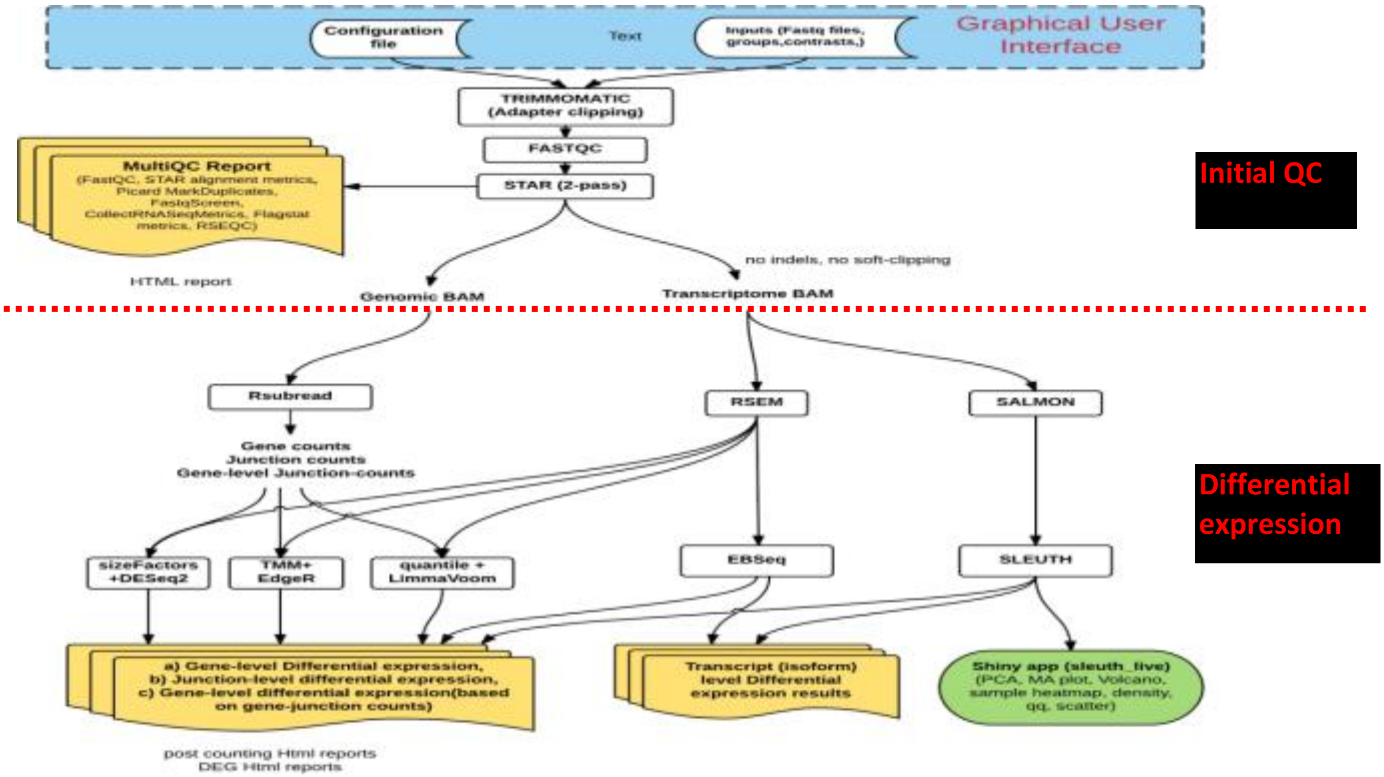
Specific Instructions for executing the RNASeq pipeline

