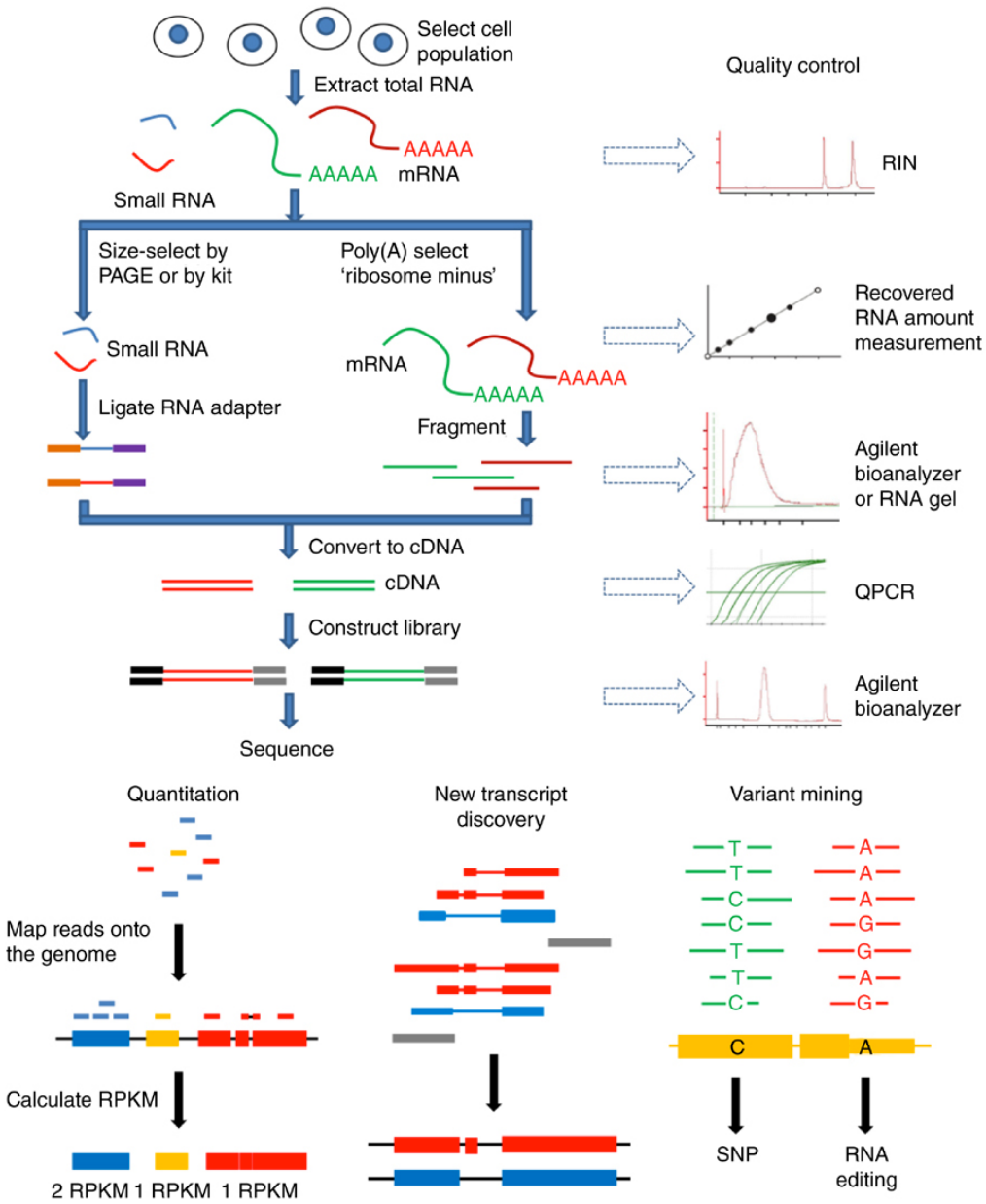
Introduction to RNASeq Analysis with BICF's Astrocyte Workflow

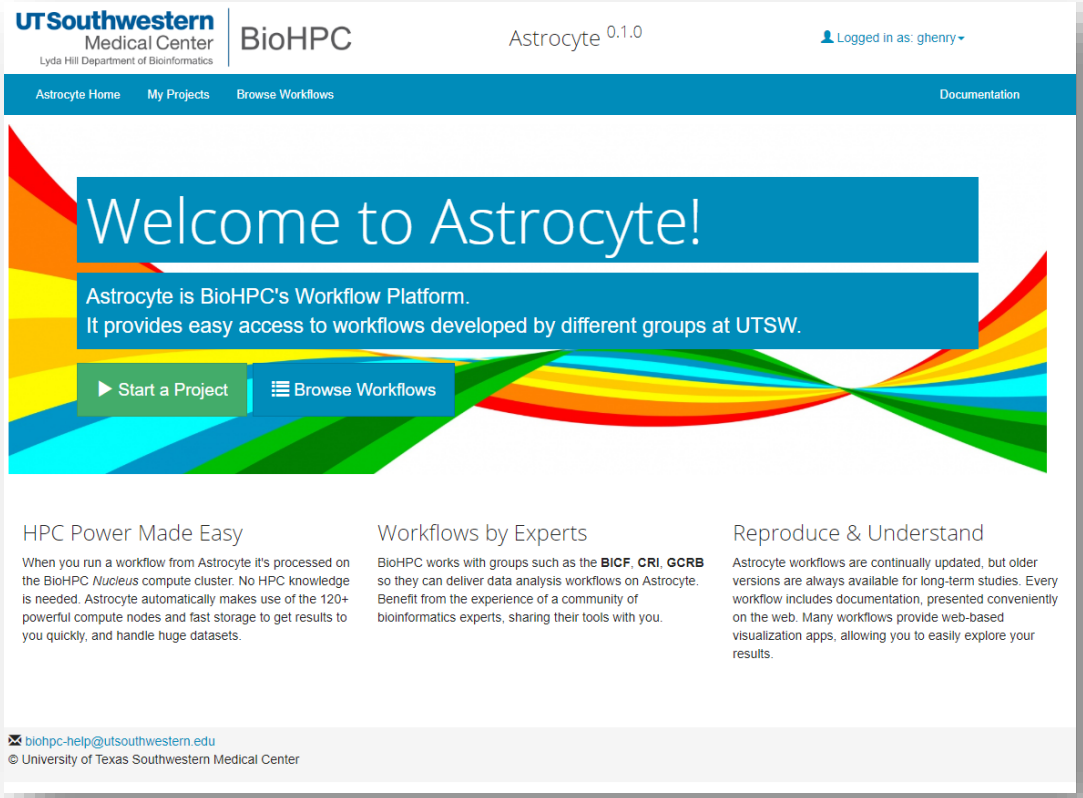
Gervaise H. Henry
Department of Urology
(BICF Fellow)


Select cell population
Extract total RNA
AAAAA
AAAAA
mRNA
Small RNA
Size-select by PAGE or by kit
Poly(A) select 'ribosome minus'
Small RNA
mRNA
AAAAA
AAAAA
Ligate RNA adapter
Fragment
Convert to cDNA
cDNA
Construct library
Sequence
Quality control
RIN
Recovered RNA amount measurement
Agilent bioanalyzer or RNA gel
QPCR
Agilent bioanalyzer
Quantitation
Map reads onto the genome
Calculate RPKM
2 RPKM 1 RPKM 1 RPKM
New transcript discovery
Variant mining
T
T
C
C
T
T
C
A
A
A
G
G
A
G
C
A
SNP
RNA editing

BioHPC, BICF and Astrocyte

astrocyte.biohpc.swmed.edu

- Allows groups to give easy-access to their analysis pipelines via the web
- Standardized Workflows
- Simple Web Forms
- Online documentation & results visualization



