Introduction to RNA-seq

myGenomics offers RNA-seq services originating from total or purified RNA. Illumina-based RNA-seq does not sequence the RNA directly, but instead cDNA is generated to create the actual sequencing library. Depending on the species and project goals, your sample will be prepared using either oligo dT primers (polyadenylated RNA-seq) or random primers (whole transcriptome). Projects using random primers must use enriched RNA as input. This is most commonly done by rRNA depletion, or less frequently from immunoprecipitation (RIP-seq).

RNA-seq is not measured in terms of coverage, since in any given cell, not all genes will be expressed, and different genes will be expressed at different levels. Because of this, coverage is meaningless when discussing RNA-seq. Instead, your project is discussed in terms of millions of reads. Each cDNA is fragmented to generate a uniformly sized library to be loaded onto the flow cell. Since the fragmentation is random, each cDNA can be expected to be fully covered and generate multiple fragments per cDNA. Because of this fragmentation, you cannot assume that you only need one cluster for each transcript in the cell. myGenomics offers paired-end or single-end sequencing. In single-end RNA-seq, each library cluster on the flow cell generates a single read from one direction. So, the number of reads you obtain will equal the number of clusters, or cDNA fragments, on the flow cell. If you request paired-end sequencing, each cluster generates 2 reads, one from each direction/end. Therefore the total number of reads will be twice the number of clusters on the flow cell.

After the sequencing is complete, myGenomics will provide you with the raw FASTQ files, and if requested additional analysis will be performed. Raw FASTQ files are NOT human readable, and you will need to have a software package to be able to utilize the files. Even if you request analysis, these files are essential if you ever need to reanalyze or use a different tool to look at your data. For routine analysis, myGenomics uses the Tuxedo suite (Tophat, Bowtie and Cufflinks) for single sample analysis and CuffDiff for differential analysis. For more details on these packages, please see refer to the webite at http://cole-trapnell-lab.github.io/cufflinks/manual/. *Please note*, Cufflinks and CuffDiff *do not* offer gene ontology or pathway analysis. If you need this analysis you will need to request custom analysis. The outputs generated by the Cufflinks and CuffDiff packages are tab-delimited files. Most computers will open these with text viewers, and the results will be very difficult to view. myGenomics recommends using Excel to view these files. To open the files you will need to first open Excel, and from the File menu select Open. In the browser window that pops up, “All Files (*.*)” must be enabled to allow the tab-delimited file to be seen. The default will only be Excel files. Once the desired file has been selected, a dialog box will pop up in which you must specify that it is a delimited file, and that tabs should be used to specify each new column. Finally be sure to select each column to be “Text” and not “General.” If any gene names resemble dates, failure to make this selection will irreversibly convert those gene names into dates. The only resolution is to close the file without saving and repeat the process being sure to make the “Text” selection.

Generally you will be looking at a .diff file if you have requested differential expression analysis. The key columns to pay attention to are the “Gene,” “Sample_#,” and usually “Log2(fold_change).” The Gene column will tell you which gene the row data relates to. The sample_# will specify which sample the respective Value_# came from. Log2(fold_change) provides the fold change calculated by CuffDiff. The individual Value_# columns provide the fragments per kilobase mapped (FPKM) in the case of paired-end sequencing since 2 reads come from a single fragment, or reads per kilobase mapped (RPKM) for single-end sequencing since each read is a unique fragment. Because of this method of calculation, each sample is internally normalized and allows you to directly compare the

values if desired. Further normalization may be done based on housekeeping genes if you desire. The status column designates those tests which had the minimum amount of data *for all conditions* in order to perform statistical analysis to generate the p and q values. So, if you have a gene that is only induced under one condition, even if many reads are detected in your induced sample, no statistical analysis will be performed since there would be insufficient data in all other conditions.

For single sample analysis the resulting files will simply show each sample in the analysis (or each replicate) and the respective FPKM or RPKM for each sample and gene/transcript/isoform.

Below is a figure that can guide you to the specific file that you will want to examine based on the level of analysis you are interested in – whole gene expression, differential splicing, isoforms (alternate start sites), etc.