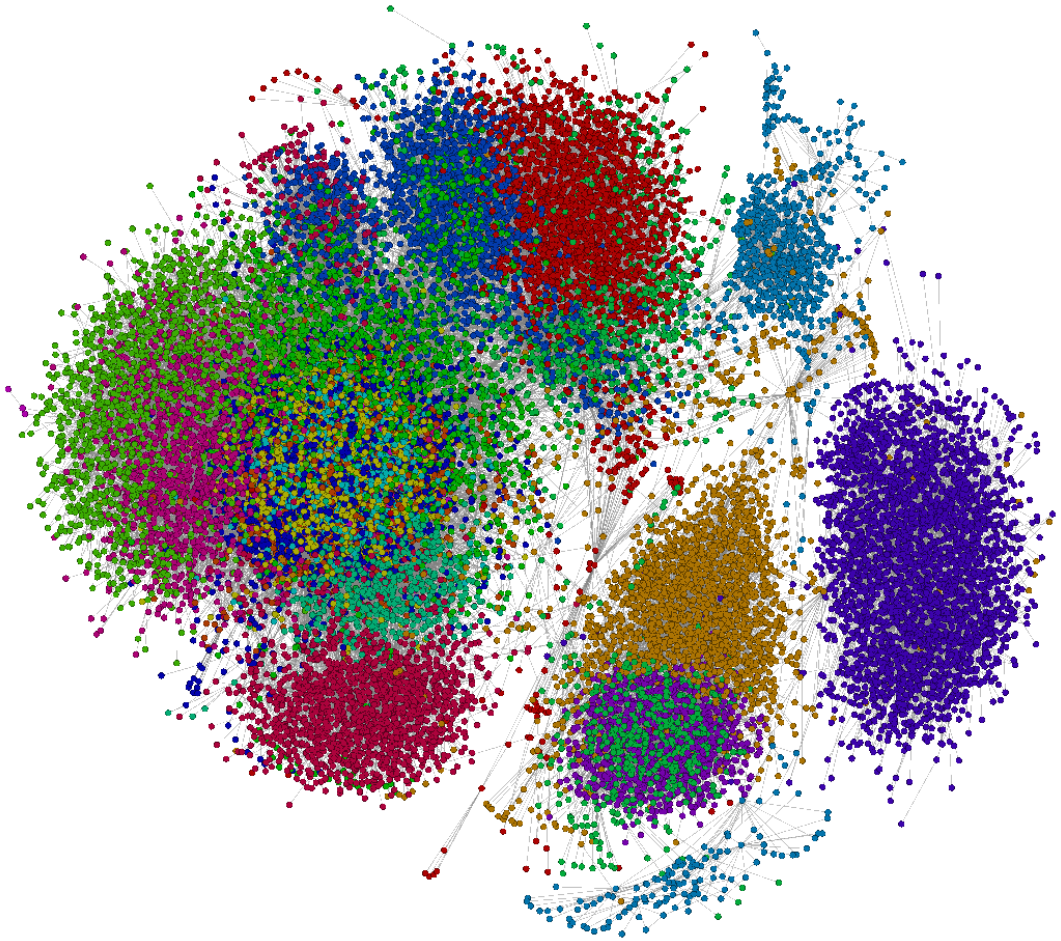

Clone8
0.0052
0.0039
0.0026
0.0013
0
Frequency
500
1000
2000
4000
8000
16000
Number of reads
2SNV
PH
True


False Positive
0.06
0.055
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Frequency
500
1000
2000
4000
8000
16000
Number of reads
2SNV
PH
True

Inferring Viral Transmissions from Intra-Host Viral Populations

NGS of HOC outbreak

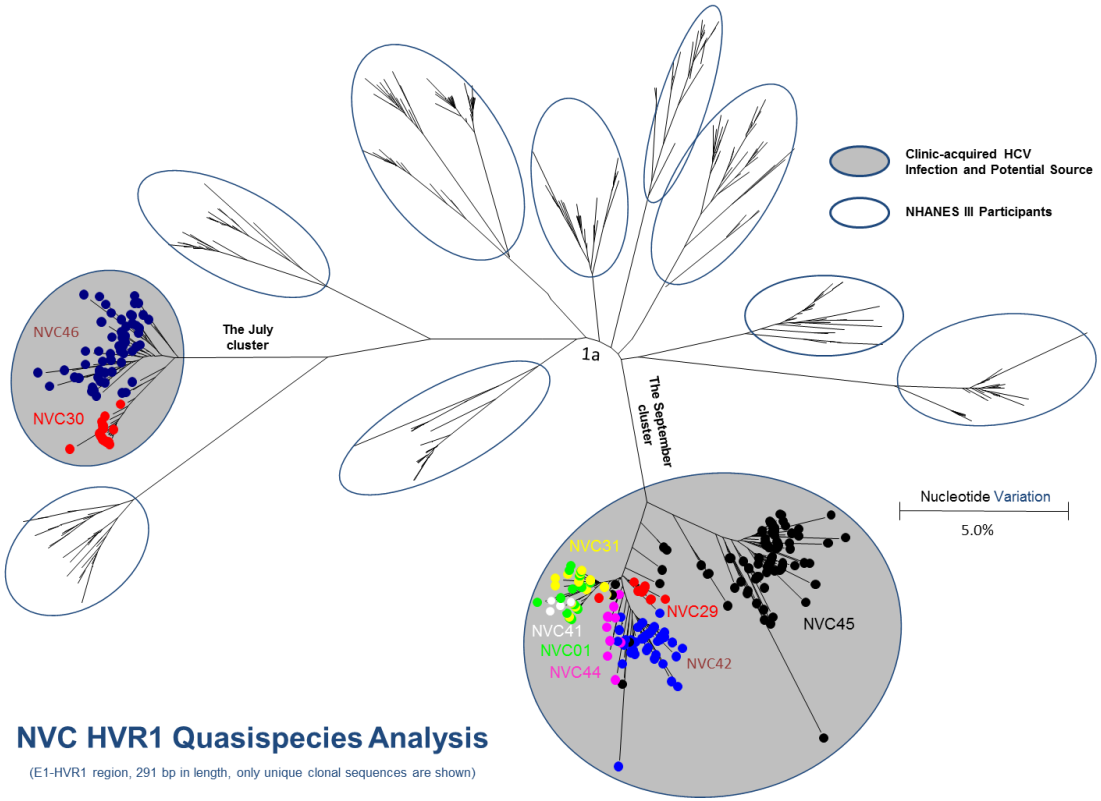
- 18 patients, 154233 reads and 33767 unique sequences.
- Each node is a unique sequence
- Different patients are shown in different colors
- Two sequences are linked if they differ in a single nucleotides



NGS of HOC outbreak

- 18 patients, 154233 reads and 33767 unique sequences.
- Each node is a unique sequence
- Different patients are shown in different colors
- Two sequences are linked if they differ in a single nucleotides

Sequences of the source patient are shown in green.



