Computational Genomics Final Project Proposal
Team Members: Boyang Zhang, Winston Lin, Jacky Lin, Srihari Mohan

Overview

We aim to detect the inter-species sequence contamination problem with two approaches: sketching and alignment. We will benchmark our tool against existing sequence decontamination tools.

Methods

Prior to runtime, reference genomes from human and other species are preprocessed in preparation for alignment and indexed in preparation for sketching.

At runtime, sequencing reads are first aligned against the human reference genome. Reads that do not map to human (unmapped reads) are sketched and compared to genomes from other species. Note that this sketching process serves as a filter. It serves to collapse a set of similar unmapped reads to a single canonical/representative read that could be mapped against other species' genomes. This likely improves the tool's efficiency by mitigating our need to check many unmapped reads against each of our reference genomes. Note that only genomes that exceed a similarity threshold when compared to the unmapped reads are kept. Then, the unmapped reads are aligned against those genomes with local alignment methods. Good alignments between the unmapped reads and genomes from other species indicate contamination. To further quantify contamination, a confidence score is assigned to each read that aligns with genomes from other species based on how prevalent their sequence patterns are seen in nature.

Method Evaluation

We can evaluate the precision and recall of contamination calling in both simulated and real data. Simulated data is generated by mixing human sequencing reads with sequencing reads from other species, with different percentages of mixture. Real genome data will be obtained from 1000 Genome Project and other available public datasets.

Milestones and Stretch Goals

Milestones:

1) Generate simulated datasets: download available real data (both human and other species) and generate simulated contaminated sequences.
2) Align contaminated sequences to the human reference genome (i.e. using bowtie2 and other methods; we could even do a benchmarking performance of several alignment tools if we have time).
3) Perform sketching (or clustering based on similarity measurement) and choose representative contaminated sequences.
4) Build a command-line tool that accepts a sequencing file and returns a list of the offsets of contaminated subsequences in the passed sequence text and overall stats on these (i.e. how many contaminated subsequences, average length, presence of any notable

repeated contaminated subsequence, the probability that each is actually a contaminant, etc.)

Stretch Goals:

1) Optimize with parallelization in our implementation. We can run our read alignments against fragments of the sequence text in parallel to potentially make the matching process more efficient. When trying to compare a fragment against genomes of different species to determine whether there is a match, we can run these matches in parallel as well.
2) Run our implementation over a distributed architecture/cluster of machines to extract as much performance gains as we can from parallelization. We could explore whether parts of our implementation are amenable to being run over tools like Apache Spark.
3) Come up with a method that assigns confidence score more efficiently and accurately than BLAST search.
    a) Explore whether sequence modeling ML architectures such as RNNs would be effective in assigning such probability scores to sequences likely to be characterized as contaminants.

References:

1. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–359. Published 2012 Mar 4. doi:10.1038/nmeth.1923
2. Jun G, Flickinger M, Hetrick KN, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet. 2012;91(5):839–848. doi:10.1016/j.ajhg.2012.09.004
3. Lu J, Salzberg SL. Removing contaminants from databases of draft genomes. PLoS Comput Biol. 2018;14(6):e1006277. Published 2018 Jun 25. doi:10.1371/journal.pcbi.1006277
4. Dittami SM, Corre E. Detection of bacterial contaminants and hybrid sequences in the genome of the kelp Saccharina japonica using Taxoblast. PeerJ. 2017;5:e4073. Published 2017 Nov 17. doi:10.7717/peerj.4073
5. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011;6(3):e17288. Published 2011 Mar 9. doi:10.1371/journal.pone.0017288
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.