Introduction

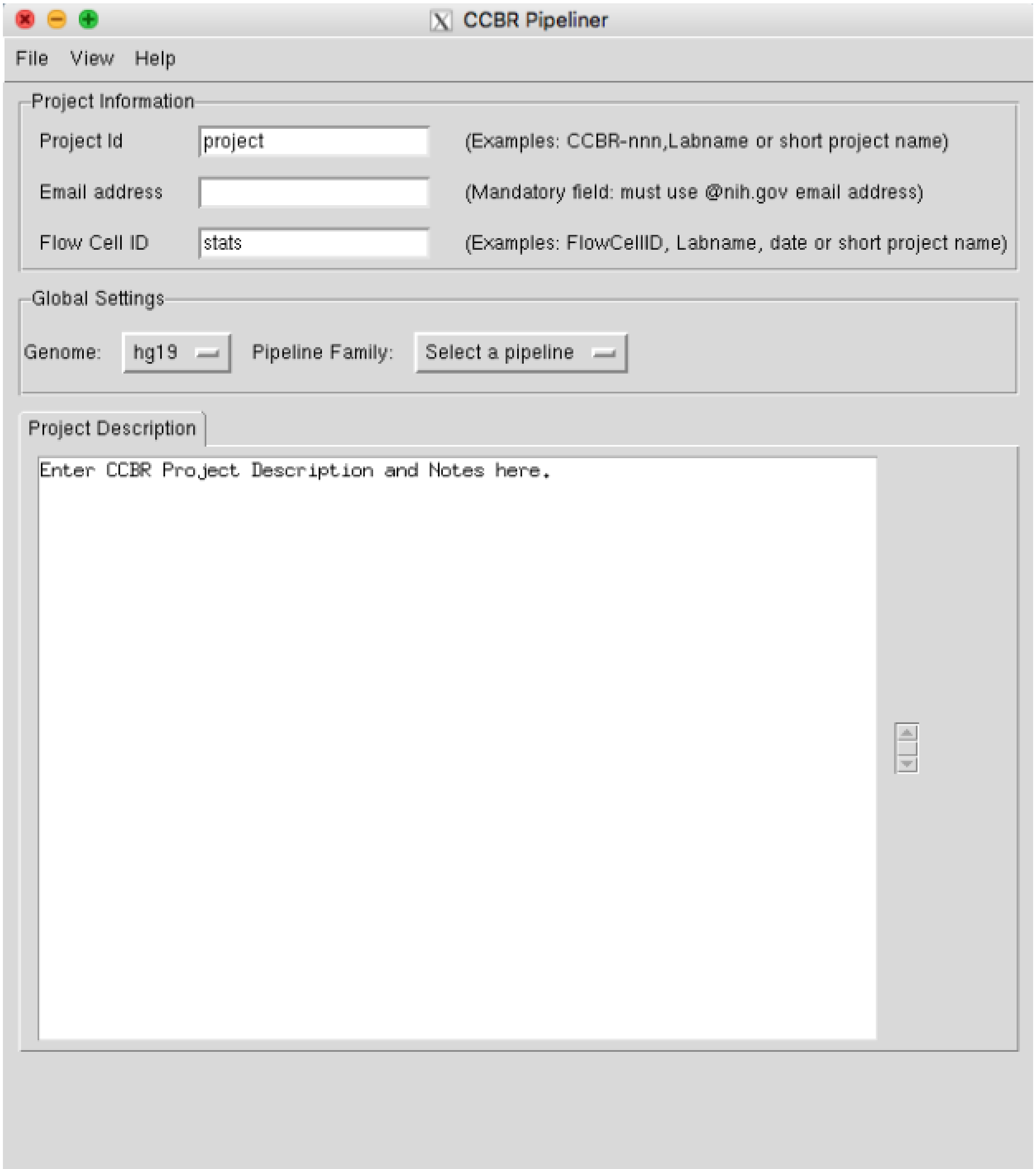
The RNASeq pipeline has been implemented to quantify gene & isoform expression and carry out tests for differential expression at both the gene- and isoform- level. The pipeline also generates results for differential expression across splice-junctions as well as differential expression at the gene-level, counting only junction-spanning reads.


Configuration file
Text
Inputs (Fastq files, groups,contrasts,)
Graphical User Interface
TRIMMOMATIC (Adapter clipping)
FASTQC
STAR (2-pass)
MultiQC Report (FastQC, STAR alignment metrics, Picard MarkDuplicates, FastqScreen, CollectRNASeqMetrics, Flagstat metrics, RSEQC)
HTML report
no indels, no soft-clipping
Genomic BAM
Transcriptome BAM
Initial QC
Rsubread
RSEM
SALMON
Gene counts
Junction counts
Gene-level Junction-counts
Differential expression
sizeFactors +DESeq2
TMM+ EdgeR
quantile + LimmaVoom
EBSeq
SLEUTH
a) Gene-level Differential expression, b) Junction-level differential expression, c) Gene-level differential expression(based on gene-junction counts)
Transcript (isoform) level Differential expression results
Shiny app (sleuth_live) (PCA, MA plot, Volcano, sample heatmap, density, qq, scatter)
post counting Html reports
DEG Html reports

The RNASeq pipeline is run in two phases. First all reads must pass through the initalQC phase of adapter-clipping, read-level QC, read-alignment and alignment-level. Second genes and transcripts counts and differential expression analysis are generated.

