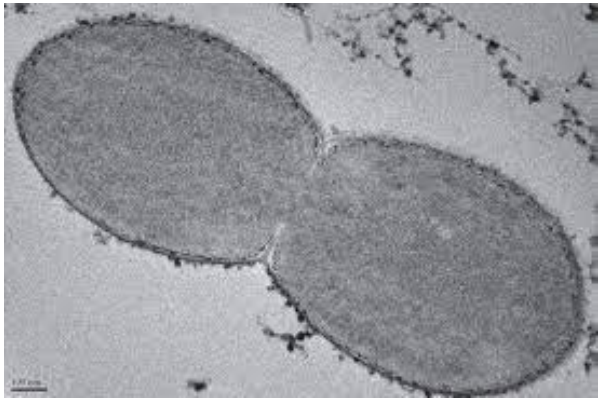
Selection and diversity in the weird and wonderful *Streptococcus pneumoniae*

Jennifer McNichol

April 3, 2024

Streptococcus pneumoniae (pneumococcus)

- A human nasopharyngeal commensal and respiratory pathogen
- Leading cause globally of pneumonia and meningitis in children
- Diverse population with over 90 serotypes
- Serotypes are associated with markedly different patterns of carriage and disease
- Vaccines are created to target specific serotypes

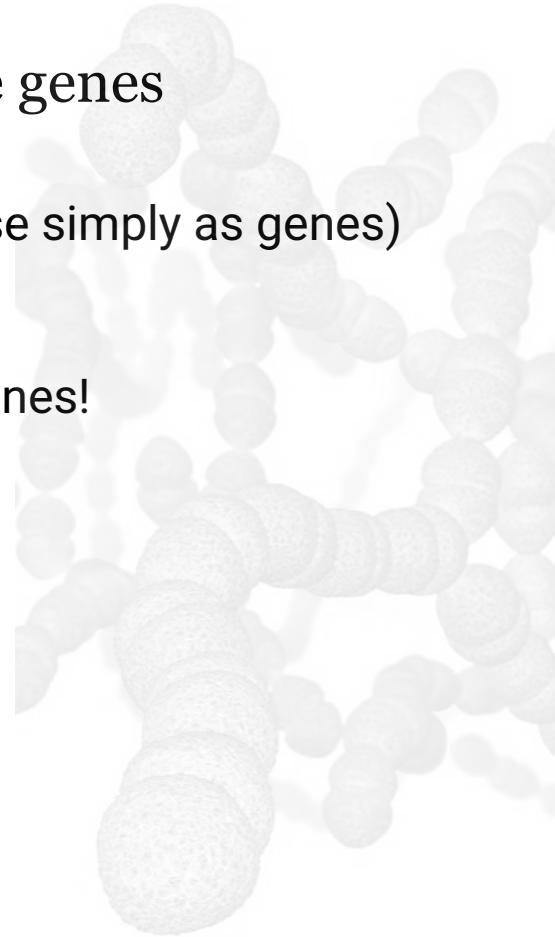


Pneumococcal vaccines

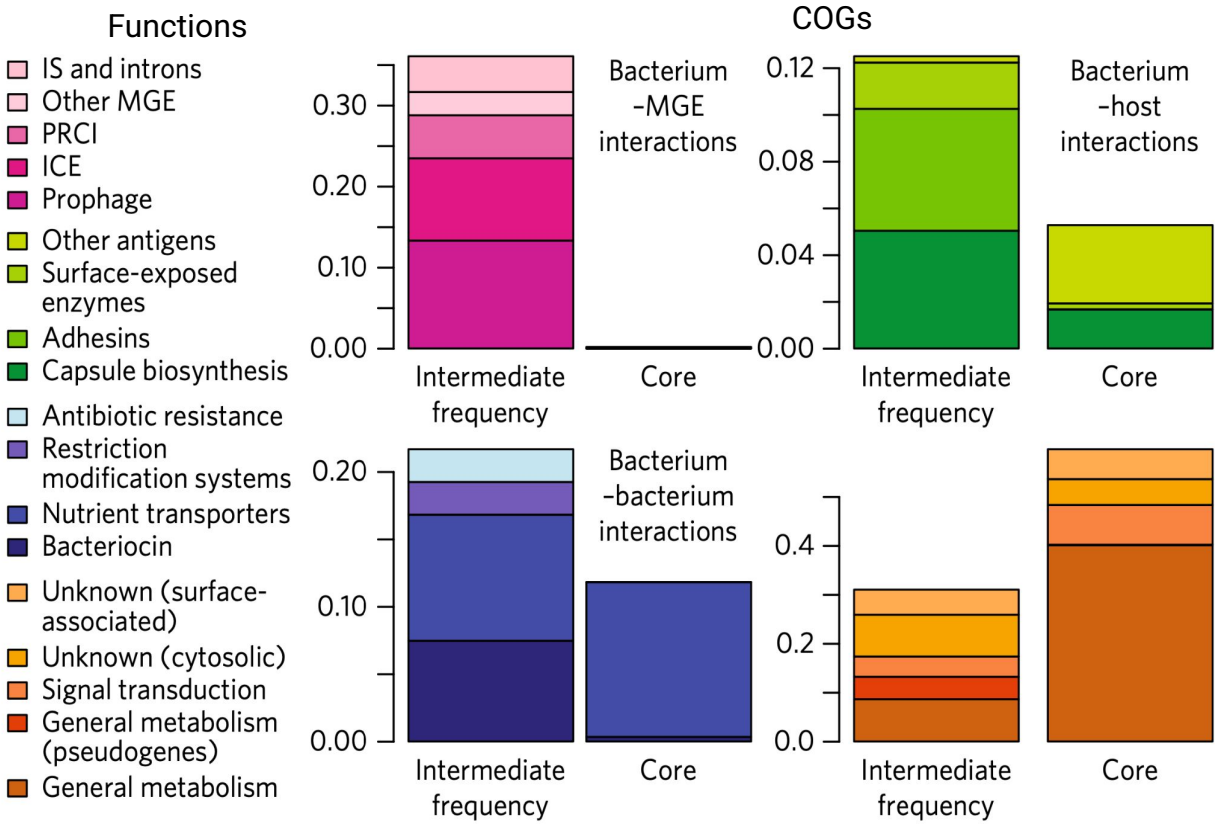
- Pneumococcal Conjugate Vaccines (PCV) target specific serotypes.
  - PCV7: approved in the early 2000s
  - PCV13: FDA approved in 2014
  - PCV20: FDA approved in 2023
- Variation in antigenic loci makes vaccine development very challenging.
- Serotype switching: allows strains to evade vaccine-induced immunity.
- Antigens become more costly when they are common because they are more frequently recognised by acquired immune responses.

