Negative frequency dependent selection in *Streptococcus pneumoniae*

Jennifer McNichol

April 24, 2024

Abstract

Streptococcus pneumoniae (the pneumococcus) is a common nasopharyngeal commensal that can lead to invasive disease. With the increased availability of pneumococcal genome sequence data in recent years, it is believed that the pneumococcus exhibits widespread negative frequency dependent selection (NFDS) in its accessory genome, however, the mechanism remains unknown. To begin to understand the NFDS at play in the pneumococcus, we investigate diversity among its loci by adapting a number of diversity measures: Shannon diversity, Phylogenetic diversity, and Tajima's D. We compute these measures for defined groups of loci exhibiting a "strong" versus "weak" effect of NFDS and find that there is no meaningful difference between the diversity of these groups of loci.

Introduction

Streptococcus pneumoniae (the pneumococcus) is a common nasopharyngeal commensal that is a primary cause of pneumonia, meningitis and sepsis globally. Pneumococci possess a 2.1 megabase (Mb) pair circular genome that consists of over 2000 predicted protein coding regions and approximately 5% insertion elements. The pneumococcus is characterised by its polysaccharide capsule (serotype) which defines its antigenic properties. For example: antibiotic resistance, virulence, invasive potential – most cases of invasive pneumococcal disease are caused by a small subset of the circulating serotypes (Hausdorff et al., 2005). Pneumococal conjugate vaccines (PCV) have

been developed to target specific serotypes. The first licensed conjugate vaccine, PCV7, targets the following seven of over 90 serotypes: 4, 6B, 9V, 14, 18C, 19F, and 23F.

S. pneumoniae exhibits a feature referred to as "natural competence", whereby the bacteria are able to actively transport DNA fragments from the environment through the cell wall, into the cell cytoplasm (Chaguza et al., 2015). These transported DNA fragments can subsequently integrate into the pneumococcal genome (i.e., genetic recombination). The pneumococcus can also acquire DNA from another pneumococcus through horizontal gene transfer, one of the most powerful forces driving the evolution of bacteria (Hiller et al., 2010). This means that the pneumococcus can both obtain new a variant (i.e., allele) of a gene and a totally new gene. Horizontal gene transfer and recombination are key evolutionary mechanisms employed by the pneumococcus to rapidly adapt to selective pressures. The rate at which the pneumococcus acquires genetic variation through recombination is much higher than the rate at which the organism acquires variation through spontaneous mutations (Chaguza et al., 2015). This allows the pneumococcus to circumvent the host innate and adaptive immune responses and escape clinical interventions such as antibiotic therapy and vaccine introduction.

S. pneumoniae can undergo recombination at the locus of its capsule protein, thereby changing its serotype–a process known as serotype switching (Brueggemann et al., 2007). Serotype switching and clonal replacement of vaccine types by distantly related non-vaccine types (NVTs) are the driving force behind the phenomena of "serotype replacement" whereby the serotypes targeted by the vaccine (vaccine types) are replaced by (NVTs) in the post-vaccine pneumococcal population (Corander et al., 2017). Serotype replacement is a major complication that can reduce vaccine impact (Colijn et al., 2020). In addition to variation in serotype, the forementioned evolutionary processes lead to extensive variation in the *S. pneumoniae* 'accessory' genome, that is, genes that are present at intermediate frequency (i.e., 5 to 95% of the population have them). These genes are not strictly necessary and perform a wide range of functions for both bacterium–host and bacterium–bacterium interactions.

It was observed in Corander et al. (2017) that accessory genes return to their pre-vaccine frequency in the post-vaccine pneumococcal population. Despite the strong correlation in accessory gene frequency pre- and post-vaccine, the frequency of serotypes were uncorrelated between the pre- and post-vaccine pneumococcal populations. A similar conclusion was drawn between locations. Corander et al. (2017) hypothesized that multi-locus-negative frequency-dependent selection

(NFDS)–selection where rare genes provide an advantage–explained these observations. The proposed model in Corander et al. (2017) differentiated between loci based on their selective strength, in particular, intermediate frequency loci with a "strong" degree of NFDS were given a weight of 0.1363 and the remaining loci were given a weight of 0.0023. The model was used to accurately predict the pneumococcal population post-vaccination.

