**Application Note Rough Draft Outline**

**Abstract**

**100 – 200 words**

* Basic info – 1 to 2 sentences:
  + Long-assemblies are much more available, which enables identification of repetitive elements in a genome. This allows us to determine the number of distinct alleles.
* Detailed background – 2 to 3 sentences:
  + Being able to determine the distinct number of alleles in 16S region of a bacterial genome allows for a new level of accuracy.
  + Short-read sequencing does not span the entire length of the 16S region, which makes it very difficult to differentiate pieces of 16S sequence from one allele to the other due to insufficient overlap.
  + While short-read sequences are a great representation of on a genus and species level, it cannot look at intragenomic variation.
* General Problem be addressed – 1 sentence:
  + We want an improved resolution of long read sequencing to compare intra(between) and inter(within) genomic variation of the 16S region in different *E. coli* strains.
* Main Results Summary – 1 sentence:
  + N/A
* Explaining Main Results – 2 to 3 sentences:
  + N/A
* General Context of Results– 1 to 2 sentences:
  + N/A
* Broad Perspective – 2 to 3 sentences:
  + Produce a python pipeline that will analyze the inter and intra genomic differences in the 16S region of closely related bacterial strains with the same species.
  + Being able to determine the degree to which intragenomic variation from intergenomic variation.
  + Being able to analyze 16S data on a greater level of resolution.

**Introduction**

**1 to 2 paragraphs**

* Biological Problem:
  + Short-read sequencing of the 16S region still remains as an impressive method due to its availability, however its resolution is only ideal for differentiating on the genus and species level. This is due the fact that shorter reads do not span the entire length of the 16S region, making it difficult to differentiate alleles. Long-read sequencing can remedy this and has the ability to look at intragenomic variation, compare alleles to one another and overall provide a better resolution than short-read sequencing.
* Brief summary of what we developed:
  + We’ve developed a python pipeline…
* GitHub Repo link:
  + https://github.com/amyrold/LongReads.git

**Implementation**

* Development Details:
  + Download all long read and hybrid assemblies for *E. coli* from Nanopore and PacBio and parse out genomes using a python script. Output should be a list of accession numbers.
  + BLAST *E. coli* genomes against a 16S rRNA database.
* Input and Output:
  + …
* Programming Languages and Functionality:
  + Unix/Linux Command line Tools
  + Python/Bio python
  + R
  + …
* Evaluation Parameters
  + Time complexity
  + Space/Memory complexity
  + Accuracy

**Results and Discussion**

* Results on test data
* Figures and Tables
* Limitations and Future Improvements

**References**

**NLM/ICMJE Style**

1. Jeong J, Yun K, Mun S, Chung WH, Choi SY, Nam YD, Lim MY, Hong CP, Park C, Ahn YJ, Han K. The effect of taxonomic classification by full-length 16S rRNA sequencing with a synthetic long-read technology. Sci Rep. 2021 Jan 18;11(1):1727. doi: 10.1038/s41598-020-80826-9. Erratum in: Sci Rep. 2021 May 19;11(1):10861. PMID: 33462291; PMCID: PMC7814050.

“We compared a 16S full-length-based synthetic long-read (sFL16S) and V3-V4 short-read (V3V4) methods using 24 human GUT microbiota samples.”

**(References 2 through 4 are ones we’ve haven’t used but can be useful)**

1. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019 Nov 6;10(1):5029. doi: 10.1038/s41467-019-13036-1. PMID: 31695033; PMCID: PMC6834636.

“We demonstrate that targeting of 16S variable regions with short-read sequencing

platforms cannot achieve the taxonomic resolution afforded by sequencing the entire

(~1500 bp) gene.”

1. Liu X, Andrews MV, Skinner JP, Johanson TM, Chong MMW. A comparison of alternative mRNA splicing in the CD4 and CD8 T cell lineages. Mol Immunol. 2021 May;133:53-62. doi: 10.1016/j.molimm.2021.02.009. Epub 2021 Feb 22. PMID: 33631555.

“While long-read technology was effective at assembling full-length alternatively spliced transcripts, the low sequencing depth did not facilitate accurate quantitation. On the other hand, short-read technology was ineffective at assembling full-length transcripts but was highly accurate for quantifying expression.”

1. Mueller RC, Ellström P, Howe K, Uliano-Silva M, Kuo RI, Miedzinska K, Warr A, Fedrigo O, Haase B, Mountcastle J, Chow W, Torrance J, Wood JMD, Järhult JD, Naguib MM, Olsen B, Jarvis ED, Smith J, Eöry L, Kraus RHS. A high-quality genome and comparison of short- versus long-read transcriptome of the palaearctic duck Aythya fuligula (tufted duck). Gigascience. 2021 Dec 20;10(12):giab081. doi: 10.1093/gigascience/giab081. PMID: 34927191; PMCID: PMC8685854.

“Our findings from a comparison between short-read and long-read reference transcriptomics contribute to a deeper understanding of these competing options.”