Accessory genome: not the all isolates have all the genes

- COG: cluster of orthologous genes (we can think of these simply as genes)
- Many intermediate frequency genes
- Horizontal gene transfer: they can gain and lose their genes!
- Highly variable accessory genome



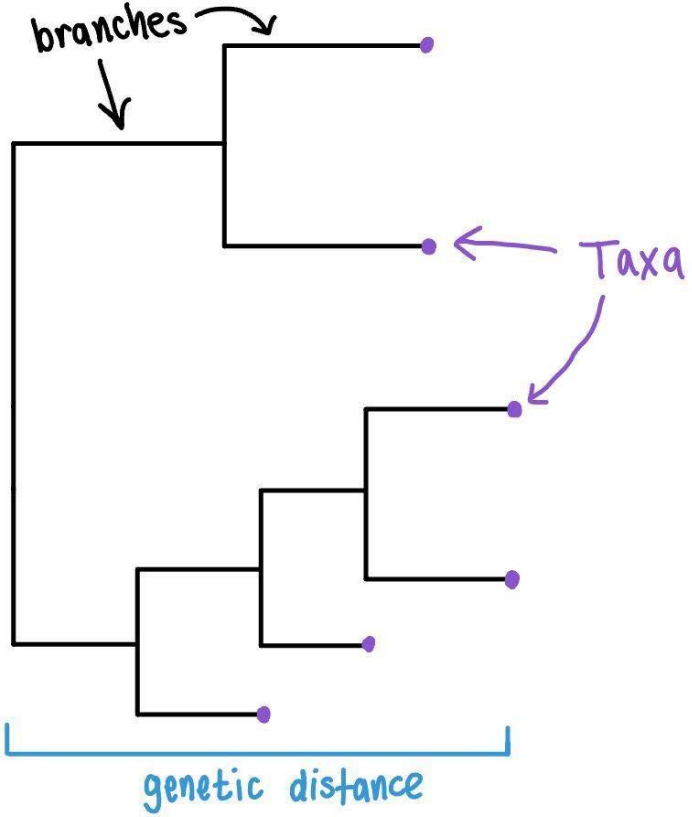
What does the accessory genome do?



Corander et al. (2017)

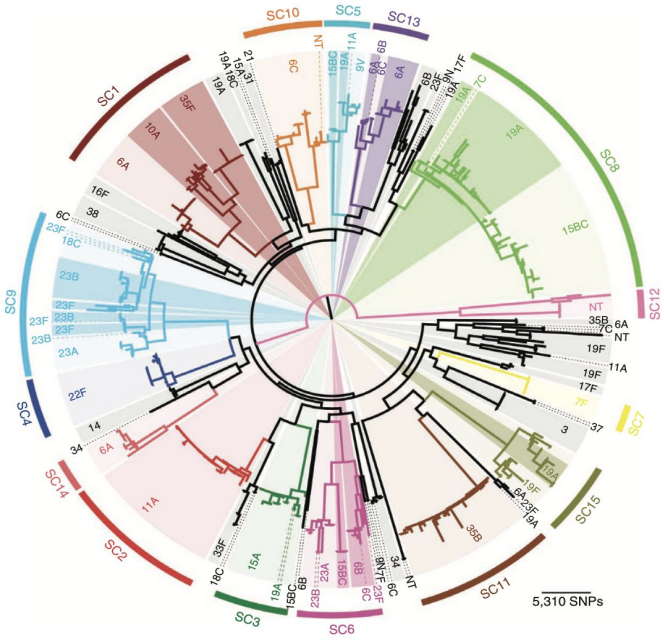
What is a phylogenetic tree?

- Depicts evolutionary descent
- Tips represent taxa
- Branch lengths represent distance
- Genetic distance in terms of substitutions per site
- Can construct phylogenetic trees with maximum likelihood methods