Available Workflows

UTSouthwestern Medical Center | BioHPC

Astrocyte 0.1.0

Logged in as: ghenry

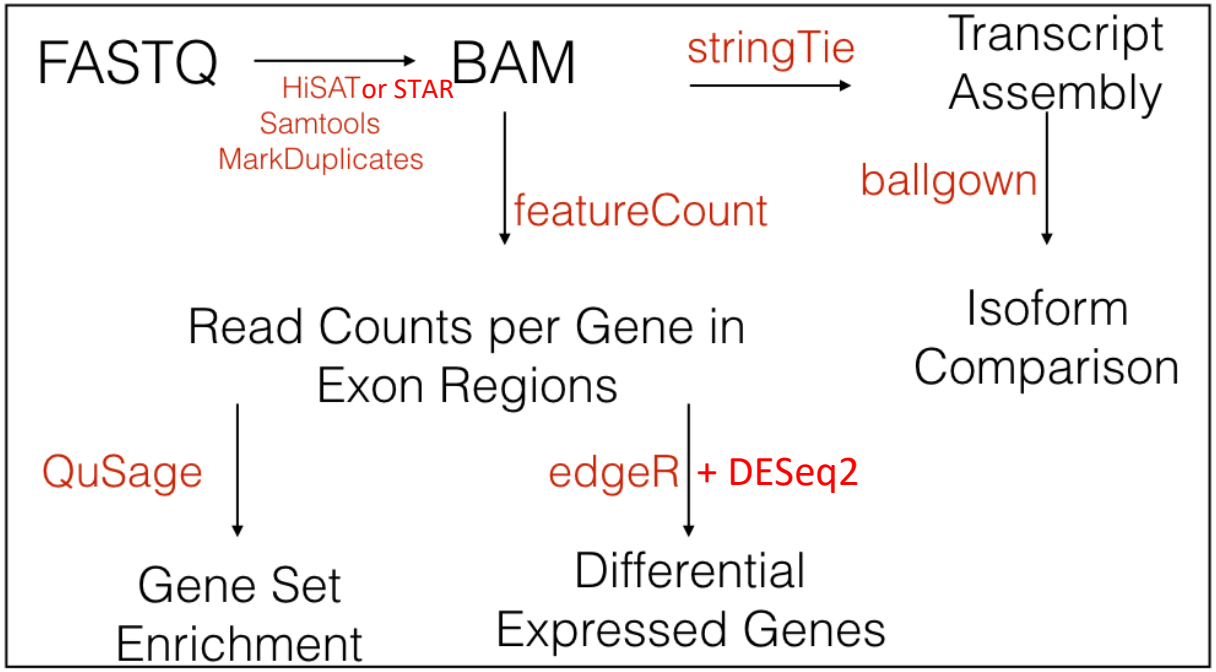
Astrocyte Home | My Projects | Browse Workflows | Documentation

Available Workflows

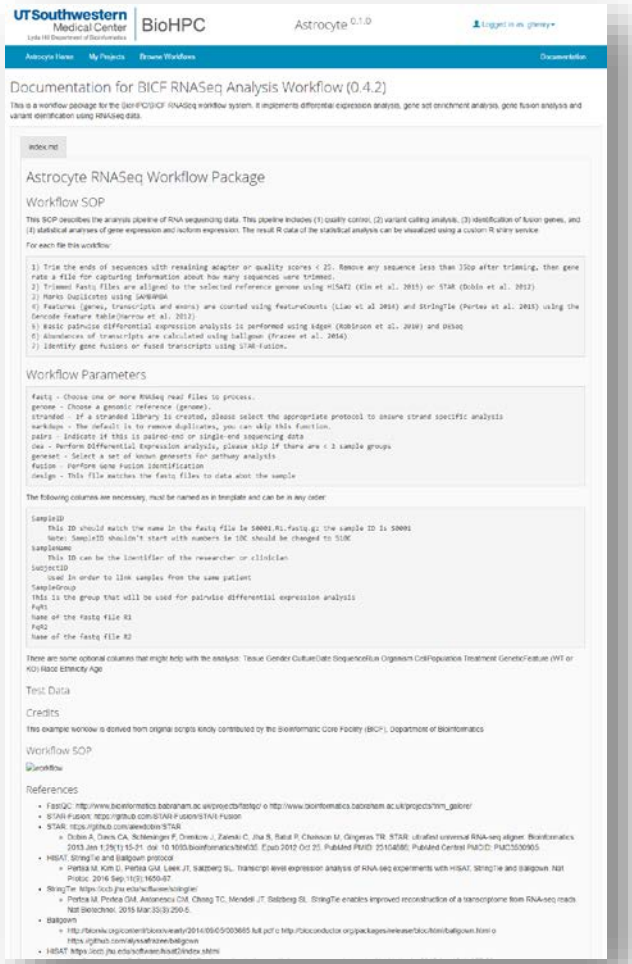
UTSouthwestern Medical Center BICF	**BICF ChIP-seq Analysis Workflow** This is a workflow package for the BioHPC/BICF ChIP-seq workflow system. It implements a simple ChIP-seq analysis workflow using deepTools, Diffbind, ChipSeeker and MEME-ChIP, visualization application.	**Current Version:** chipseq_analysis_bicf - 0.0.12 **Author:** Beibei Chen **Contact:** biohpc-help@utsouthwestern.edu	Run Workflow Documentation All Versions
UTSouthwestern Medical Center BICF	**BICF 16S rRNA Microbiome Analysis Workflow** This is the BioHPC/BICF 16S rRNA Sequence Analysis for Microbiome. It implements the mothur miSeq SOP see the documentation.	**Current Version:** rna_16Sanalysis - 0.1.0 **Author:** Brandi Cantarel **Contact:** bicf@utsouthwestern.edu	Run Workflow Documentation All Versions
RESEARCH INSTITUTE	**Astrocyte Example ChIPSeq Workflow** This is an example workflow package for the BioHPC astrocyte workflow system. It implements a simple ChIPSeq analysis workflow using BWA and MACS, plus a simple R Shiny visualization application.	**Current Version:** astrocyte_example - 0.0.5 **Author:** David Trudgian **Contact:** biohpc-help@utsouthwestern.edu	Run Workflow Documentation All Versions
UTSouthwestern Medical Center BioHPC	**Example Wordcount Workflow** This is a minimal test workflow package that counts the occurances of words in a text file. It can be used as a template to develop workflows, and as to test the astrocyte platform.	**Current Version:** example_wordcount - 0.0.4 **Author:** David Trudgian **Contact:** biohpc-help@utsouthwestern.edu	Run Workflow Documentation All Versions
UTSouthwestern Medical Center BioHPC	**CompOmics Protein ID Workflow** This workflow performs protein identification using 4 search engines, the SearchGUI and PeptideShaker tools from the CompOmics group.	**Current Version:** compomics_proteinid - 0.0.2 **Author:** David Trudgian **Contact:** biohpc-help@utsouthwestern.edu	Run Workflow Documentation All Versions
RESEARCH INSTITUTE	**CRI flowcytometry workflow** This workflow implements R package cytofkit pipeline to analyse	data generated by flowcytometry	**Current Version:** Flowcytometry Workflow - 0.1.0 **Author:** Long Lu **Contact:** long.lu@utsouthwestern.edu	Run Workflow Documentation All Versions
UTSouthwestern Medical Center BICF	**BICF RNASeq Analysis Workflow** This is a workflow package for the BioHPC/BICF RNASeq workflow system. It implements differential expression analysis, gene set enrichment analysis, gene fusion analysis and variant identification using RNASeq data.	**Current Version:** maseq_bicf - 0.4.2 **Author:** Brandi Cantarel **Contact:** biohpc-help@utsouthwestern.edu	Run Workflow Documentation All Versions
UTSouthwestern Medical Center BICF	**BICF RNASeq Variant Analysis Workflow** THIS WORKFLOW IS OBSOLETE! The Main BICF workflow includes variant analysis and differential expression analysis as one easy to use workflow.	**Current Version:** maseq_variant_bicf - 0.0.11 **Author:** Brandi Cantarel **Contact:** biohpc-help@utsouthwestern.edu	Run Workflow Documentation All Versions
UTSouthwestern Medical Center BICF	**BICF Somatic Mutation Calling** This is a workflow package for the BioHPC/BICF Somatic Mutation workflow system. It implements a simple Somatic Mutation analysis workflow.	**Current Version:** somatic_bicf - 0.0.3 **Author:** Brandi Cantarel **Contact:** biohpc-help@utsouthwestern.edu	Run Workflow Documentation All Versions
UTSouthwestern Medical Center BICF	**BICF Germline Variant Analysis Workflow** This is a workflow package for the BioHPC/BICF Germline Variant workflow system. It implements a simple germline variant analysis workflow using Trimmomatic, BWA, Speedseq, GATK, Samtools and Platypus. SNPs and indels are integrated using BAYSIC, then annotated using SNPEFF and SnpSift.	**Current Version:** germline_bicf - 0.2.2 **Author:** Brandi Cantarel **Contact:** biohpc-help@utsouthwestern.edu	Run Workflow Documentation All Versions
RESEARCH INSTITUTE	**Astrocyte CRI ChIPSeq Workflow** This is an CRI chipseq workflow package for the BioHPC astrocyte workflow system. It implements a simple ChIPSeq analysis workflow.	**Current Version:** CRI_chipsec - 0.1.3 **Author:** CRI **Contact:** biohpc-help@utsouthwestern.edu	Run Workflow Documentation All Versions
UTSouthwestern Medical Center	**Astrocyte GCRB ChIPSeq Workflow** This is an GCRB chipseq workflow package for the BioHPC astrocyte workflow system. It implements a simple ChIPSeq analysis workflow.	**Current Version:** gcrb_chipseq - 0.1.3 **Author:** GCRB **Contact:** biohpc-help@utsouthwestern.edu	Run Workflow Documentation All Versions
RESEARCH INSTITUTE	**Normalized data profiling and visualization** This workflow uses R and Cluster to visualize the profile of a tabular dataset such as RNA-seq data, microarray data, proteomic data, and so on. We assume that the data have been normalized. If not, the workflow can be used to visualize some features of the original data but functions such as clustering and heatmap might be much less meaningful.	**Current Version:** Normalized_data_profiling - 0.0.51 **Author:** Liang Shi **Contact:** liang.shi@utsouthwestern.edu	Run Workflow Documentation All Versions
	aMAP Pipeline execution Workflow This workflow automates various steps involved in image registration and segmentation process. The input is stitched tif image files from tissuecyte microscopes. The workflow performs the task of converting tif to nifty files and then mapping with the reference atlas to perform nifty registration and segmentation. The output nii files are converted back to tif stack.	**Current Version:** aMAP_Workflow - 0.0.5 **Author:** Apoorva Ajay **Contact:** Apoorva.Ajay@utsouthwestern.edu	Run Workflow Documentation All Versions

