IGB Microbial Genomics Working Group

January 19, 2024
Slides and recording will be available

Mission

To bring together **microbiologists and computational scientists**
to address new complex questions in **microbial function and ecology**
that are **stretching the limits** of existing computational tools.

Last semester (Gaulke)

Table 1. Number of cancer samples downloaded from TCGA for re-analysis, from bladder urothelial carcinoma (BLCA), head and neck squamous cell carcinoma (HNSC), and breast invasive carcinoma (BRCA). The last column shows the number of reads that did not align to the human genome in the TCGA raw BAM files. WGS: whole-genome shotgun sequencing.

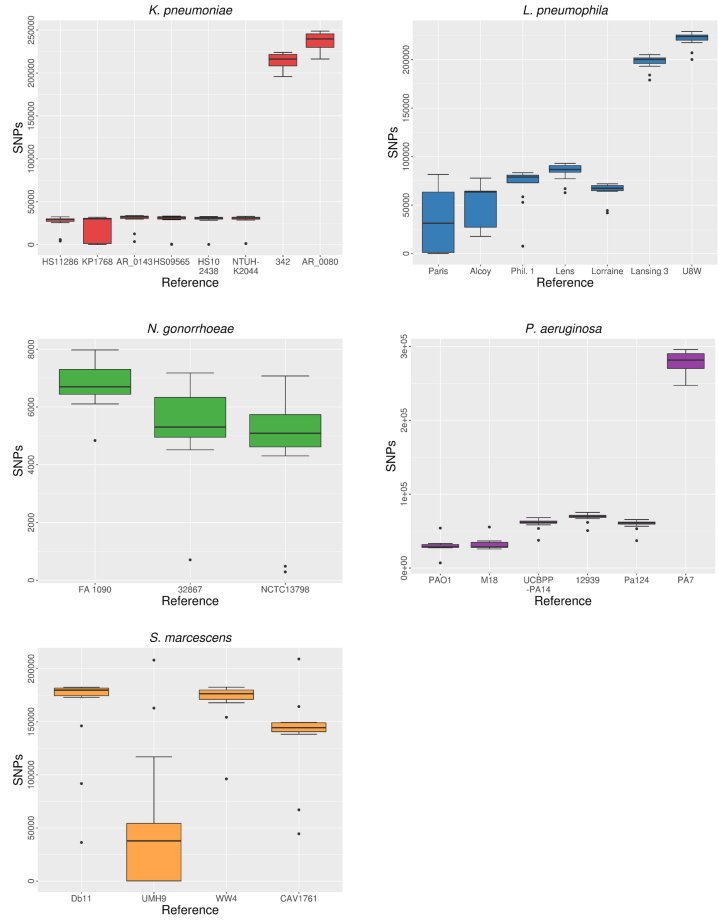
Cancer type	Number of samples downloaded			Total read count	Initially unmapped reads
	WGS	RNA-seq	Total		
BLCA	277	406	683	205,521,556,080	5,032,358,291 (2.4%)
HNSC	334	0	334	258,961,253,944	3,573,898,240 (1.4%)
BRCA	238	0	238	324,824,097,837	1,532,210,153 (0.5%)

The TCGA read data were analyzed with the Kraken program (11), a very fast algorithm that assigns reads to a taxon using exact matches of 31 basepairs (bp) or longer. **The Kraken program is highly accurate, but it depends critically on the database of genomes to which it compares each read. Poore et al. used a database containing 59,974 microbial genomes, of which 5,503 were viruses and 54,471 were bacteria or archaea, including many draft genomes.** Notably, their Kraken database did not include the human genome, nor did it include common vector sequences. This dramatically increased the odds for human DNA sequences present in the TCGA reads to be falsely reported as matching microbial genomes. This problem can be mitigated by including the human genome and by using only complete bacterial genomes in the Kraken database.