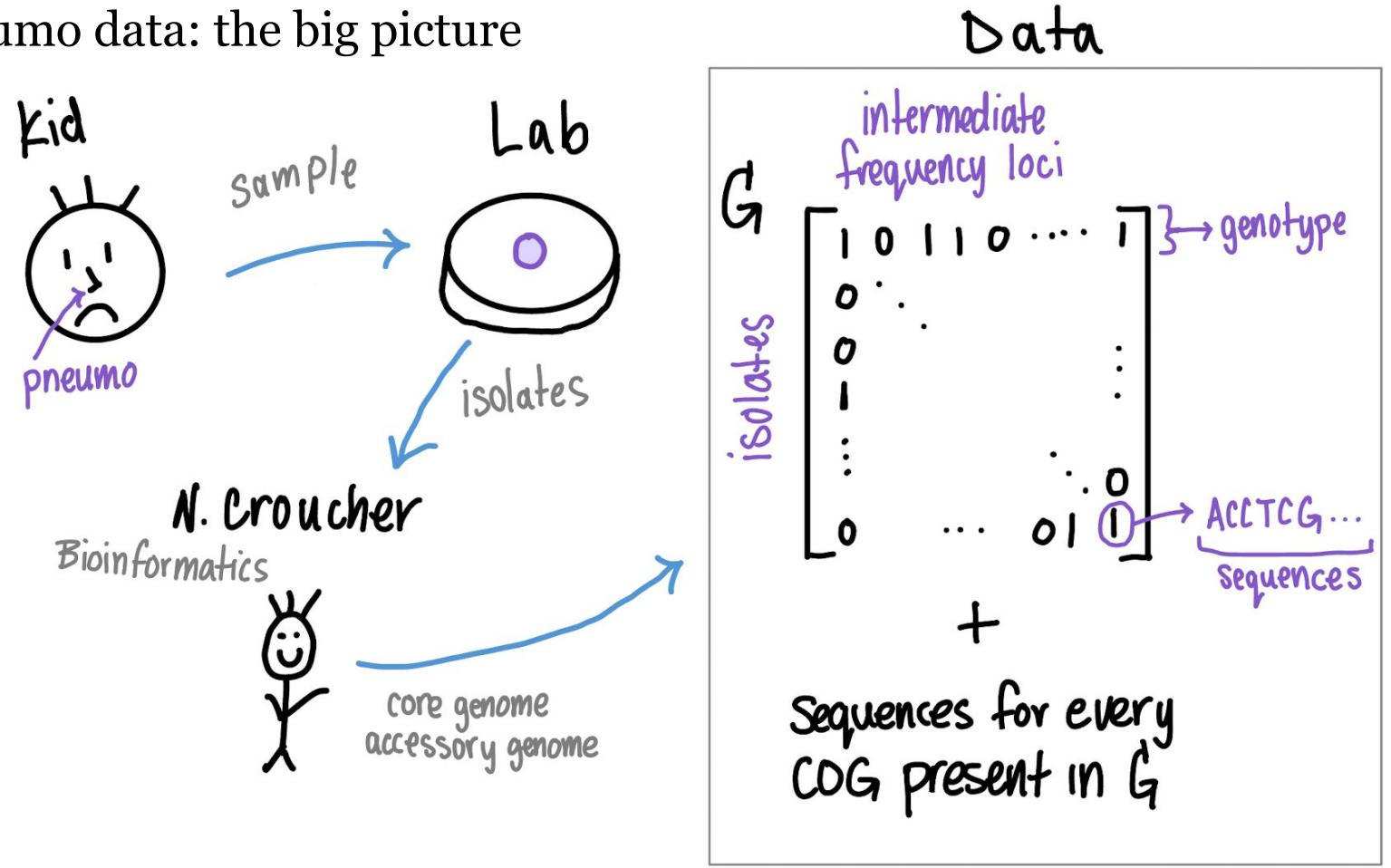
Summary of Croucher et al (2013)

- Sequencing of samples collected from Massachusetts children before and after vaccine
- Key findings:
  - Closely related things had similar accessory genomes, and and more distantly related things had different accessory genomes, with little in between.
  - The pneumococcal population was heavily disrupted by the vaccine, with almost complete loss of the seven vaccine serotypes and their rapid replacement with non-VT strains.



Maximum likelihood phylogeny (Croucher et al 2013)

