Fat Fly Single Cell RNA-seq Pipeline

**Overview**

Single cell RNA-seq analysis is a novel method in next generation sequencing (NGS), bioinformatics, and computational biology. The difference between bulk RNA-seq (RNA-seq) and single-cell RNA-seq (scRNA-seq) is that a transcriptomic library is built and barcoded from one cell compared to lysing a bunch of different cells and mixing their different transcripts. The benefit of doing scRNA-seq is that you can get a snapshot in time of what a particular cell of interest is doing given certain conditions.

**Context**

Due to the novelty, different library preparations are done to provide various results depending on the biological question at hand. In the NGS experiment at hand, a library preparation method called CEL-Seq2 (add citation) was employed to produce sensitive results to analysis differential expression. It utilizes UMI technology with ligation-free adapter addition in order to improve mapping rates by retaining more transcripts. To further complicate the analysis is that no gold standard bioinformatics pipeline has yet to emerge. This can cause a decrease in confidence in the interpretation the data.

While there are different ways of handling the pre-processed data, the authors of CEL-Seq2 have optimized a raw data pre-processing pipeline that coincides with their library preparation methods. The pipeline takes multiplexed FASTQ and demultiplexes the files based on their illumine barcodes and UMIs, aligns the sequences to a reference genome with the assistance of a gene annotation file, and constructs the gene expression matrix. Once the expression matrix is formed, further that is used to conduct further statistical analysis such as clustering and differential expression.

The methods used for clustering and differential expression for scRNA-seq are relatively the same as bulk RNA-seq, however, there are certain algorithms that are more pertinent to single cell analysis. Unsupervised clustering approaches are more common in practice which will set up analysis for differential expression. The differential expression analysis is largely based on various statistical methods that have pros and cons to each method depending on the biological question. Generally the statistical methods employed are Negative Binomials, Zero-inflated Negative Binomial, and Poisson-Beta. These methods will need to be compared and evaluated in order to make sense of the question at hand.

**Goals & Non-Goals**

The goals for this project will be split into multiple stages:

* Pre-processing raw data
  + Evaluate raw data processing packages that perform demultiplexing, UMI handling, alignment, and expression matrix construction.
    - Starting input is a multiplexed FASTQ file
  + The two pipelines selected are CEL-Seq2 analysis pipeline and zUMI analysis pipeline
    - Need to determine alignment quality control, and matrix construction are accurate, and precise.
    - If time, possibly modify the pipelines to include Kallisto, pseudo-aligner, which can introduce more resolution at the transcript level and will include different statistical packages to use.
* Clustering and Differential Expression Analysis
  + Evaluate packages that perform the clustering and differential expression using statistical methods that are relevant to the biological question
    - May need to evaluate specific packages if a pseudo aligner is introduced.
  + Determining that the results are interpretable, accurate, and precise
    - Requires testing with known truth sets to ensure proper flags and parameters are in place.
    - Determine the output format in which will be most readable
  + NEED TO CHOOSE THESE PACKAGES (POSSIBLY R OR PYTHON); may need to piece together various languages
* User friendly design
  + Build a docker image container that will self-contain all necessary dependencies to be for easy transportation to different remote systems if needed.
  + Well-documented, concise, and readable code for future maintenance and reproducibility

The non-goals for this project are:

* Developing novel statistical methods for single-cell RNA-seq analysis
* Modifying existing open-source packages to handle data generated by platforms other than CEL-Seq2 (i.e. 10xGenomics, SMART-seq)
* Coercing any statistics to match the biological question or hypothesis

**Milestones Sprints**

* Milestone 1: 3/24/2019
  + Data processing software is compared, tested, and chosen.
  + Downstream Analysis software or design is chosen and finalized to begin implementation. Need to research and strongly consider the different statistical methods used.
  + Construct custom scripts to help glue certain pieces together and to make the pipeline more user friendly upstream.
  + Unit testing this part of the pipeline for bugs and setting up important logging information to capture details about the pipeline and various metrics.
  + Determine QC steps and metrics on output files from the pre-processing steps
  + If development is going well at this point, consider adding Kallisto as an additional pseudo-aligner to implement a specific package that looks at gene level counts and transcript level counts.
* Milestone 2: 4/7/2019
  + Downstream analysis software is fully implemented and connected to upstream processes.
  + Unit testing of the pipeline to make sure appropriate flags, bugs, and edge cases are caught. Need to document the testing performed.
  + Ensure the entire pipeline is running end to end with minimal problems, fixing any bugs or edge cases.
* Milestone 3: 4/14/2019
  + Rigorously test the statistical methods chosen and comparing reference truth sets to ensure the statistics are correct and that the pipeline is not biasing results.
  + Rigorous unit testing to catch more bugs and edge cases.
  + Begin to dockerize the package.
* Milestone 4: 4/21/2019
  + Heavy end to end testing of the final product with different reference sets to ensure accuracy and reliability of the pipeline.
  + Heavy documentation of testing, and report writing of the results to include in github.
* Milestone 5: 4/30/2019
  + Project Complete

**Proposed Solution**

*This proposed solution could change depending on results found during the development process.*

The proposed solution is encapsulated in a chart and will consist of testing multiple software solutions in mostly python and R language. The solution will be the following:

* Compare CEL-Seq2 and zUMI software for raw data processing.
  + Inputs are the FASTQ Read 1 and Read 2
    - Read 1 contains the UMI and barcodes
    - Read 2 contains the transcript
  + Processing includes demultiplexing, quality control, alignments, UMI and read counting
    - May need to modify some code to include quality control on the reads and trimming if not done by the software
    - Need to add QC step to the alignments and to the expression matrix construction
    - Consider adding a pseudo-aligner
  + Outputs from this step
    - Gene Expression Matrix
    - BAM files
* Compare and test various down stream analysis software packages post expression matrix
  + Compare DESeq2, EdgeR, MAST, Scanpy. This is still open for suggestions on which software to use depending on the results and statistics.
    - Need to determine Cell QC metrics
    - Normalize and Dimension Reduction
    - Determine clustering algorithm, typically UMAP
    - Perform the Differential Gene Expression
  + Determine gene expression validity and output format, along with the necessary logs to track statistics and results
* Dockerize this protocol for easy use and transferring between remote machines or the cloud.

