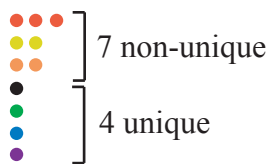


The theory behind LIBSHUFF

A library represents a sample of a larger population. For example, imagine a sample of colored marbles. In this example, each color represents a different 16S rRNA gene sequence. Closely related sequences are represented by similar colors.



To calculate the coverage of the population by the sample, determine the number of unique marbles in the sample.



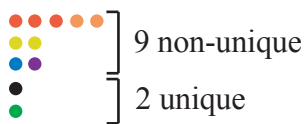
Thus, coverage of the population by your sample would be:

$$C = 1 - (\text{number unique}/\text{total number})$$

$$C = 1 - (4/11)$$

$$C = 63.6\%$$

If we broaden our definition of "unique" such that close colors are considered the same (e.g. red/orange; blue/violet), then our coverage changes.



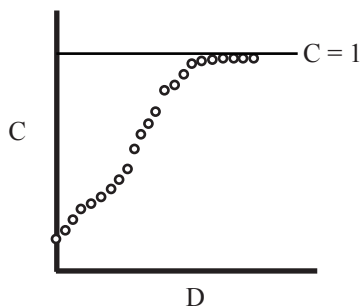
Thus, coverage of the population by your sample would now be:

$$C = 1 - (2/11)$$

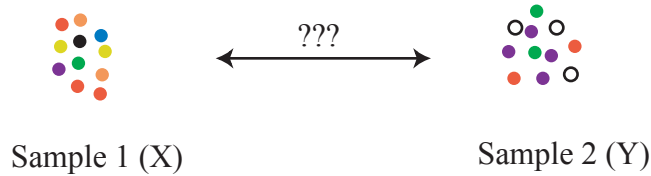
$$C = 81.8\%$$

The 16S rRNA gene equivalent of this is to increase the allowable evolutionary distance that determines whether two samples are the same. For example, where previously a distance of 0 might be required for two sequences to be the same (i.e. 100% sequence similarity), now a distance of only 0.01 is required (i.e. ~99% sequence similarity).

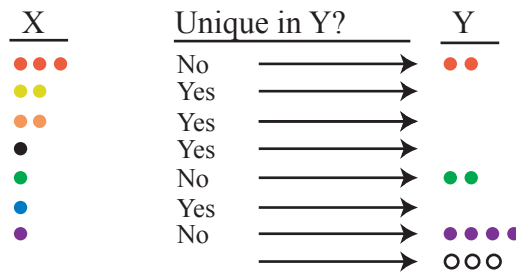
Plotting the coverage (C) vs. Distance (D), gives a coverage curve for the sample (called a "homologous" coverage curve, or C_x)



If we have two samples, from two populations, we can compare them and determine if they are significantly different.



To compare the samples, we first compare every marble in X individually to the entire sample Y to see if it would be unique if it was in library Y.

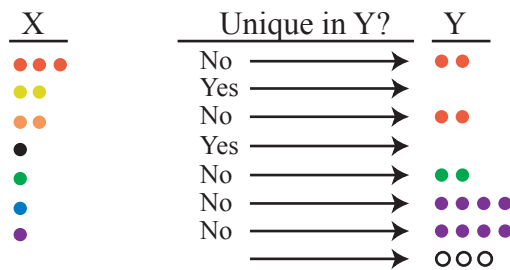


In this example, 6 out of 11 marbles from X (2 yellow, 2 orange, 1 black, and 1 blue) were not found in Y, thus these marbles would be unique. Then, our calculated "heterologous" coverage of X by Y would be:

$$C = 1 - (6/11)$$

$$C = 45.5\%$$

Once again, relaxing the criterion for uniqueness such that red/orange and blue/violet are considered the same, we get a different coverage.

