Data Management Checklist for Bulk RNA-Seq Projects

**** Important note about new data arriving for an existing project***
In some cases, you will receive new sequencing data for a project that already exists. Each new analysis should have its own hbc code and its own directory under the PI name. Often it will be the existing code followed by "_x", where x is the data set number (e.g. hbc03877, hbc03877_1, hbc03877_2, etc.).

☐ Check in with Maria to ensure you have the correct hbc code if you are unsure.

☐ Check you have the right project name (It chould be in Trello checklist, if not ping your manager)

☐ Check you have the right environment to work with. Check Platform page.

Before you run nf-core

☐ Set up your project following Platform best-practices

☐ Set up your RNAseq templates following Platform code

☐ Create a github directory using the project name and add the metadata file for nf-core and a read me

☐ Create a folder in DropBox under the PI name and set up a similar directory structure (clone from github)
***If the PI directory does not exist, create it using all lowercase letters*

Downloading data (if not done by Maria)

☐ Download data to the "data" directory on O2. Check the md5 checksums if available.

☐ Check permissions. They should be set to group readable and writeable. Chmod 775

☐ Make sure your data is transferred to the S3 bucket if necessary. Contact someone from platforms for this.

After nf-core finishes

☐ Make sure that output folder is copied from scratch to /n/data1/cores/bcbio/PIs/…

☐ Add salmon.merged.gene_counts.rds and tximport-counts.csv to a "counts" folder on DropBox.

After QC is done

☐ Create a “QC” folder on DropBox. Add the following files to this folder:

- QC.Rmd file
- Knitted QC.html
- Any params files necessary to run the rmd
- Multiqc_report.html

☐ Add QC.Rmd file and any params files (e.g., params_qc_nf-core.R) to GitHub

After DE is done

☐ Create a “DE” folder on DropBox. Create subdirectories if you have multiple comparisons. Add to these folders:

- DEG.Rmd and DEG.html files. Also add any params files necessary to run the rmd.

☐ Add DEG.rmd and any params files to the github.

☐ Create a “data” folder inside the “DE” folder on DropBox. Add to this folder:

- Normalized counts for all genes x all samples (csv format)

☐ Create a “results” folder inside the “DE” folder on DropBox. Add to this folder:

- Full results for all genes (csv format) with annotation columns appended. ***Separate files will be created for each individual contrast* .
- Significant genes results file (subset of annotated full results by chosen p-value and LFC). *Separate files will be created for each individual contrast*.
- Significant genes results file as described above, but additionally append columns containing normalized count values for each sample.
- Make sure to append the gene symbols to these tables so the researcher can interpret the results.

For multiple comparisons, do a single report. In the results folder, have the comparison names in the files.

After client is satisfied (i.e. analysis phase is finished)

☐ Add any other bash scripts or custom R code generated to GitHub.

☐ Create a README.md defining each custom script in GitHub and its usage.

☐ Any custom R and .Rmd files should also go in DropBox. Make sure all .Rmd code in DropBox is accompanied by the knitted html version if it exists.

☐ Add any custom figures to DropBox and additional result files that were generated as part of the custom analyses.

☐ Write up methods as soon as you know the analysis is complete. Add in methods you wrote, manuscript versions from clients, etc. Keep updating until they submit the manuscript.

During manuscript preparation

☐ Create a "Manuscript" folder on DropBox.

Returning raw and processed data to clients

☐ Review data and delete duplicate data or analyses that were incorrect.

☐ Make sure the client has their raw data files (fastq) and raw counts. We can also share bam files upon request.

☐ Set up a Globus share on scratch that symlinks to the right directories.

☐ Confirm with Shannan that the project is complete.

☐ Delete the data after confirming the client has all their data.