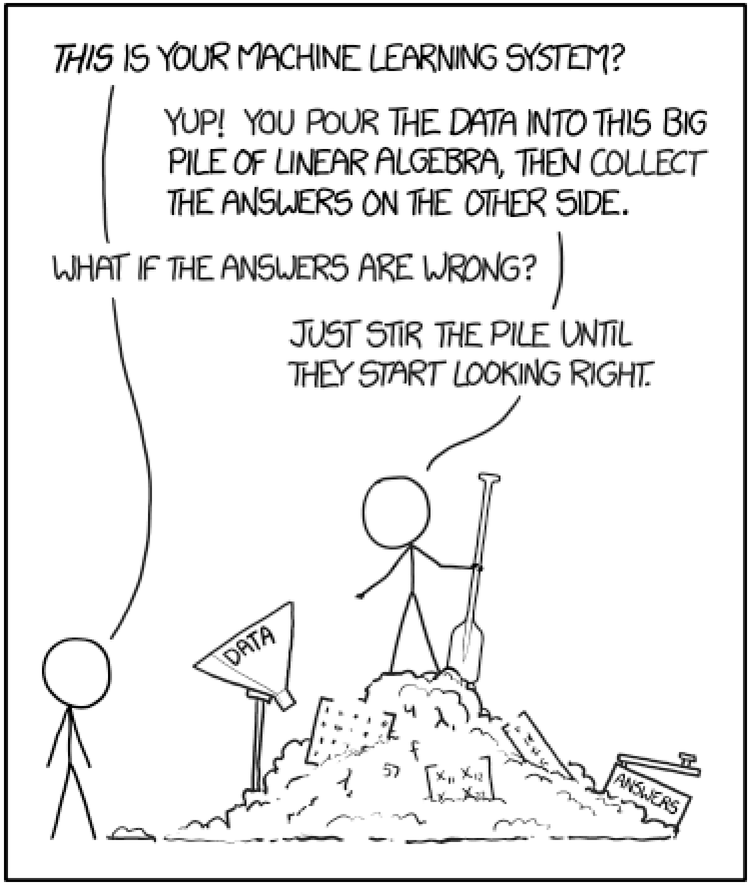
MIMM4750G
Genetic clustering


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

What is a cluster?

- A cluster is a subset (group) of objects that are more similar to each other than objects outside the cluster.
- Clustering is subjective. Our brains are wired to see patterns where none exist.

Why is clustering useful?

- Clustering is a means of finding useful patterns.
- To reduce a large database to a representative subset.
- For infectious disease research:
    - Clustering can be used to define bacterial "species" (limited morphology, extensive horizontal gene transfer).
    - To define strains or "subtypes" of a virus.
    - To track the spread of an infectious disease.

Clustering methods

- There are an enormous number of methods (algorithms) for clustering data.
- It is easiest to talk about different categories of clustering methods.
- Clustering is used in so many contexts that it can be confusing when different methods are used on different kinds of data in the same study!

Supervised and unsupervised clustering

- Terms associated with machine learning.
- *Supervised* clustering means that you have assigned some data to clusters yourself, and leave the rest to the machine.
- *Unsupervised* clustering means that the machine has to figure it all out itself.

Agglomerative and dissociative

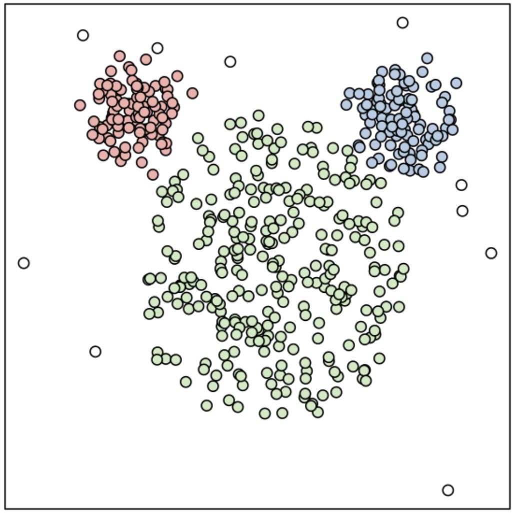
- *Agglomerative* (bottom-up) clustering begins with every object in its own tiny cluster, and starts lumping the closest together.
- *Dissociative* (top-down) clustering begins with every object in one huge cluster, and starts cutting.