biohpc-help@utsouthwestern.edu

BICF RNA-Seq Analysis Workflow



Using BICF RNA-Seq Analysis Workflow on Astrocyte



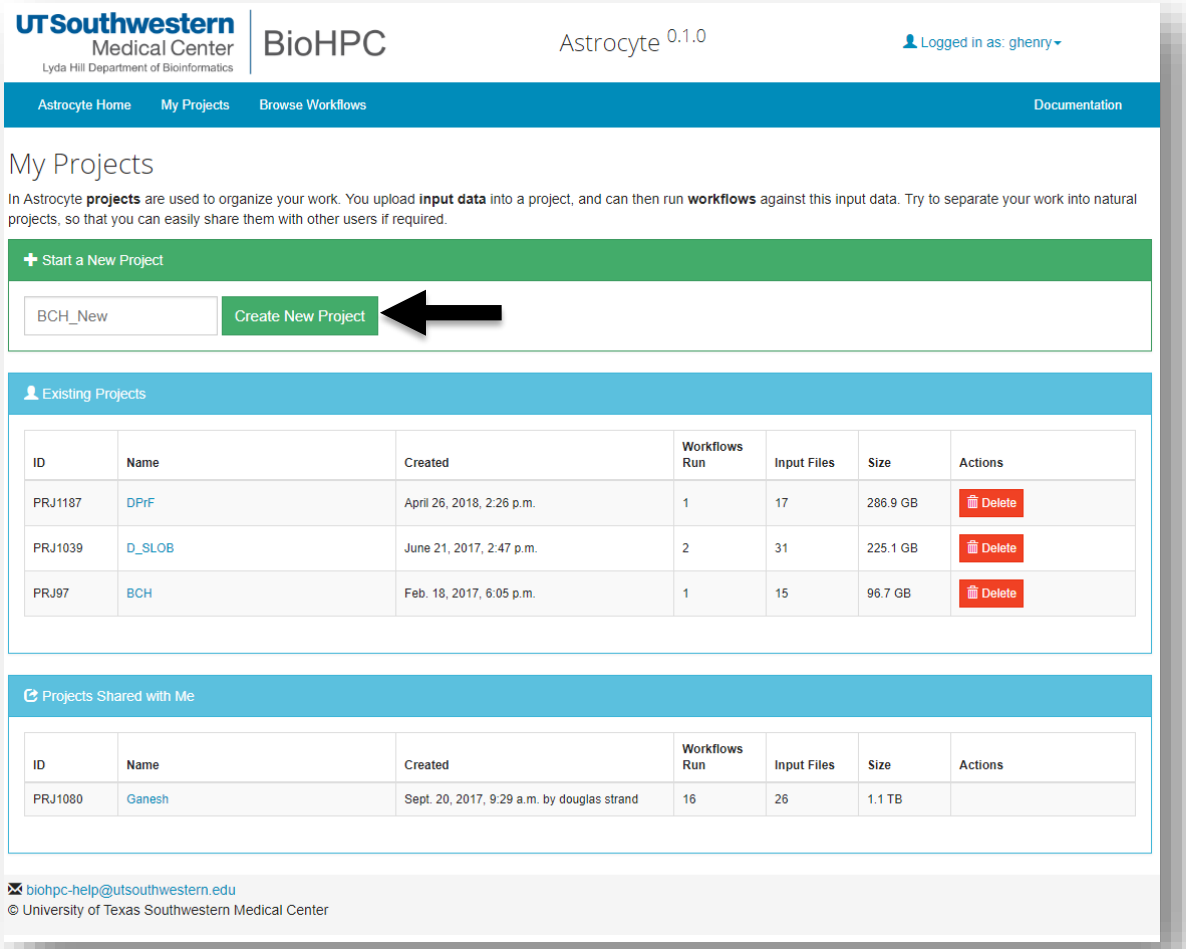
The following columns are necessary, must be named as in template and can be in any order:

```
SampleID
    This ID should match the name in the fastq file ie S0001.R1.fastq.gz the sample ID is S0001
    Note: SampleID shouldn't start with numbers ie 10C should be changed to S10C
SampleName
    This ID can be the identifier of the researcher or clinician
SubjectID
    Used in order to link samples from the same patient
SampleGroup
This is the group that will be used for pairwise differential expression analysis
FqR1
Name of the fastq file R1
FqR2
Name of the fastq file R2
```

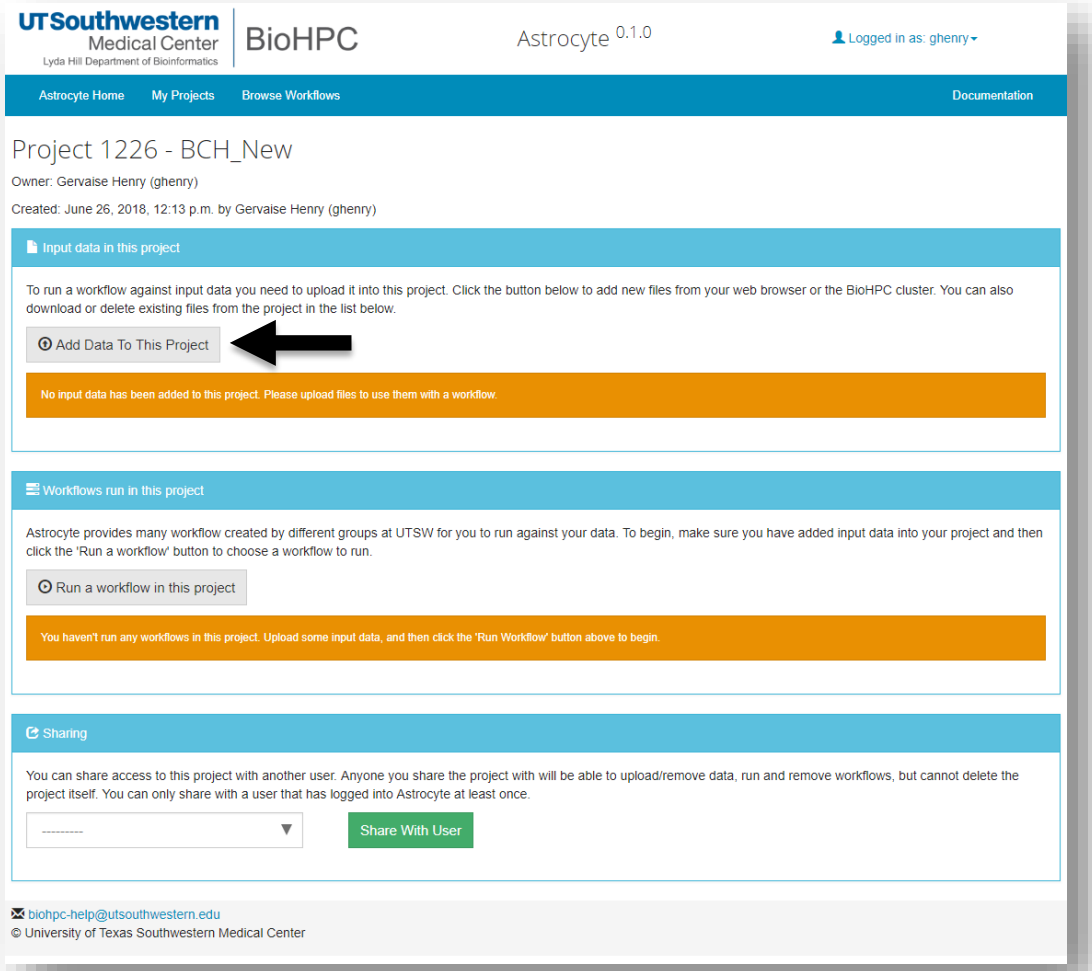
SampleID	SampleName	SubjectID	SampleGroup	FqR1	FqR2
S110BE	HS420	110BE	BCH	S110B.fastq.gz	
S117BE	HS420	117BE	BCH	S117B.fastq.gz	
S132INBE	HS420	132INBE	BCH	S132INB.fastq.gz	
S085BE	HS420	085BE	NoBCH	S085B.fastq.gz	
S089BE	HS420	089BE	NoBCH	S089B.fastq.gz	
S114BE	HS420	114BE	NoBCH	S114B.fastq.gz	
S118BE	HS420	118BE	NoBCH	S118B.fastq.gz	

- Use tab as delimiter
  Excel save as “Text (tab delimited)”
- If no SubjectID, use same number/character for all rows
- SampleID and SampleName
- If no FqR2, leave them empty
- For all contents, no “-”
- For all contents, no spaces
- Columns names MUST be exactly the same as documented

Using BICF RNA-Seq Analysis Workflow on Astrocyte



Using BICF RNA-Seq Analysis Workflow on Astrocyte



Using BICF RNA-Seq Analysis Workflow on Astrocyte

Upload files from the web

You can upload any size file via your browser, but large files may take a long time to complete. Do not navigate away from this page before an upload is complete.

