**MCINTYRE LAB DOCUMENTATION PROTOCOL (last updated Feb 2018)**

This details the layout we use for our documentation system. Here we use a spreadsheet (Excel, LibreOffice Calc, etc.) to document all the steps used in the analysis of a given project, what each step is doing, where it was performed, the inputs and outputs, and any notes we want to include that pertain to each step.

A very good example of our documentation system can be found in the cegs\_sm\_sd\_paper project folder on the share.

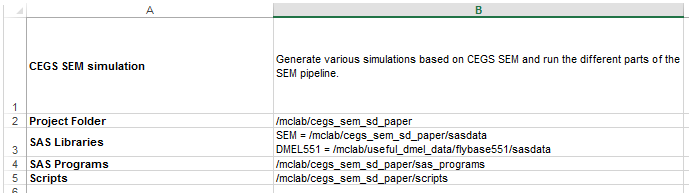
1. **Documentation folder**

In each project folder should be folder called “documentation”. This folder contains the spreadsheets used to document all the steps taken in the analysis of a project, as well as other files that document some other aspect of the project. For example, this might contain information about how sequencing plates were set up, JuPyteR notebooks documenting specific decisions made in an analysis, and any other pieces of information that document part of the experiment.

Each distinct piece of a project should have its own documentation spreadsheet. For example, for RNA-seq differential expression analysis there will be separate spreadsheets for QC, alignments and analysis. This distinction will be dependent on the nature of the project.

1. **Headers**

The first few lines of the spreadsheet should provide detail on what is documented in the spreadsheet, the project folder, location of scripts, SAS programs and libraries used. For example:



First line contains the title of document (“CEGS SEM simulation”) and a description of what is being documented in the file.

**Project folder**: this contains the path to location of the project folder. This should point towards the location on the share drive, as well as its location on HPC (if applicable)

**SAS Libraries** (if applicable): This should consist of the library name (8 alphanumeric characters) as well as the path to its location (share and HPC, if applicable)

**SAS Programs** (if applicable): this should point towards the path to the folder containing SAS programs (share and HPC, if applicable)

**Scripts**: this should point towards the path to the folder containing shell scripts, R and python programs, HPC submission scripts, etc. (share and HPC, if applicable).

1. **Columns in the documentation spreadsheet**

Each documented step consists of several parts, detailing what the step is, briefly what is being done (and why), where the step was located, programs and scripts used, the inputs and outputs, what flags or variables were created, and any results or notes pertaining to that step.

Step: This is a brief “title” of the step and should convey what is being done at this step in a few words. For example, if this step concerns alignments of RNAseq reads to a genomic reference sequence, then this step might be called “Align to genome”

Location: This is the location of where the step was performed. Values here may include:

HPC: Step was performed on Hipergator

Share: Step was performed on the share drive

Local: Step was performed locally (large datasets)

SAS: Step was performed in SAS

Brief description: This should be a brief description (no more than 2-3 sentences) detailing what is being done at the step, and if applicable, why.

Program: This can contain any programs (R code, python code, SAS programs, binaries, etc.) used at this step. If applicable, include the version of the program used.

For example, if I am using a python program to parse alignments and I am doing this in Python 2.7.6, I would include the following:

“python/2.7.6

parse\_alignments.py”

Submission script: This should contain the SBATCH submission script (for HPC), shell script (for share/local) or SAS makefile used. Examples:

If using a SAS makefile to run several SAS programs sequentially: makefile\_my\_sas\_analysis.sas

If using a bash/shell script to run a program locally/on the share: run\_my\_program.sh

Note: Outside of some specific examples (e.g. SAS programs), every step should have a submission or shell script of some kind, as **we do not run open code or commands directly on the command line**. This ensures that anything that is run can be re-run with the exact same set of parameters, inputs, outputs, etc.

Input files: This should contain the input files or datasets used in this step. For SAS programs this will be in the format “library\_name.dataset\_name” and not the path. For all other inputs (CSVs, SAM/BAM files, etc.), include the path name.

Output files: Similar to input files, this should contain the output files or datasets used in this step. For SAS programs this will be in the format “library\_name.dataset\_name” and not the path. For all other output files (CSVs, SAM/BAM files, etc.), include the path name.

Created flags/variable names: This should list any variables or flags that were created in this step. For example, if I am deciding (flagging) genes that are considered “expressed”/”on” at a level of APN>0, then I would list here the variable name created that flags genes considered “on”, e.g “flag\_fusion\_apn\_gt0”.

Logs: Path to any log files created during this step.

Results and notes: Brief results output (e.g. counts summary, ODS output table from SAS) or brief note pertaining to the analysis step should be included here. This might be a note about some part of the analysis (e.g. “output file is 20GB and will not be needed after the next step, so I have saved this locally” or “Looking at the distribution, to achieve a 5% TIER, I need a BIC difference of 12.44, ie the diff value where cumulative frequency of 95%”), or a brief summary of some results (“15670 BLAST hits imported, 1889 hits passed my filtering criteria”). This is important if your code is being checked, as then whoever is rerunning your code knows what to expect.