Chinese CDC workshop

Outbreak detection in real time with genetic clustering

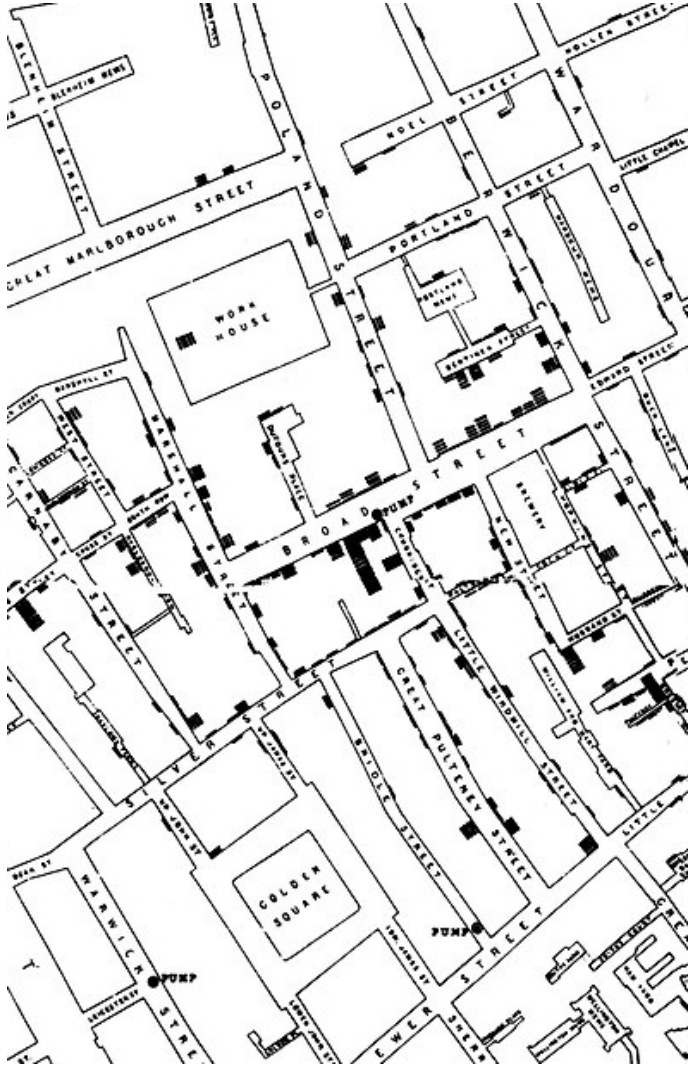
Art Poon

Department of Pathology & Laboratory Medicine
Western University

Background

What is an outbreak?

- A cluster of cases in space (location) and time
- Implies a common underlying cause
- Classic example: 1854 Broad Street cholera outbreak, spatial clustering of cases implicated a contaminated water pump
- In fact John Snow's map was drawn *after* the outbreak (Brody et al. 2000).

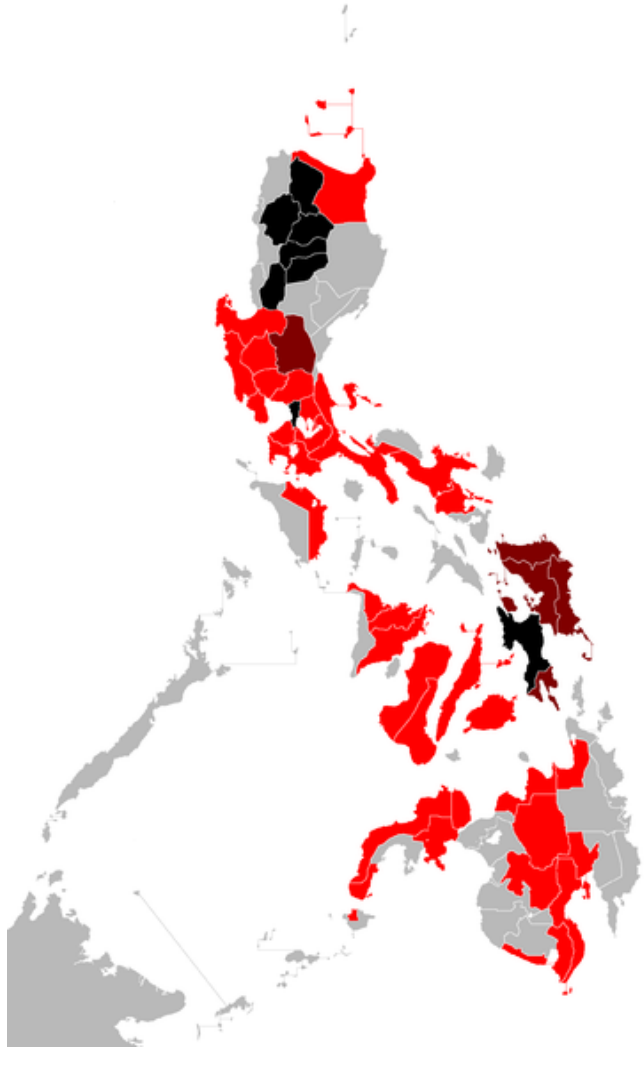


Brody et al. (2000) Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. Lancet 356; 64-68.

Background

Space-time clustering

- Disease surveillance systems*:
- space-time interaction methods (Mantel test)
- scan statistics, *e.g.*, cylindrical space-time scan
- model-based methods: GLMMs, Bayesian sampling



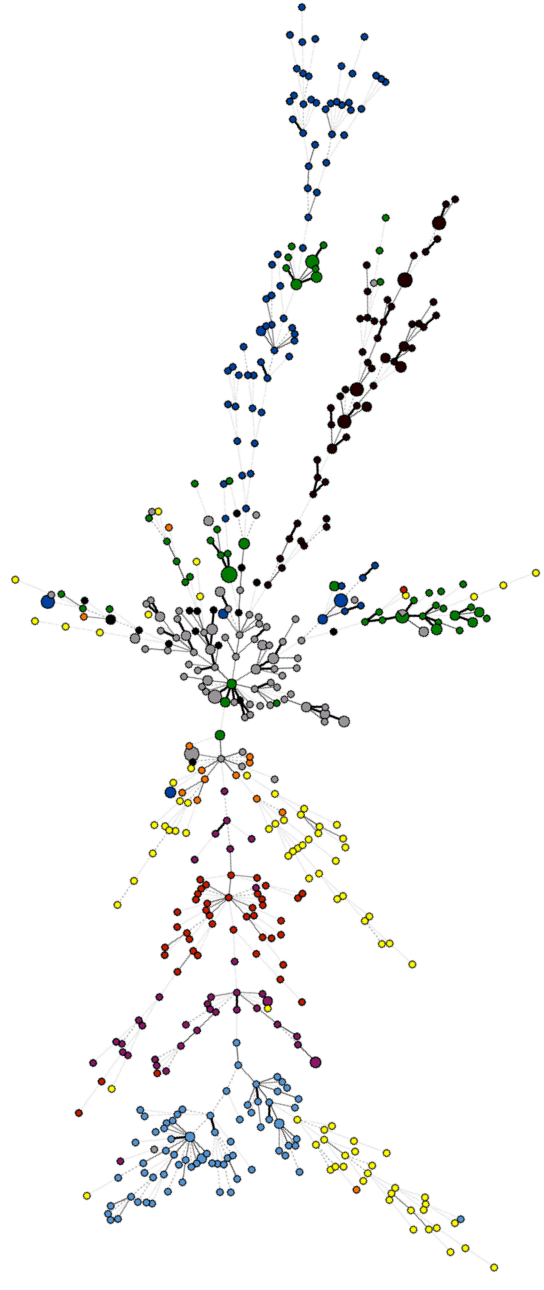
*C Robertson et al. (2010) Review of methods for space-time disease surveillance. Spatial and Spatiotemporal Epidemiology 1(2-3): 105-116.

Background

Genetic clustering

- A subset of infections that are more genetically similar than other infections.
- Early uses of genetic clusters for HSV-1, TB.
- Use genetic "space" to approximate geographic space.
- Needs evolution on same time scale as transmission.