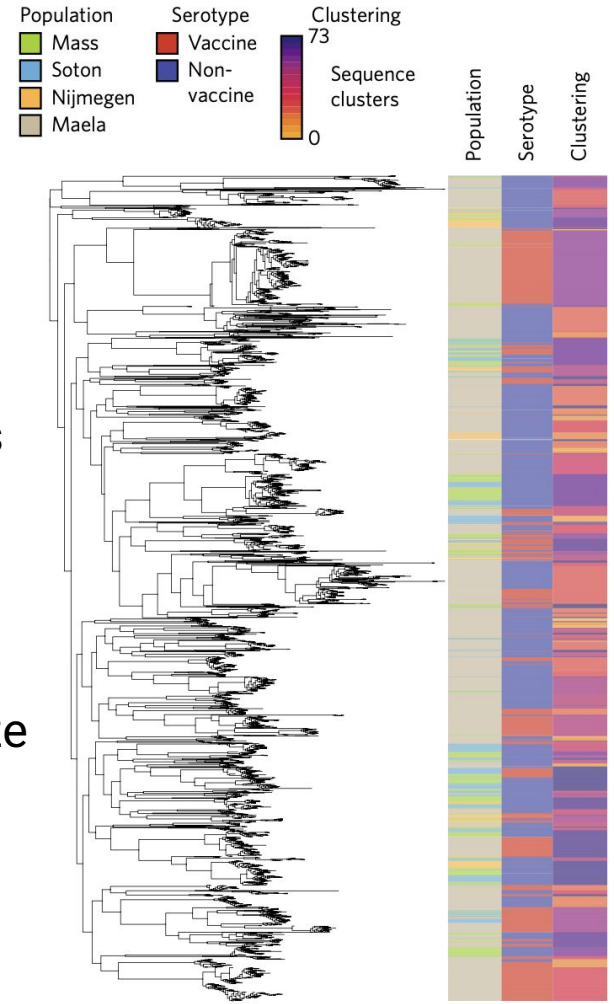
Pneumo data: the big picture



Summary of Corander et al (2017): The same lineages are in all populations

Comparative population genomics

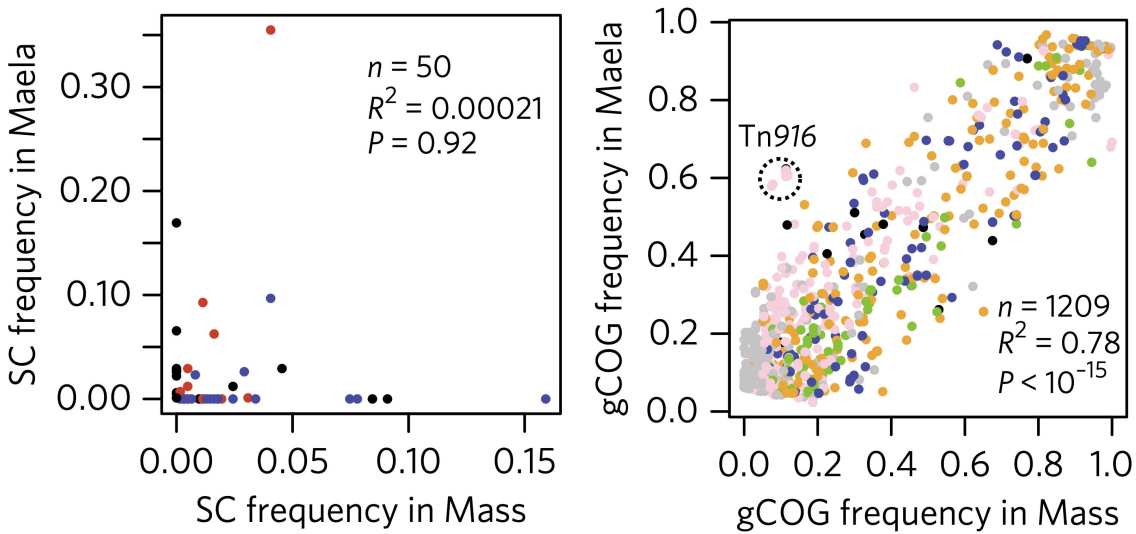
- Sequence clusters (SCs): groups of similar genes in the core genome
- SCs across sites were not correlated
- There is population structure such that different combinations and proportions of SCs characterize each sample



Corander et al 2017

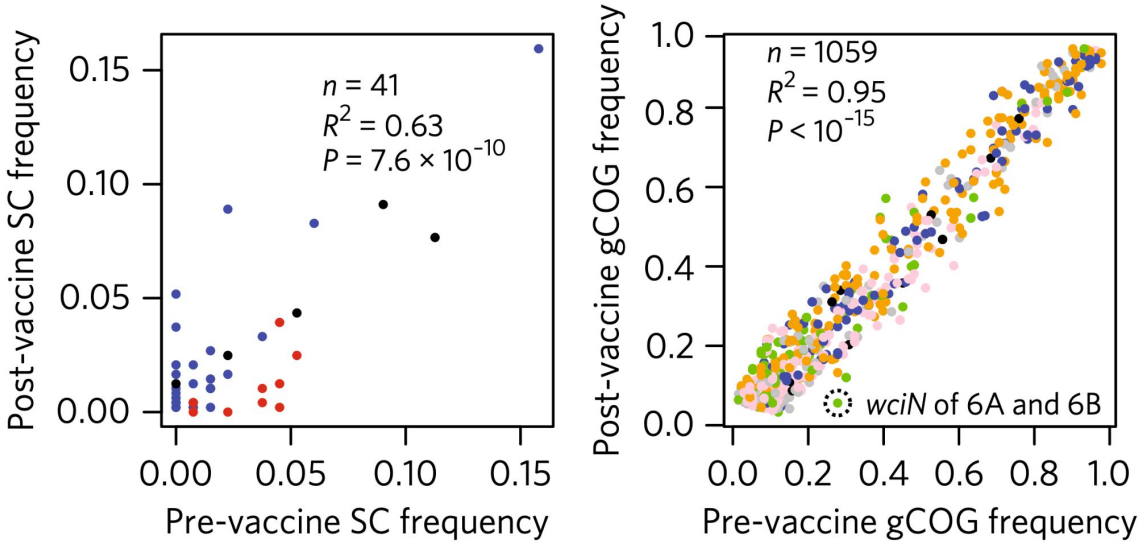
Correlated accessory genomes despite uncorrelated SCs

- If each SC has its own combination of accessory genes, then we would expect the frequencies of these to vary among the samples depending on which SCs were present.



Corander et al 2017

COGs return to their pre-vaccine frequency



Corander et al 2017

- ❏ Major selective pressure on vaccine type serotypes
- ❏ The frequencies of accessory genes before and after vaccination were even more tightly correlated!

Negative frequency dependent selection

Corander et al (2017) suggests this is frequency dependent selection.

- The accessory genes are likely to be mobile elements, or immunogenic, and as such might experience an advantage when rare and be selected against when common.
- *Negative frequency dependent selection* (NFDS): being rare has an advantage.

Corander et al (2017) NFDS model

- Multilocus NFDS model in a discrete-time Wright-Fisher framework
- Heterogeneous-rate multilocus NFDS model had parameters to represent the strength of weaker NFDS acting on a fraction of the samples and stronger NFDS acting on the other fraction of samples.
  - Genotypes with many COGs *below* equilibrium frequency go *up*
  - Genotypes with many COGs *above* equilibrium frequency go *down*
- The model makes predictions in time

