Supplementary Materials

Supplementary Methods

1. Organizing bacterial genome data

For training the model we needed to identify the phyla of the bacterial genomes. To ensure that there was no bias, we separated bacterial sequences used in training and testing based on their phyla. Using the taxonomic ids we were able to get phyla information of the organisms. Taxonomic classification of the genomes were downloaded from the GenBank Taxonomic database. Summary files for every genome was downloaded from Refseq that had information on the taxonomic id. We extracted the taxonomic id from these files using the script *taxid.py* (see project code). Once the taxonomic id of the genomes were obtained, we obtained the phyla level information of these genomes using the code in *phylum.py*.

2. Establishing Metrics for Comparison

Table 1.1. Labeling different scenarios of classification

S. No.		Input sequence	Assigned to / Classified
	BWA		
A		Human reads	Human genome
B		Bacterial reads	Human genome
C		Human reads	Unmapped
D		Bacterial reads	Unmapped
	Kraken2		
E		Unmapped bacterial	Classified
F		Unmapped bacterial	Unclassified

Table 1.2. Definitions of metrics used for comparison of mapping against the deep learning model.

S. No.	Metric	Constituents
1	Correctly classified human reads	A
2	Correctly classified bacterial reads	E
3	Incorrectly classified bacterial reads	B + F
4	Incorrectly classified human reads	C