# Data Management Checklist for 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 should be in the Trello checklist, if not ping your manager)

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

## Before you run analysis

□ Set up your project following Platform best-practices

□ Previous step helps you to set up the github repository 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.

## Before sharing with client

□ Make sure the report and data files (tables/figures) you are sharing can be reproduced in a clean environment:

- Remove all objects from your R/python environment
- Restart the session
- Run your Rmd or code to reproduce the reports/tables/figures you are sharing with clients
- This should run at once w/o errors

  - Use R/python objects to skip compute/memory intensive processes
  - Use qs to store and save objects (faster than RD)
- DO NOT share reports/tables/figures if the code is not independent of manual steps

□ 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 draft 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.

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

□ Finalize methods as soon as you know the client is ready to share results with collaborators or for paper publication. 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.