Selection strength in the multilocus model

Selection strength:

$$\pi_{i,t}(G,Y) = \sum_{l=1}^{L} w_l G_{i,l}(e_l - f_{l,t}(Y))$$

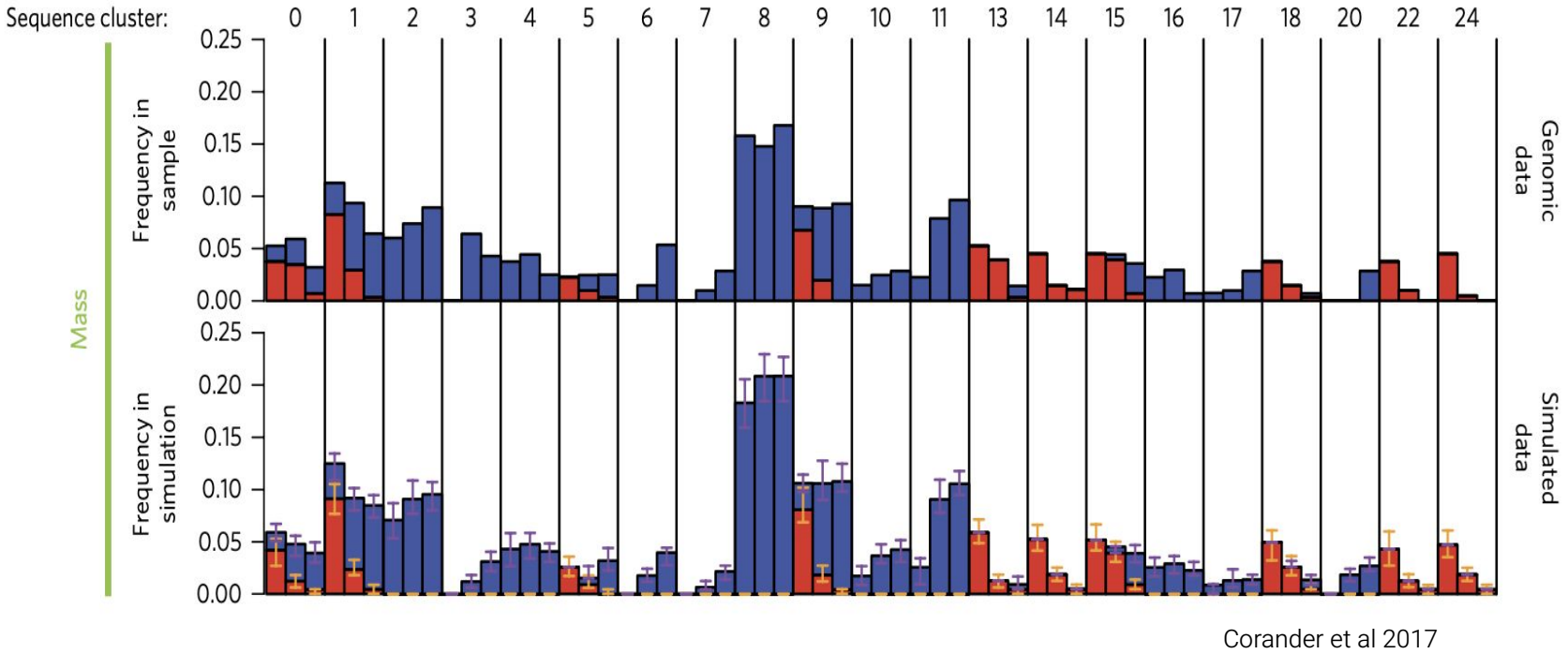
$G_{i,l} = 1$ if genotype i has locus l, 0 if not.

Weights

COG frequencies

- If $f_l < e_l$ for many l in a genotype, then $\pi > 0$. I.e., positive selection, increase in frequency.
- A proportion of intermediate frequency loci experience *strong* NFDS are given one weight and the rest, experience *weaker* NFDS are given a different weight.

Model results



Corander et al 2017

What might be the mechanism behind NFDS here?

Hypothesis: Loci that are more diverse may enjoy NFDS.

- ❏ NFDS is inherently a *diversity promoting* form of selection.
- ❏ Genes under NFDS are rare and thus have an advantage.

or

Loci that are *less* diverse may enjoy NFDS.

- ❏ As a form of selection, there must be a function that the gene does to experience NFDS.

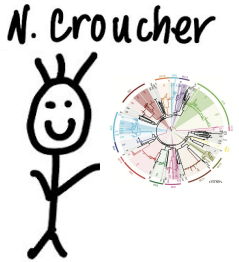
However, the functions that would do this probably are not on these accessory genes that are gained and lost.

My project

- Could diversity be different among loci experiencing strong vs. weak NFDS?
- Compare diversity between strong weight group and weak weight group defined in Corander et al (2017).
- Can measure diversity based on which COGs are present or diversity in the sequences
- Diversity methods:
  - Shannon entropy
  - Phylogenetic diversity
  - Tajima's D

Recall: data from Croucher et al (2013)

- Samples from children in Massachusetts
  - 3 time points: Pre-vax, post-vax, and long after vax when then pneumo population returned to equilibrium
- 5,442 COGs
- ‘Core’ genome = 1,194 COGs that were present in a single copy in all genomes
- 616 isolates (rows of G)
- Data corresponding to G: vax type, time