NVC HVR1 Quasispecies Analysis

(E1-HVR1 region, 291 bp in length, only unique clonal sequences are shown)

The main challenge:

- Finding consensus sequence is not enough
- It is crucial to get the **whole viral quasispecies spectrum** (all sequences and their relative frequencies), since minor variants can be responsible for viral transmission

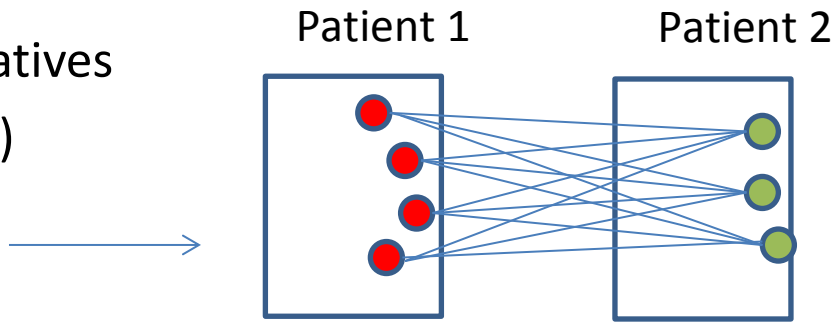
Advanced Molecular Detection of viral transmissions and outbreaks

- Phylogenetic analysis
- Threshold-based methods
- Random processes
- Nonparametric methods

Threshold-based methods

Outbreak detection and display

- Step 1: calculate distances among patients
- We can measure distances among patients in different ways
  - Distance between representatives (consensus or most frequent)
  - Average distance
  - Minimal distance



- Step 2: Link populations with distances smaller than a cutoff

Distance between consensuses

Wertheim et al, **The Global Transmission Network of HIV-1**,
The Journal of Infectious Diseases 2014;209:304–13

Cutoff: 1.5% (approximate level of intrahost diversity early in infection known from a literature)

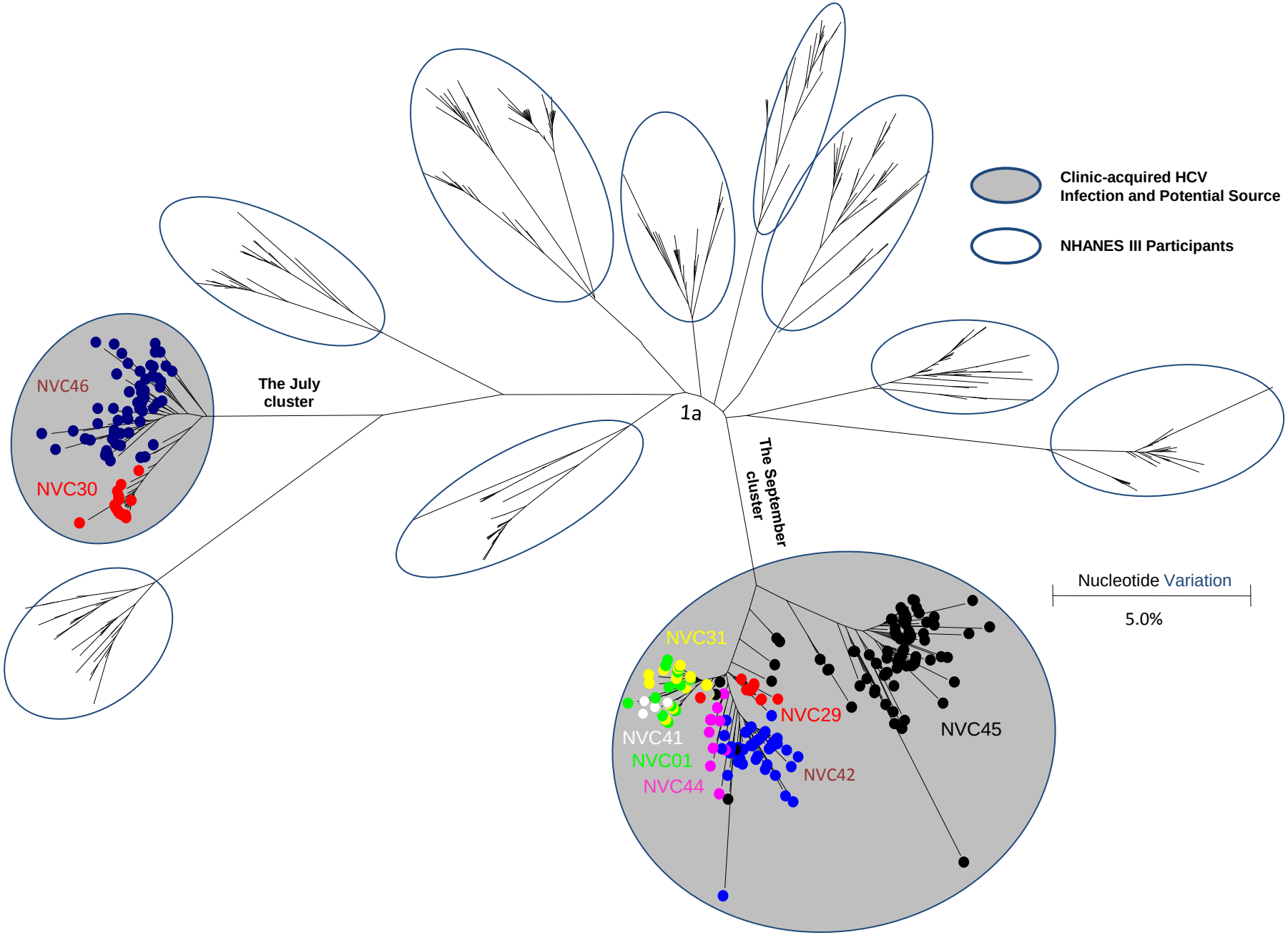
Pros:

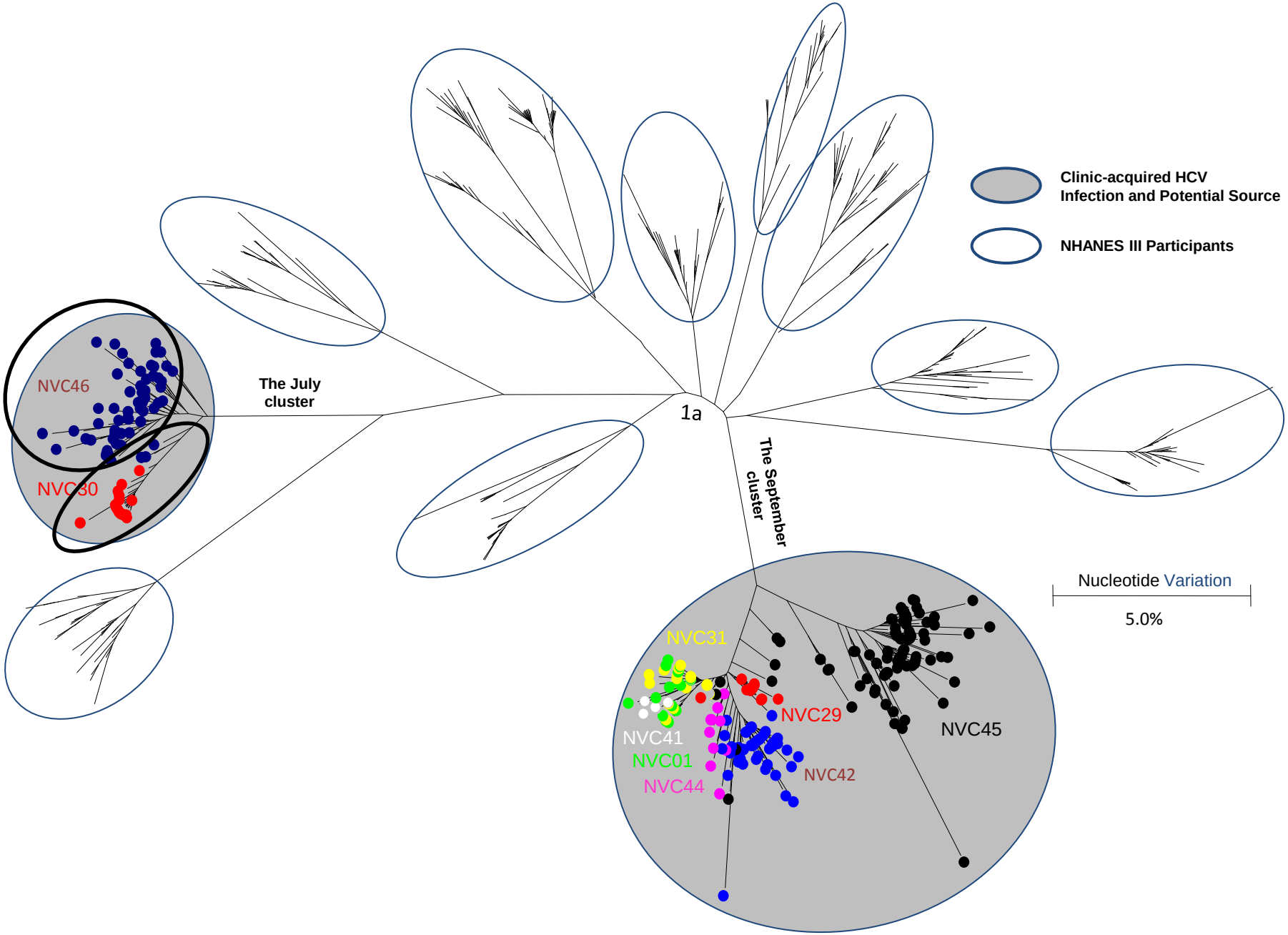
- Easy to automate
- Simple and computationally efficient (linear)

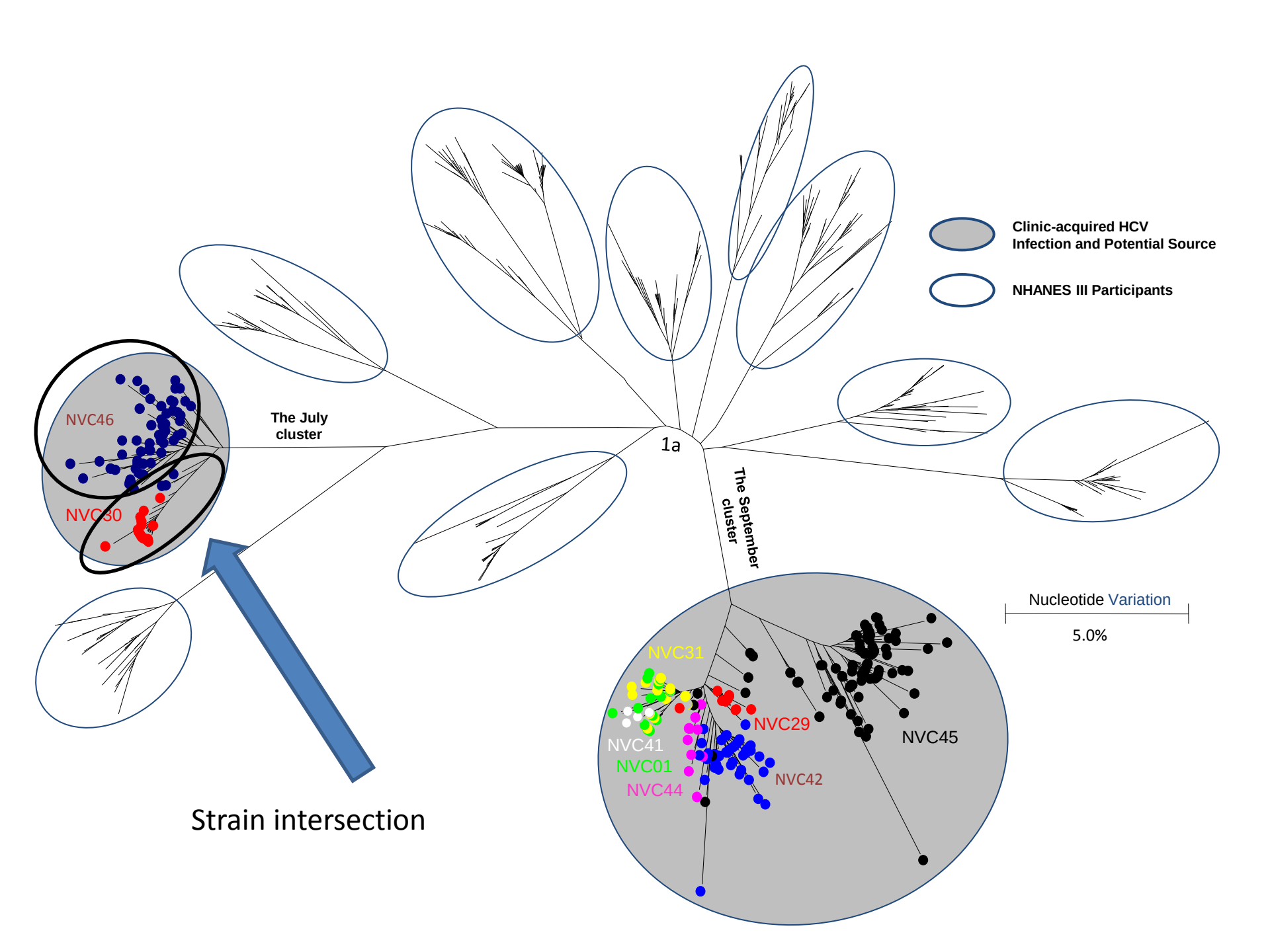
Cons:

- Does not take into account structure of quasispecies population
- Does not allow detection of directions of transmissions
- May not detect transmissions of minor viral subpopulations

Nonparametric detection of transmissions


Clinic-acquired HCV Infection and Potential Source
NHANES III Participants
NVC46
NVC30
The July cluster
1a
The September cluster
Nucleotide Variation
5.0%
NVC31
NVC41
NVC01
NVC44
NVC29
NVC42
NVC45


Clinic-acquired HCV
Infection and Potential Source
NHANES III Participants
NVC46
NVC30
The July
cluster
1a
The September
cluster
Nucleotide Variation
5.0%
NVC31
NVC41
NVC01
NVC44
NVC29
NVC42
NVC45

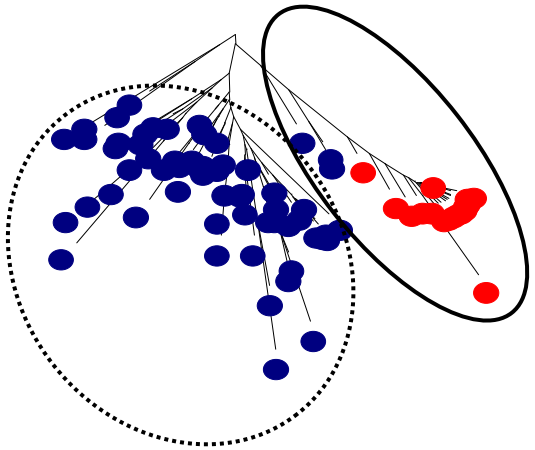

Clinic-acquired HCV
Infection and Potential Source
NHANES III Participants
NVC46
NVC30
The July
cluster
1a
The September
cluster
Nucleotide Variation
5.0%
NVC31
NVC29
NVC45
NVC41
NVC01
NVC44
NVC42
Strain intersection

Strain intersections

Given: two viral populations P_1 and P_2

1) partition the union $P_1 \cup P_2$ into clusters $C_1,\ldots,C_k$
2) $P_1 \overline{\cap} P_2 = \bigcup_{i \in B} C_i$, where $B = \{i \in \{1, \ldots, k\}: C_i \cap P_1 \neq \emptyset, C_i \cap P_2 \neq \emptyset\}$

$P_1 \overline{\cap} P_2$ is the union of clusters that contain sequences from both P_1 and P_2



Relatedness depth

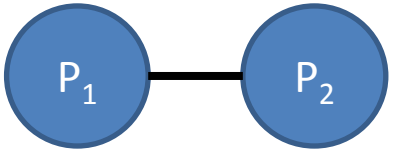
$$d(P_1, P_2) = \begin{cases} 0, \text{if } I = P_1 \,\overline{\cap}\, P_2 = \emptyset \\ +\infty, \text{if } P_1 \,\overline{\cap}\, P_2 = P_1 \cup P_2 \\ 1 + d(P_1|_I, P_2|_I), \text{otherwise} \end{cases}$$

Input Two sets of viral sequences P_1, P_2.
Output Separation coefficient $d(P_1, P_2)$

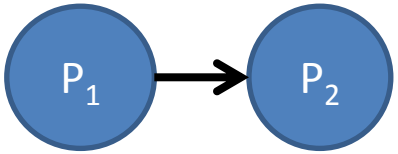
$d \leftarrow 0$
$k \leftarrow 2$
$I \leftarrow P_1 \overline{\cap} P_2$
while $I \neq \emptyset$ and $k \leq |P_1| + |P_2|$ **do**
    $d \leftarrow d + 1$
    if $I \neq P_1 \cup P_2$ **then**
        $P_1 \leftarrow P_1|_I$, $P_2 \leftarrow P_2|_I$
        $k \leftarrow 2$
    else
        $k \leftarrow k + 2$
    end if
    $I \leftarrow P_1 \overline{\cap} P_2$
end while

Relatedness depth

Two populations P_1 and P_2 are **genetically related**, if $d(P_1,P_2) > 0$



Direction of transmission: if at some iteration of Algorithm 1
$P_1 \setminus I \neq \emptyset$, $P_2 \subseteq I$

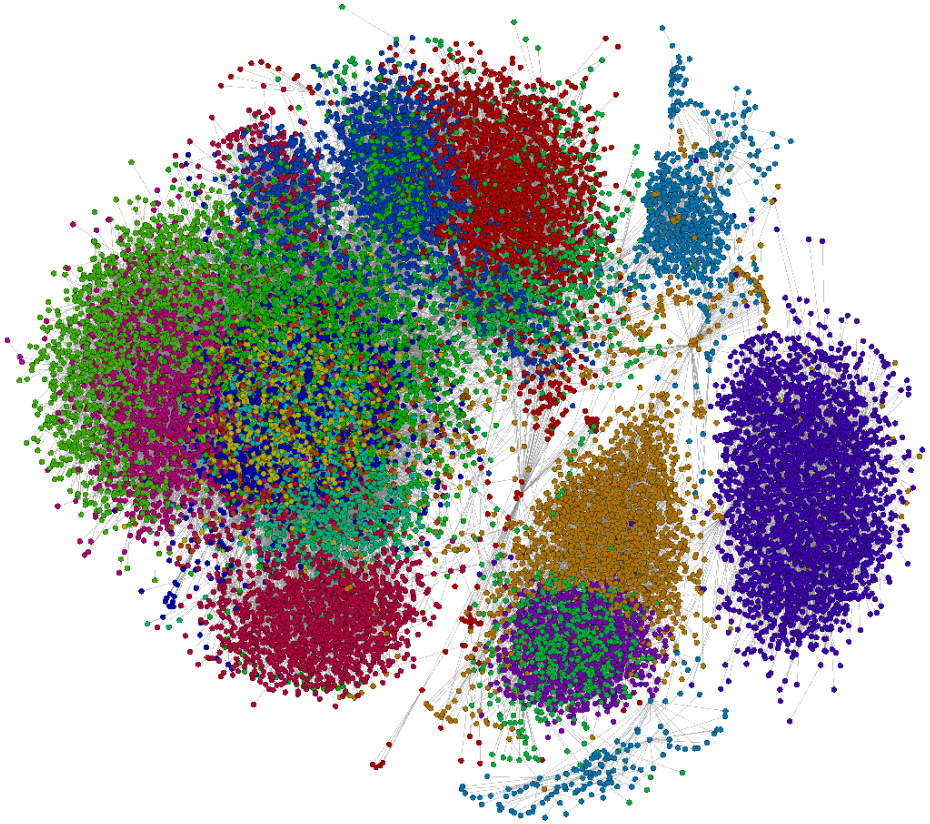


Clustering: hierarchical clustering
Transmission clusters: weakly connected components
Sources: vertices with highest eigenvector centrality