Select file to upload... | Finished uploading files

Upload Progress

Select a file to upload

Import from incoming directory

Copy your files into **/project/apps/astrocyte/astrocyte_incoming/ghenry** on BioHPC to import them into your project directly.

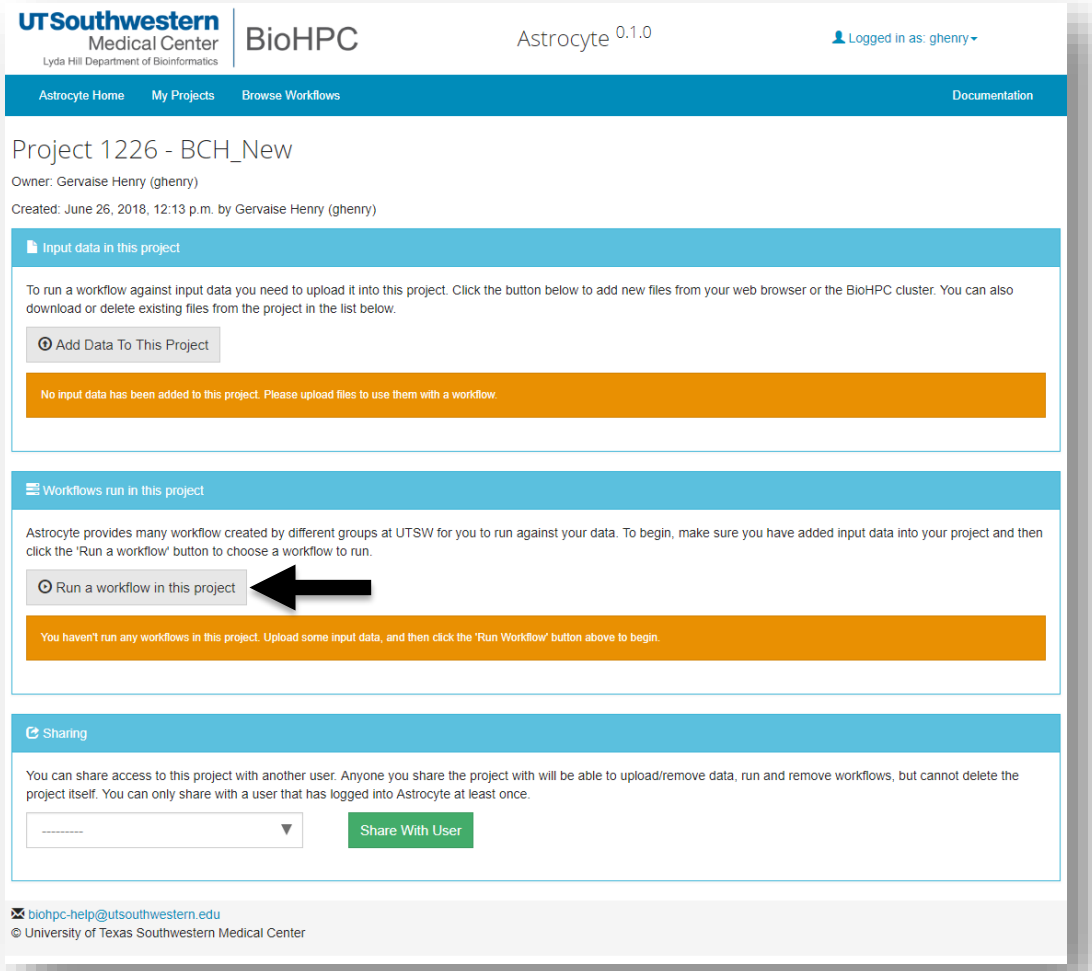
Import Selected Files | Finished importing files

Show 10 entries

Search:

	File	Size
☐	BCH/S085B.fastq.gz	690.9 MB
☐	BCH/S132INB.fastq.gz	1.3 GB
☐	BCH/S114B.fastq.gz	846.1 MB
☐	BCH/S110B.fastq.gz	1.5 GB
☐	BCH/S117B.fastq.gz	854.8 MB
☐	BCH/design.txt	345 bytes
☐	BCH/S118B.fastq.gz	863.2 MB
☐	BCH/S089B.fastq.gz	849.7 MB
☐	20170615_CPL127_4055/S228F_OE.fastq.gz	801.7 MB

Using BICF RNA-Seq Analysis Workflow on Astrocyte

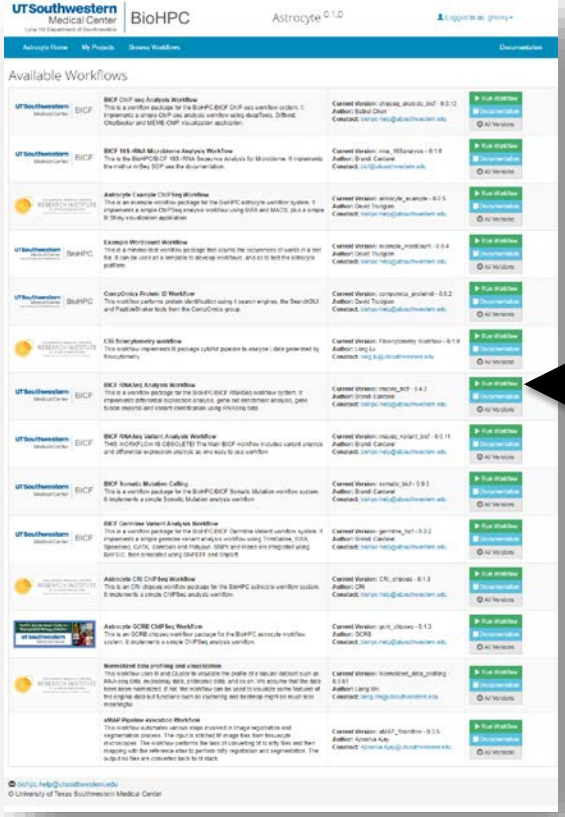


Available Workflows

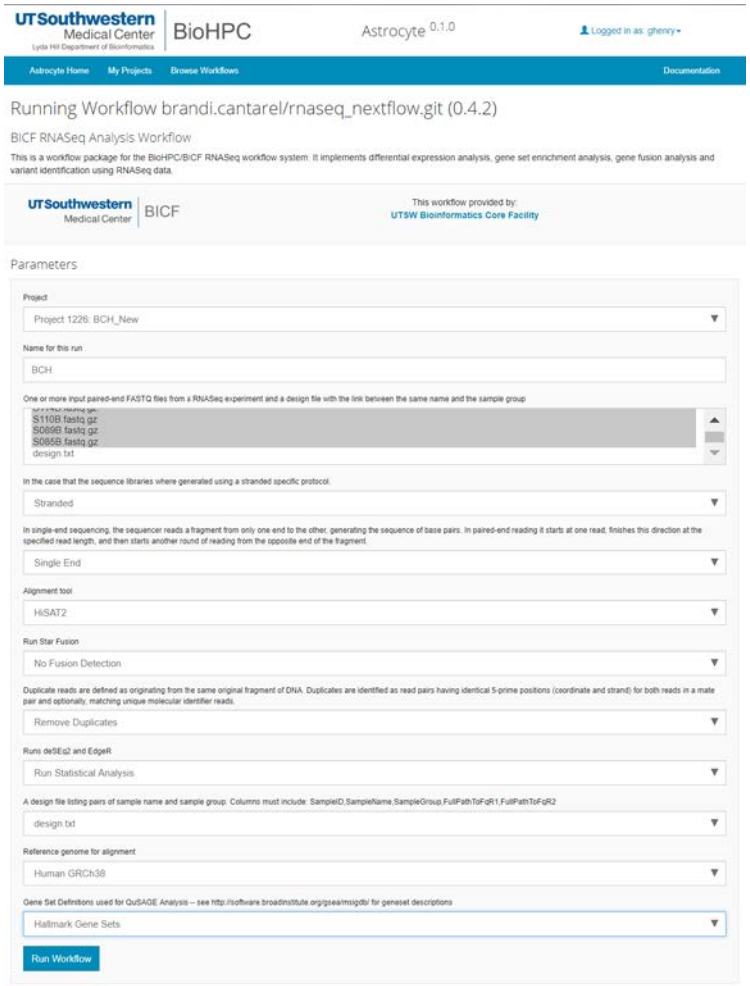
UTSouthwestern Medical Center | BICF

BICF RNASeq Analysis Workflow

This is a workflow package for the BioHPC/BICF RNASeq workflow system. It implements differential expression analysis, gene set enrichment analysis, gene fusion analysis and variant identification using RNASeq data.