Shannon entropy

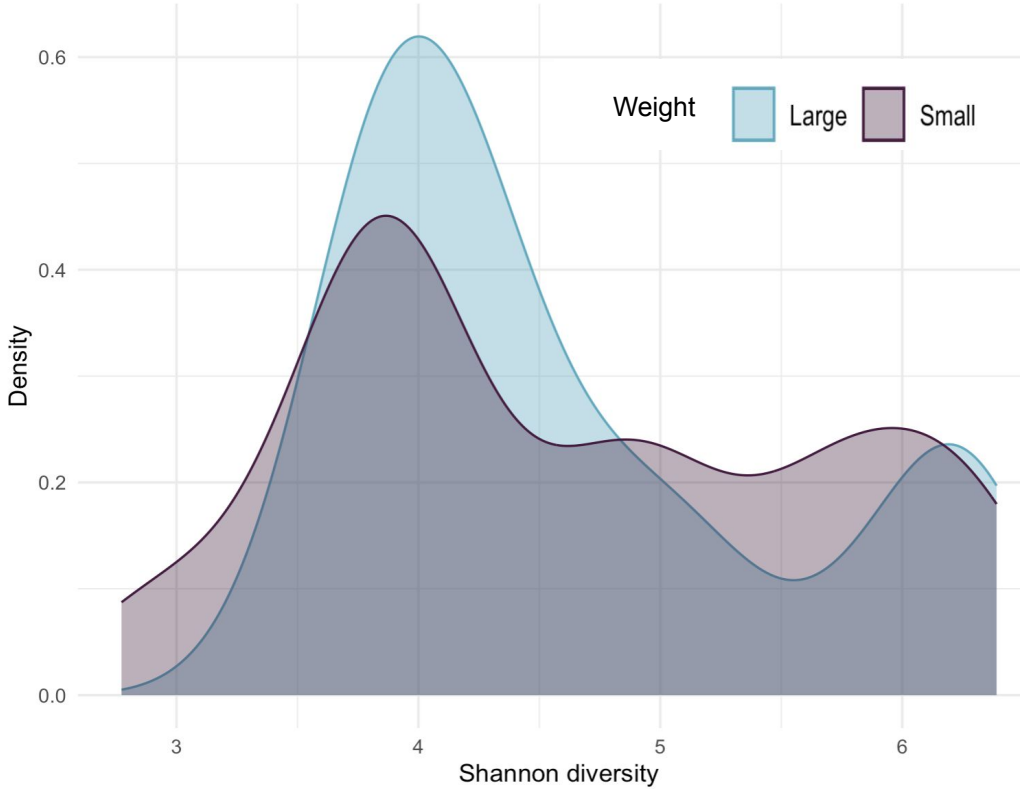
- Shannon entropy is a measure of the average level of uncertainty inherent to the possible outcomes of a random variable.
- As a measure of genetic diversity, Shannon entropy (or Shannon diversity) is defined as

$$H = -\sum_{i=1}^{m} p_i \log(p_i)$$

where p_i is the proportion of isolates that have gene i.

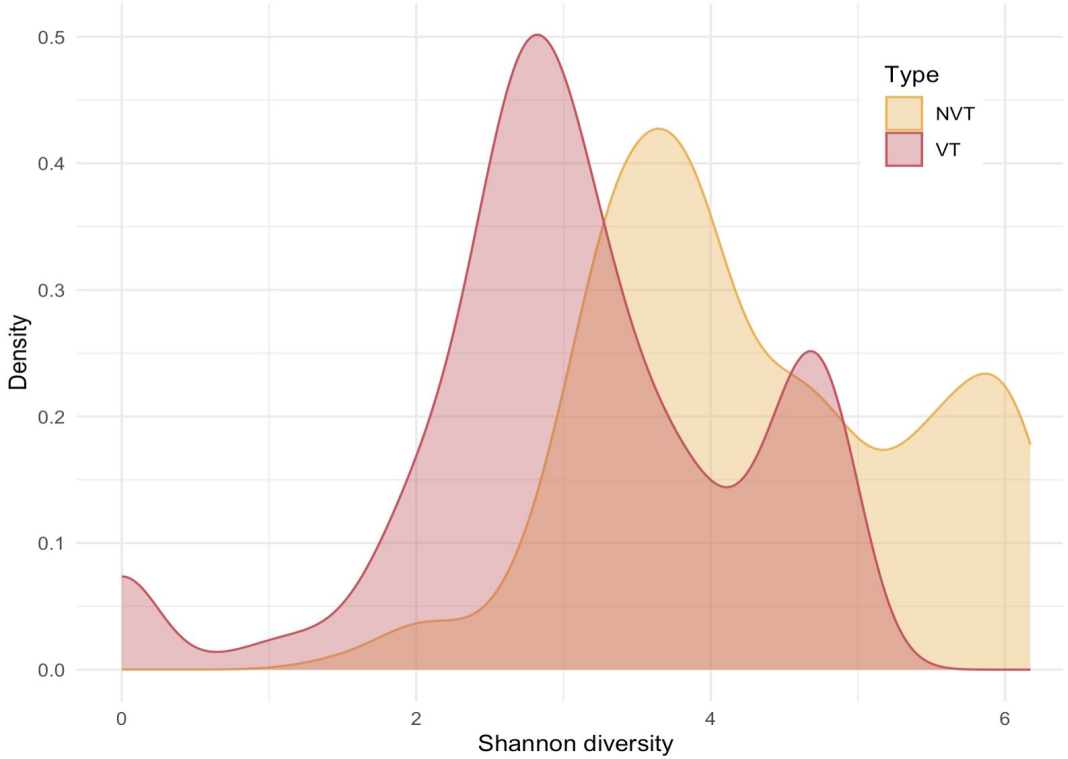
Shannon diversity of presence/absence of COGs by weight

- “Small” or “large” weight based on selection strength from Corander et al (2017)
- p-value = 0.074