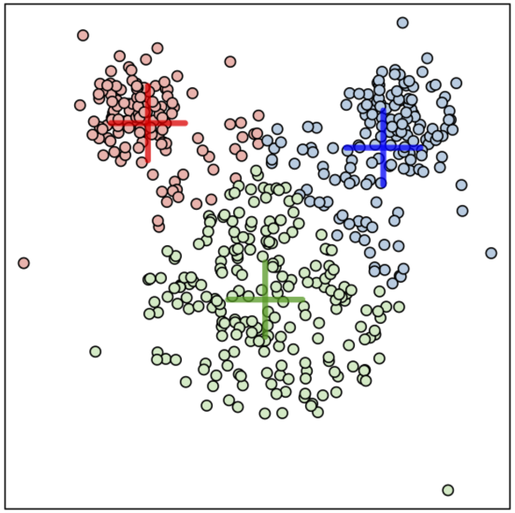
Non-parametric and parametric

- A *non-parametric* clustering method uses the observed distribution of one or more characteristics to cluster the data.
- For example, if we look at cars on a one-lane road, we can build up clusters from any two cars closer than some cut-off distance of each other.
- A *parametric* clustering method fits a model to the data to define clusters.
-

k-means clustering

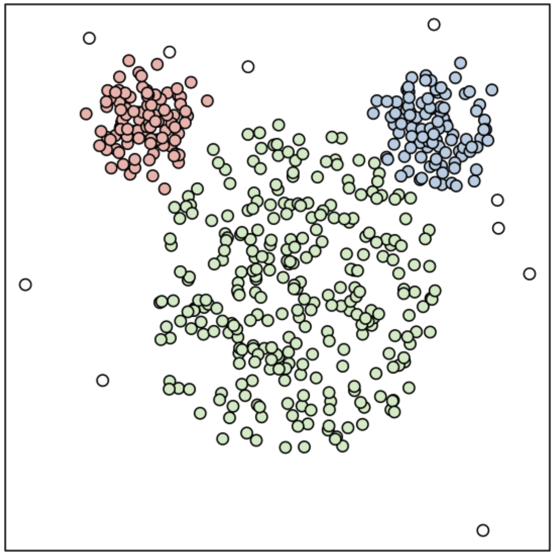
- A popular clustering method (unsupervised, dissociative, nonparametric)
- *k* refers to the number of clusters defined by "means".
- Assign each point to the closest mean, while locating the optimum locations of means.
- (top) A simulated dataset with three clusters, called *mouse.*
- (bottom) A k-means clustering of *mouse* with *k* set to the true value.

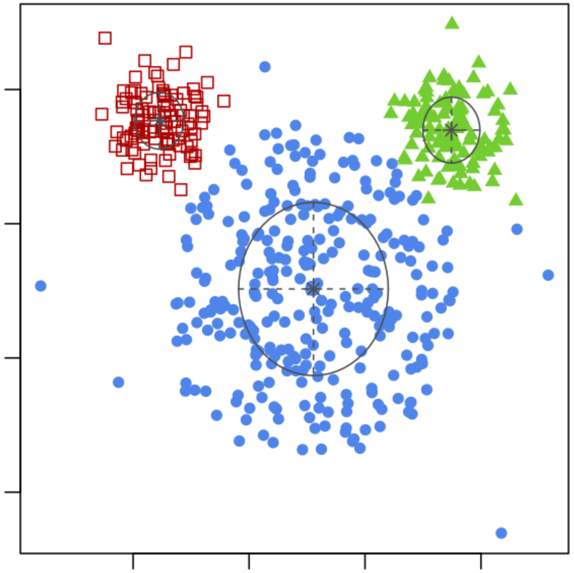




Gaussian mixture models

- Another popular clustering method (unsupervised, **parametric**)
- Find the assignments of each data point to one of *k* Gaussian distributions.
- Also find the mean and variance of each Gaussian that maximizes likelihood.
- Method can determine for itself the optimal number of clusters.
- (bottom) Gaussian mixture model applied to *mouse* data.