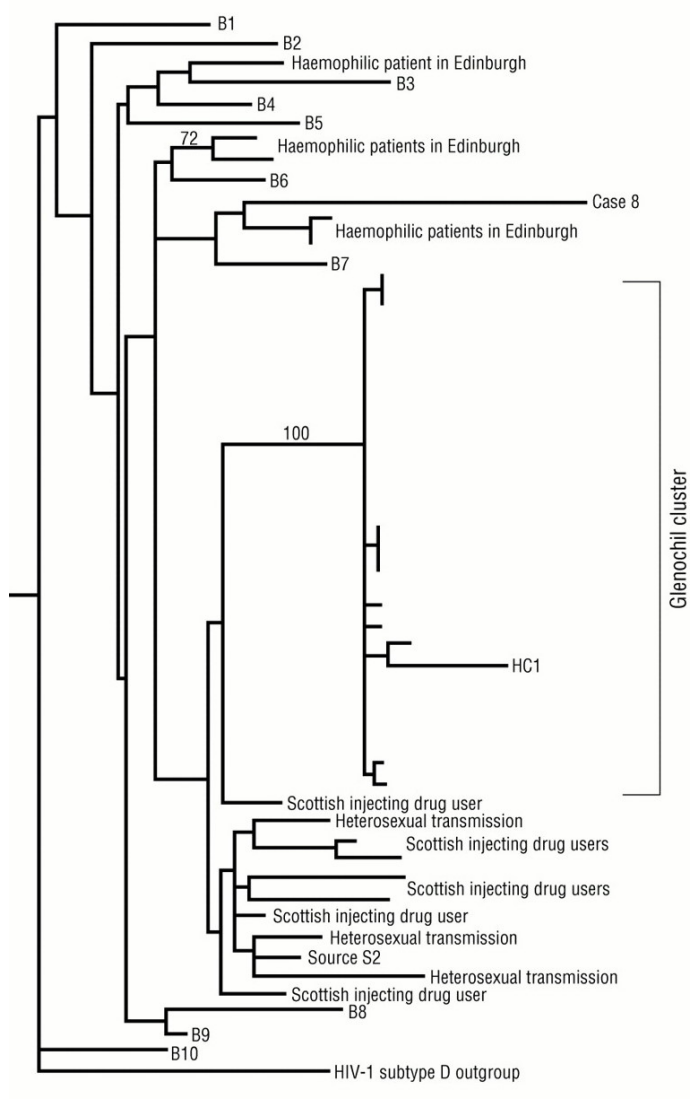
Tuberculosis MIRU-VNTR minimum spanning tree from Reynaud et al. PLOS ONE 2017;e0171584



Background

Genetic clustering and HIV

- HIV-1 outbreak in Glenochil prison, Scotland
- Blood samples collected from 14 inmates positive for HIV infection
- One of the earlier examples of clustering applied to HIV-1
- Now a very popular method for studying HIV transmission



Clustering methods

How do we make clusters?

- Enormous number of *ad hoc* approaches to genetic clustering, no "best" method.
- Pairwise clustering
- Subtree clustering
- Both are *non-parametric* - clusters are informed by empirical distribution(s), not by a model.

Clustering methods

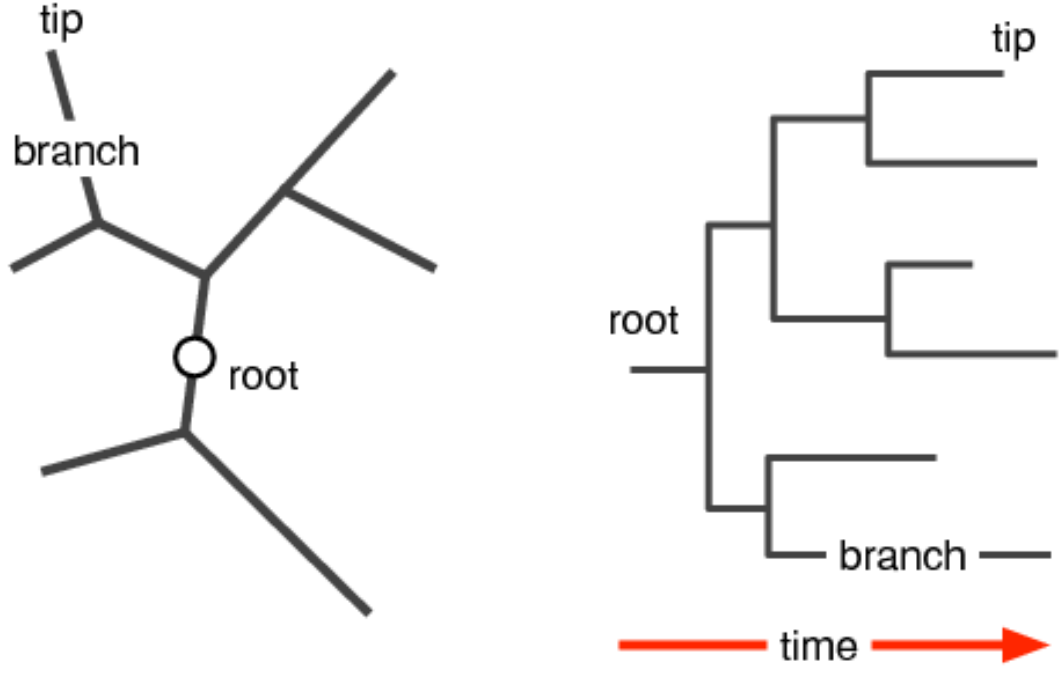
Pairwise distances

- A genetic distance is a function that maps two sequences to a number, $d(s_1, s_2) \rightarrow \mathbf{R}$.
- This distance is often used to approximate their *divergence time*.
- A simple example is the Hamming distance (number of differences between two aligned sequences).
- Build up clusters from pairs of sequences with a distance below some threshold.

Clustering methods

Phylogenies

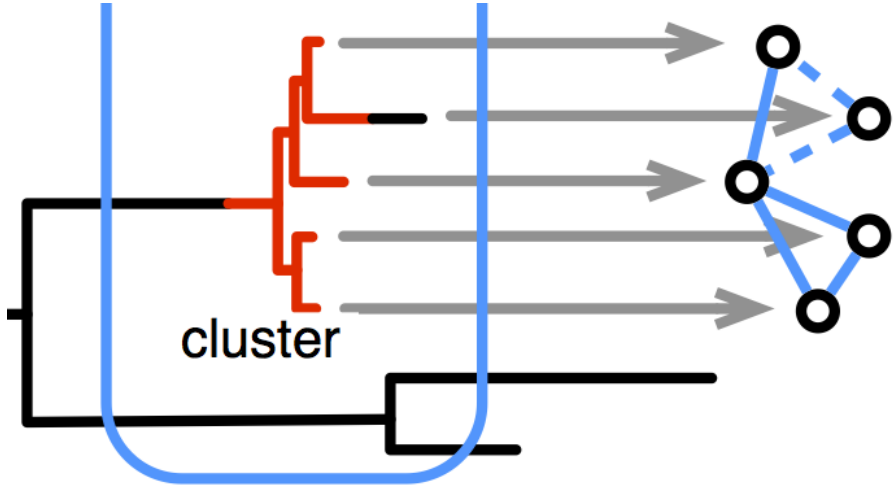
- A molecular phylogeny is a tree-based model of how sequences are related by common ancestors



Clustering methods

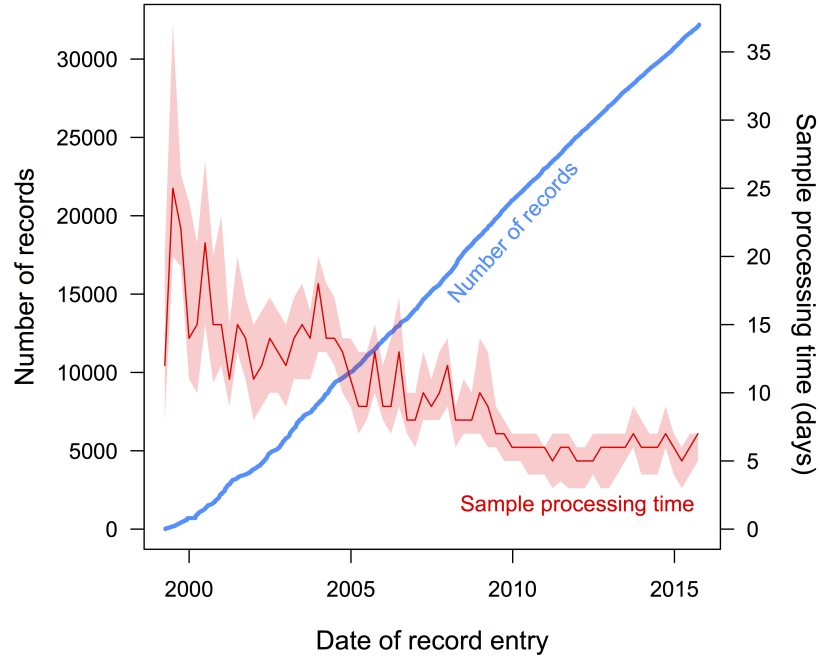
Subtree clustering

- Several different criteria can be used to label a subtree as a cluster
- *e.g.*, mean branch length within subtree
- Bootstrap support - a measure of confidence in subtree given the data



