**Prediction of laboratory origin of sequencing runs by machine learning**

**Introduction**

**Data**

300 paired-end sequencing runs were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). 100 of the 300 datasets were submitted to the SRA by the CDC in May 2017. The remaining datasets were submitted by the CDC in May 2017. All of the datasets used Illumina MiSeqs to sequence *Escherichia coli*.

The SRA Toolkit (version 2.8.2-1) was used to download SRA files from the Sequence Read Archive. The same toolkit was used to convert the SRA files into FASTQ files. Custom Python scripts (version 2.7) were used to preprocess the data into a useable format as input into the machine learning algorithms.

Reference where the experiment data came from

**Methodology**

**Features**

We used three different feature types to infer the laboratory origin of sequencing runs. The first set of features was the proportion of every 4-mer in the reads of a sequencing run. Prior to sequencing, genomic DNA must randomly sheared into smaller pieces. There are several types of methods for shearing DNA, which include physical, enzymatic, and chemical. Since these methods fragment DNA differently, it is possible these methods have biases at where they fragment DNA. In addition different sequencing methods may preferentially sequenced particular subset of the generated DNA fragment. These biases may be ascertained by comparing the proportion of different 4-mers across sequencing runs. To calculate the proportion of 4-mers in a sequencing run, we slid a four DNA base size window down each read. At every position in the reads, we then would increment the corresponding 4-mer. This resulted in our first set of 256 features. (need better transition to last sentence…say stored in a 256 length vector?)

The second set of features we used is related to the quality of the sequencing reads. Depending on the protocol used and the skill of the technician processing the sample, there can be differences in the quality of the sequencing run. Each base in a sequencing read

Mean, Sd, Full distribution, Average decrease in quality as a function of base

The third set of features used as input into the machine learning algorithms is the fragment size distribution of each sequencing run.

**Machine Learning**

60% train, 20% test, 20% validation

Logistic regression – 97% accuracy

Feature set

Do you include N as a valid base or throw out

**Results**

**Future Work**

Additional datasets

Datasets from different experiments from the same laboratory

Datasets with different organisms

**References**

<http://www.nature.com/nrg/journal/v11/n10/full/nrg2825.html>

<https://academic.oup.com/nar/article/42/21/e161/2903156/svaseq-removing-batch-effects-and-other-unwanted>

<https://www.nature.com/articles/srep39921>

<https://genomemedicine.biomedcentral.com/articles/10.1186/gm208>

<https://cancergenome.nih.gov/researchhighlights/tcgainaction/AkbaniBatchEffectsCaseStudy>

<https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-014-0087-z>