Running the RNASeq pipeline

Please refer to the CCBR Pipeliner overview documention for specific instructions to launch the CCBR pipeliner. Here we assume you succeed to launch the GUI as shown below



Project Information

Project Id: project (Examples: CCBR-nnn,Labname or short project name)

Email address: (Mandatory field: must use @nih.gov email address)

Flow Cell ID: stats (Examples: FlowCellID, Labname, date or short project name)

Global Settings

Genome: hg19 Pipeline Family: Select a pipeline

Project Description

Enter CCBR Project Description and Notes here.

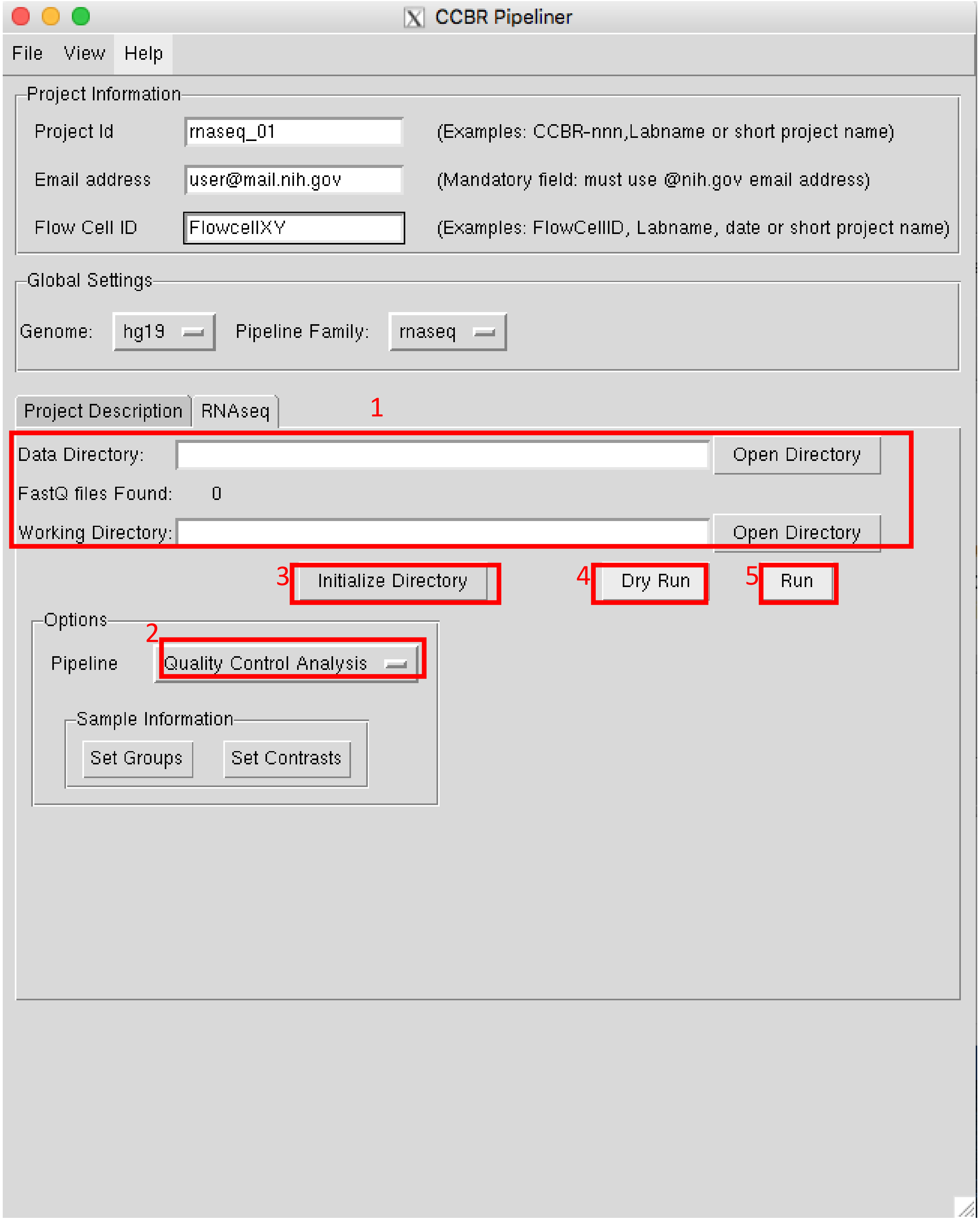
Phase1: Initial QC

After entering information on the 'Project Information and 'Global Options' tab, select the 'rnaseq' pipeline from the pipeline family list box and set the RNASeq Pipeline Options as follows:

1) First, select or type the full path for the data and working directory. The working directory name **should be new** and the **system will create it automatically**
2) Select the sub-pipeline "**Quality Control Analysis**" and ignore the sample information that are not used for the Initial QC phase.
3) Click on button "initialize Directory" to initialize the working directory
4) Click on button "Dry run" to check the initialqc phase tasks and settings.
5) **If there is no error message in the "Dry run"**, then click on the button "Run" to run this pipeline QC phase on Biowulf cluster.

NOTE: Once the pipeline job is submitted to Biowulf cluster, you will be notify by email and you can close the GUI. On the command prompt, you can check the status of your jobs by using this command:

- sjobs (or squeue –u <username>)



CCBR Pipeliner

Project Information

Project Id: rnaseq_01 (Examples: CCBR-nnn,Labname or short project name)

Email address: user@mail.nih.gov (Mandatory field: must use @nih.gov email address)

Flow Cell ID: FlowcellXY (Examples: FlowCellID, Labname, date or short project name)

Global Settings

Genome: hg19

Pipeline Family: rnaseq