Real-time clustering
Rationale

- Abundant data accumulating at centers of HIV research and care.
- Routine genotyping HIV *pol* for drug resistance screening.
- With sufficient testing, it may be possible to monitor outbreaks in real time (Little et al. 2014*).



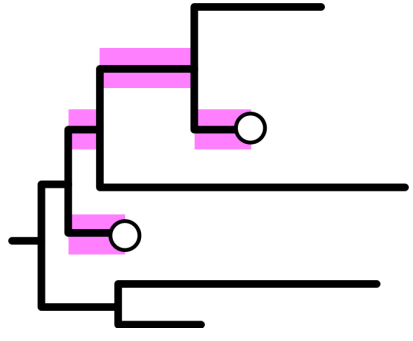
Accumulation of HIV-1 genotypes and faster sample processing time at the BC Centre for Excellence in HIV/AIDS.

*SJ Little et al. *Using HIV networks to inform real time prevention interventions.* PLOS ONE 2014;9(6):e98433.

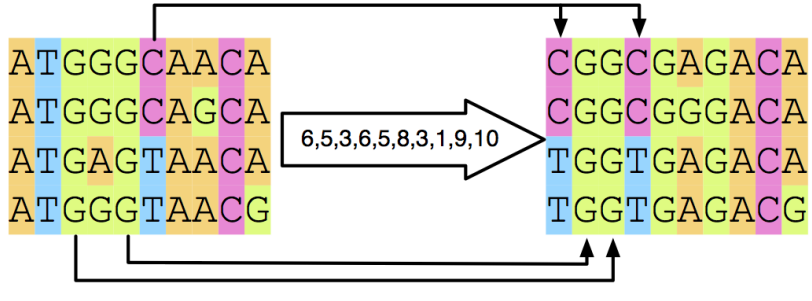
Real-time clustering

Yet another clustering algorithm

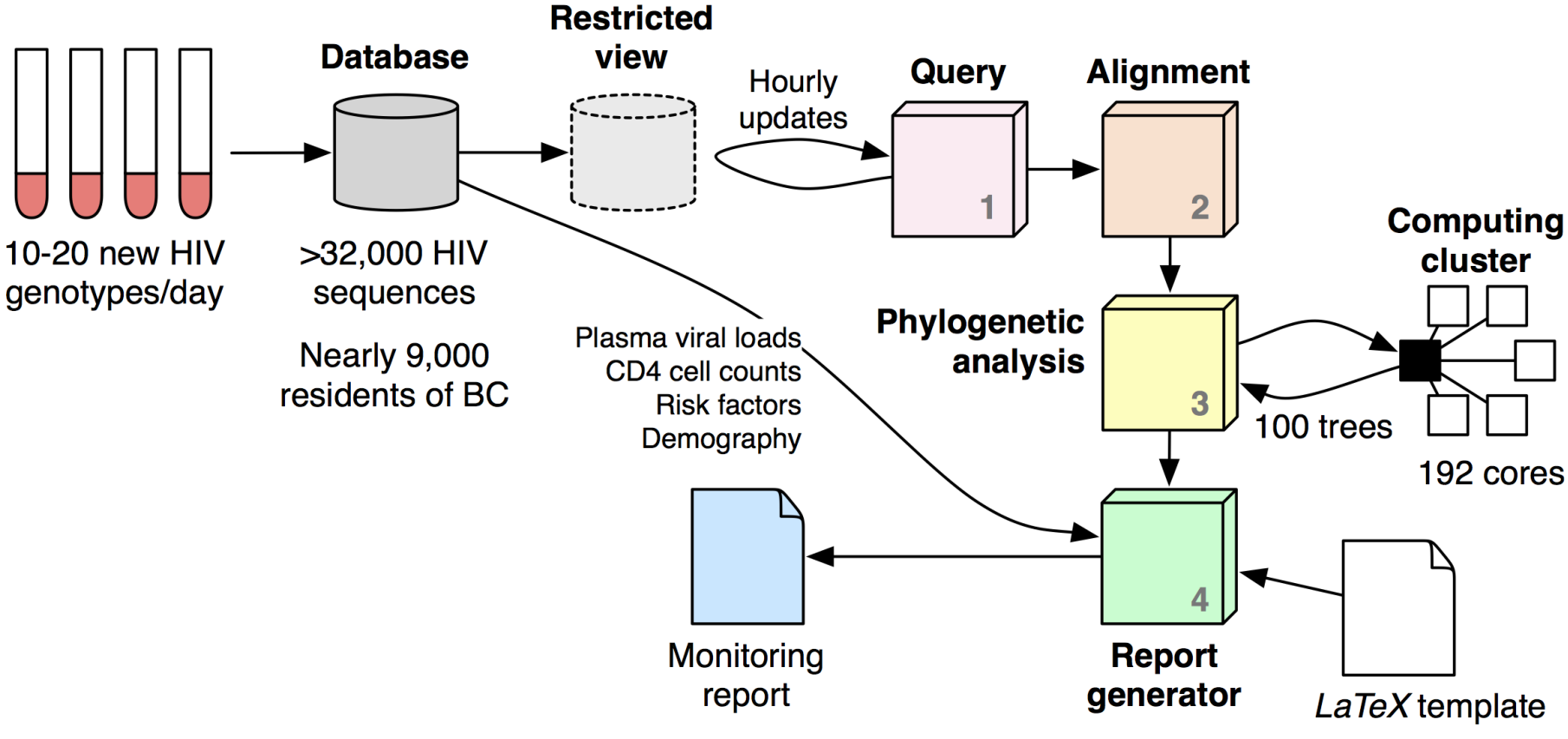
- Patristic distance: the total branch length from one tip to another in the tree (custom Python script).



- Averaged distances over 100 bootstrap trees:



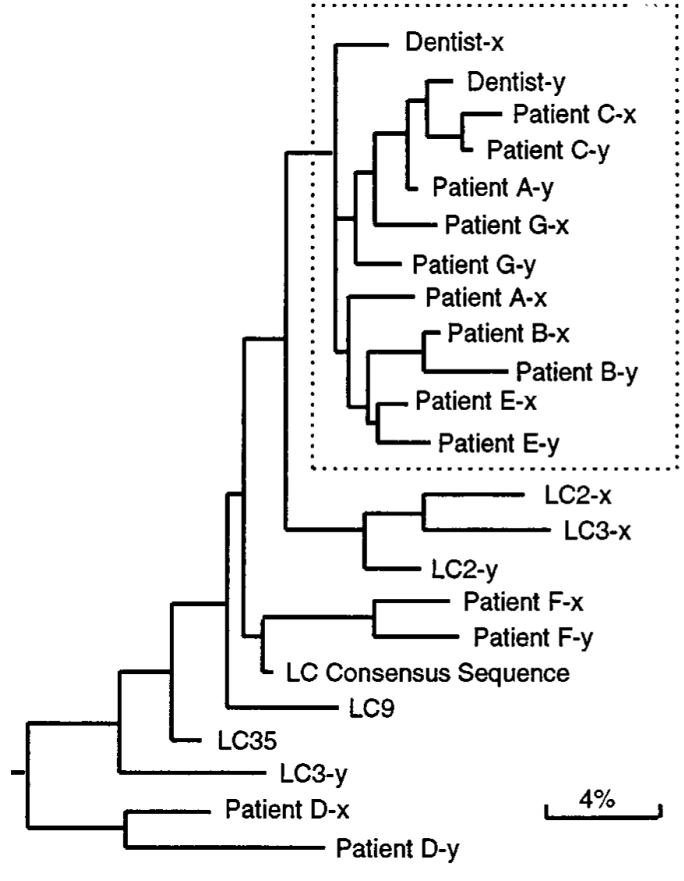
Real-time clustering



Real-time clustering

Ethical and legal concerns

- Long history of using HIV clustering in court cases.
- In Canada, people found guilty for HIV transmission without disclosure can be found guilty of aggravated sexual assault.
- Up to **life imprisonment** and registration on the Sexual Offender Registry.



