RNASeq MiSeq Data Analysis Work Flow Using Galaxy Mason Account

After your run is completed successfully on the MiSeq, you need to download your data from BaseSpace. Go to BaseSpace, open your run, click on "Download" near the upper left corner of run window, choose "Unaligned data (fastq). The data will be in Fastq format and all zipped.

Then once you have your data on your computer, you will need to unzip them using the Zipeg program (Zipeg program is free to download for Macs, use 7Zip for PCs) These unzipped files are ready to be used on Galaxy now. This workflow is for the Galaxy instance on the Mason supercomputer at Indiana University. If you are in the Williams lab, you can use Weam's account, Chris White-Ziegler has an account for her lab, and Louie has an account for all others. Other accounts can be obtained for people with NSF funding. Contact Louie for help.

1. Upload your unzipped fastq files on Galaxy from **Import Data**; go to **Upload file** from your computer.
    a. (*This might take a couple hours to days and some time weeks, depend of how crowed is the Galaxy in use*) (If larger than 2GB, see Louie. She will load it onto her mason supercomputer account, link it to galaxy in a history, and share that history with you. From here it can be copied to any history you would like by choosing "copy dataset" from the gear icon in the upper right hand corner of your galaxy window.)

2. The first step is to groom your files, changing the offset from Illumina to Sanger (I don't know either it just needs to be done for the rest of the programs to be able to read your files). Under **Quality Control,** you will choose **FASTQ Groomer**, select one of your fastq files from the drop down menu, Select your input as Sanger (I know it is Illumina but for some reason this only works when you choose Sanger, so choose Sanger). Forget about advanced options and hit execute. Do this for all fastq sequence files.

3. Once groomed, you can check the quality of your sequences. Still under Quality Control, choose **FastQC: Reads QC.** Select one of your groomed files for "Short read data from your current history" and hit execute. Do this for each groomed reads file.
    a. After the program has run you can check the quality of your reads by clicking on the eye icon to the right of the data name. Scroll through the graphs to ensure that the quality is good. The most important graph to look at is Per base sequence content, which will tell you where there is noise in your sequence, usually just at the beginning and end in each groomed read. With this information, you can trim the ends as little as necessary to give you good clean sequence.

For the cleanest sequence (most important for de novo assembly and not quite as important for other applications), go to **FASTQ Trimmer**, again under **Quality Control**, and one at a time enter your groomed file for "FASTQ File". Define Base Offsets as:" Absolute Values" and then set the number of basepairs to be trimmed from each end (usually 9 on the 5' end and 2 or 3 on the 3' end) based on your FASTQC report.

4. For de novo assembly of RNA, the best program is **Trinity**. If you do not have a reference genome, you will need to do a de novo assembly and use that as your reference for downstream analysis. The more reads you can put into an assembly, the more complete the assembly will be. Trinity only accepts one single read file or two paired end files. Thus, if you have more than this, you must concatenate your data sets in order to get them all into the assembly. Under **Data Manipulation**, choose **Concatenate datasets tail-to-head**. For single end reads, simply select your trimmed reads from the drop down menu, adding new data sets until all are in. For paired end reads, you must concatenate all of your read 1s in one concatenated dataset, and all of your read 2s (IN THE SAME ORDER AS YOUR READ 1S) in another.

5. Now you are ready for your assembly**.** Under **De novo Assembly,** choose **Trinity – Executes on Mason** De novo assembly of RNA-Seq. Choose paired or single from the first drop down menu, and then your concatenated reads. Use your bioanalyzer data to give the maximum fragment length expected and it is fine to leave everything else as default. The job will need at least 10 hours, may want to choose 24 just in case. This will cause mason to put you farther in back of the line but if you choose 10 and it doesn't get done, the program will fail and you will have to set it up again, losing a lot more time.

   a. The remainder of the functions in this workflow are under **<u>Mapping and Alignment</u>** under a sub-folder called **RNA-SEQ**

6. First choose **TopHat for illumina**, this finds splice junctions using RNA-seq reads, to establish transcripts that will be assembled in the next step**..** Perform this for each trimmed read file (not concatenated ones) under the RNA-Seq FASTQ file drop down. For each you will want to specify the same reference sequence. Your choices are to use a built in genome, then choose from the drop down menu of genomes that are in galaxy, or use a genome from history, and then choose your trinity assembly, or if there is a published one you want to use that is not in galaxy, simply download it onto your desktop and then upload the file by going under **Import Data** and choosing **Upload File**. Choose whether library has paired reads or not and enter appropriate paired trimmed reads if appropriate. Try not to use default parameters. Use bioanalyzer data and any info you have on the genome to fill in parameters as accurately as possible. Use Microexon Search. Give it at least 10 hours to run.

7. Next choose **Cufflinks.** This will assemble the transcripts assigned by Tophat and assign gene names if you have reference annotation. If not, you will have to blast the sequence of every differentially expressed gene of interest when you get the list of significant genes with differential expression from **CuffDiff.** Run Cufflink on each "Tophat accepted hits" file. Read the parameters list at the bottom for help with the parameters. The more accurate you can be with parameters, the better the output will be.

8. Next choose **Cuffmerge.** Now you want to merge all of your data into one dataset for the differential expression analysis, which will put all of the normalized transcript counts together that you want compare between conditions in **CuffDiff**. So choose a "Cufflinks assembled transcripts" file from the drop down menu, and "Add new Additional GTF Files" until all Cufflinks assembled transcript files are entered. Add annotation files if you have them. You may need to make several merged files depending on the comparisons you want to make in **CuffDiff.**

9. Next choose **Cuffdiff.** Enter the Cuffmerge file under "Transcripts"and the "TopHat accepted hits" files for each replicate under each of your conditions. On the same page you need to change *Dispersion estimation method*: to **a blind** if you have no replicates (otherwise leave as pooled). Normalization method is up to you. I like classic, but you may want to try it with different ones. Normalizing is essentially normalizing the number of reads for each length of assembled transcript so that longer transcripts don't artificially have more reads just because they are longer. I would start with a false discovery rate of .01 (can go back to .05 if you don't get enough significant genes) Use multiread and bias corrections. And go to additional parameters and again use your bioanalyzer data to give as accurate info as possible. And look at the parameter list for help.

10. Then from this **cuffdiff** showing the differential expression you can use the MeV program to visualize the gene expression (one option) or use the visualization environment on Galaxy??? We have a commandline pipeline that is a possibility, and a CummerBund workflow (see this under "PowerPoints" on the CMB website. Ask Louie for help with these.