Using BICF RNA-Seq Analysis Workflow on Astrocyte



Using BICF RNA-Seq Analysis Workflow on Astrocyte

UTSouthwestern Medical Center | BioHPC

Lyda Hill Department of Bioinformatics

Astrocyte 0.1.0

Logged in as: ghenry

Astrocyte Home | My Projects | Browse Workflows | Documentation

Run 'BCH' in Project 'BCH_New'

Run Information

Running Workflow	BICF RNASeq Analysis Workflow brandi.cantarel/rnaseq_nextflow.git / 0.4.2
Status	QUEUED
Created	June 26, 2018, 3:32 p.m. by ghenry
Size	0 bytes

Parameters

Parameter	Value
genome	/project/shared/bicf_workflow_ref/GRCh38
design	design.txt
align	hisat
fusion	skip
pairs	se
fastqs	S132INB.fastq.gz
fastqs	S118B.fastq.gz
fastqs	S117B.fastq.gz
fastqs	S114B.fastq.gz
fastqs	S110B.fastq.gz
fastqs	S089B.fastq.gz
fastqs	S085B.fastq.gz
dea	detect
stranded	1
markdups	picard
geneset	h.all.v5.1.symbols.gmt

Input Files

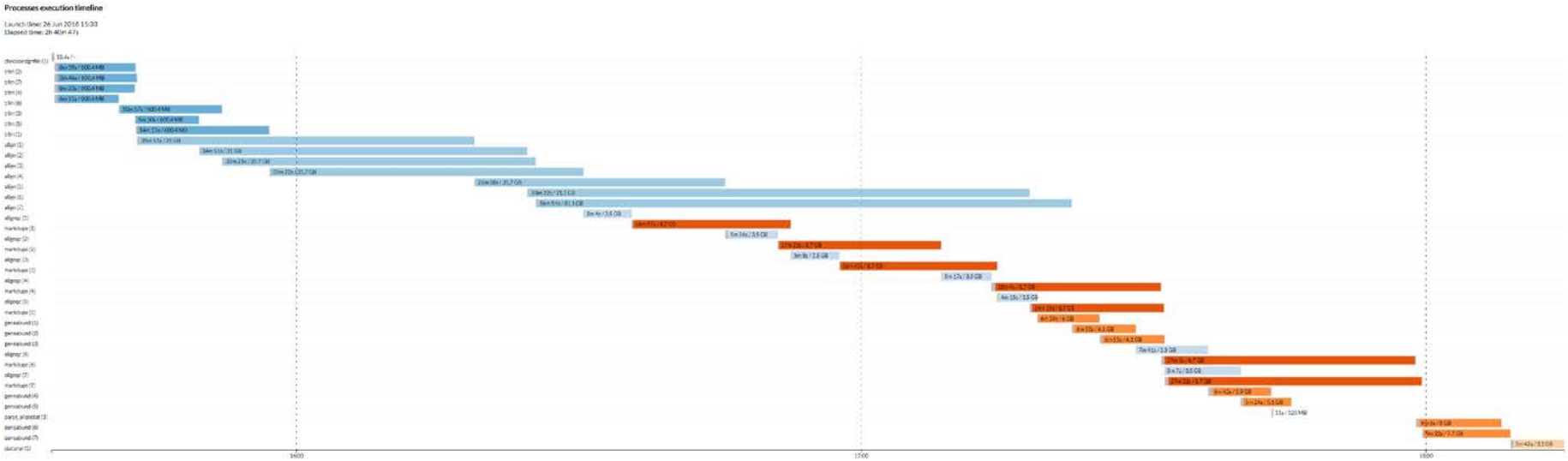
Filename	Size
design.txt	345 bytes
S132INB.fastq.gz	1.3 GB
S118B.fastq.gz	863.2 MB
S117B.fastq.gz	854.8 MB
S114B.fastq.gz	846.1 MB
S110B.fastq.gz	1.5 GB
S089B.fastq.gz	849.7 MB
S085B.fastq.gz	690.9 MB

Workflow Progress

View Workflow Timeline (opens in new window)

Task Number	Task Name	SLURM Job	Status	Submitted	Duration

Using BICF RNA-Seq Analysis Workflow on Astrocyte



Results

Workflow Output / Visualization

You can **download** an archive file containing all output of the workflow, or **export** it directly to a location on the BioHPC cluster storage for further work.

Note - Mac OSX cannot extract zip files >4GB. A tar file download will be added shortly.

Download Workflow Output: Download as .zip file

Export Output: Export to /project/apps/astrocyte/astrocyte_outgoing/ghenry

The **Visualization App** (vizapp) allows you to explore the results of your workflow on the web. Use the buttons below to start/stop and connect to a vizapp session. It takes 30s for the vizapp to start, or longer if there is a queue on the BioHPC cluster. Please stop the vizapp when you are finished using it, as it occupies a slot on the BioHPC cluster.

Vizapp Status: Start Vizapp

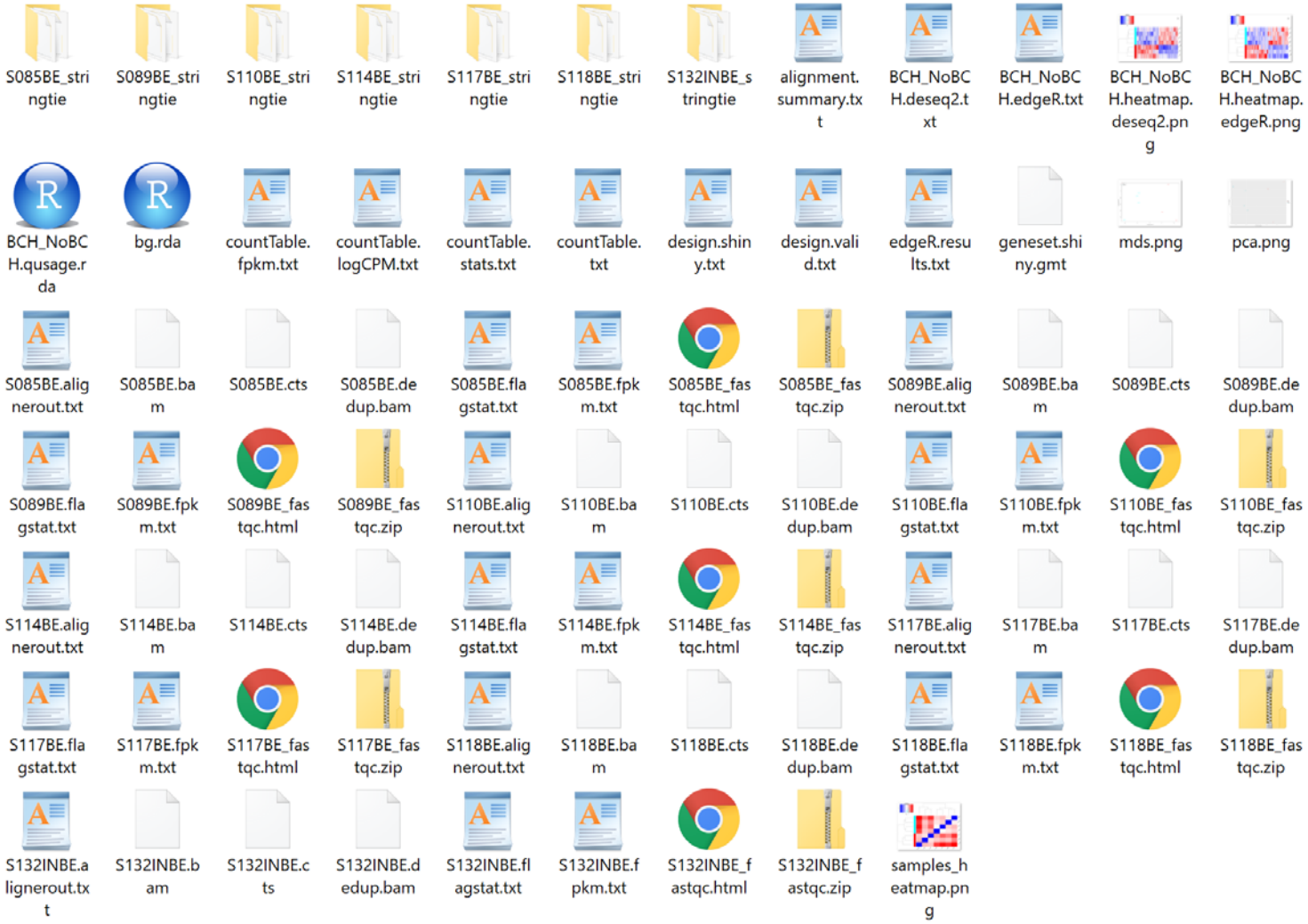
Output Browser

Click the 'Generate Direct Link' button to obtain a direct web link you can use with external tools, such as the UCSC Browser, that need to access the file directly. These links are valid for 24 hours.