Since NFDS is a inherently a diversity promoting form of selection, that is, rare genotypes will be selected for and thus being rare has an advantage. If the pneumococcus experiences an alteration in its genome, say to switch its serotype to one not targeted by the vaccine, that is to say that it is now more rare in the population and thus not only does the genetic diversy of that serotype increase, then that serotype will undergo NFDS. In this project, we hypothesize that loci purported to be under strong NFDS (high weights in the Corander et al. (2017) model) will be more genetically diverse than loci under "weak" NFDS (low weights). We compare diversity between these groups of loci using a number of diversity measures: Shannon entropy, phylogenetic diversity, and Tajima's D.

Methods

In this analysis, we use the 616 *S. pneumococcus* isolates and corresponding clusters of orthologous genes (COGs) from Croucher et al. (2013). Samples were collected from children in Massachusetts, USA between 2001 and 2007. COGs are essentially groups of proteins that have the same or very similar function and the sequences of their genes are very similar, that is, there are not enough differences between sequences to account for an entirely new gene. Thus, we can think of a COG simply as a gene. There were 1,112 COGs are present in 5% to 95% of isolates (intermediate frequency loci) – these COGs define the accessory genome. Of these, we utilize 221 COGs exhibiting "strong" NFDS according to the Corander et al. (2017) model, and 221 COGs in the group of "weak" NFDS loci. The code used to apply the methods and produce the results in this report is available at `https://github.com/jmcnichol/pneumo-diversity`.

Shannon diversity

For $m = 1, ..., 616$ isolates and $l = 1, .., 442$ loci, let $\mathbf{G}$ be an $m \times l$ matrix of 0s and 1s, with rows corresponding to genotypes for each isolate and columns representing each locus. If $g_{m,l} = 1$, then

isolate m has locus l (and we have the corresponding COG sequence), otherwise, $g_{m,l} = 0$ and isolate m does not contain locus l. A simple method we may use to quantify the diversity of a locus is the Shannon diversity (Shannon, 1997). Originally proposed in the field of information theory as way to quantify the amount of uncertainty in a random variable, the Shannon diversity has become a widespread tool in ecology (Konopiński, 2020). We calculate the Shannon diversity for each locus in $\mathbf{G}$ using

$$H_l = -\sum_{i=1}^{m} p_i \log(p_i), \tag{1}$$

where p_i is the proportion of isolates that have gene i at locus l. For a higher resolution, we calculate the Shannon diversity of a COG by letting p_i in Equation 1 represent the proportion of unique sequences i (for $i = 1, ..., m_l$ where m_l is the number of isolates containing a gene at locus l) in the sample of isolates.

We compare the distributions of Shannon diversity values between the two groups of COGs using the two sample Kolmogorov-Smirnov (KS) test (Kolmogorov, 1933). The KS statistic is

$$\mathrm{KS}_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

where $F_{1,n}(x)$ and $F_{2,m}(x)$ are the empirical cumulative distribution functions of each group, and n and m are their sample sizes, respectively. The null hypothesis is rejected at level α if

$$\mathrm{KS}_{n,m} > \sqrt{-\log\left(\frac{\alpha}{2}\right)\frac{1+\frac{m}{n}}{2m}}.$$

Phylogenetic diversity

A branch is a segment of a tree, lying between two nodes, and with no other nodes along it; the nodes can be terminal nodes, i.e., "tips" representing taxa, or internal branching points. The general definition of phylogenetic diversity (Faith, 1992) is as follows. Let $\mathcal{T}$ denote the tree for the complete set of N taxa, and s denote a subset of the taxa. The *minimum spanning path* for s is made up of the smallest assemblage of branches from $\mathcal{T}$ such that, for any two members of s, a path along $\mathcal{T}$ connecting the two can be found that uses only branches in the assemblage. The phylogenetic diversity of s is equal to the sum of the lengths of all those branches that are members of the corresponding minimum spanning path. Clearly, in this work, the phylogenetic diversity of a COG, computed from its multiple sequence alignment, is the sum of the branch lengths of the entire

tree for that COG since all of the taxa that have a gene at the locus of that COG are represented. Note that for a COG, the branches of the tree are indicated in units of nucleotide substitutions per site, that is, the number of changes divided by the length of the sequence.

Since the lengths of the aligned sequences vary between COGs, we scaled the sum of branch lengths for a given COG by a factor of $2\sum_{i=1}^{n-1}\frac{1}{i}$ (where n is the number of sequences in that COG) according to Wakeley (2009). We then test of a difference in the distribution of phylogenetic diversities among COGs between each group using the two sample KS test defined in the previous section.