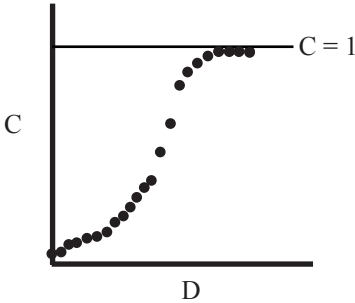


Now, only 3 of 11 marbles in X were not found in Y, and would be considered unique. Thus, our "heterologous" coverage of X by Y under this definition of "unique" would be:

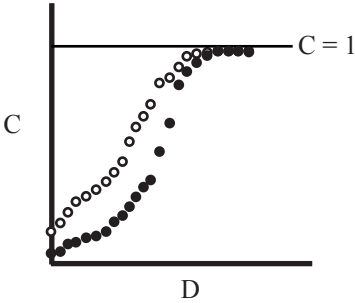
$$C = 1 - (3/11)$$

$$C = 72.7\%$$

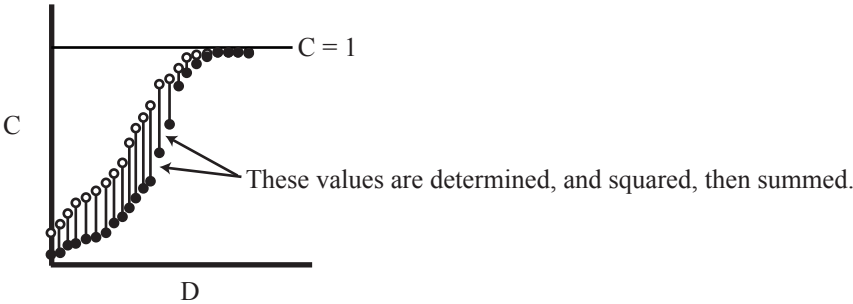
Similarly to the homologous coverage curve, a "heterologous" coverage curve, or C_{XY} can be created.



When the two curves (C_x and C_{XY}) are plotted on the same graph, we begin to see how well sample Y represents sample X. When two samples are not significantly different, the curves are similar. However, in this example, the two curves do not appear similar. But are they different enough to consider the samples significantly different?



To determine this, we first determine the value ΔC using the Cramér von-Mises type statistic. Simply, the difference between equivalent points in the two curves is determined, that value is squared, and these squares are summed to give a single number (ΔC).



Once a ΔC has been determined for the original samples, the samples are shuffled together.

In order to illustrate the concept of the shuffling, consider this simple comparison:



If we shuffle these two samples together . . .



. . . and then randomly separate them into NEW samples of the same size as the originals . . .



. . . then most of the time the NEW samples will look something like this, a mixture of the two original samples. In this example, it is obvious that the original samples (X and Y) are far more different to each other than the NEW samples are to each other. If ΔC values were calculated for the original samples and the shuffled samples, the ΔC value would be much greater for the original samples.

If we shuffle the originals a large number of times (say 999), it is likely that the original samples will be more different (higher ΔC) than the VAST MAJORITY of shuffled samples (lower ΔC s). Thus the original samples would appear to be significantly different.

For the opposite extreme:



Sample X



Sample Y

If we shuffle these two samples together . . .



. . . and then randomly separate them into NEW samples of the same size as the originals...



. . . probability determines that the shuffled samples will look very similar to the original samples. If we shuffle these samples a large number of times (say 999), chances are that a fair number of shuffled sample will appear MORE different (higher ΔC s) than the original samples. Thus, the original samples would NOT appear significantly different.

Obviously, complex samples will not be as obvious as these examples. To determine the significance of difference between original samples and shuffles of those samples, a p-value is determined by ranking ΔC values.

For example, imagine that two samples have a ΔC of 0.251, and we shuffle the sample 999 more times to obtain a total of 1000 ΔC values. After sorting the entire list of ΔC values, we can determine where our actual ΔC value ranks . .

ΔC	Rank
⋮	⋮
0.253	450th
0.251	451st
0.250	452nd
0.249	453rd
0.245	454th
⋮	⋮

← Thus, the p-value would be 0.451, the samples would not be significantly different.

However, if the libraries were not similar, the ΔC value may rank differently...

ΔC	Rank	
⋮	⋮	
0.253	4th	
0.251	5th	← In this example, the p-value would be 0.005, and the samples would be considered significantly different.
0.250	6th	
0.249	7th	
0.245	8th	
⋮	⋮	

Because 16S rDNA clone libraries may have hundreds of sequences (signified in these examples by marbles), with hundred of different sequences (different colors), the calculations are much more complex than these simple examples. However, these examples should illustrate some of the basic concepts underlying the LIBSHUFF analysis.

Thank you for your interest in our analysis.