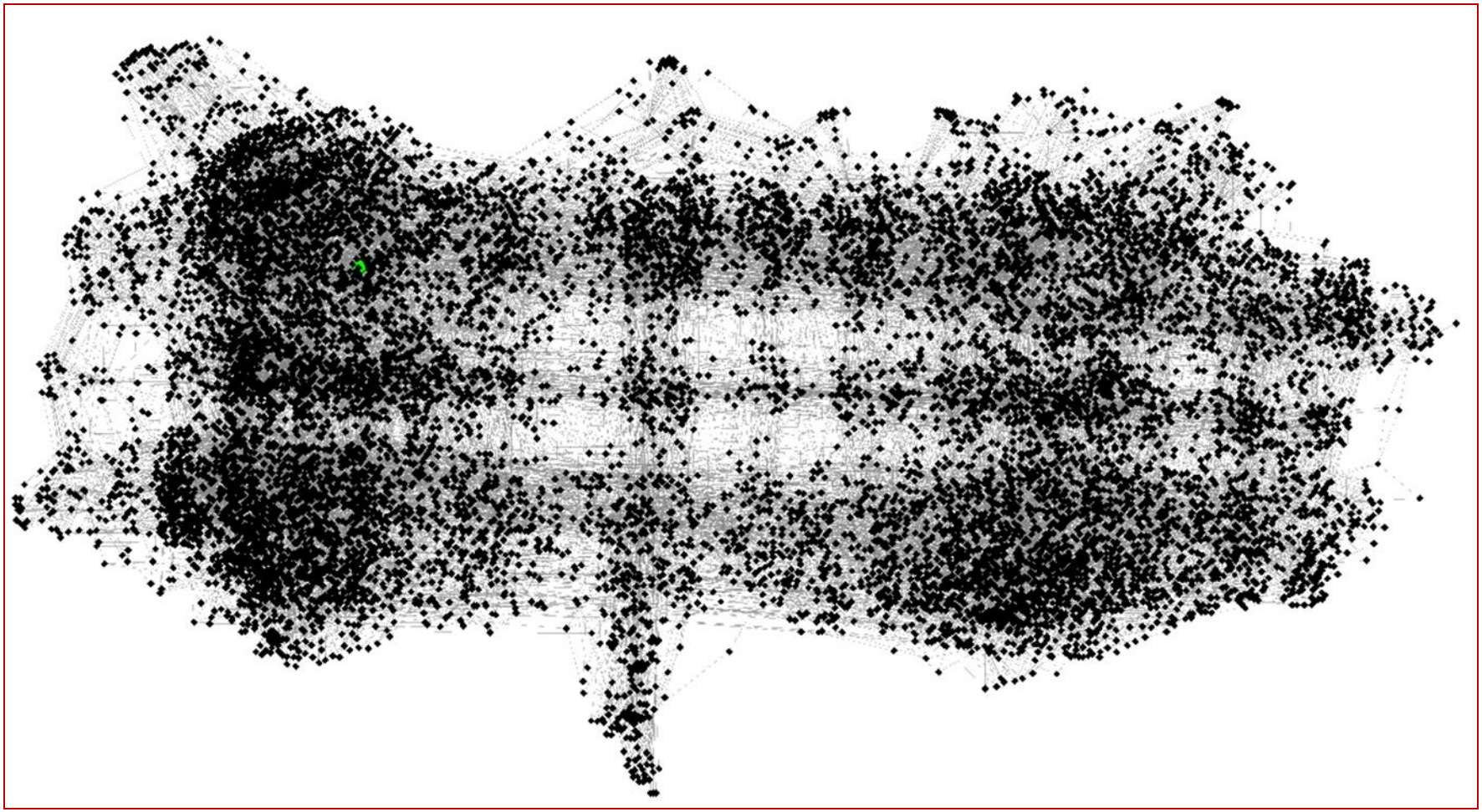
Simulation/Random Walk Method

Motivation:

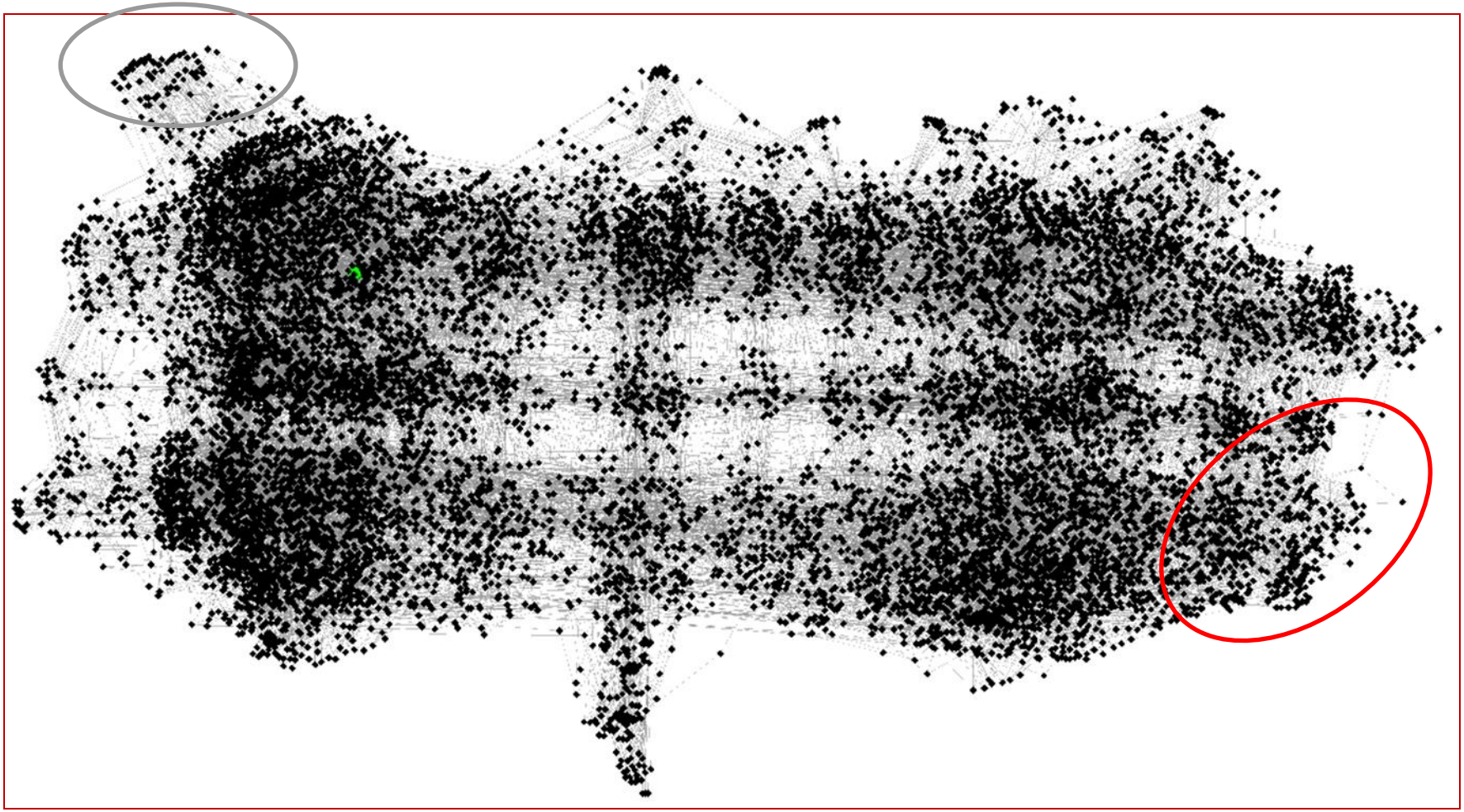
- To take full advantage of the knowledge of populations structures
- To estimate time of transmission