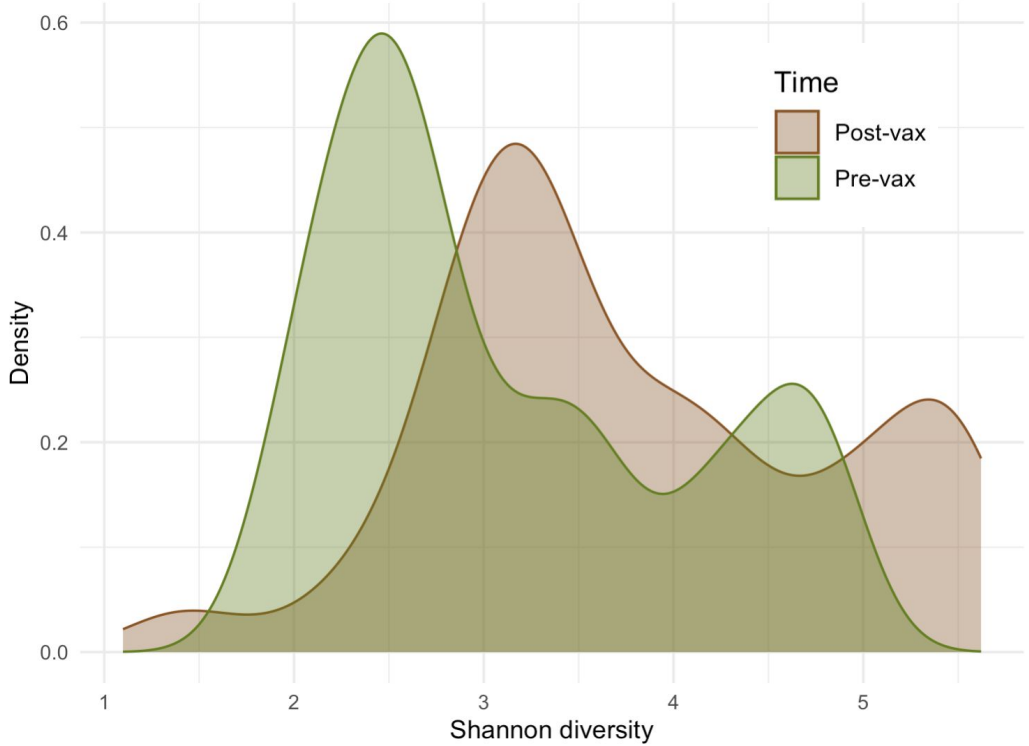
Shannon diversity of presence/absence of COGs by type

- Recall that VT represent 7 serotypes targeted by the PCV7 vaccine.
- p-value < 2.2e-16



Shannon diversity of presence/absence of COGs by time

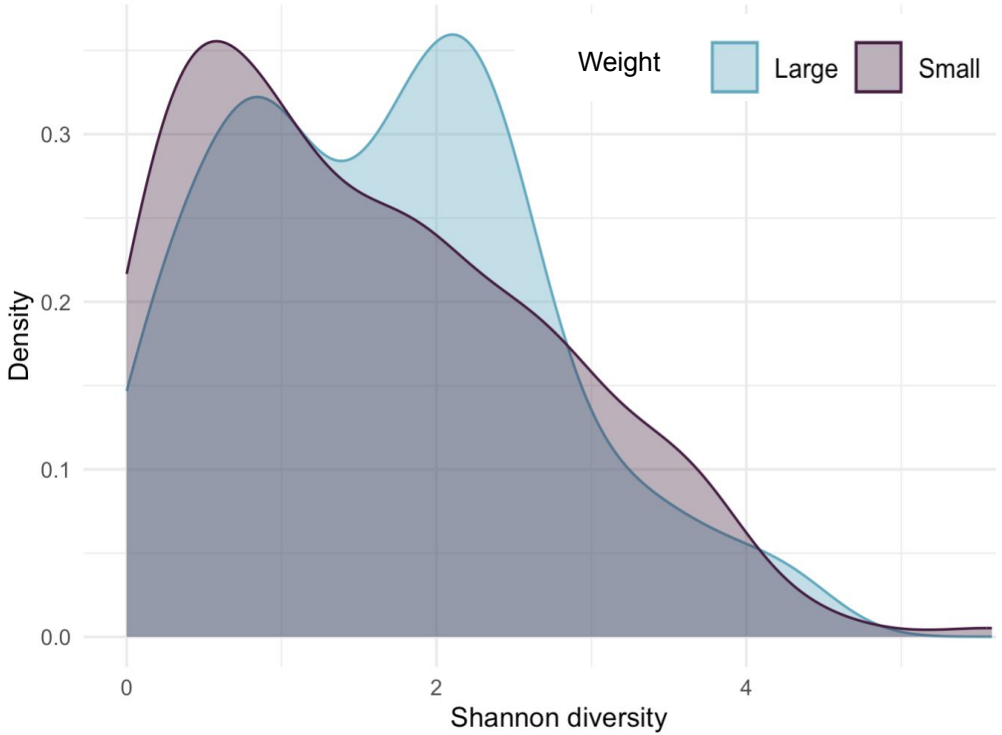
- Unexpected result
- Is this because there are more vaccine types and they are naturally more diverse?
- p-value < 2.2e-16



Shannon diversity of unique sequences by weight

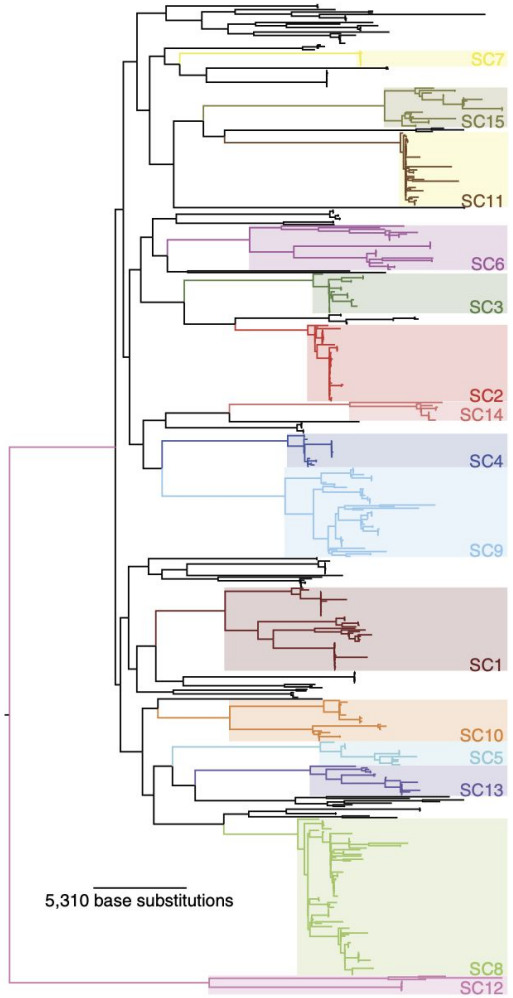
- $$H = -\sum_{i=1}^{m} p_i \log(p_i)$$

  where p_i is the proportion of unique sequences
- Comparable to result from presence/absence data
- p-value = 0.118



