Discussion

Nate Olson

2017-11-08

1 Discussion

- Using a two-sample-titration mixture design to assess the 16S rRNA metagenomic measurement process. * Comparing pipelines * Relative abundance * log fold-change

By using mixtures of environmental samples we were able to generate benchmarking datasets with the diversity, relative abundance dynamic range, and sequencing artifacts of a real dataset. There were two primary limitations of the study that were a product of the experimental design. Only features that were differentially abundant between the pre- and post-exposure were used in the assessment. Using samples from the vaccine trial provided a specific features, *E. coli* that could be used during method development. However, only a limited number of features were differentially abundant between the pre- and post-exposure samples resulting in a smaller set of features that could be used in our assessment. Generating mixtures of samples with less similarity would increase the number of features used in the assessment. Additionally, using samples from other environments would increase the taxonomic diversity of features used in the assessment and potentially allowing for a more rigorus evaluation of the relationship between the assessment metrics and phylogeny. The second limitation of the experimental design was the difference in the proportion of bacterial DNA between the pre- and post-exposure samples. We were able to use an assay targeting the 16S rRNA gene to detect changes in the concentation of bacterial DNA across titration but we were unable to estimate the proportion of bacterial DNA in the unmixed samples using the qPCR data. Using the 16S sequencing data we inferred the proportion of bacterial DNA from the post-exposure sample in each titration. However, the uncertainty and accuracy of the inferrence method is not known resulting in an unaccounted for error source. A better method for estimating the proportion of bacterial DNA in the unmixed samples would increase the accuracy of the error metrics.

Evaluated the performance of four different bioinformatic pipelines, open-clustering, de-novo clustering, sequencing inference, and unclustered. Running these pipelines on the same dataset resulted in a range of total feature abundance and features per sample. Despite the wide range in number of features the pipelines all generated datasets with similar levels of sparsity. As the dataset is highly redundant, 180 samples derived from 10 envirnomental samples lower sparsity was expected. The qualitative assessment results indicate that the sequence inference method had a high rate of false negative features. This high false negative rate likey resulted in higher sparsity. For the other pipelines the high sparsity is attributed to features that were artifacts of the sequencing process, false positives. The high rate of false positive features have been observed in benchmarking studies using mock communities. The 16S region sequenced in the study is larger than the region the de-novo and open clustering pipelines were initially developed for. The larger region has a smaller overlap between the forward and reverse reads as a result in our study the merging of the forward and reverse reads did not allow for the sequence error correction that occurs when their is greater overlap.

As the qualitative assessment results were pipeline dependent the implications for 16S metagenomic studies vary by pipeline. For de-novo and open-reference clustering methods any conclusions made based on low abundance features require additional justification. Specifically, how do you know whether the feature is a measurement artifact or represents a member of the microbial community. This is especially relevant for studies characterizing the rare biosphere. A study exploring the microbial ecology of **SOME BIRD** used a hard filter for low abundance features, but also compared the results with and without the filter ensuring that any conclusions were not biases by using the artibrary filter or including the low abundance features that are likely predominantly measurement artifacts. For 16S metagenomic studies using DADA2, missing low abundance features are more likely to impact presence/absence diversity analysis. Though a user can be more confident that an observed feature represent a member of the microbial community and not a measurement artifact. It is unlikely that the number of features in a sample accurately reflects the true richness of a

sample though how well the results datasets are able to detect real differences in richness between samples is unknown.

- Why quantitative analysis is biological replicate dependent?
  - What dependency means for 16S gene surveys, when does bioinformatic pipeline matter and when does it not matter?
  - Differences in proportion of bacterial DNA between the pre- and post-exposure samples drives individual specific results.
  - The proportion of non-prokaryotic DNA in a sample is not taken into considering for nearly all 16S studies.
  - How do differences impact inferences drawn from statistical analyses?
- Relative abundance
  - Outliers
  - Need to summarise across replicates
  - Noisy data
- log fold-change - outlier features
- Relationship between factors impacting quant and qual analysis

2 Conclusions

- How this dataset can be used to evaluate and characterize bioinformatic pipelines and clustering methods.
- Given study results
  - How would you analyze 16S sequencing data assuming current methods?
  - How would you like to analyze 16S sequencing data?
  - What are the limitations of current methods?
  - What would you like to see a clustering method/ pipeline be able to do?
    - What should be improved?