In this work we use IQ-TREE (Nguyen et al., 2015), a command line interface for maximum likelihood tree optimization to build a phylogentic tree for each COG. The basis of building a maximum likelihood tree involves a DNA substitution model (e.g., Jukes and Cantor, 1969; Felsenstein, 1981; Kimura, 1969) to find a likelihood given a tree and maximize over all such trees. There are many ways to do this, we will describe one such way and underlying substitution model to convey the general idea.

Computing the likelihood of a tree

A maximum likelihood tree is found by computing the probability of a particular set of sequences on a given tree and maximizing this probability over all evolutionary trees. The probability of obtaining a given set of sequences at the tips of a given tree can be computed if we have a model specifying the probability that sequence, say s_1 changes to sequence s_2 during evolution along a branch of the tree of length t. One such evolutionary model is Felsenstein's substitution model (Felsenstein, 1981). We will describe this model and use it to develop a likelihood in the following sections.

Substitution model It is assumed in Felsenstein (1981) that nucleotide changes at different sites in the sequence are independent, thus we can compute the probability of a given set of sequences arising on a given tree site by site. Consider a single site and let $P_{ij}(t)$ be the probability that a lineage which is initially in state i will be in state j after t units of time have elapsed, where i and j take values 1, 2, 3, and 4 corresponding to the four bases A, C, G, and T. We assume $P_{ij}(t)$ is a continuous-time Markov process, so, the probability of a base changing may depend on its current state, but not on its past history. We assume that in a small interval of time of length dt, there is a probability μdt that the current base at a site will change, where μ is the rate of base substitution

per unit of time. If a base is replaced, its replacement is A, C, G, or T with probabilities π_1, π_2, π_3, or π_4. It follows that the number of base substitutions during a particular time interval is independent of the history of changes outside of this interval, so each time increment is independent and at time $t = 0$, clearly, the number of base substitutions will be 0.

It should now be clear that the process of changing from the current base to another base in a time interval dt can be modelled as a Poisson Process. Thus if we let $N(t)$ be the number of transitions from base i to base j at time t, then $P(N(t) = k) = e^{-\mu t}\frac{(\mu k)^k}{k!}$ for $k = 0, 1, 2, ...$ So $P(N(t) = 0) = e^{-\mu t}$ and $P(N(t) > 0) = 1 - e^{-\mu t}$. Finally, let X be the time of the first event. The probability that the base does not change is $1-P(X \leq t) = 1-P(N(t) > 0) = 1-(1-e^{-\mu t}) = e^{-\mu t}$. Therefore the transition probabilities in Felsenstein (1981) are:

$$P_{ij}(t) = e^{-\mu t}\delta_{ij} + (1 - e^{-\mu t})\pi_j,$$

where δ_{ij} is 0 if $i \neq j$ and 1 if $i = j$ (the Kronecker delta function).

Likelihood Now we describe the expression for the likelihood of the tree, namely the probability of the data given the tree, according to Felsenstein (1981). If we knew the states (bases) of all the internal nodes of our tree the likelihood of the tree would be the product of the probabilities of change in each tree segment times the prior probability of the state. Since we do not know the states of the internal nodes in practice, we sum over all possible assignments of bases to the internal nodes which results in a likelihood with 2^{2n-2} terms for a tree with n tips (Felsenstein, 1981). To get the idea we will use a toy example of the tree shown in Figure 1. Let s_m be the sequence at node $m = 0, 1, .., 2n - 1$. For our toy example, there are 7 nodes and s_5 is at node number 5. The likelihood is composed of the likelihood for each branch along all the paths from the root to each tip:

$$\mathcal{L} = \sum_{s_0=1}^{4} \pi s_0 \sum_{s_5=1}^{4} P_{s_0 s_5}(v_5) P_{s_5 s_1}(v_1) P_{s_5 s_2}(v_2) \sum_{s_6=1}^{4} P_{s_0 s_6}(v_6) P_{s_6 s_3}(v_3) P_{s_6 s_4}(v_4). \tag{2}$$

where v represent the branches between nodes and $P(v)$ is the substitution model.