https://www.biorxiv.org/content/10.1101/2023.07.28.550993v1

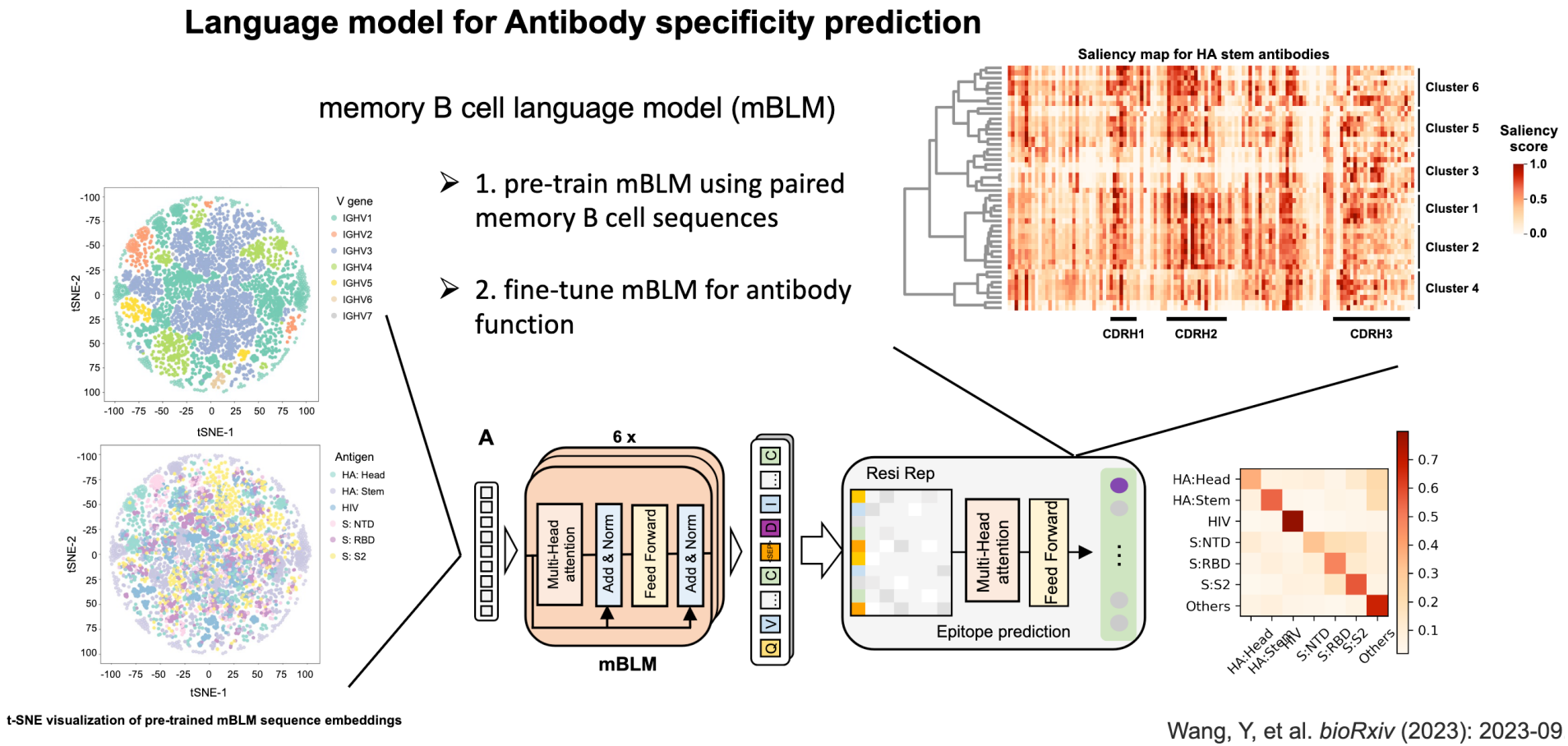
Last semester (David Vereau and Katy Heath)

In this work, we evaluated the effect of reference choice on short-read sequence data from five clinically and epidemiologically relevant bacteria (*Klebsiella pneumoniae*, *Legionella pneumophila*, *Neisseria gonorrhoeae*, *Pseudomonas aeruginosa* and *Serratia marcescens*). Publicly available whole-genome assemblies encompassing the genomic diversity of these species were selected as reference sequences, and read alignment statistics, SNP calling, recombination rates, d*N*/d*S* ratios, and phylogenetic trees were evaluated depending on the mapping reference. **The choice of different reference genomes proved to have an impact on almost all the parameters considered in the five species.** In addition, these biases had potential epidemiological implications such as including/excluding isolates of particular clades and the estimation of genetic distances.



https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008678

Last semester (Yiquan Wang, Wu lab)



Wang, Y, et al. *bioRxiv* (2023): 2023-09

Plans for this semester

- Meetings Fridays at noon in IGB 612 with lunch

  February 16, 2024: Chris Fields
  March 29, 2024: Ilan Shomorony
  April 26, 2024: TBD

- White paper on problems and potential approaches? PNAS Perspectives?

- Collaborative projects? Training? Genomics Corps?