Distance-based clustering

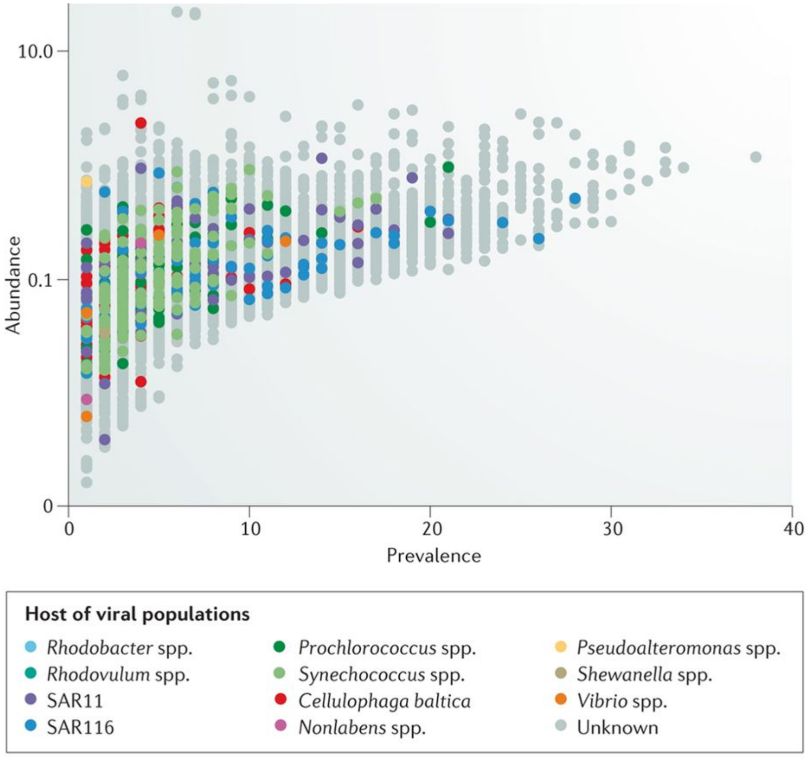
- A simple nonparametric clustering method that is popular for sequence data.
- Look at the distribution of all distances between pairs of objects.
- The distance may be a function of one or more features, *e.g.*, Euclidean distance, $\sqrt{(x_1^2 + x_2^2 + \ldots + x_n^2)}$.
- Pick a threshold - any pair below the threshold forms a cluster.

Genetic distance clustering

- Recall from last lecture, a *genetic distance* is used to quantify the difference between two sequences.
- The Tamura-Nei (1993, TN93) distance is the most complex distance that can be written as a closed-form expression.
- The International Committee on the Taxonomy of Viruses allows the definition of a new virus species based on genetic clustering, although this remains controversial.