This gives rise to the popular method of ML tree building called *Felsenstein's pruning algorithm* (Felsenstein, 1981) which defines the conditional likelihood, $\mathcal{L}^x_{(s_m)}$, to be the likelihood based on the sequence s_m at or above (above meaning from that point outward to the tips) a node x on the tree given the node x is known to have state s for the site under consideration. Felsenstein's pruning algorithm starts at the tips of the tree working down the tree (hence "pruning"). For node

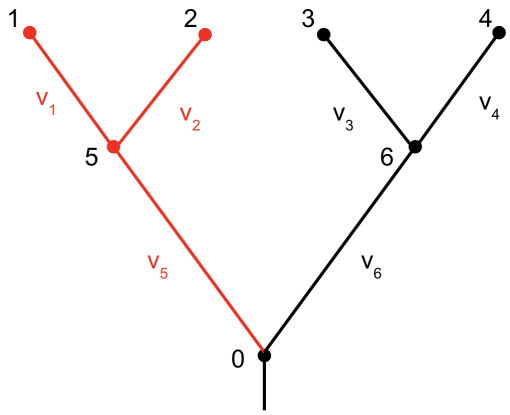

Figure 1: Toy example of a phylogenetic tree, labelled in the usual way, where nodes are given integer values starting at the tips, with root indicated by 0, and the branches are labelled from tip to root. The red section of the tree is included for easy interpretation of the likelihood in Equation 2.

x whose immediate descendants are y and z, we can compute all values of s_x (the probability that each descendant of x is in one of the four states):

$$\mathcal{L}_{s_x}^{(x)} = \sum_{s=1}^{4} \left[\sum_{s_y=1}^{4} P_{s_x s_z}(v_y) \mathcal{L}_{s_y}^{(y)} \sum_{s_z=1}^{4} P_{s_x s_z}(v_z) \mathcal{L}_{s_z}^{(z)} \right].$$

We continue computing conditional likelihoods where the decedents of the tips in the previous step become the new "tips" in the following step until we run out of nodes or reach the root of the tree. When we reach the root of the tree, we will have computed the four conditional likelihoods for the state of the root, call this $\mathcal{L}_{s_0}^{0}$. Then the overall likelihood is

$$\mathcal{L} = \sum_{s_0}^{4} \pi_{s_0} \mathcal{L}_{s_0}^{0}.$$

Once the likelihood is found, algorithms such as the one implemented in IQ-TREE are employed to seek the best tree. In particular for IQ-TREE, the algorithm will find the best tree with the best fit substitution model according to the Bayesian information criterion. For each COG, we let IQ-TREE determine the best fit model to build the phylogenetic tree from which we calculate the sum of its branch lengths, and hence phylogenetic diversity.

Tajima's D

Tajima's D (Tajima, 1989) is a test to distinguish between a DNA sequence evolving randomly (under neutral evolution) and a sequence evolving under a non-random process such as selection or

hitch-hiking. Tajima's D is computed as the difference between two measures of genetic diversity: the mean number of pairwise nucleotide differences and the number of segregating sites, each scaled so that they are expected to be the same in a neutrally evolving population of constant size.

The number of segregating sites, S, at a locus is the total number of polymorphic positions. It has been shown (Watterson, 1975) that

$$\mathrm{E}(S) = a_1 M \quad \text{and} \quad \mathrm{Var}(S) = a_1 M + a_2 M^2$$

where $M = 4N_e u$ ($2N_e u$ for haploid) and N_e is the effective population size, i.e., the number of individuals that participate in producing the next generation in an idealized population, and u is the mutation rate per generation per sequence. Finally, $a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$ and $a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}$ where n is the number of sequences. Note that M may be estimated by $\hat{M} = \frac{S}{a_1}$.

Let $\hat{k}$ represent the average number of pairwise nucleotide substitutions Tajima (1983) showed that

$$\mathrm{E}(\hat{k}) = M \quad \text{and} \quad \mathrm{Var}(\hat{k}) = b_1 M + b_2 M^2,$$

where $b_1 = \frac{n+1}{3(n-1)}$ and $b_2 = \frac{2(n^2+n+3)}{9n(n-1)}$.

In order for the above equations to hold, Tajima (1989) assumes a random mating population of N diploid individuals. Note that here, since we are working with bacteria, this assumption translates to assume that all individuals have the same probability of reproducing and passing on their genes at the same frequency. Tajima (1989) also assumes there is no selection and no recombination between sequences and that the number of sites is so large that when a new mutation occurs, it occurs at a new site (i.e., infinite site model; Kimura (1969)) – here, there is no recombination within a COG.