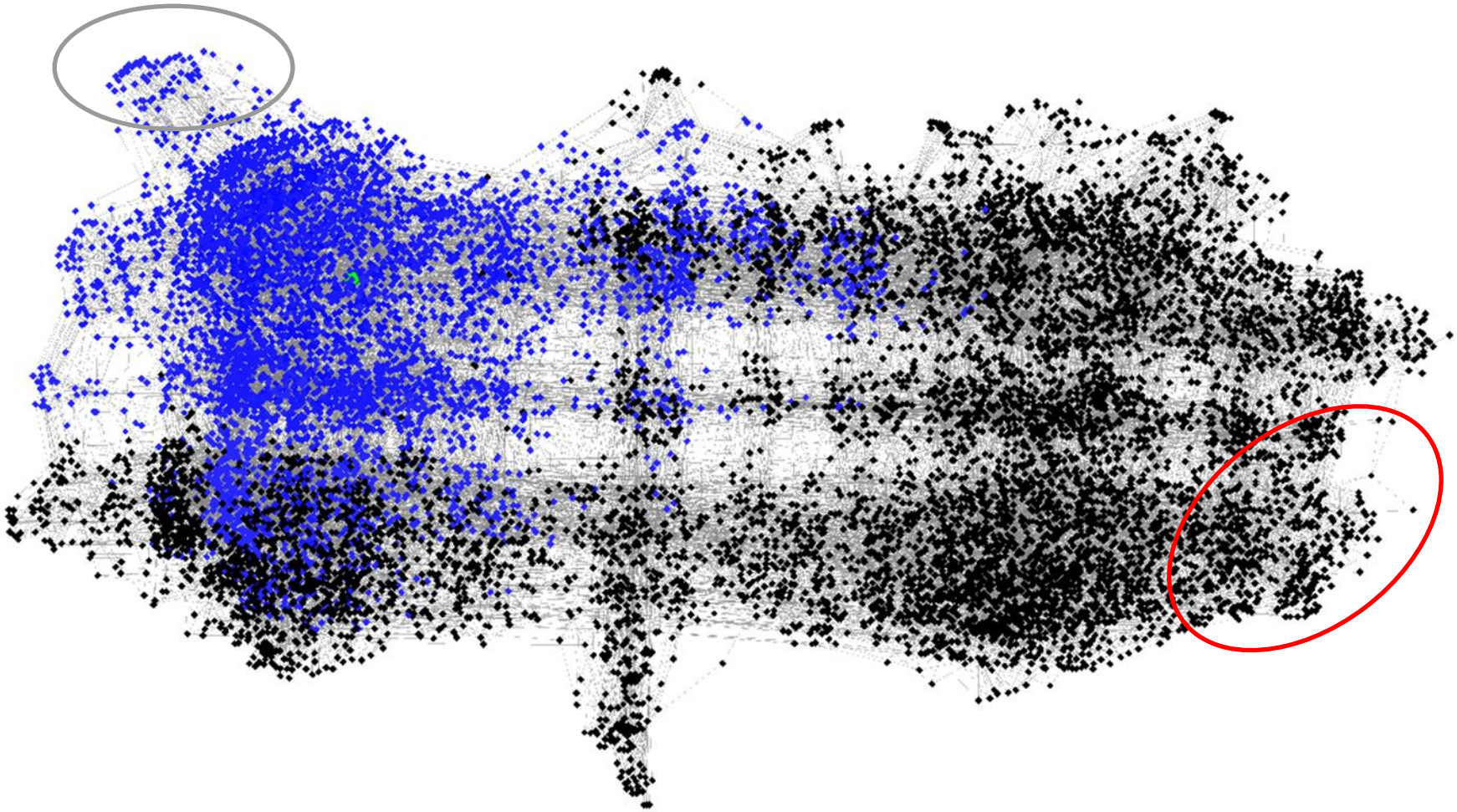
Method: simulate viral evolution



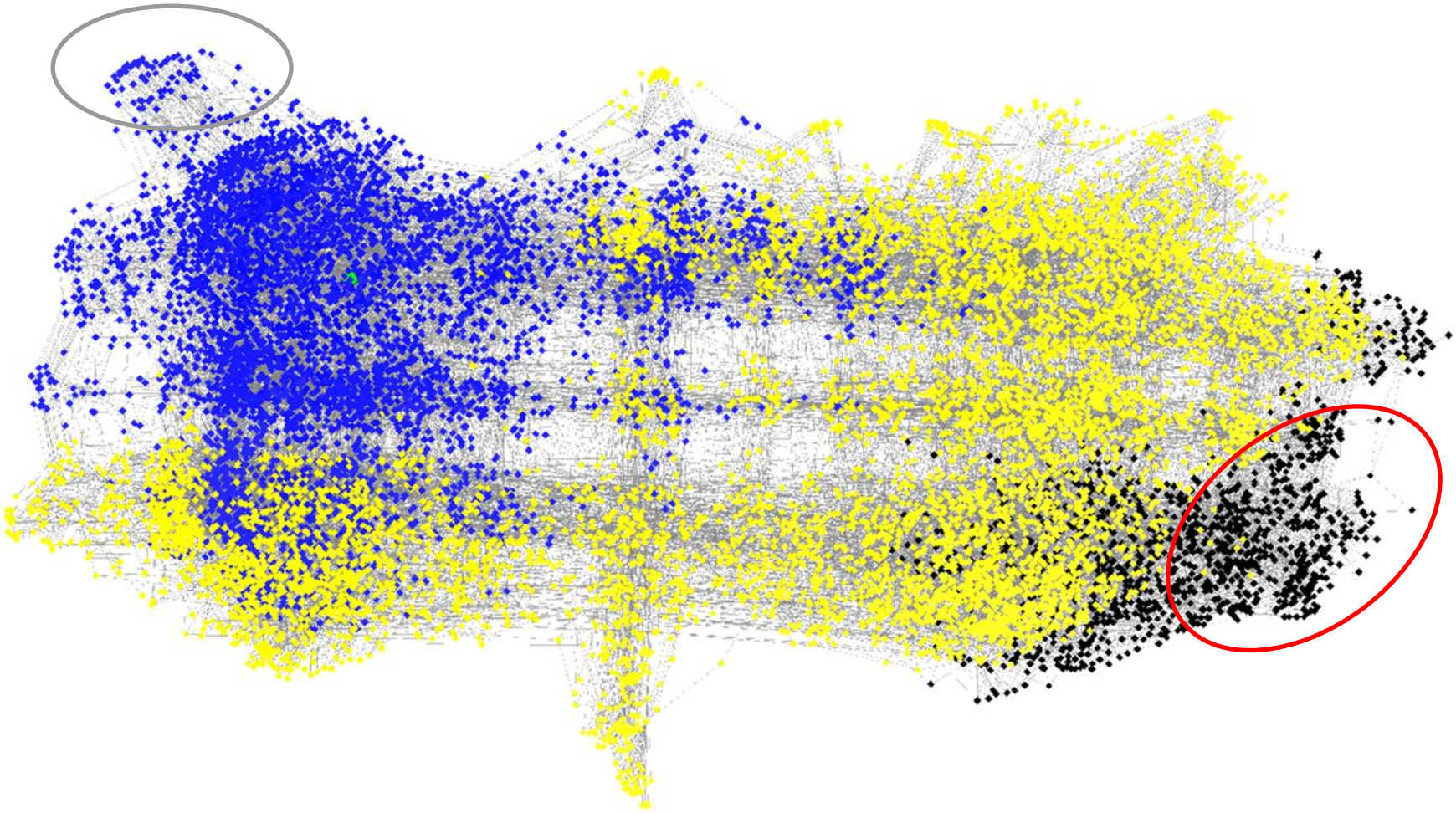
0 generations



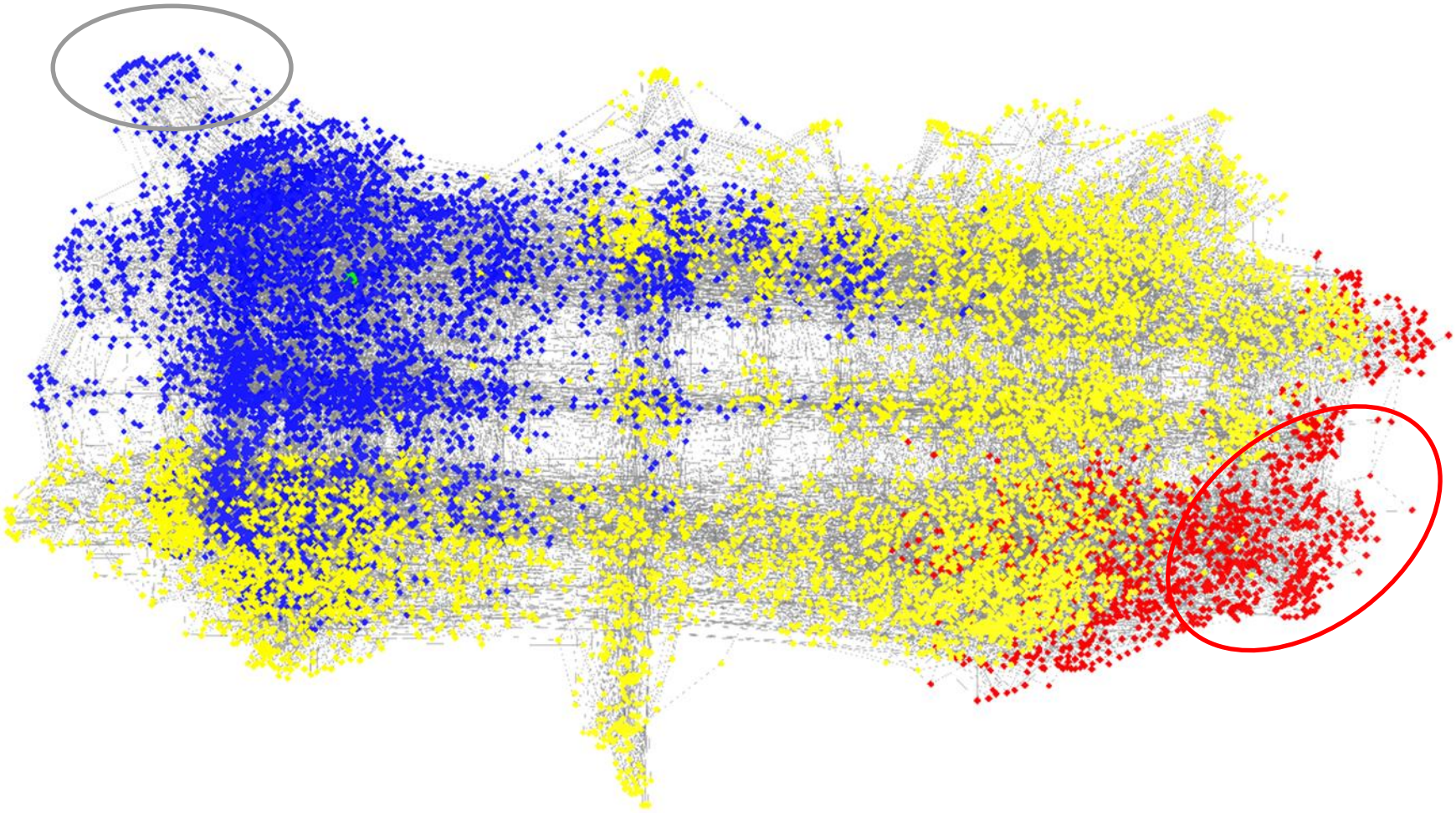
1 to 25 generations



26 to 699 generations



700 to 2500 generations



Sequence Space Traversing (SST) Algorithm

Input: two viral populations P_1 and P_2, $P_1 = \{v_1,\ldots,v_n\}$, $P_2 = \{v_{n+1},\ldots,v_{n+m}\}$

Output: time distance $t(P_1,P_2)$

1. For a sequence set $P_1 \cup P_2$ construct median-joining network $G=(V,E)$

2. Simulate viral evolution using ODE model:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \left(1 - \frac{\sum_{j=1}^{|V|} x_j}{M}\right)\left(rx_i + q\sum_{ji\in E} x_j\right), \qquad i = 1,\ldots,|V|$$