Data Directory:

FastQ files Found: 0

Working Directory:

Options

Pipeline: Quality Control Analysis

Results from Initial QC

Once the pipeline completes successfully, the working directories will contain the following folders/files:

- **QC folder**: contains Fastqc results on the adapter-clipped fastq files
- **trim folder**: contains temporary adapter-clipped fastq files (both paired and unpaired)
- **STAR 1st pass output files**:
  - *.Aligned.out.bam
  - *.SJ.out.tab
  - *.Log.progress.out
  - *.Log.final.out
- **STAR 2nd pass output files:**
  - *.p2.Aligned.toTranscriptome.out.bam
  - *.p2.ReadsPerGene.out.tab
  - *.p2.Aligned.sortedByCoord.out.bam
  - *.p2.SJ.out.tab
  - *.p2.Log.progress.out
  - *.p2.Log.final.out
  - *.star_rg_added.sorted.bam
  - *.star_rg_added.sorted.dmark.bam
  - *.star_rg_added.sorted.dmark.bai

- **Alignment QC files:**

  a) *.RnaSeqMetrics.txt: Output from Picard CollectRNASeq Metrics

  b) *.flagstat.concord.txt: Output from samtools flagstat

  c) RSEQC output:
  - *.inner_distance_plot.pdf
  - *.inner_distance_plot.r
  - *.inner_distance_freq.txt
  - *.inner_distance.txt
  - *.GC.xls
  - *.GC_plot.pdf
  - *.strand.info
  - *.Rdist.info

  d) ProjectID_FlowCellID_Summary.txt : comprehensive Alignment QC metrics from RSEQC

  e) ProjectID_FlowCellID.xlsx: QC-metrics and barcharts in excel format

f) FQscreen folder: contains results from Fastq-Screen

- **Reports folder:**
  a) aggregate_fastqc_report.html: quick view of PASS/WARN/FAIL flags using color codes, across all samples and all metrics
  b) **multiqc_report.html**: comprehensive and interactive HTML page aggregating QC metrics across several tools

Phase2: Differential expression workflow (after Initial QC)

This workflow is to be executed in the same working directory as that specified in the InitialQC phase, so that the pipeline can identify the output files from the InitialQC run, that are being used as input files to the differential expression workflow.