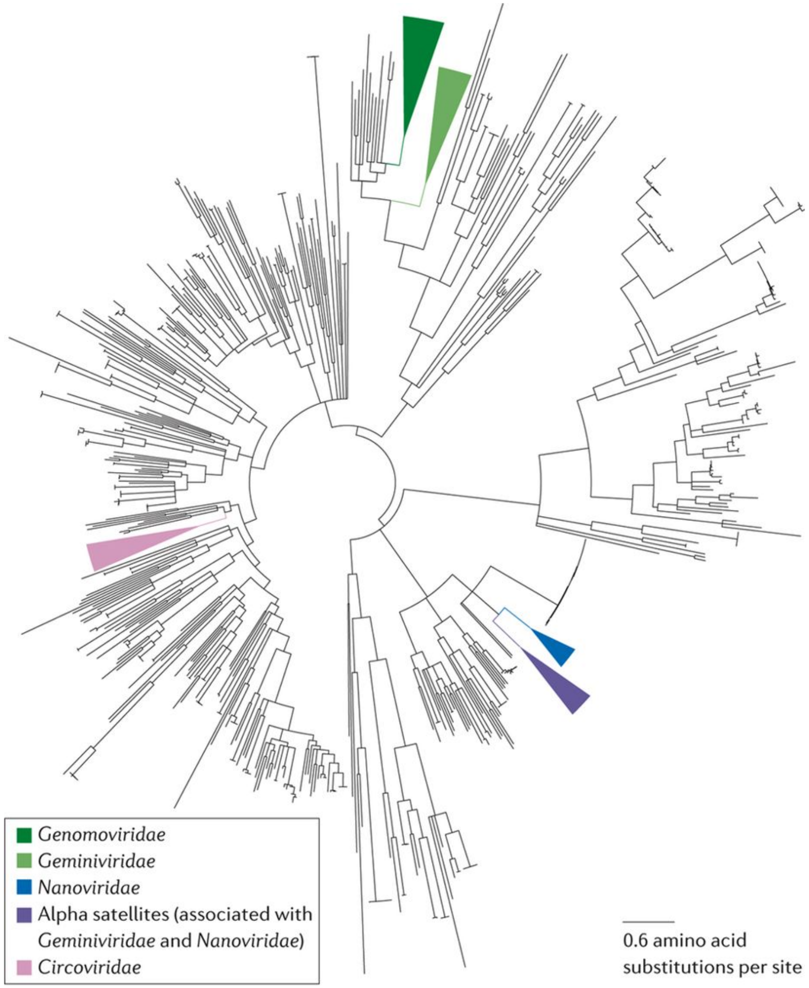
Defining new virus species

Prevalence and abundance of marine viruses



Nature Reviews | Microbiology

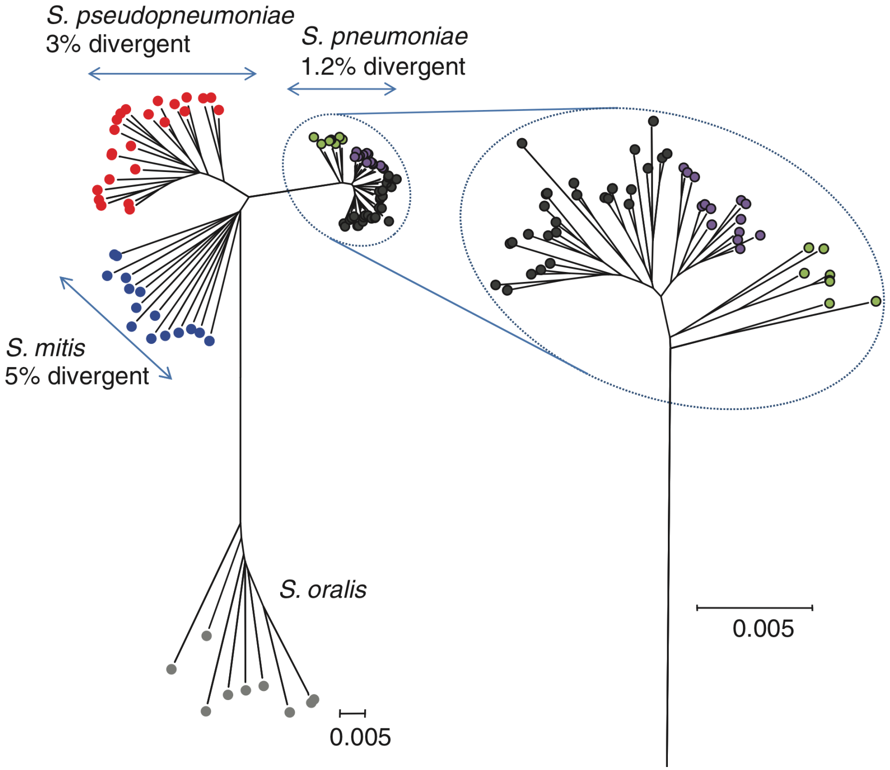
Phylogeny of known and novel circular REP-encoding ssDNA viruses


■ *Genomoviridae*
■ *Geminiviridae*
■ *Nanoviridae*
■ Alpha satellites (associated with *Geminiviridae* and *Nanoviridae*)
■ *Circoviridae*
0.6 amino acid substitutions per site

Nature Reviews | Microbiology

Defining bacterial species

- Specific loci are frequently used to measure bacterial diversity (*e.g.*, 16S sRNA)
- Horizontal transfer of genes between different bacteria makes it difficult to define species.
- This problem can be overcome by multilocus sequence analysis (MLSA): using conserved "housekeeping" genes to generate a phylogenetic tree.



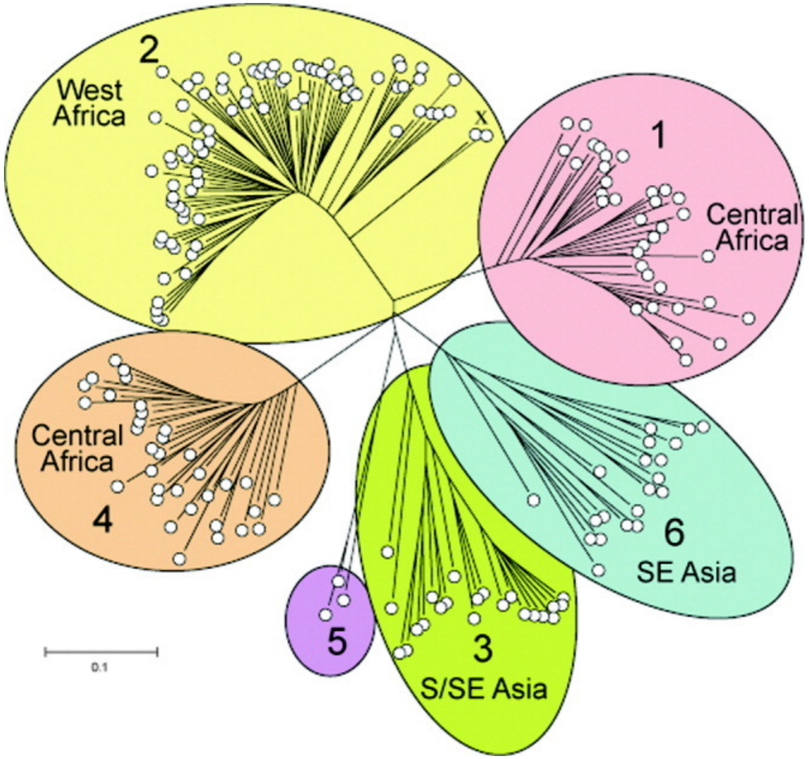
MLSA phylogenetic tree of *Streptococcus* from C Fraser *et al.* (2009) Science 323: 741.