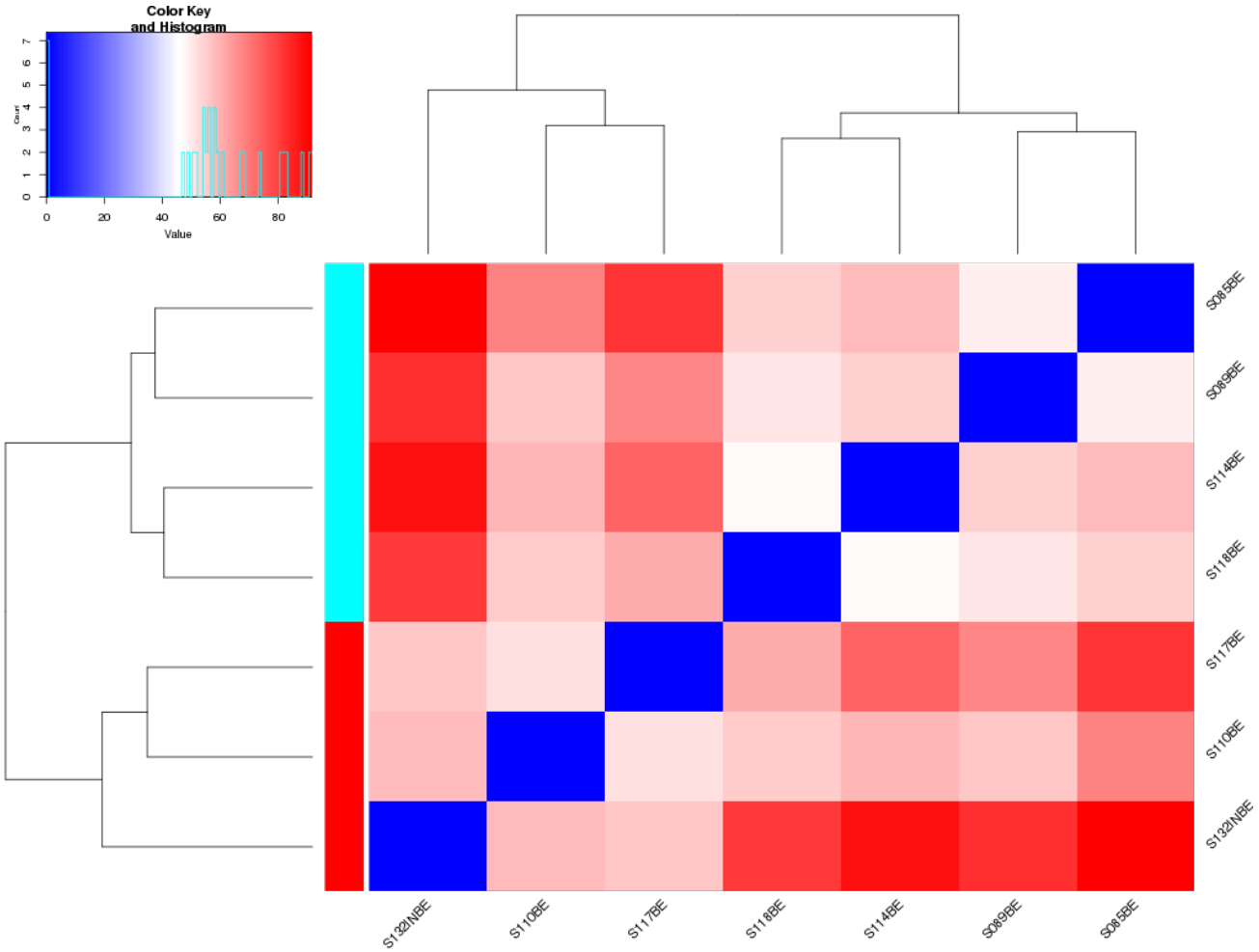
Current Directory: (/)

- S110BE_stringtie
- S085BE_stringtie
- S114BE_stringtie
- S132INBE_stringtie
- S117BE_stringtie
- S089BE_stringtie

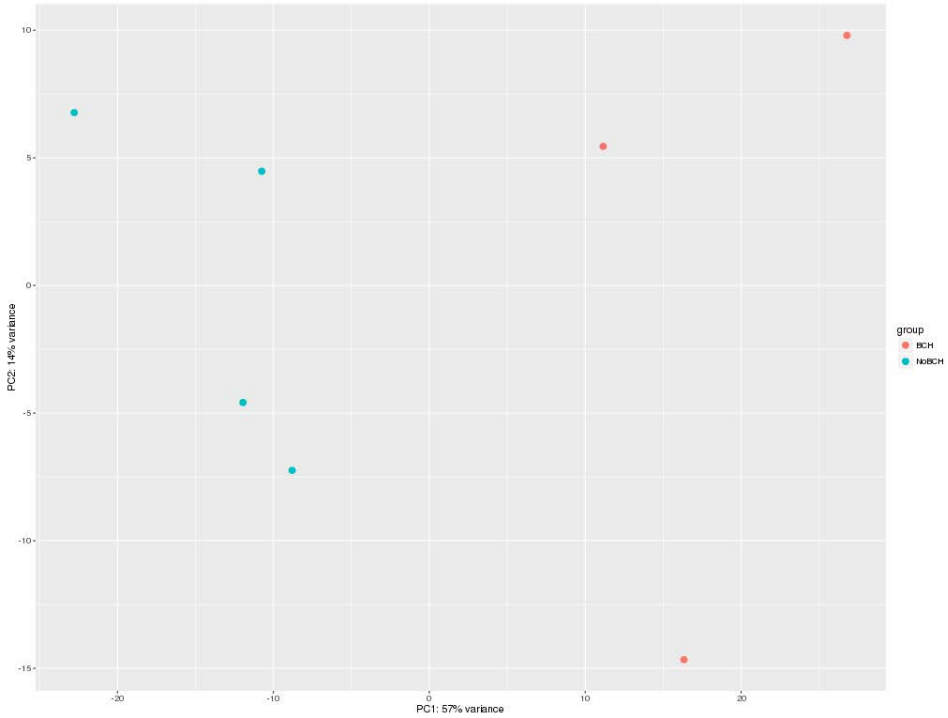
Results

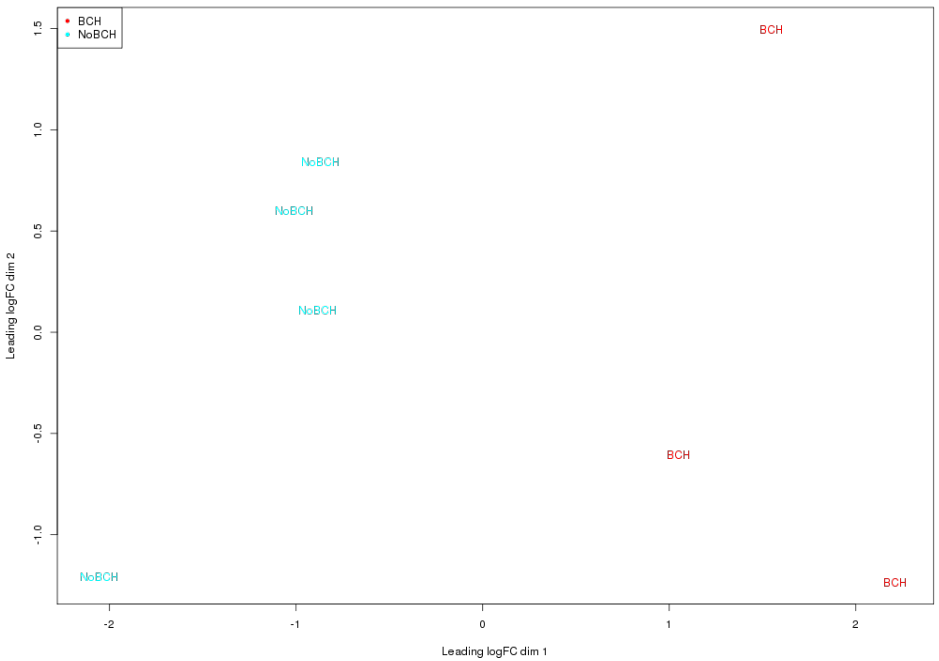

S085BE_stringtie
S089BE_stringtie
S110BE_stringtie
S114BE_stringtie
S117BE_stringtie
S118BE_stringtie
S132INBE_stringtie
alignment.summary.txt
BCH_NoBCH.deseq2.txt
BCH_NoBCH.edgeR.txt
BCH_NoBCH.heatmap.deseq2.png
BCH_NoBCH.heatmap.edgeR.png
BCH_NoBCH.qusage.rda
bg.rda
countTable.fpkm.txt
countTable.logCPM.txt
countTable.stats.txt
countTable.txt
design.shiny.txt
design.valid.txt
edgeR.results.txt
geneset.shiny.gmt
mds.png
pca.png
S085BE.alignerout.txt
S085BE.bam
S085BE.cts
S085BE.dedup.bam
S085BE.flagstat.txt
S085BE.fpkm.txt
S085BE_fastqc.html
S085BE_fastqc.zip
S089BE.alignerout.txt
S089BE.bam
S089BE.cts
S089BE.dedup.bam
S089BE.flagstat.txt
S089BE.fpkm.txt
S089BE_fastqc.html
S089BE_fastqc.zip
S110BE.alignerout.txt
S110BE.bam
S110BE.cts
S110BE.dedup.bam
S110BE.flagstat.txt
S110BE.fpkm.txt
S110BE_fastqc.html
S110BE_fastqc.zip
S114BE.alignerout.txt
S114BE.bam
S114BE.cts
S114BE.dedup.bam
S114BE.flagstat.txt
S114BE.fpkm.txt
S114BE_fastqc.html
S114BE_fastqc.zip
S117BE.alignerout.txt
S117BE.bam
S117BE.cts
S117BE.dedup.bam
S117BE.flagstat.txt
S117BE.fpkm.txt
S117BE_fastqc.html
S117BE_fastqc.zip
S118BE.alignerout.txt
S118BE.bam
S118BE.cts
S118BE.dedup.bam
S118BE.flagstat.txt
S118BE.fpkm.txt
S118BE_fastqc.html
S118BE_fastqc.zip
S132INBE.alignerout.txt
S132INBE.bam
S132INBE.cts
S132INBE.dedup.bam
S132INBE.flagstat.txt
S132INBE.fpkm.txt
S132INBE_fastqc.html
S132INBE_fastqc.zip
samples_heatmap.png

