**Q3.1**

To identify which bacteria from our rRNA database are present and how abundant each species/strain is based off of these metagenomic data there is biological issues of the experiment I think we should address. rRNA genes have regions that are hypervariable and other regions are ultra-conserved. It is likely that some of our reads fall within these ultra-conserved regions. If we have reads from the ultra-conserved regions these reads will not be informative for identifying species/strain of bacteria or estimating abundance, and they will just slow down future analyses. To remove these ultra-conserved reads, I will try to align all 1000 of the reads to a single ultra-conserved region reference. Any reads that successfully align to it I will not include in future analyses.

Using just the remaining hypervariable reads, I will align each read to a reference of all the rRNA sequences of our rRNA database. I will be able to identify the species/strain represented in microbiome sample by simply looking at which species/strains from the rRNA database had hypervariable reads align to them. This should be a relatively fast process, since we will have fewer than 1000 reads to align and a relatively small reference of 10,000 rRNA genes. This method should give us an accurate depiction of what species/strains are represented in our sample, but could be biased by low read counts.

The method I would use to estimate the abundance of get bacteria is to count the number of hypervariable reads that aligned to each species/strain. There are a lot of variables that could affect the accuracy of our abundance estimate. First, is that cells generally produce a lot of rRNA, so there is not a 1 to 1 correlation between rRNA molecules and bacterial cells. Along the same lines, there is no reason to assume that all of our bacteria produce the same amount of rRNA. With such small read counts, it could be that a few bacteria are greatly biasing our counts. It would also be important to account for PCR biases when looking at this data by removing duplicate reads, but that is a common QC step in NGS analyses.

**Q3.2**

To identify which of species/strains of bacteria are associated with IBS I would like to use a neural network approach, because I do not think the bacterial rRNA abundance data will cluster our IBS and control samples into two convex sets. Therefore, I think a more complicated classifier system will be required, which is a strength of using a neural network. My reasons for thinking that these data will require a more complicated classifier is that there has been a lot of work attempting to correlate bacterial species/strains with IBS, but there has been little success at the species/strain level.2 I believe this is due to multiple different combinations of bacterial species being associated with IBS. These different combinations may only lead to IBS under specific environmental factors, likediet, or due to a biochemical component that are associated with many different species/strains, but only leads to IBS under certain bacterial species and strain combinations.

In order to have a densely populated feature space I plan to define my feature space as the abundance counts for the subset of the species/strains that I think it is likely I can get read counts for in all or the vast majority of the IBS and control samples. I can define this by only include the species/strains that had a high read count from my Q3.1 experiments, but I would prefer to define this based on a larger sample size, and deeper sequenced data. My training set will be the species/strain abundance data for 50 randomly selected IBS samples and 50 randomly selected control samples. Thus, my test set will be the remaining 50 IBS samples and 50 control samples. This 1:1 cross validation will make sure I do not overfit these data. I am concerned about overfitting because even 100 test and 100 control samples may be too small to be representative of the human population. Also, if I am severely limited in the number of species/strains I can use due to our ability to reliably get read counts, then it is possible we are not looking at the abundance of a species/strain that is very important to IBS due to not having sequenced it well in our Q3.1 study.

Q3.1

First, I will use the time and space efficient BLAST tool to do approximate optimal local alignment against the sequences in the database. Then the reads will be separated into two groups based on their similarity scores. If the similarity scores of the reads are lower than a threshold (ex: 99%), we will consider them derived from new species, while the reads with higher scores will be mapped to the microbial species with the highest similarity. Then multiple sequence alignments will be carried out for the reads in the lower-scored group including constructing phylogenetic trees and aligning all the reads. Similarity, reads with similarity score higher than 99% will be clustered into the same species. To quantify the relative abundance of each species, we will use the read count for each species divided by the total read count to represent the abundance for all the species which the reads from the bowel sample are mapped to in our methods (including the mapped species in the database and the species in our putative clustering groups).

