1. **Evaluation of Alignment Free Approaches for Transcript Abundance Quantification**

Recently, several alignment free approaches have been proposed to speed up the process of transcript abundance quantification via RNA-Seq. These methods include Sailfish, RNA-Skim, Kallisto, Salmon, and Rapmap. However, most of the evaluations neglect RNA-Skim in their analyses. A thorough comparison including RNA-Skim will be valuable. Compare the performance of one alignment-based method (such as Tophat2 or Star), one latest alignment-free method (Kallisto, Salmon, or Rapmap), and RNA-Skim. There are several aspects to be considered in the evaluation, including, but not limited to

1. Comparing the running time and accuracy among these three methods
2. Comparing the accuracy among these three methods on pseudogenes (genes with low sequence complexity).

Suggested Dataset:

Rat TGx - rat liver response to chemicals data: <http://camda2016.bioinf.jku.at/doku.php/data_download>

Human BodyMap:

<http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/samples/>

Suggested Simulated Data Generators:

polyester: https://github.com/alyssafrazee/polyester

RSEM: http://deweylab.biostat.wisc.edu/rsem/README.html

1. **Selecting the Best Size of *k-mer* for Transcript Abundance Quantification**

All existing alignment free approaches use a fixed size of *k-mer* to represent the signatures of transcript sequences. However, selecting the best *“k”* is challenging. The value of *k* can not be too small nor too big to be effectively and accurately recognize the RNA-Seq reads. Evaluate the effect of different *k* in terms of accuracy for all k-mer based methods.

1. **Using *k-mers* of Variable Sizes for Transcript Abundance Quantification**

RNA-Skim partitions the transcriptome into a set of non-overlapping clusters. *Sig-mers* are selected to represent the “signatures” of each cluster. One way to select *sig-mers* of variable sizes is to utilize the properties and structure of a suffix tree. The occurrences of these *sig-mers* in each read can be efficiently computed using the Aho-Corasick algorithm.

1. The most memory efficient implementation to handle genomic sequences is the Sadakane’s compressed suffix tree. Develop an approach to further improve the construction time or memory footprint.
2. Given a list of 6-9 millions *sig-mers* of variable-length, leverage the genomic sequence properties to build a memory efficient automaton for the Aho-Corasick algorithm.
3. Can BWT be used to further speedup the algorithm?
4. **Leveraging Replicates to Improve the Accuracy for Transcript Abundance Estimation**

RNA-Seq experiments are often conducted with biological or technical replicates. Current alignment free approaches are limited to process one sample at a time, without considering the replicates. Design and implement a model that can leverage the replicates information to improve the accuracy of abundance estimation at transcript-level.

1. **Base-Calling Method for Oxford Nanopore Long Read Sequencing**

The third generation sequencing produces long reads with read length up to 40-50 kb. Instead of reading out one nucleotide (letter) at a time, the Oxford Nanopore technology reads out 4 or 5 nucleotides at a time, referred to as 5-mers based-calling. Several probabilistic methods have been proposed to decode the DNA sequence, including the HMM model4 and ProSeq5. Compare and discuss the limitation of these two methods, or any other two existing methods for Nanopore long read base-calling.

References:

4. Timp, W. "DNA Base-Calling from a Nanopore Using a Viterbi ... - NCBI." 2012. <<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3353060/>>

5. Szalay, T. "De novo sequencing and variant calling with ... - Nature." 2015. <<http://www.nature.com/nbt/journal/v33/n10/full/nbt.3360.html>>

Suggested Simulated Data Generators:

NanoSim: <https://github.com/bcgsc/NanoSim>

ReadSim: <https://sourceforge.net/p/readsim/wiki/manual/>

1. **Phenotype Prediction Using Metagenomic Reads**

Metagenomic analysis has revealed a distinct microbiome composition in human diseases and health conditions. Develop an approach to classify patients into different disease status using their metagenomic sequencing reads.

References:

6. Pasolli, E. “Machine Learning Meta-analysis of Large …. - PLoS Comp. Biol.” 2016.

<<http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004977>>

Suggested Datasets:

Type 2 Diabetes:

<http://www.ebi.ac.uk/ena/data/view/ERA000116&display=html>

Obesity:

<http://www.ebi.ac.uk/ena/data/view/PRJEB4336>

More datasets can be found in the paper listed above

1. **MetaSUB Inter-City Challenge from CAMDA**

The MetaSUB International Consortium aims to create the world’s only longitudinal metagenomic map of mass-transit systems and other public spaces across the globe.

1. Compare organism fingerprints from public places across cities. Investigate organism sequences and biodiversity vs location.
2. Which computational tools have the highest sensitivity and specificity for species detection?

References:

<http://camda.info/>