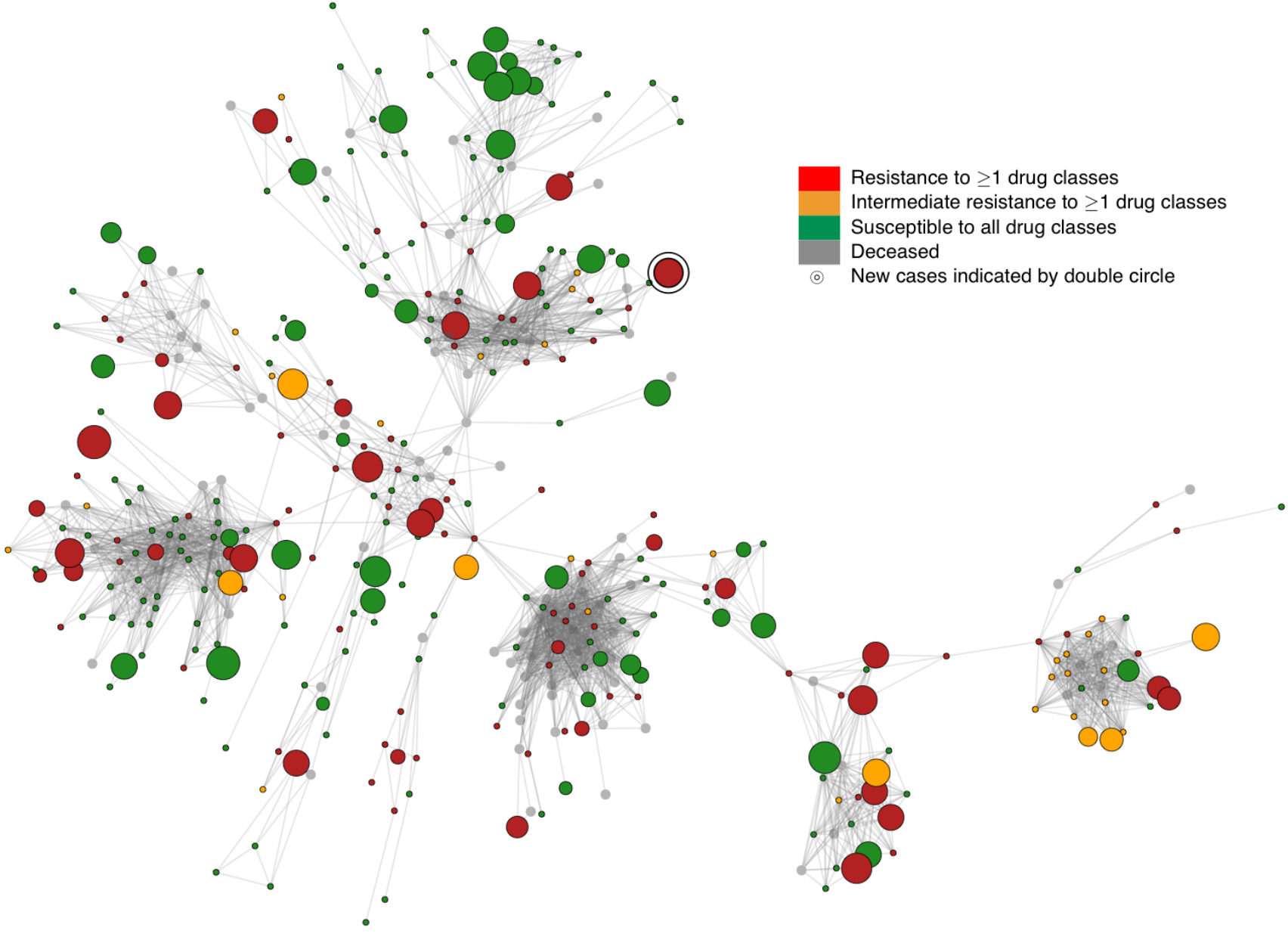
Phylogeny from case of "Florida dentist" from Ou *et al.* (1992, Science).

Real-time clustering

Data security

- Entire network behind dual firewall (no outside access)
- Anonymous patient IDs replaced with randomized labels.
- Sequence data randomized (bootstrap) while preserving evolutionary information.
- All data securely erased from storage media (Guttman method) immediately after analysis.

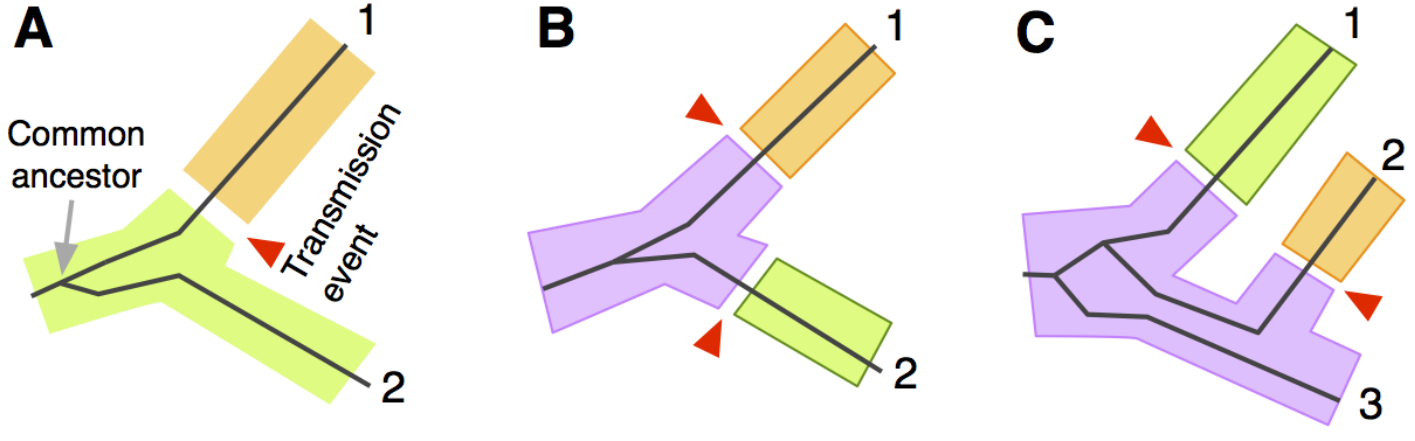
Real-time clustering



Real-time clustering

Reading network diagrams

- Many applications of network diagrams, *e.g.*, contact tracing.
- Public health officials tended to interpret network "ties" as contacts or transmission events.
- Genetic similarity is not sufficient data to infer transmission.