Defining subtypes: HCV

- Clinical significance of virus subtypes (genotypes): differences in pathogenesis, response to treatment.
- Hepatitis C virus is a flavivirus that can cause fatal liver disease if not cleared by the immune system.
- About 71 million people worldwide have chronic HCV infection.
- Rapid evolution: HCV in individuals infected from the same source can become >35% divergent over 17 years (McAllister *et al.* 1998).

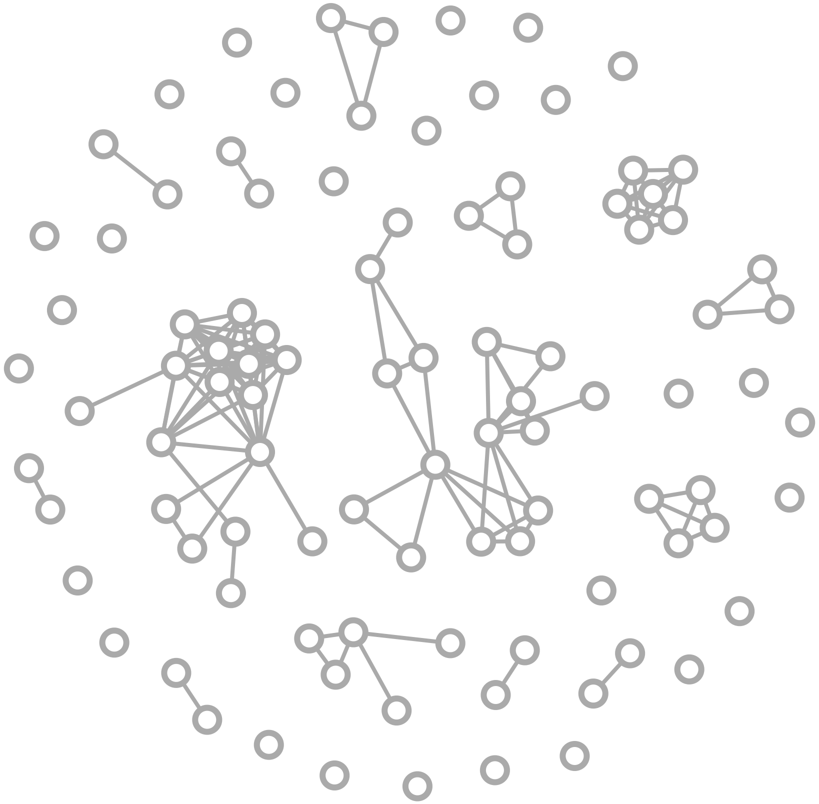
HCV genotype and subtype clustering

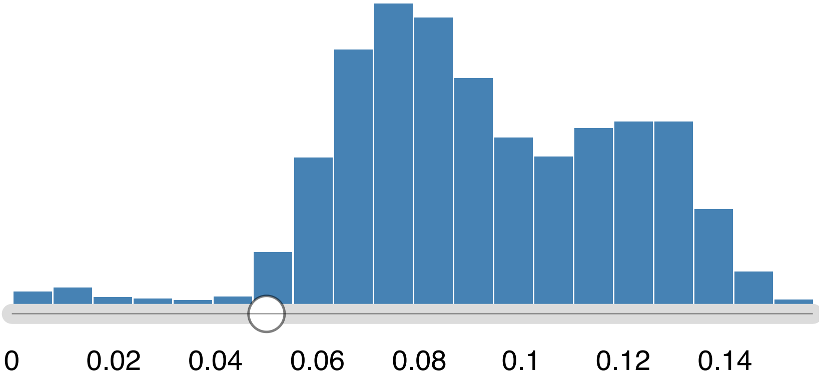
- Early system used distance clustering.
- Used p-distance of aligned sequence to make tentative assignment.
- Next generate a tree by clustering on K2P distances.