Results

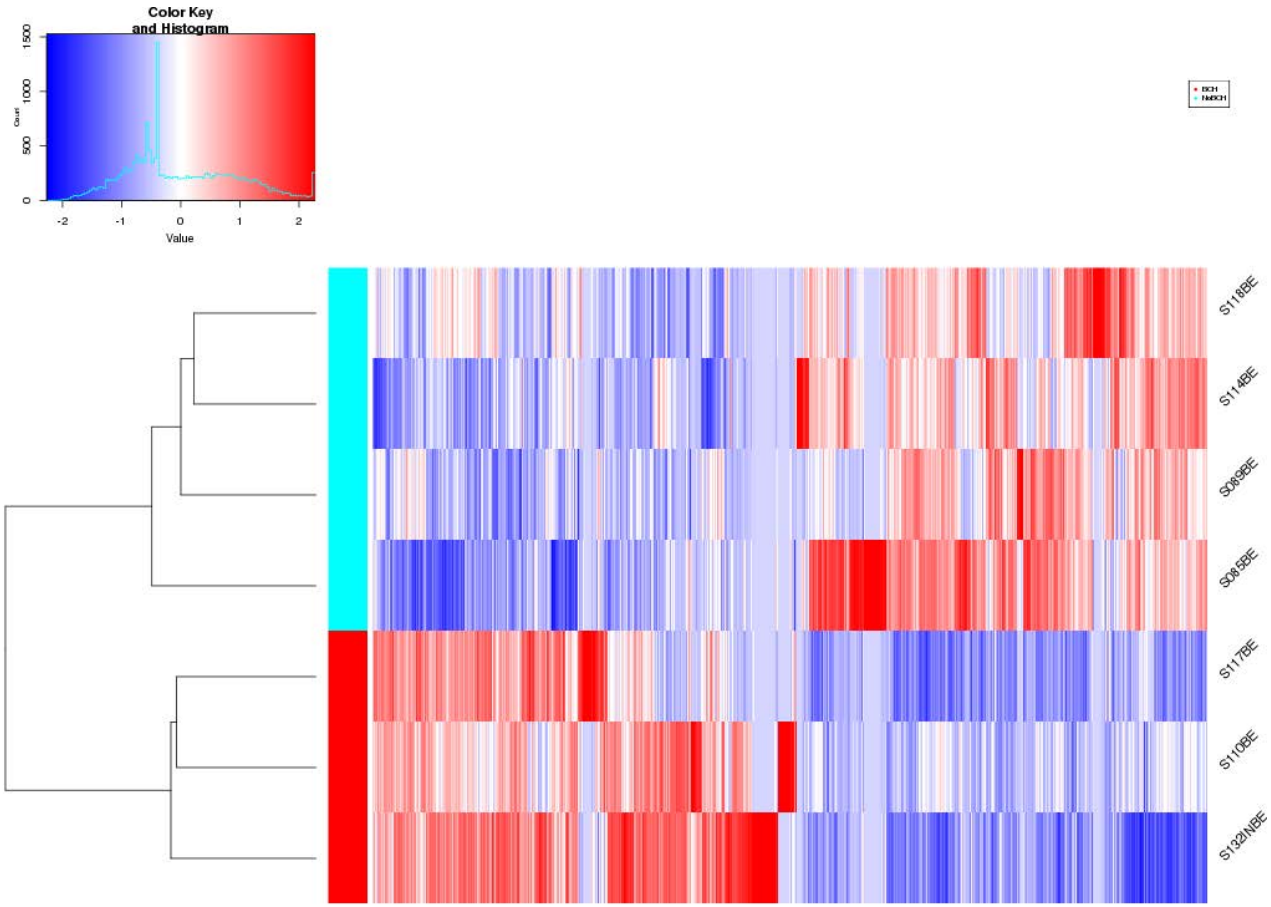

Color Key
and Histogram
Count
0
1
2
3
4
5
6
7
0
20
40
60
80
Value
S085BE
S089BE
S114BE
S118BE
S117BE
S110BE
S132NBE
S132NBE
S110BE
S117BE
S118BE
S114BE
S089BE
S085BE