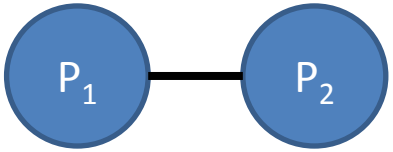
$$x_i(0) = \begin{cases} x_0, i = 1,\ldots,n \\ 0, i = n+1,\ldots,|V| \end{cases}$$

where $r = (1-\varepsilon)^L, q = (\varepsilon/3)(1-\varepsilon)^{L-1}$

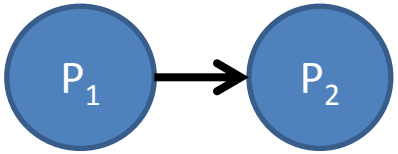
3. $t(P_1,P_2) = \min\{t : x_i(t) \geq x_0 \text{ for } i = n+1,\ldots,n+m\}$

Sequence Space Traversing Algorithm

Two populations P_1 and P_2 are **genetically related**, if $t(P_1,P_2) \leq T^*$



Direction of transmission: if $t(P_1,P_2) \leq t(P_2,P_1)$



Transmission clusters: weakly connected components
Sources: vertices with highest eigenvector centrality

Experimental Results: Data

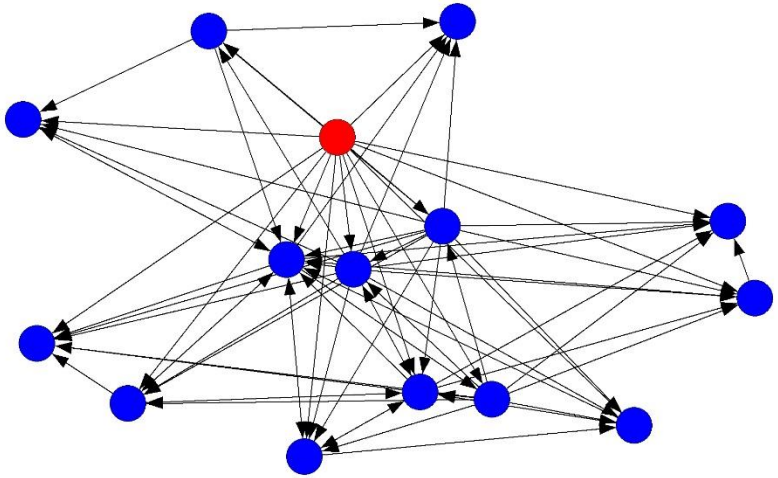
- **Epidemiologically related samples.** 142 HCV HVR1 samples from 33 epidemiologically curated outbreaks reported to CDC in 2008-2013. Sources are known for 10 outbreaks as a result of epidemiological investigations

- **Unrelated samples.** 193 HCV HVR1 samples from infected individuals without any known epidemiological relationship

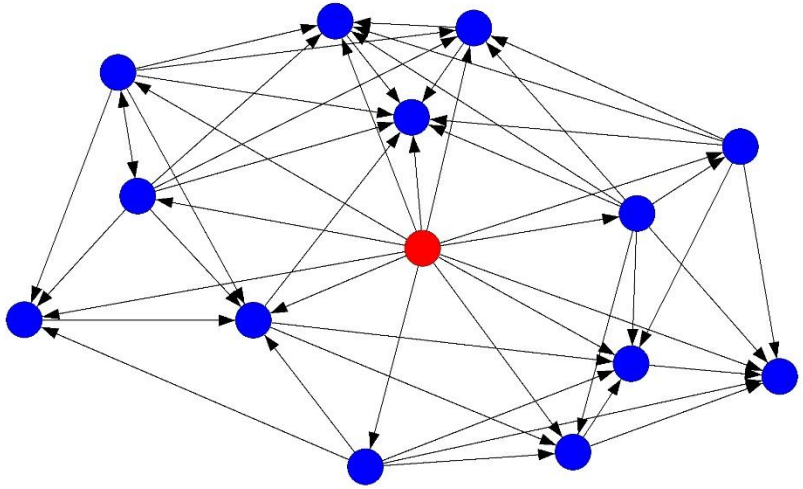
Algorithms for comparison

- Relatedness Depth (ReD)
- Sequence Space Traversing (SST)
- Consensus with 4.5% cutoff
- Consensus with 6.5% cutoff

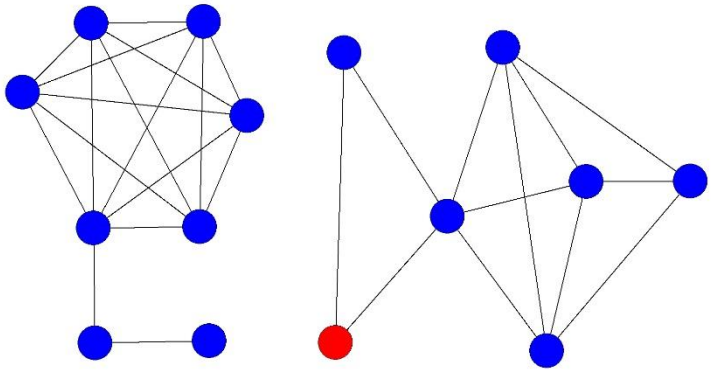
Source estimation accuracy (related samples)



ReD



SST



Consensus (4.5%)

TABLE I

COMBINED RESULTS FOR RELATED SAMPLES (33 CLUSTERS) AND UNRELATED SAMPLES (193 SAMPLES), WITH TPR

Methods	Related samples				Unrelated samples		
	# predicted clusters	TPR	FPR	Source identification accuracy	# predicted clusters	TPR	FPR
ReD	37	98.96%	0%	90%	192	100%	0.01%
SST	37	96.03%	0%	90%	193	100%	0%
CBC[4.5%]	43	81.84%	0%	0%	193	100%	0%
CBC[6.5%]	38	96.66%	0%	10%	171	100%	1.37%

True Positive Rate (TPR) = % of truly related pairs predicted as related

False Positive Rate (FPR) = % of truly unrelated pairs predicted as related

CONCLUSIONS

- Molecular analysis is one of the major tools used for investigations of viral outbreaks and inference of transmission networks. It generates novel bioinformatics problems and challenges

- Replacement of simple consensus-based approaches and expert phylogenetic analyses with novel automatic algorithms is a major advancement in molecular surveillance of viral infections.

- Superior performance of the new algorithms over the traditional consensus-based methods indicates importance of full-edged quasispecies analyses for viral molecular surveillance and outbreaks investigation.

Acknowledgements

Alex Artyomenko
Olga Glebova

Pavel Skums
David S. Campo
Zoya Dimitrova
Nana Li
Seth Sims
Yury Khudyakov

Serghei Mangul
Eleasar Eskin
Ren Sun

Leonid Bunimovich

Nicholas Wu