After entering information on the 'Project Information' and 'Global Options' tab, select the 'Differential Expression Analysis' pipeline from the pipeline family list box and set the RNASeq Pipeline Options as follows:

1) First, select or type the full path for the data and working directory. The working directory should the same one you specified in the initial QC phase
2) Select the sub-pipeline "**Differential Expression Analysis**"
3) choose option "yes, Report Differentially Expressed genes" if you have a set contrasts or choose the default "no, Do not Report Differentially Expressed genes" if only interested in genes or trancripts counts
4) set your criteria for gene filtering: minimum reads counts shared by at least a number of samples. The minimum read count is used to compute a minimum count per millon (cpm) threshold based on the maximum library size. This filtering removes genes that are not expressed in any sample (all counts are zeros), as well as filters out genes with very low cpm below the computed cpm threshold.

5) click on the button "Set Groups" to define the list of samples to include in this phase with the the group (condition) and label (for plots) information. You can type all information needed. After filling in the information, make sure to click on 'Save' to save this information (a file groups.tab will be created). You can also load (read) the data from a predefined file "groups.tab". **No field name is required**.
Here one example:

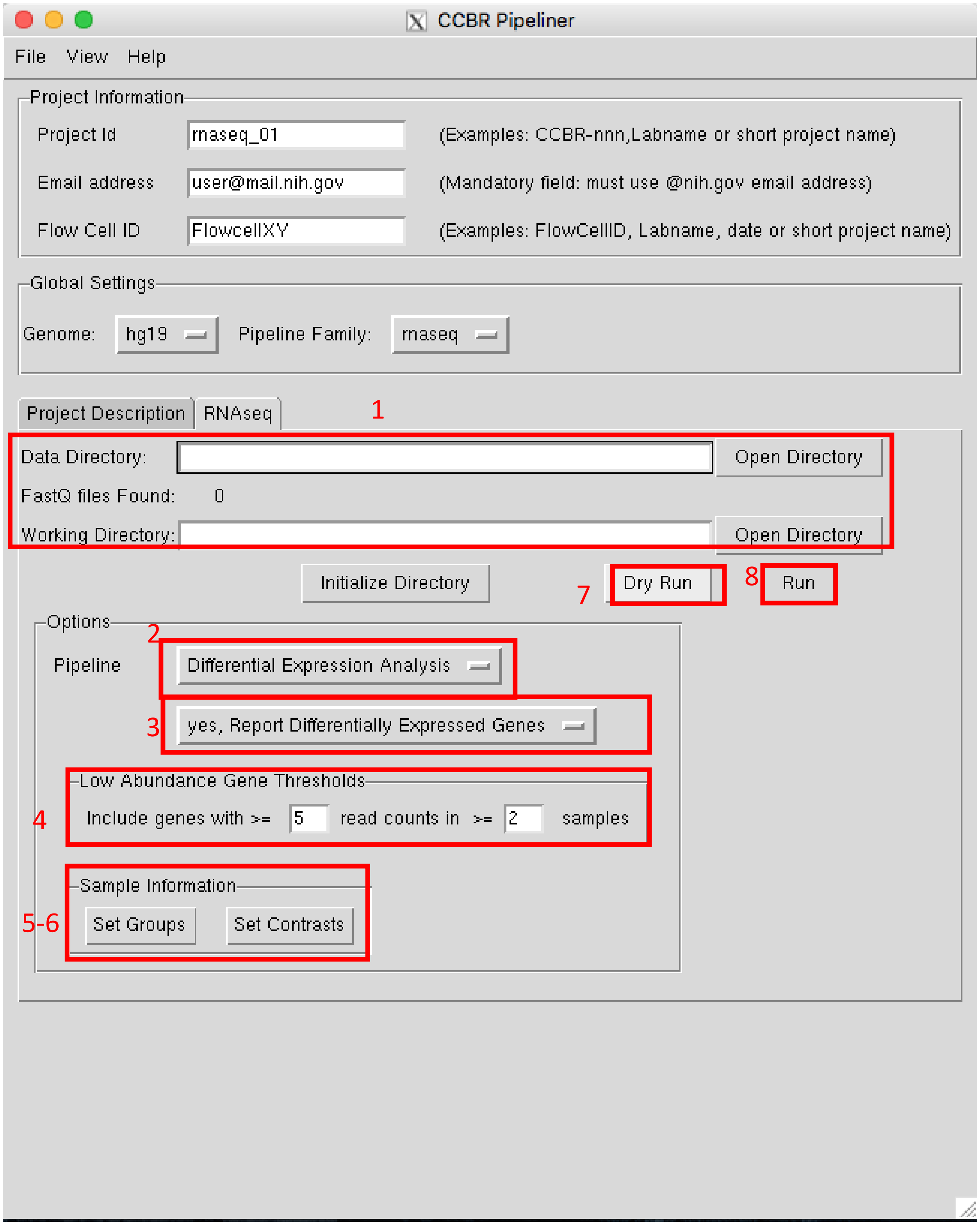
<sample>	<condition>	<label>
SS.SRR950078	G1	s78
SS.SRR950079	G1	s79