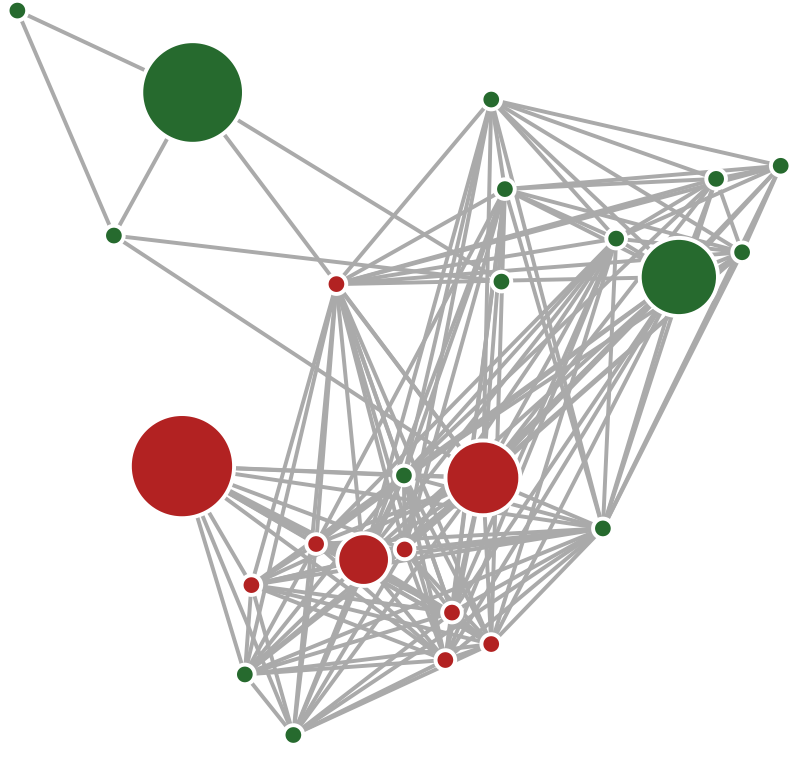
Real-time clustering

What is an actionable cluster?

- How do decide which clusters are actionable for public health?
- A popular indicator was the number of cases in the last month / 3 / 6 months: **genetic-time clustering**.

Index	Total	New	Coastal				Island				Fraser			
			total	6mo	3mo	new	total	6mo	3mo	new	total	6mo	3mo	new
65	66	2	55	7	3	1	1	1	1	1	3	1		
32	29	1	25	1	2		12	1	1	1	1			
54	16	1	9				1				1	1	1	2
188	17	1	3				9	1	1	2	2			
122	11	1	4	1	1	1	1				2			
180	8	1	1				1				1			

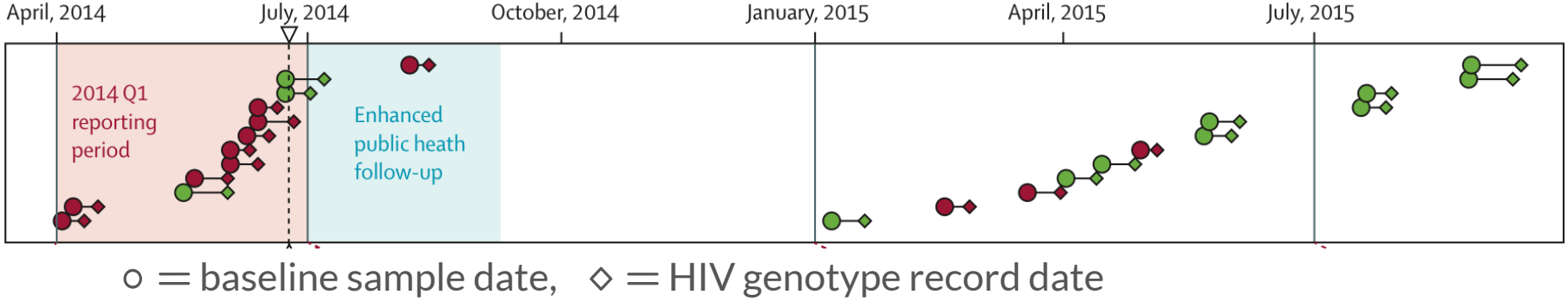
Real-time clustering



Cluster 55

Real-time clustering

Cluster 55 timeline



- 8 new cases in 3 months detected (June 24, 2014)
- Public Health Officer authorizes formal outbreak investigation (June 27, 2014)
- Investigation followed by public health follow-up (July 2, 2014)

Real-time clustering

Retrospective

- Perhaps the first example of automated real-time genetic monitoring directly leading to a public health response.
- Not feasible to prove effect of intervention without an unethical controlled study.
- Similar systems are now being built around the world.

Limitations of clustering

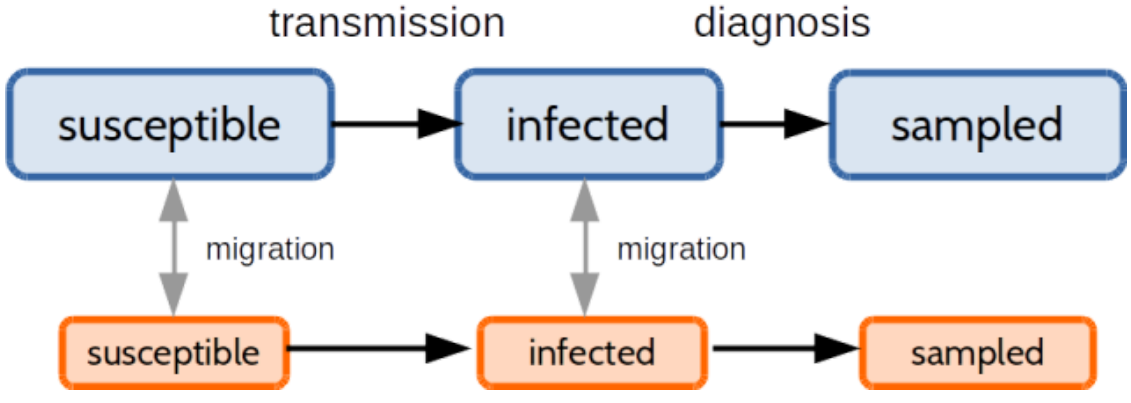
What do clusters really mean?

- Inherent assumption that a cluster represents an outbreak.
- Infections may be genetically similar only because they were rapidly diagnosed.
- Explains strong association between clusters and acute infections (Volz et al. 2012).

EM Volz et al. (2012) *Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection*. PLOS Comput Biol 28:e1002552.

