Background

In the broad scope of bioinformatics I am mostly interested in RNA based analysis. RNA sequencing has many different formats and the first step to do an analysis of most of these types of sequencing is alignment. Alignment is mapping the rna-seq fastq files to a reference genome and counting how many reads map to each gene providing a measure of its expression level in the sample. There are many different alignment programs based on the sequencing type from STAR alignment for Bulk RNA-Seq to more modern technologies like CellRanger Count for 10X's single cell RNA-Seq. For my final project I would like to make a STAR alignment web interface for Bulk RNA-Seq. STAR

Toll Functionality

For this project the front end would ask for the user input of: fastq files, species, and library prep method for the fastqs. Currently I am thinking of only using a single sample for the fastq input which would involve 2 fastq files, a R1 and R2. The user would also specify the species which would either be human or mouse, both reference genomes can be stored in the mySQL database. Lastly the user would specify the library prep used to generate the fastq files because different library preps require different arguments for STAR alignment. Based on those selections from the front end, the backend will use python cgi to run a STAR alignment using subprocess calls. For bulk RNA-Seq other functions should be run both pre and post alignment. I was planning on running pre-alignment functions like FastQC for checking the quality of the fastqs and Cutadapt for adapter trimming. For post alignment functions Picard MarkDuplicates, Samtools Sort, MultiQC and final counts matrix generation. The final output would be a FastQC.html, CountsMatrix.txt, and a MultiQC.html.

Inputs	R1 fastq (file), R2 fastq (file), Species (dropdown menu), Library prep (dropdown menu)
Commands	FastQC, Cutadapt, STAR --quantMode, Picard MarkDuplicates, Samtools Sort, MultiQC
Outputs	FastQC.html, CountMatrix.txt, MultiQC.html

Tool Description

MySQL: Since this is an analysis tool MySQL will be used to store files. The reference genomes will be stored here and for each project run we can store all of the project information. Each run will have a projectID along with the species and library prep used. I can also store all of the data for each project like the fastqs, BAM files, FastQC and MultiQC html files.

Python CGI: The python cgi will read in the input fastqs and other input data and run subprocess calls for all of the functions mentioned before. This will also read and write to the mysql database.

CSS/HTML/JavaScript GUI: I will have a very simple and elegant user interface where the inputs will be file selectors and dropdown menus. The output will be displayed as downloadable files.

Software Needed

Some of the software that would be needed on the bfx server for this project is cutadapt, picard, STAR, and multiqc. FastQC and samtools are already installed on the bfx server.

Questions

One possible upgrade is to allow functionality for 10X scRNA-seq with CellRanger Count. Cellranger. If you think single cell alignment would be a cool addition, cellranger would need to be installed too, cellranger download. Another question I have is whether it's better to have downloadable outputs from the webpage or to have the outputs automatically saved to mysql and make the user have to retrieve them.