It is important to draw attention to the fact that the difference between S and $\hat{k}$ is the effect of selection and both S and $\hat{k}$ provide estimates of M. Let $d = \hat{k} - \frac{S}{a_1}$ it is shown in Tajima (1989) that

$$\mathrm{Var}(d) = c_1 M + c_2 M^2$$

where $c_1 = b_1 - \frac{1}{a_1}$ and $c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}$. The variance of d can be estimated by

$$\hat{V}(d) = e_1 S + e_2 S(S-1),$$

where $e_1 = \frac{c_1}{a_1}$ and $e_2 = \frac{c_2}{a_1^2 + a_2}$. Thus the Tajima's D statistic is defined as:

$$D = \frac{d}{\sqrt{\hat{V}(d)}} = \frac{\hat{k} - \frac{S}{a_1}}{\sqrt{e_1 S + e_2 S(S-1)}}.$$

The "neutral mutation hypothesis" i.e., the null hypothesis, assumes that M is estimated by $\frac{S}{a_1}$. Recall that $\hat{k}$ is also an estimate of M. Thus under the null hypothesis, the expected value of the two quantities $\hat{k}$ and $\frac{S}{a_1}$ should be equal. To test the hypothesis, Tajima (1989) showed that the distribution of D with mean 0 and variance 1 approximately follows a beta distribution.

Tajima (1989) points to the fact that a test may be derived for multiple loci by utilising the sum of independent beta distributions. We construct a test to compare Tajima's D for each group of COGs. Since under the null hypothesis, the values of D are assumed to follow a beta distribution scaled to have mean 0 and variance 1, we deploy the Central Limit Theorem and assume that the sum of all values of D will follow a standard normal distribution, under the null. We then construct a test where at significance level α, we reject the null hypothesis that the mean value Tajima's D differs between the two groups of COGs if

$$Z_{\frac{\alpha}{2}} < \frac{\bar{D}_{\text{large}} - \bar{D}_{\text{small}}}{\sqrt{\sigma^2_{\text{large}}/n_{\text{large}} + \sigma^2_{\text{small}}/n_{\text{small}}}}.$$

We also perform the Tajima's D test in the traditional sense as described above, with a Bonferroni correction for multiple testing.

Results

We computed the Shannon diversity for each COG, the distribution of diversity values for both weight groups is shown in Figure 2. The mean Shannon diversity of the large group was 1.68 (sd = 1.04) and 1.57 (sd = 1.15) for the small group. Visually, from Figure 2 the distribution for the large group looks to be bi-modal. Furthermore, there are a few more values in the tail of the distribution in the small group, this may contribute to the distributions looking different while they have similar means. The p-value of the KS test between the two distributions is 0.118. Thus at $\alpha = 0.05$ we conclude that there is not a significant difference between the Shannon diversity among COGs in the large weight group versus COGs in the small weight group.

We computed the scaled phylogentic diversity for each COG, the distribution of diversity values