Q3.2 First, I will create a matrix with rows representing each patients and columns representing all the microbial species identified in the 200 samples. The K-fold cross-validation (K=5) will be used to create models, test our models and get the results in the following steps. To create the model, 200 samples will be randomly separated into 5 parts and 160 samples (160 rows) will be randomly chosen as the training dataset. The original value will be first put into each cell (the value is zero if certain species is not present in any samples), and we will use quantile normalization to normalize all the read count values. Next, we will perform logistic regression using normalized read counts for each species as variables and the two phenotypes (IBS patients/ control) as the predictive values. Then a logistic regression model can be generated and the key species that might be associated with IBS can be extracted by evaluating the significance of each variable. To test the model, the rest 40 samples will be used as the testing data. A matrix with 40 rows will be generated with the normalized read counts put into each cell, which is similar to the method used for training data. Then accuracy of the model can be evaluated by dividing the count of samples with right prediction by the total sample count. The training and testing process will be carried out for 5 times (5-fold cross-validation) and the mean accuracy can be calculated for evaluating our models. We will use union of the identified key species in each cycle as our resulting key species that are associated with IBS.

Q3.1

To identify the species and strains present in the bowel sample, we should first consider that while the 16S rRNA is a very highly conserved gene, it has regions that are highly variable between species, and within those regions, has sequence regions that are highly variable between strains. These variable regions are so species/strain-specific that a mere 58 nucleotides can distinguish between many bacterial species within infections [Chakravorty et al. 2007]. Thus, for computational efficiency, we can first reduce our search space by filtering our 100 bp reads by the simple absence/presence of these less oft-conserved regions, excluding regions that are so highly conserved that they can be used, as Chakravorty et al. describe, as “universal primers”. We can do same of those regions in the 16S-rRNA database using the “universal primer regions” this study describes, reducing the database string length for each of our queries. Then, we can map the remaining, more variable regions to the 10,000 fragmented database regions using local alignment, having cut down our search space on both sides. Quantifying relative abundance of each species (and especially of each strain) could be difficult given that we have 1000 reads each containing small fragments of presumably overlapping microbial individuals. However, in mapping our 16S-rRNA reads to the database strings, if we increase “observed count” only if the second read to a certain species from the same (or a highly similar) unique string of the database 16S-rRNA, we should be able to get relative abundances of these species. Though it is not likely that this database also has information about every strain of every species, if we do see multiple hits to the same species, we can compare the sequences of reads mapping to the same species to see if there are any noticeable strain differences.

Q3.2

Assuming we derived useful relative abundance information from our analysis in Q3.1, we could train a supervised, Support Vector machine(SVM) to predict the IBS status of humans based on the 16S-rRNA contents of their gut microbiota. Specifically, our feature space would consist of a “relative abundance” variable for each of the observed species (or, if our Q3.1 results showed discernible strains, for each of the observed strains). We would train our SVM using 50 healthy and 50 IBS-positive subjects as our training set. Though these data could be quite sparse considering that many strains/species could be absent (either for some clinically meaningful reason or because of drift alone) from some patients, our SVM should still be able to use these data to differentiate between . We will use the remaining 50 healthy and 50 IBS patients’ 16S information to validate our SVM’s output. That is, we will see if the SVM classifier is able to accurately sort the subjects in the test set into IBS/Non-IBS categories based on the same features it trained on. Though an SVM would be very useful for incorporating many relative abundance features into a binary classifier, a major downside of this method could be a lack of interpretability— unlike a simpler method such as PCA or regression, to my knowledge, you can’t examine the loadings and easily identify the highest-contributing features to the classifier. With unlimited computational power (ha ha) we could permute on the features we include(or more likely, leave out) in training the SVM, and see which features’ loss causes the predictions to suffer the most. References Q3 Chakravorty S, Helb D, Burday M, Connell N, Alland D, A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria, Journal of Microbiological Methods, 69(2) (2007)

https://en.wikipedia.org/wiki/Support\_vector\_machine