HCV distance-based tree from P Simmonds *et al.* (2005) Hepatology 42: 962.

TN93 clustering



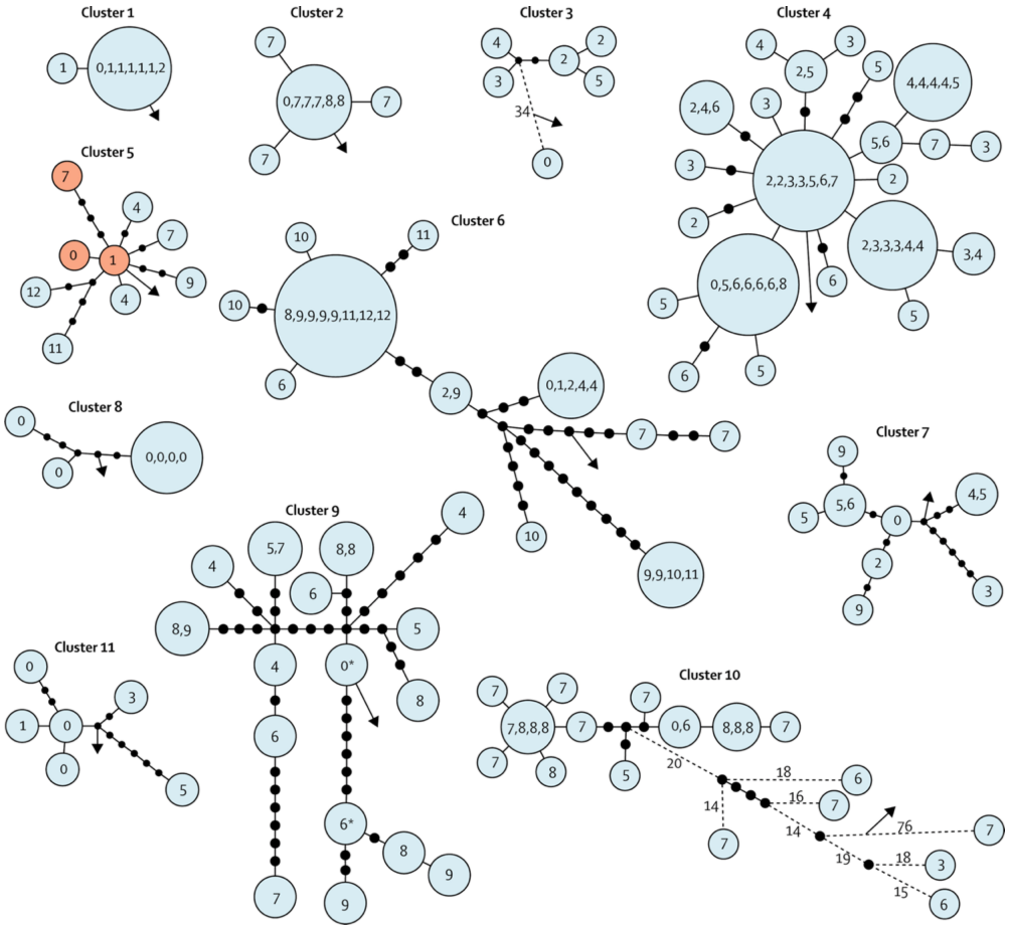


Clustering for epidemiology

- In public health, a cluster of cases in space and time implies a common source.
- A *genetic* cluster of infections similarly suggests that they are related by recent and rapid transmission events.
- Genetic clustering requires measurable evolution on a similar time scale as transmission.

Tuberculosis

- TB is one of top 10 causes of death worldwide
- Caused by lung infection by *Mycobacterium tuberculosis.*
- Clustering of whole-genome sequence data can idenfity high-risk groups and detect undiagnosed cases.



Cluster diagram by TM Walker *et al.* (2013) Lancet Inf Dis 13: 137.

Suggested readings

- Consensus statement: Virus taxonomy in the age of metagenomics
- ICTV: Comments to proposed modification to code rule 3.21 (defining virus species)