SS.SRR950080	G2	s80
SS.SRR950081	G2	s81

6) click on the button “Set Contrasts” if only you have to report Report differentially expressed genes. You need to enter the groups to compare (contrasts). You can type a contrast per line where the first group (group1) will be compared to the second group (group2 as control). Make sure to save this information (a file contrasts.tab will be created). You can also load (read) the data from a predefined file “contrasts.tab”. No field name is required.
Here one example:
G2 G1

7)Click on button “Dry run” to check this phase tasks and settings

8) **If there is no error message in the “Dry run”**, then click on the button “Run” to run this pipeline differential expression analysis phase on Biowulf cluster.



Project Information

Project Id: rnaseq_01 (Examples: CCBR-nnn,Labname or short project name)

Email address: user@mail.nih.gov (Mandatory field: must use @nih.gov email address)

Flow Cell ID: FlowcellXY (Examples: FlowCellID, Labname, date or short project name)

Global Settings

Genome: hg19 Pipeline Family: rnaseq

Data Directory:

FastQ files Found: 0

Working Directory:

Options

Pipeline: Differential Expression Analysis

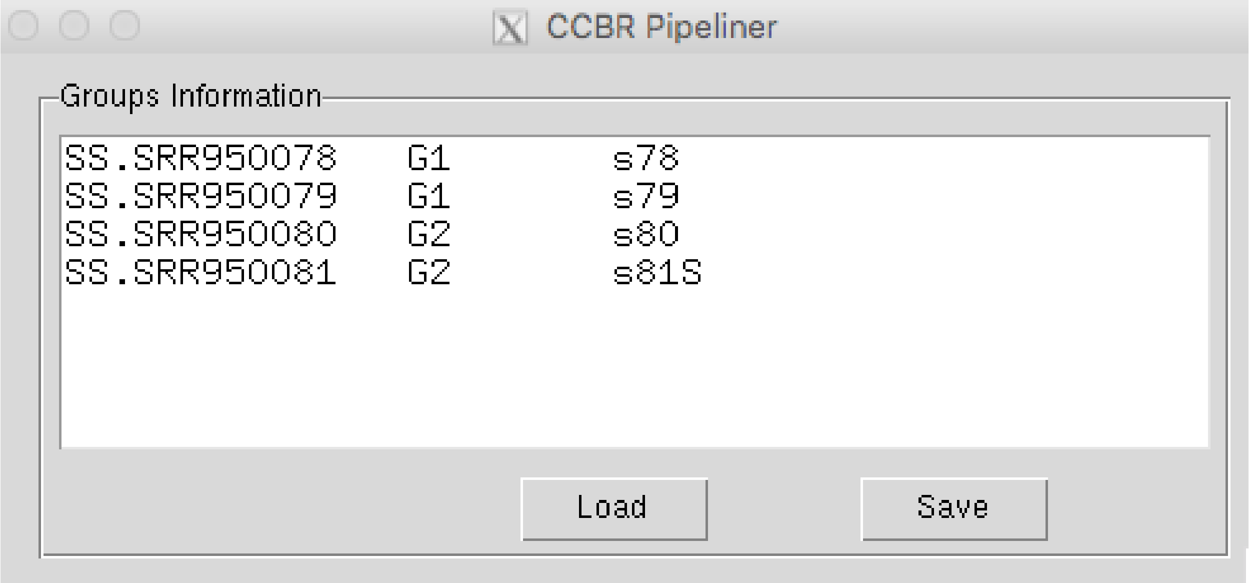
yes, Report Differentially Expressed Genes

