Progress Report #1

Jeremy Fisher, Simon Levine,Tomas Matteson and Siddharth Reed

(1) A brief statement of the problem in your own words

We need to build a bioinformatic pipeline that isolates non-host/viral sequences from an RNA-assay and identifies their functional and evolutionary origins. While this pipeline should process a dataset provided during development time, it should be able to be invoked on unknown 'mystery' data. Our pipeline should accept an arbitrary RNA-Seq data set from a human sample and determine which viruses are present in the sample and find relevant viral ORFs or structural elements and known or predicted functional annotations.

(2) A list of relevant papers and software and a 1 sentence description of why they might be relevant

Pipeline Tooling

- **Snakemake** is a superset of Python that provides a declarative language for steps in a pipeline, intuitively providing useful features such as caching, cluster execution and configuration (https://snakemake.readthedocs.io/en/stable/)
- **Docker** is an operating system level virtualization solution that, in our case, helps with computational reproducibility; **Conda** is a package management solution for resolving compatibility issues and specifying toolsets (https://www.docker.com; https://docs.conda.io/en/latest/)

Quality Control

- **Trimmomatic** is a command line tool that removes adapter sequences from illumina sequencing data (Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.)

Alignment

- **STAR** is an alignment tool optimized for large, genome-wide targets that can be used for determining if a sequence derives from the human genome (Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635)
- **Bowtie2** is a similar alignment tool but is typically used for smaller targets, allowing us to check if a sequence derives from a viral genome (Langmead, Ben,

and Steven L Salzberg. “Fast Gapped-Read Alignment with Bowtie 2.” Nature methods 9.4 (2012): 357–359. Web.)

Statistical modeling

- **Scikit-learn** is a python library for dimensionality analysis and provides the interface for hidden-markov models implemented by, for example, hmmlearn (https://scikit-learn.org/ ; https://hmmlearn.readthedocs.io/en/stable/)
- **Pytorch** is an autogradient library that provides for shift-invariant deep learning models such as transformers, recurrent neural networks or convolutional neural networks (https://pytorch.org/)

Structural Element Search

- **BLAST** is a biological sequence database (i.e., proteins and nucleotide polymers) that enables efficient querying of biological origin given only a search sequence (https://blast.ncbi.nlm.nih.gov/Blast.cgi)

(3) Questions about the problem (do not include questions that are easily answered by doing your own research)

Can we expect the unknown sequence(s) for inference purposes to be similar to the given training set?

Do we know that the mystery sequences will contain a human virus sequence?

Are there memory/wall time performance requirements for running the pipeline during inference?

(4) Assignments of roles to each member of the team (who will lead the presentation, the write-up, the user manual, the validation of the software, and be the lead developer)

1. program manager – Tomas Matteson
2. technical lead – Jeremy Fisher
3. lead technical writer – Siddharth Reed
4. communications lead – Simon Levine