Phylogenetic diversity

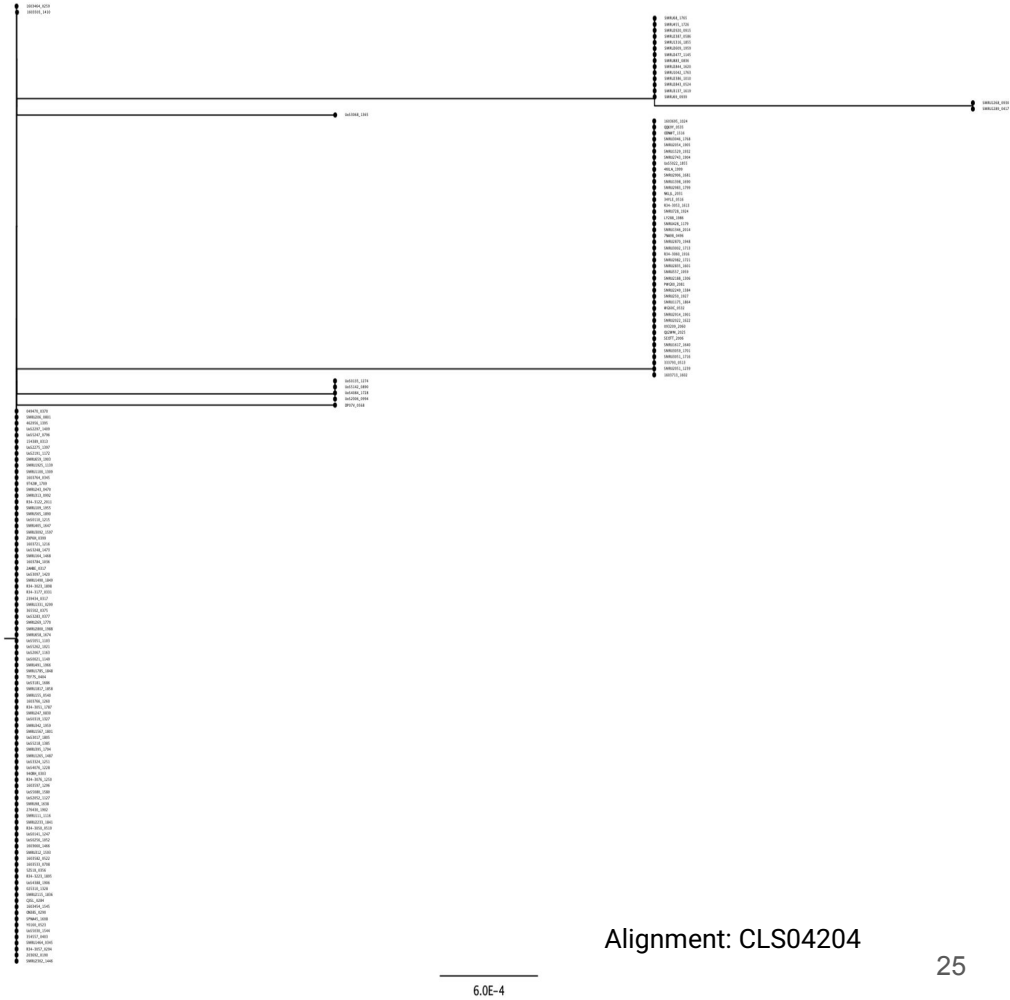
- Phylogenetic diversity is the branch lengths of the phylogenetic tree.
- Longer branch represents more opportunity for evolutionary change
- In our case the edges of the tree are in units of genetic distances (substitutions per site).



Croucher et al 2014

Phylogenetic diversity example

- Unrooted phylogenetic tree in units of substitutions per site using iqtree.
- Notice this is a particularly sparse tree



Alignment: CLS04204

Tajima's D

- *Segregating site*: positions where polymorphisms occur in a sequence alignment.
- *Effective population size*, N_e, is the number of individuals that effectively participate in producing the next generation, i.e., the size of the idealized population.
- Computes a standardized measure of the total number of segregating sites in the DNA and the average number of mutations between pairs in the sample.

Next steps

- ❏ Finish computing diversities
- ❏ Measure diversity among different groups of loci:
    1. Group based on drug resistance
    2. Group based on functional groups
- ❏ Use the weight groupings but restrict diversity measurements to when the isolate was sampled, i.e., pre- or post-vaccine; or wether the isolate is vaccine-type or non-vaccine type.

References

- Corander, J., Fraser, C., Gutmann, M. U., Arnold, B., Hanage, W. P., Bentley, S. D., Lipsitch, M., & Croucher, N. J. (2017). Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nature Ecology & Evolution, 1(12), 1950–1960.
- Croucher, N. J., Coupland, P. G., Stevenson, A. E., Callendrello, A., Bentley, S. D., & Hanage, W. P. (2014). Diversification of bacterial genome content through distinct mechanisms over different timescales. Nature Communications, 5, 5471.
- Croucher, N. J., Finkelstein, J. A., Pelton, S. I., Mitchell, P. K., Lee, G. M., Parkhill, J., Bentley, S. D., Hanage, W. P., & Lipsitch, M. (2013). Population genomics of post-vaccine changes in pneumococcal epidemiology. Nature Genetics, 45(6), 656–663.