**RNA-SEQ PROJECT OVERVIEW**

Folder Structure:

Folder structure for the project is as follows:

1. **data** – the raw data and the metadata sheet
   1. **raw data** will usually be in the form of count files. This is the *direct output* from transcript assembly and is *not altered*.
   2. **Metadata** sheets take lots of forms – but these are the various forms that will be read into the pipeline. There could be many iterations and many versions, but the basic format is:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sample | Annotcol1 | … | AnnotcolN | Comp1 | … | CompN |
| Samp1 | Red | xx | male | 1 |  | 0 |
| Samp2 | blue | xx | 2 | 0 |  | 1 |

* 1. **Docs** can also be here – where are various documents that inform the experiment, and will also likely be the source data for the metadata

1. **Code** – the location for the scripts used for these analyses. Note that these are likely NOT STAND ALONE SCRIPTS, but rather master scripts that source in many functions from other helper files. These will need to be provided on an as needed basis
2. **Presentations** – this is where powerpoint summaries will go and various other presentation-focused analyses will go
3. **Manuscript** – where the manuscript docs and figures will go when needed
4. **Output** – all of the outputted analyses, more on this below
5. **Archive** – old analyses, docs, data, etc. Things that are considered old and most likely obsolete, but are kept just in case.

Interpreting Output

The output folder has several subfolders, each of which have their own subfolders. There is a lot of output, but it all serves a purpose, I will do my best to highlight important outputs here.

NOTE – there may be SUBOUTPUT FOLDERS AS WELL. It is common to exclude and add samples, metadata, etc. and each of these may get a different output folder.

1. **Processing** – this folder contains all of the upfront processing of the data.
   1. **Concat\_count\_files.txt** – the individual count files concatenated together in an unaltered form
   2. **Pre/postfiltered\_samp\_readcount\_hist** – filtering pre and post of our minimum readcount per sample tfor a SAMPLE to be included in the analysis
   3. **Pre/postfiltered\_gene\_rowmeans\_hist** – filtering pre and post of our minimum number of reads per gene per sample for a GENE to be included in the analysis
   4. **Filtrawcounttab.txt** – the raw count table after the minimum gene and read count filters – with the genes and sample excluded
   5. **Metatable\_filt.txt** – should be the same as the inputted metadata – but without the samples that were filtered out due to readcount filtering
   6. **read\_distribution\_bar\_chart** – a bar chart showing the top 50 represented reads in the project, on a per sample basis
   7. **raw\_count/norm\_count/norm\_ratio\_per sample\_grid1/2/3.pdf** – this is a visualization of the pre and post DEseq normalization of the read counts.
      1. **The Grid 3** figure is a nice view to visualize “blowing up” of any samples in which normalization had an unintended effect and overcorrected a potentially outlying sample
   8. **Normcounttab.txt** – the NORMALIZED count table
   9. **UQ\_raw/norm…grid.pdf** – normalization done via upper quartile normalization. These results should be similar to DESeq – and are done as a sanity check to insure that the normalization is consistent
   10. **UQ\_normcounttab**.txt– normalized data after UQ normalization
   11. **Normcounttab\_v12/v13**.**gct** - .GCT files of the normalized counts. .GCT files include metadata about rows and columns with the count data. See the format below.
       1. <http://software.broadinstitute.org/cancer/software/genepattern/file-formats-guide#GCT>
       2. <https://clue.io/connectopedia/gct_format>
   12. **Pca\_plots** – PCA plots of the data – colored by each annotation column provided in the metadata. VERY USEFUL for getting a snapshot sorting of your data
   13. **Pca\_plots\_normalized** - PCA plots of the NORMALIZED data – colored by each annotation column provided in the metadata. VERY USEFUL for getting a snapshot sorting of your data.
   14. **Outlier\_detection\_density\_curves** – a very crude way to see outliers by their gene counts. Curves that look drastically different from the others should be observed for read bias
2. **Plotting** – heatmaps and other figures for global views of the data
   1. **Hm1\_heatmap\_Allmetadata.pdf** – shows the top varied genes with all annotation data, and the genes and samples are clustered
   2. **hmXX\_heatmap\_sorted\_XX.pdf** – heatmap per annotation column that shows a heatmap that clusters genes, but MANUALLY SORTS columns by each annotation column provided
3. **deseq** – the DESeq output (differential gene expression)
   1. **comp\_XXXX** – there is a FOLDER PER COMPARISON COLUMN which will contain the DEseq results, and then a volcano plot and heatmap highlighting the most varied genes.
   2. **Pvalue/padj\_XXXXX\_DEseq\_summary** – these folders contain summaries of results at the given pvalue and log2fc
      1. **XXX\_summary.pdf** – barchart showing the number of genes found at the given log2fc and pvalue
      2. The rest of the tables show the genes that meet the cutoffs for EACH COMPARISON, in various forms – one with the gene name, one with the comparison name, one with log2fc, one with the pvalue, all of which are useful for different purposes.
4. **Gsea** – geneset analysis output per comparison – contents of the comparison folders are listed below
   1. **Comp\_XXXX\_gsea\_GO/HALL/KEGG.csv** – this is the GSEA output table for the analysis specified by the GO/HALL/KEGG
   2. **Comp\_XXXX\_gsea\_GO/HALL/KEGG\_plot.pdf** – this is the GSEA output summary barplot for the analysis specified by the GO/HALL/KEGG
   3. **Comp\_XXXX\_hypergeo\_GO/HALL/KEGG.csv** – this is the hypergeometric test output table for the analysis specified by the GO/HALL/KEGG
5. **GOI** – stands for Genes of Interest – these are plots highlighting SPECIFICALLY REQUESTED plots showing particular genes or pathways of interest
6. **Ssgsea** – single sample geneset analysis output – with a FOLDER PER ANNOTATION FILE (Kegg, GO, hallmark, etc.)
   1. **XXX\_ssgsea.pdf** – plot showing the ssGSEA for the top differentially expressed pathways output from the BULK GSEA done previously (described above)
   2. **Most everything else in the folder is not important**
   3. **Normcounttab\_v12\_ssGSEA-combined.gct** – is the super table for all of the ssgsea analysis. This will include the columns that show the enrichment scores, pbalues, fdrs, etc..