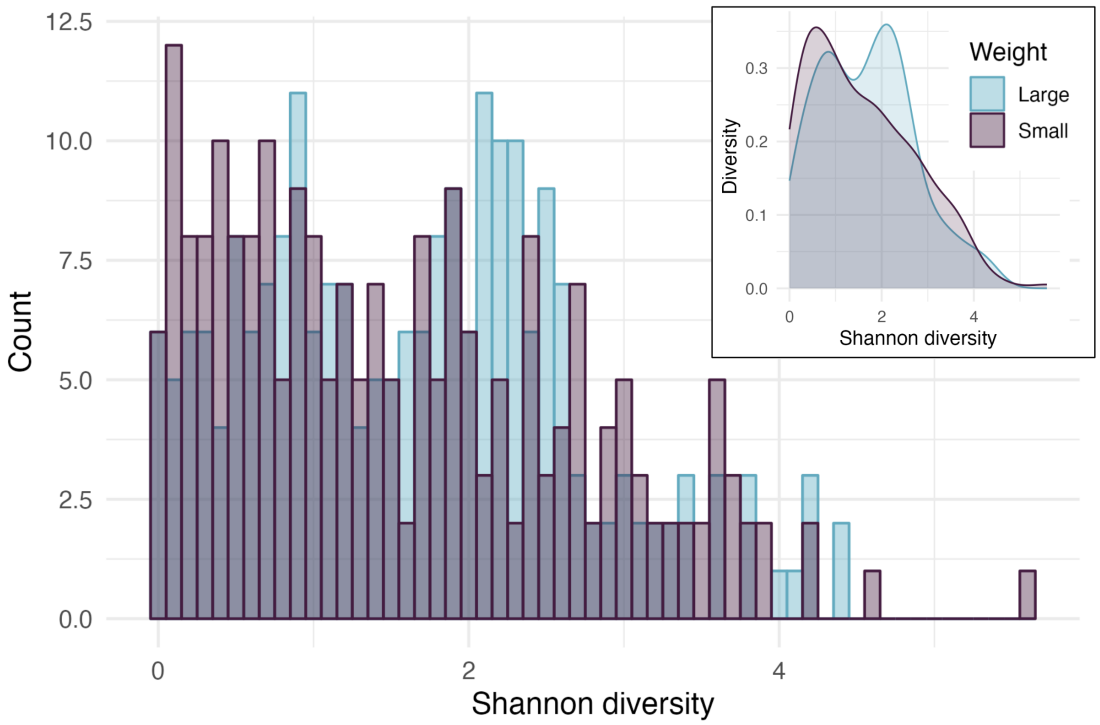

Figure 2: Histogram and density of values of Shannon diversity for each COG in the large (blue) and small (purple) group.

for both weight groups is shown in Figure 3. We observe that there appears to be more extremely small values in the distribution of the large weight group, additionally there was an outlier in the large group with a scaled phylogenetic diversity of 0.232. With this value included, the mean phylogenetic diversity of the large group was 0.00986 (sd = 0.0204) and without it the mean was 0.00885 (sd=0.0138). The mean for the small group was 0.00737 (sd = 0.0143). The p-value of the KS test between the two distributions is 0.1308. Thus at $\alpha = 0.05$ we conclude that there is not a significant difference between the phylogenetic diversity among COGs in the large weight group versus COGs in the small weight group. We also compared the phylogenetic diversity within each group, pre- and post-vaccine. We found these to be very similar to the overall distribution and no significant differences.

We computed Tajima's D for both groups of COGs, the distributions of which are given in Figure 4. Again, visually the distributions look similar. The mean value of D in each group was $\bar{D}_{\text{large}} = -1.43$ and $\bar{D}_{\text{small}} = -1.63$. For reference, the corresponding standard deviations are 1.81 and 1.73, respectively. Assuming that the mean D values follow a standard normal distribution, the value of our test statistic for comparing two means is 2.04. At significance level 0.05, the corresponding p-value is 0.0492. Thus, we can conclude there is a difference between the means of