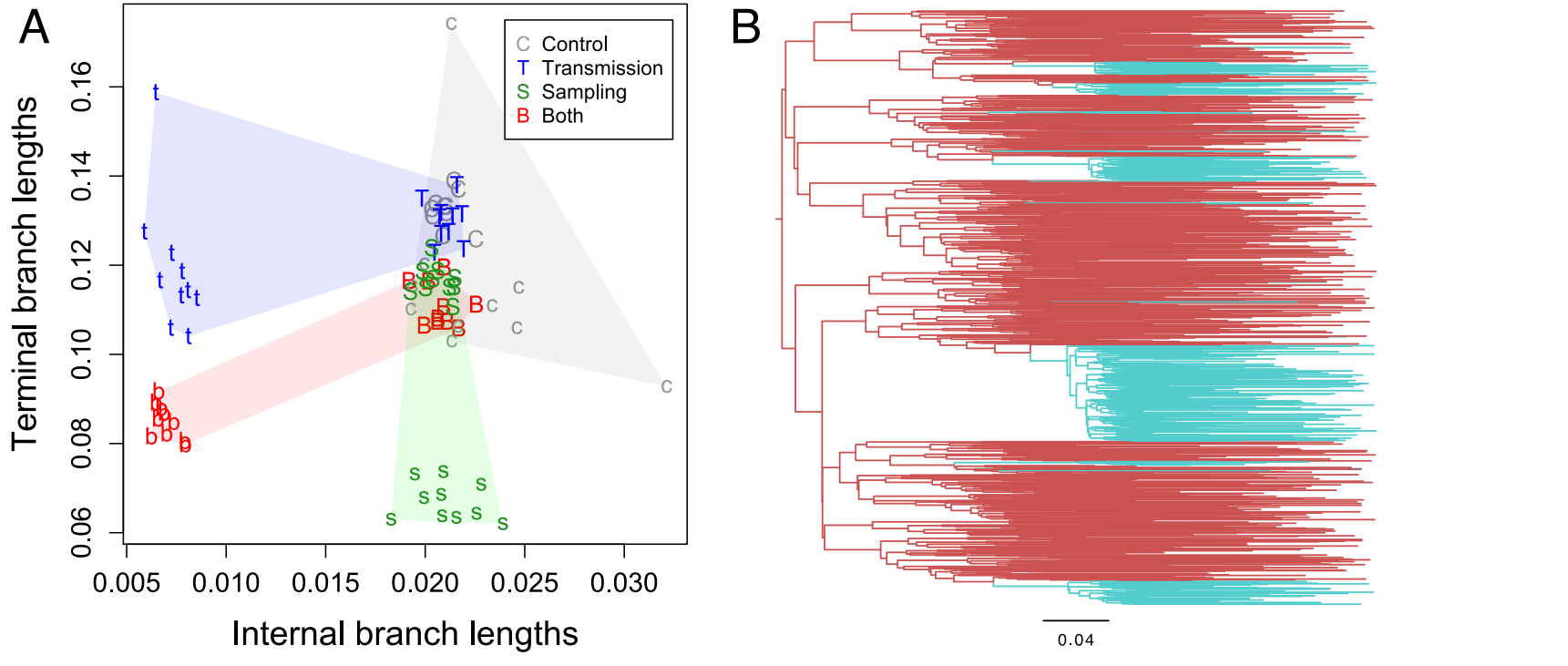
Limitations of clustering

Simulation experiments

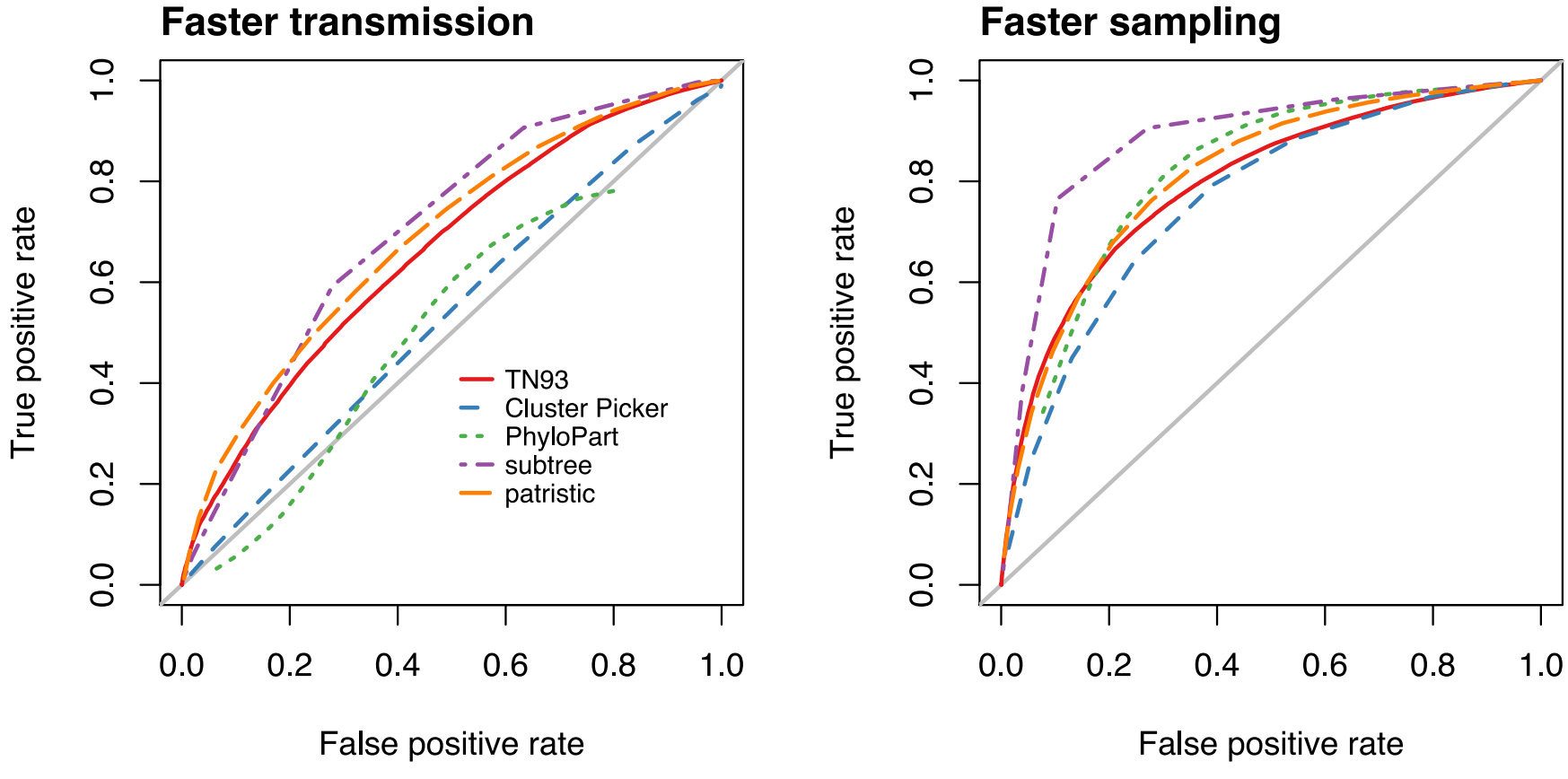


- Epidemic in a risk-structured population.
- Compartmental SIR dynamics.

Limitations of clustering
Simulation outputs



Current methods tend to pick clusters of sampling.



Poon (2016) *Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks.* Virus Evol 2(2):vew031.

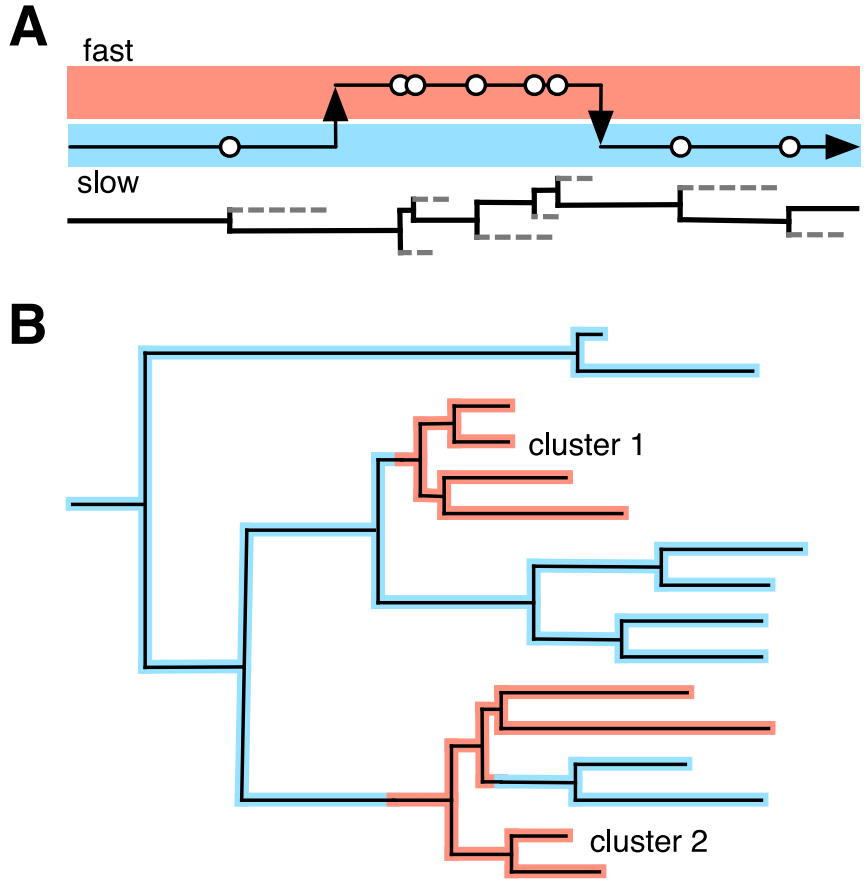
Model-based clustering
A new approach

- Instead of looking at genetic similarity, what we focus on what we really care about?
- Coalescence rate is more related to the rate of transmission than the number of infections (Volz et al. 2009).
- We develop a model to examine HIV transmission through branching rates (coalescence).

EM Volz et al. (2009) *Phylodynamics of infectious disease epidemics*. Genetics 183: 1421.