Low Abundance Gene Thresholds

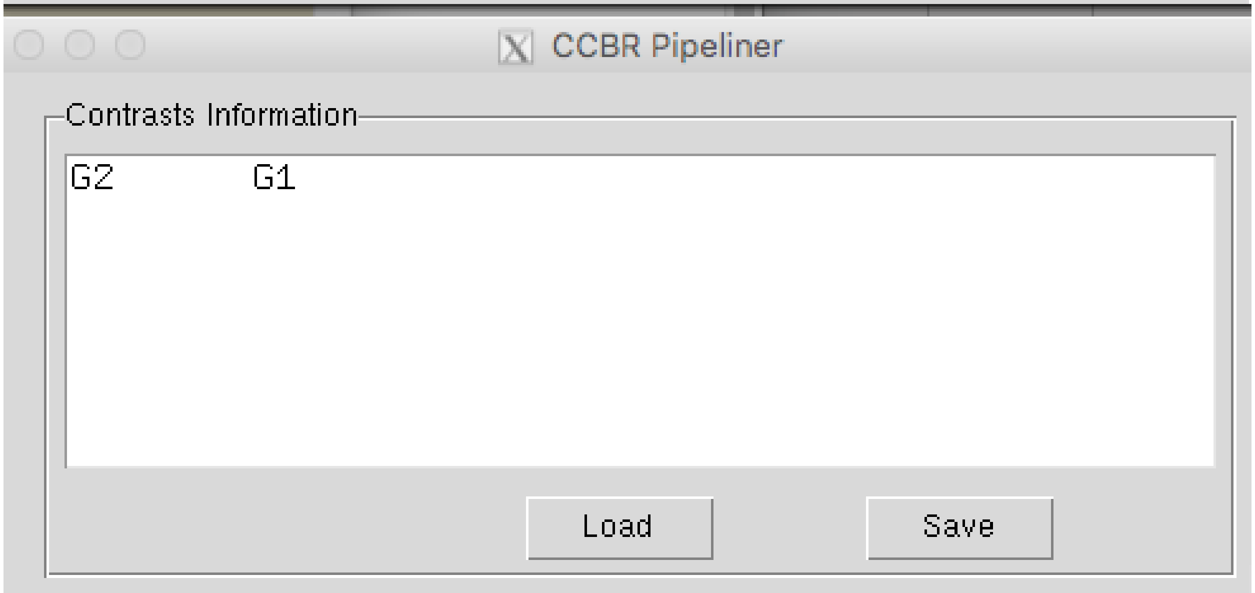
Include genes with >= 5 read counts in >= 2 samples

Sample Information

Here is a snapshot of how the “Groups” and “Contrasts” text-box would look after filling in the information:



```
SS.SRR950078    G1          s78
SS.SRR950079    G1          s79
SS.SRR950080    G2          s80
SS.SRR950081    G2          s81S
```



```
G2        G1
```

Results from Differential expression analysis

Once the pipeline completes successfully, the working directory should contain the following files/folders:

- **Folder DEG_genes**: contains results of differential expression at the gene-level based on FeatureCounts method raw counts (Subread) from three DE algorithms: Limma, DESeq2 and EdgeR

- **Folder DEG_rsemgenes**: contains results of differential expression at the gene-level based on RSEM raw counts from three DE algorithms: Limma, DESeq2 and EdgeR

- **Folder DEG_geneJunctions**: contains results of differential expression at the gene level, counting only exon-exon junction-spanning reads (useful if the library is total RNA, where we see a large proportion of intronic reads, originating from unspliced RNA). All 3 DE methods are applied to find differentially expressed genes.

- **Folder DEG_junctions**: contains results of differential expression at the junction-level.Here we are simply quantifying reads across each known splice-junction in the chosen genome. So the output contains read counts for each exon-exon junction in the genome (not collapsing by gene). Then all 3 methods are applied to find differentially expressed junctions.

- **RSEM results**:
  - *.rsem.genes.results: gene-level counts from RSEM
  - *.rsem.isoforms.results: isoform-level counts from RSEM
- **EBSeq results**:
  - *.ebseq: differential expression results from EBSeq at the isoform- level (from RSEM counts)

- **Folder salmonrun**: Contains one folder for each sample, containing salmon quantification files as well as differential expression results from Sleuth at the isoform- level