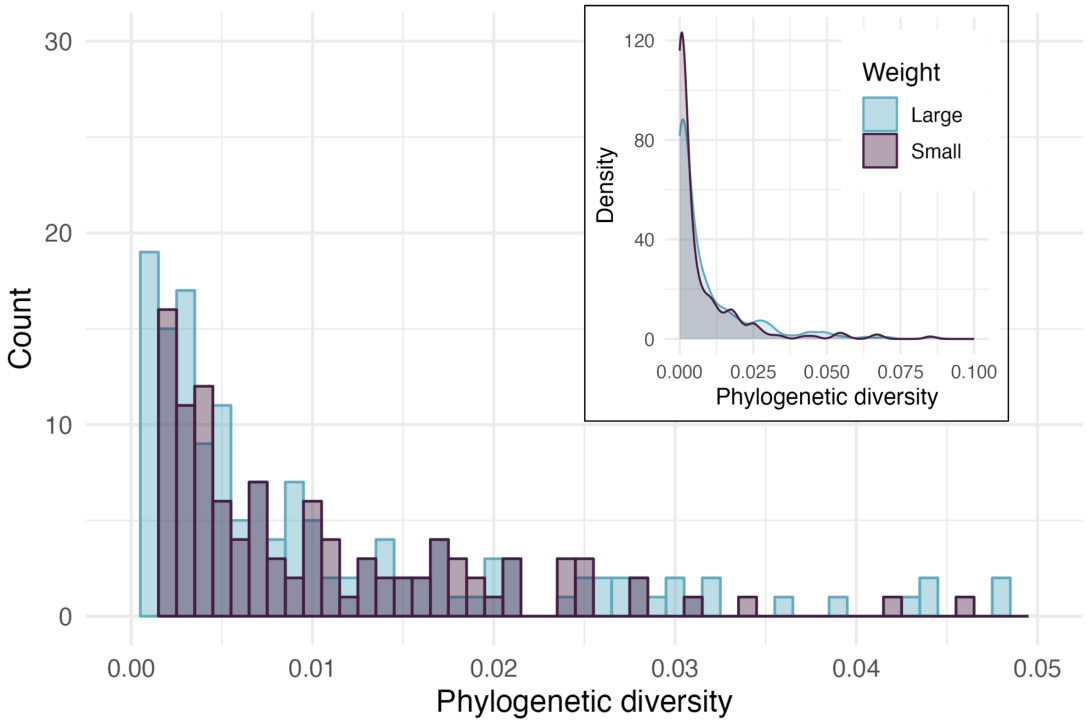

Figure 3: Histogram and density for scaled phylogenetic diversity for each COG in the large (blue) and small (purple) group.

the two groups of COGs.

Finally, we obtained p-values for Tajima's D test for each locus with multiple testing correction. Of these p-values, 43% were significant at level 0.05 in the large group and 47% were significant in the small group. Recall that rejection of the null hypothesis in Tajima's test implies that there is a process of selection acting on that locus, that is, random mutation alone does not account for the genetic variation. These results suggest that non-neutral evolutionary patterns are common in the intermediate-frequency loci of *S. pneumoniae* and that there is not much evidence of a different between the groups.

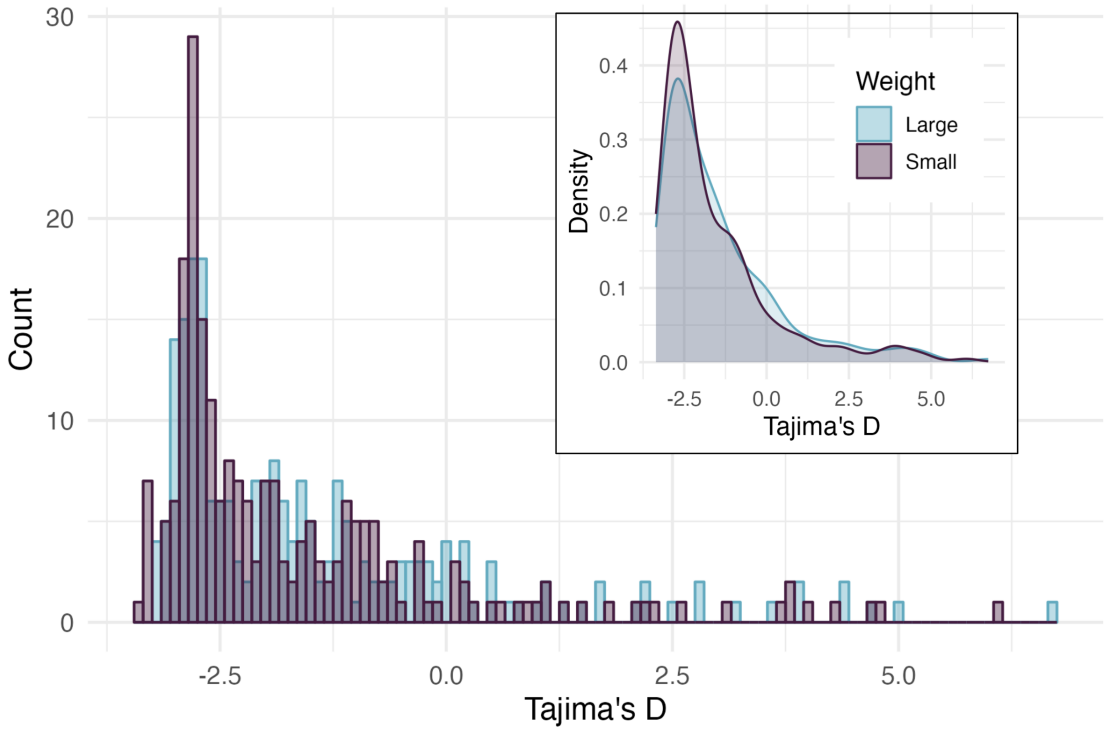

Figure 4: Histogram and density of values of Tajima's D for each COG in the large (blue) and small (purple) group.