Model-based clustering

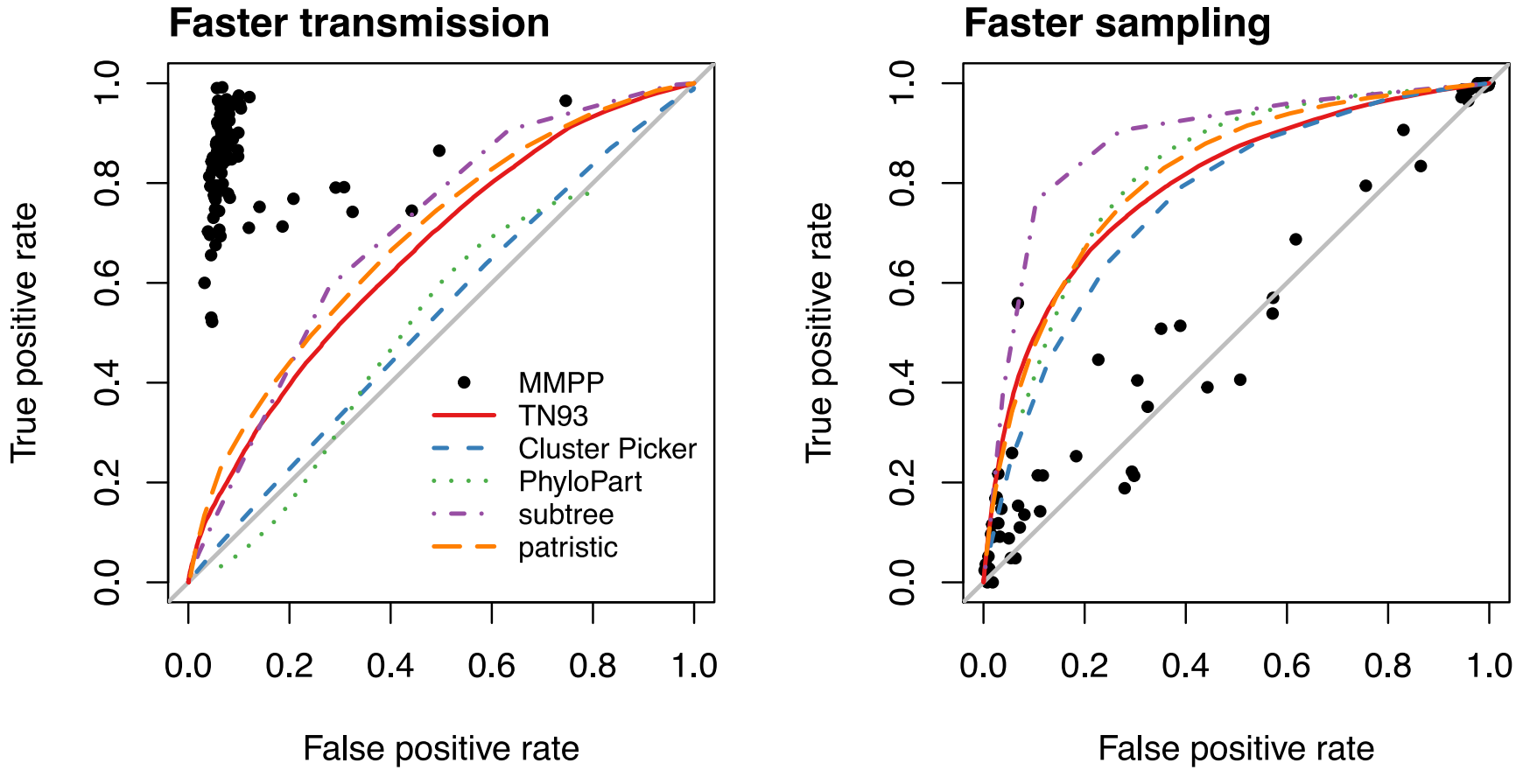
Markov-modulated Poisson process



McCloskey and Poon (2017) PLOS Comput Biol 13(11): e1005868

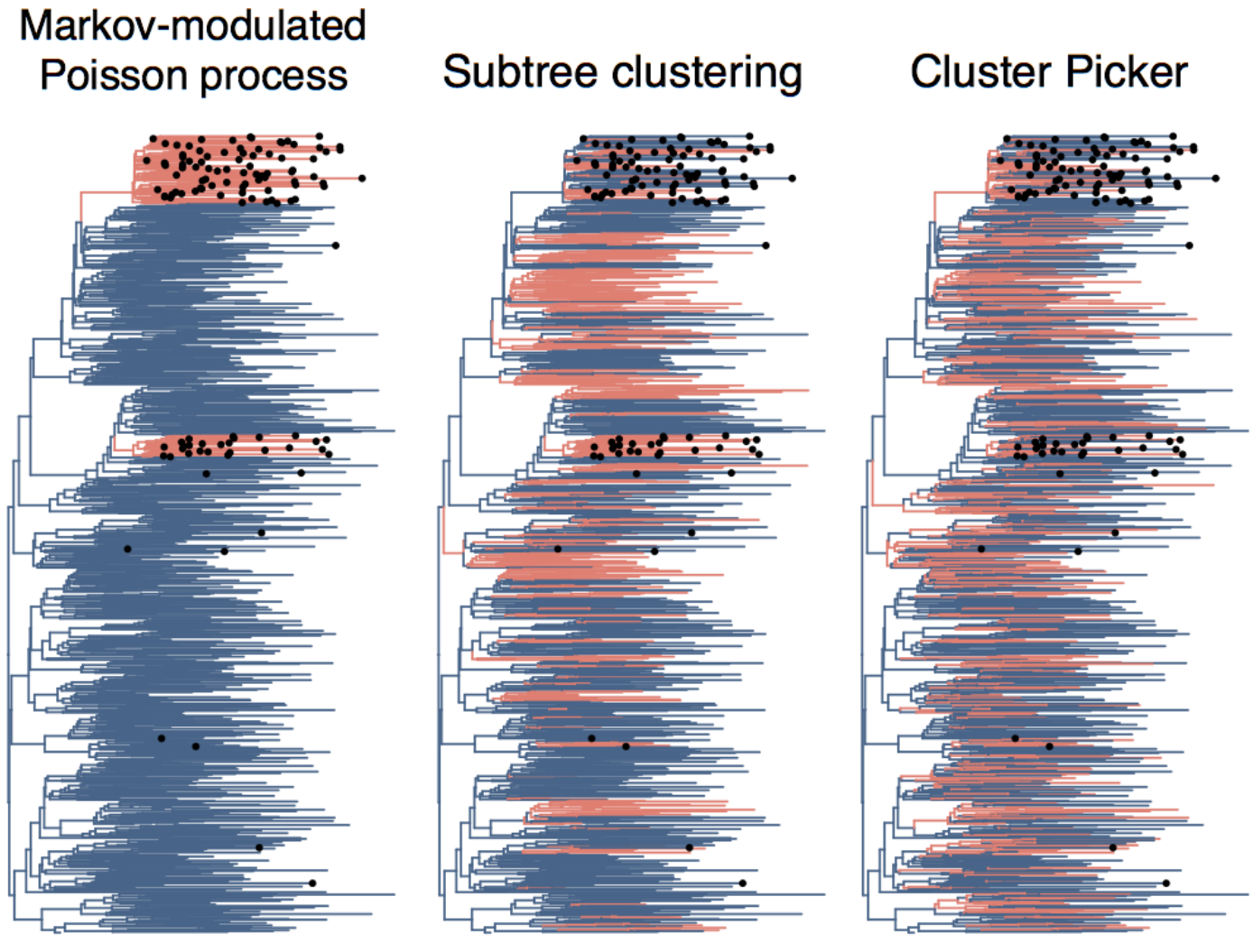
Model-based clustering

MMPP vs. nonparametric methods



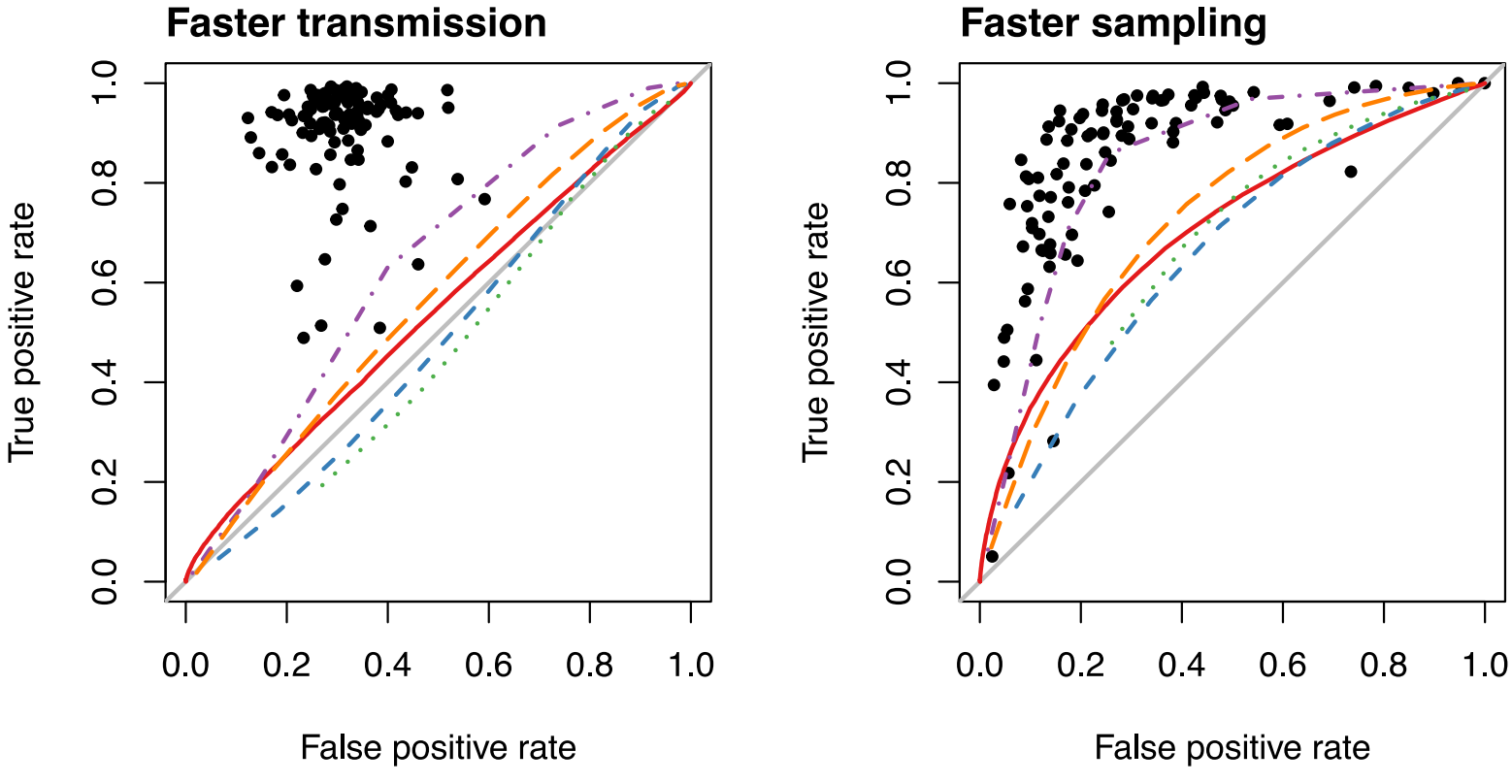
McCloskey and Poon (2017) PLOS Comput Biol 13(11): e1005868

Model-based clustering
Detailed comparison



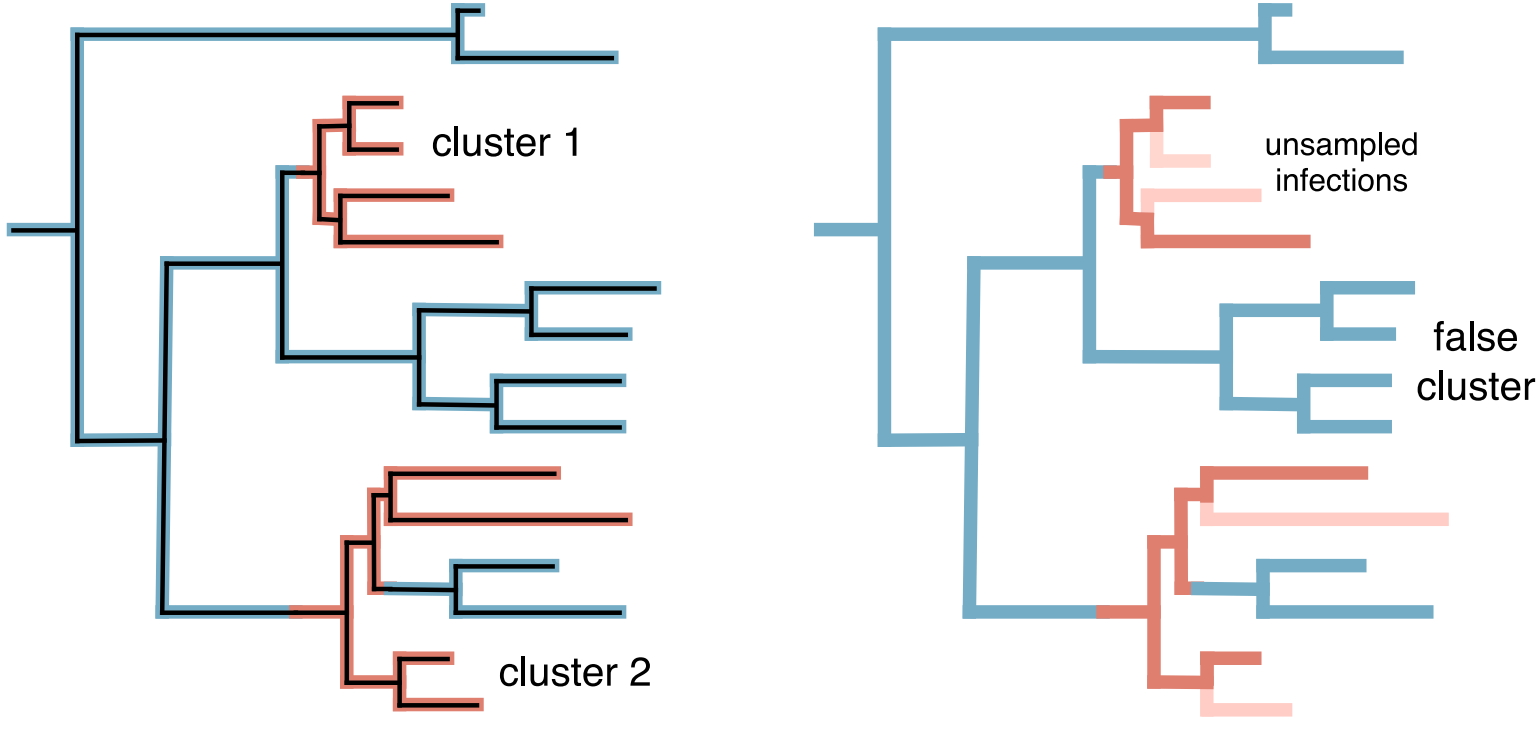
Model-based clustering

Incomplete sampling is still a problem



Results from simulations with sampling reduced from 98% to 50%.

If high-risk groups are less sampled (under-diagnosed) then **clustering may systematically bias public health away from groups most at risk.**



clmp: an R module (Clustering with MMPP)

http://github.com/PoonLab/clmp

```
> require(clmp)
Loading required package: clmp
Loading required package: ape

> t1 <- read.tree('examples/test.nwk')  # a simulated tree with 1000 tip
> res <- clmp(t1)  # returns an ape::phylo tree object
log likelihood for 2 state model is 2238.543290
rates: 495.368085 1305.115860
Q: [    *   2.526691 ]
   [ 23.309483   *    ]

> index <- match(t1$tip.label, names(res$clusters))
> labels <- grepl("_1_", t1$tip.label)  # extract truth from the tip lab
> table(labels, res$clusters[index])

labels    0   1
  FALSE 860   3  # false positive rate, 3/(3+860)=0.34%
  TRUE   13  98  # true positive rate, 98/(98+13)=88.2%
```

Concluding remarks

- There are many ways to define clusters - we still do not know which method is most effective for real-time prevention (but see Wertheim *et al*. 2018).
- Model-based clustering seems to confer greater accuracy for detecting clusters of transmission.
- *All* methods suffer from high false positive rate due to incomplete sampling.

JO Wertheim et al. (2018) *Growth of HIV-1 Molecular Transmission Clusters in New York City.* J Infect Dis, in press.

Thanks!


Ontario Genomics


GenomeCanada


NIH
NIAID


CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada


NSERC
CRSNG