Discussion

In this project we investigated the genetic diversity of *Streptococcus pneumonia*, a microbe whose accessory loci are thought to experience widespread negative-frequency-dependent selection (Corander et al., 2017). Our intuition was that, since NFDS is a diversity promoting form of selection, if a locus experiences a strong NFDS effect, then it will be more genetically diverse than loci not experiencing NFDS. By applying three diversity measures to COG sequences, our aim was to determine if the loci claimed to be experiencing a strong effect of NFDS in Corander et al. (2017) were more diverse than those loci that were experiencing a weak effect of NFDS in Corander et al. (2017). Our three measures of diversity were in agreement; there does not appear to be a difference in diversity between these two groups of loci except when we compared group means of Tajima's D, we got very weak significance.

One drawback of this analysis was that we computed diversity independently for each COG. We did this because, due to the nature of the accessory genome of the pneumococcus, not all COGs contain sequences of the same taxa, which means that creating a multiple sequence alignment for all the COGs in one of our NFDS groups would limit the analysis to loci found in a subset of taxa, which would require throwing away a high volume of data since the accessory genome is so variable.

Additionally, the groupings of loci experiencing NFDS according to Corander et al. (2017) may be suspect. In Corander et al. (2017), the space of all possible locus weightings is vast (2 times the number of loci) and assignment of COGs to groups depending on NFDS is complicated. There may be other groupings that are more effective. An interesting future direction would be to use only the loci that were deemed to be significant in Tajima's test in the Corander et al. (2017) model and evaluate its performance.

References

A. B. Brueggemann, R. Pai, D. W. Crook, and B. Beall. Vaccine escape recombinants emerge after pneumococcal vaccination in the united states. *PLoS Pathog.*, 3(11):e168, Nov. 2007.

C. Chaguza, J. E. Cornick, and D. B. Everett. Mechanisms and impact of genetic recombination in the evolution of streptococcus pneumoniae. *Comput. Struct. Biotechnol. J.*, 13:241–247, Apr. 2015.

C. Colijn, J. Corander, and N. J. Croucher. Designing ecologically optimized pneumococcal vaccines using population genomics. *Nat Microbiol*, 5(3):473–485, Mar. 2020.

J. Corander, C. Fraser, M. U. Gutmann, B. Arnold, W. P. Hanage, S. D. Bentley, M. Lipsitch, and N. J. Croucher. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat Ecol Evol*, 1(12):1950–1960, Dec. 2017.

N. J. Croucher, J. A. Finkelstein, S. I. Pelton, P. K. Mitchell, G. M. Lee, J. Parkhill, S. D. Bentley, W. P. Hanage, and M. Lipsitch. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.*, 45(6):656–663, June 2013.

D. P. Faith. Conservation evaluation and phylogenetic diversity. *Biol. Conserv.*, 61(1):1–10, Jan. 1992.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376, 1981.

W. P. Hausdorff, D. R. Feikin, and K. P. Klugman. Epidemiological differences among pneumococcal serotypes. *Lancet Infect. Dis.*, 5(2):83–93, Feb. 2005.

N. L. Hiller, A. Ahmed, E. Powell, D. P. Martin, R. Eutsey, J. Earl, B. Janto, R. J. Boissy, J. Hogg, K. Barbadora, R. Sampath, S. Lonergan, J. C. Post, F. Z. Hu, and G. D. Ehrlich. Generation of genic diversity among streptococcus pneumoniae strains via horizontal gene transfer during a chronic polyclonal pediatric infection. *PLoS Pathog.*, 6(9):e1001108, Sept. 2010.

T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, Jan. 1969.

M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903, Apr. 1969.

A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

M. K. Konopiński. Shannon diversity index: a call to replace the original shannon's formula with unbiased estimator in the population genetics studies. *PeerJ*, 8:e9391, June 2020.

L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, 32(1): 268–274, Jan. 2015.

C. E. Shannon. The mathematical theory of communication. 1963. *MD Comput.*, 14(4):306–317, 1997.

F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2): 437–460, Oct. 1983.

F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, Nov. 1989.

J. Wakeley. *Coalescent Theory*. W.H. Freeman, 2009.

G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, 7(2):256–276, Apr. 1975.