Results

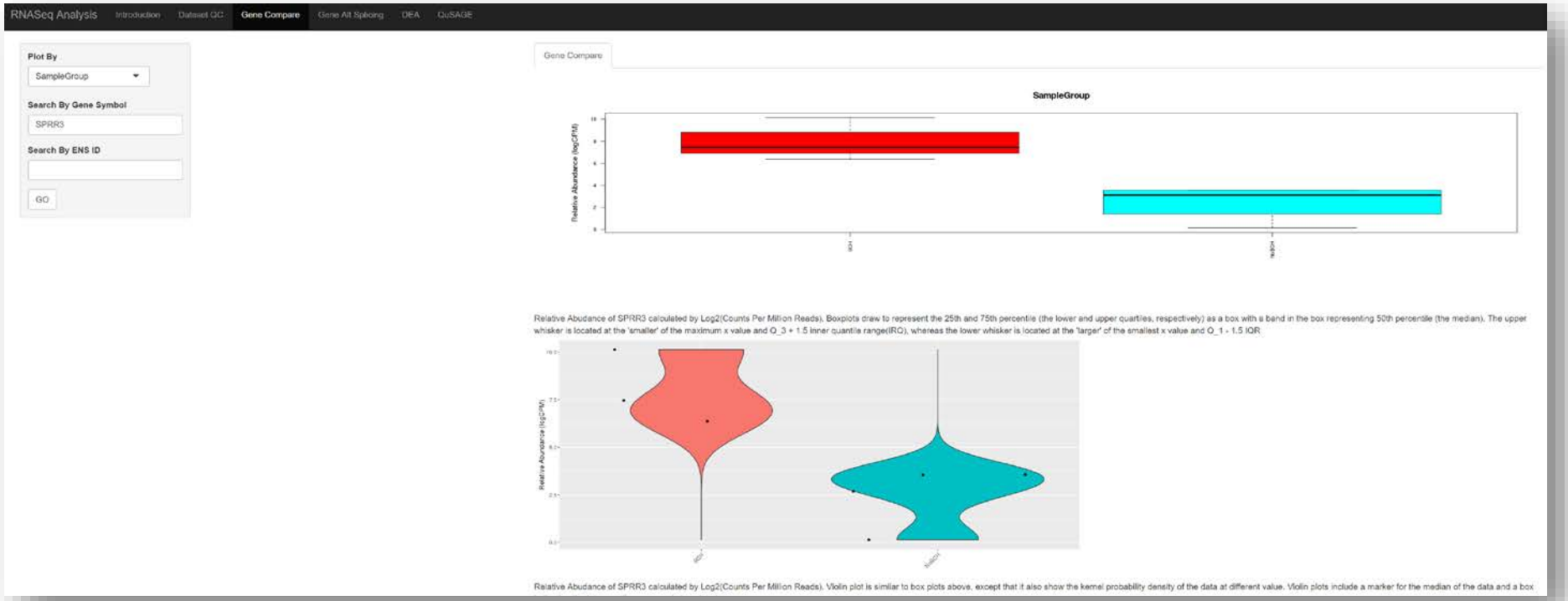

PC2: 14% variance
PC1: 57% variance
group
BCH
NoBCH


BCH
NoBCH
Leading logFC dim 2
Leading logFC dim 1

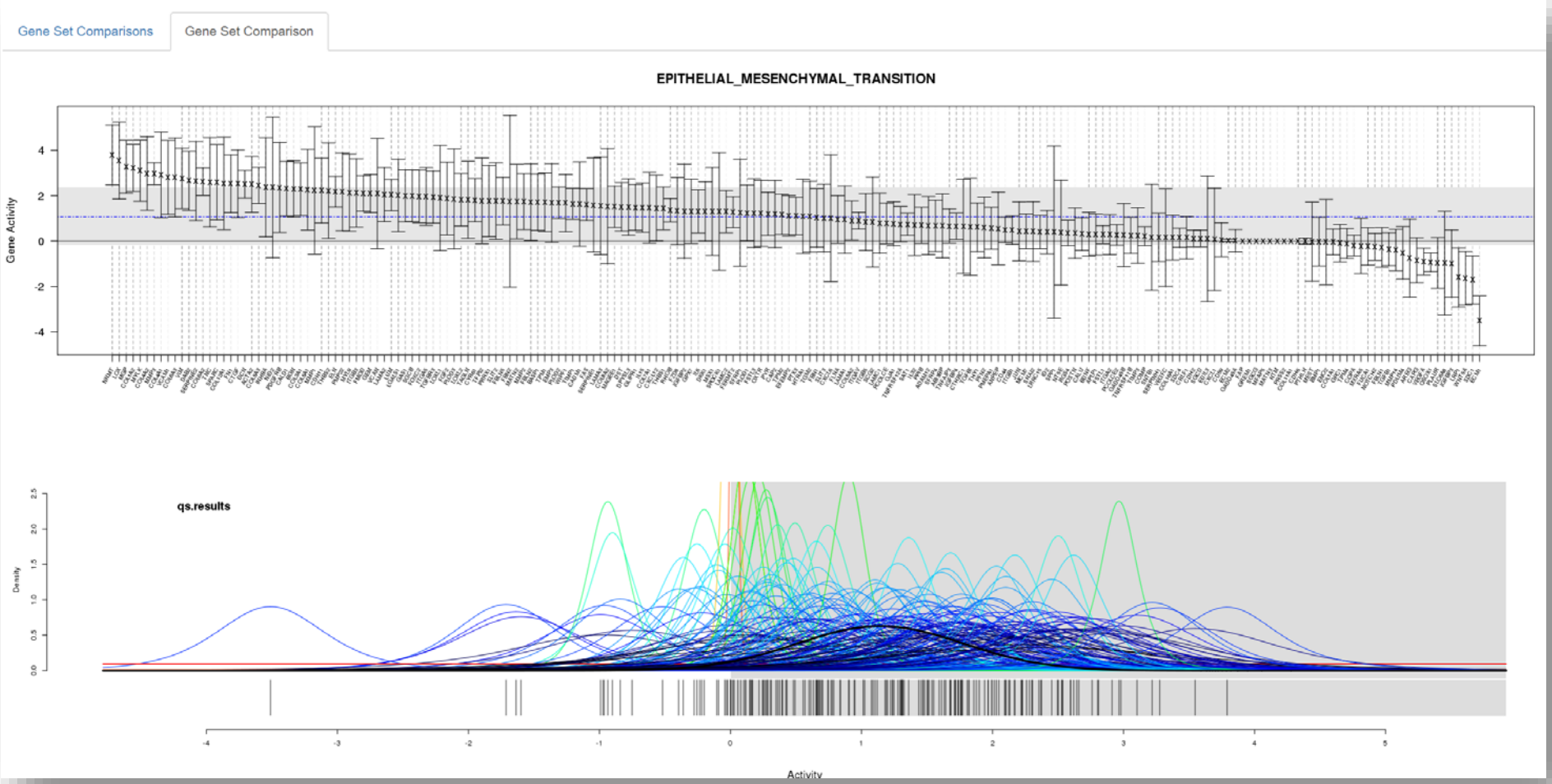
Results


Color Key
and Histogram
Count
0
500
1000
1500
-2
-1
0
1
2
Value
BCH
NoBCH
S118BE
S114BE
S083BE
S085BE
S117BE
S110BE
S132NBE

Results


RNASeq Analysis
Introduction
Dataset QC
Gene Compare
Gene Alt Splicing
DEA
QuSAGE
Plot By
SampleGroup
Search By Gene Symbol
SPRR3
Search By ENS ID
GO
Gene Compare
SampleGroup
Relative Abundance (logCPM)
Relative Abundance of SPRR3 calculated by Log2(Counts Per Million Reads). Boxplots draw to represent the 25th and 75th percentile (the lower and upper quartiles, respectively) as a box with a band in the box representing 50th percentile (the median). The upper whisker is located at the 'smaller' of the maximum x value and Q_3 + 1.5 inner quantile range(IRQ), whereas the lower whisker is located at the 'larger' of the smallest x value and Q_1 - 1.5 IQR
Relative Abundance (logCPM)
Relative Abundance of SPRR3 calculated by Log2(Counts Per Million Reads). Violin plot is similar to box plots above, except that it also show the kernel probability density of the data at different value. Violin plots include a marker for the median of the data and a box

Results


Gene Set Comparisons
Gene Set Comparison
EPITHELIAL_MESENCHYMAL_TRANSITION
Gene Activity
qs.results
Density
Activity

Under the hood

```
library(edgeR)
library(DESeq2)
library("RColorBrewer")
library("gplots")
library(qusage)
```

Under the hood

```
##################### Read in Data ##################################
genenames <- read.table(file="genenames.txt",header=TRUE,sep='\t')

tbl <- read.table('countTable.txt',header=TRUE,sep="\t")
tbl2 <- read.table('countTable.logCPM.txt',header=TRUE,sep="\t")
ct <- tbl[,4:length(tbl)]
row.names(ct) <- tbl$ENSEMBL

samples<- names(ct)
```

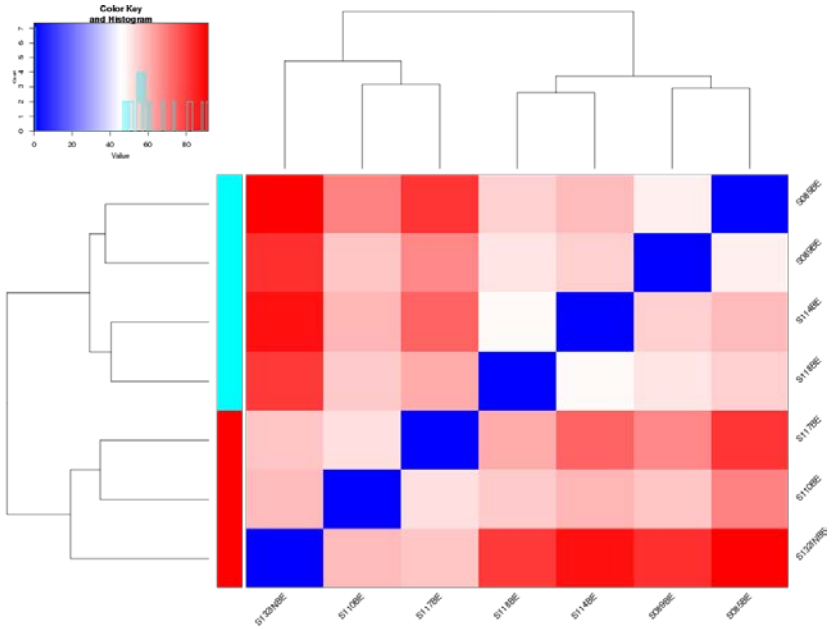
Under the hood

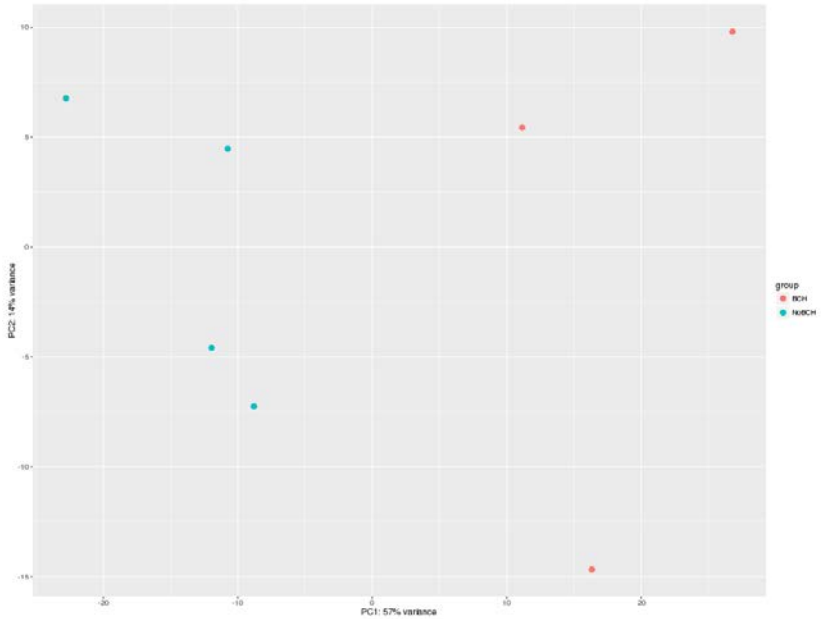
```
##################### Run DESEQ2 ################################
dds <- DESeq(dds)
rld <- rlogTransformation(dds, blind=TRUE)
sampleDists <- dist(t(assay(rld)))
```

Under the hood

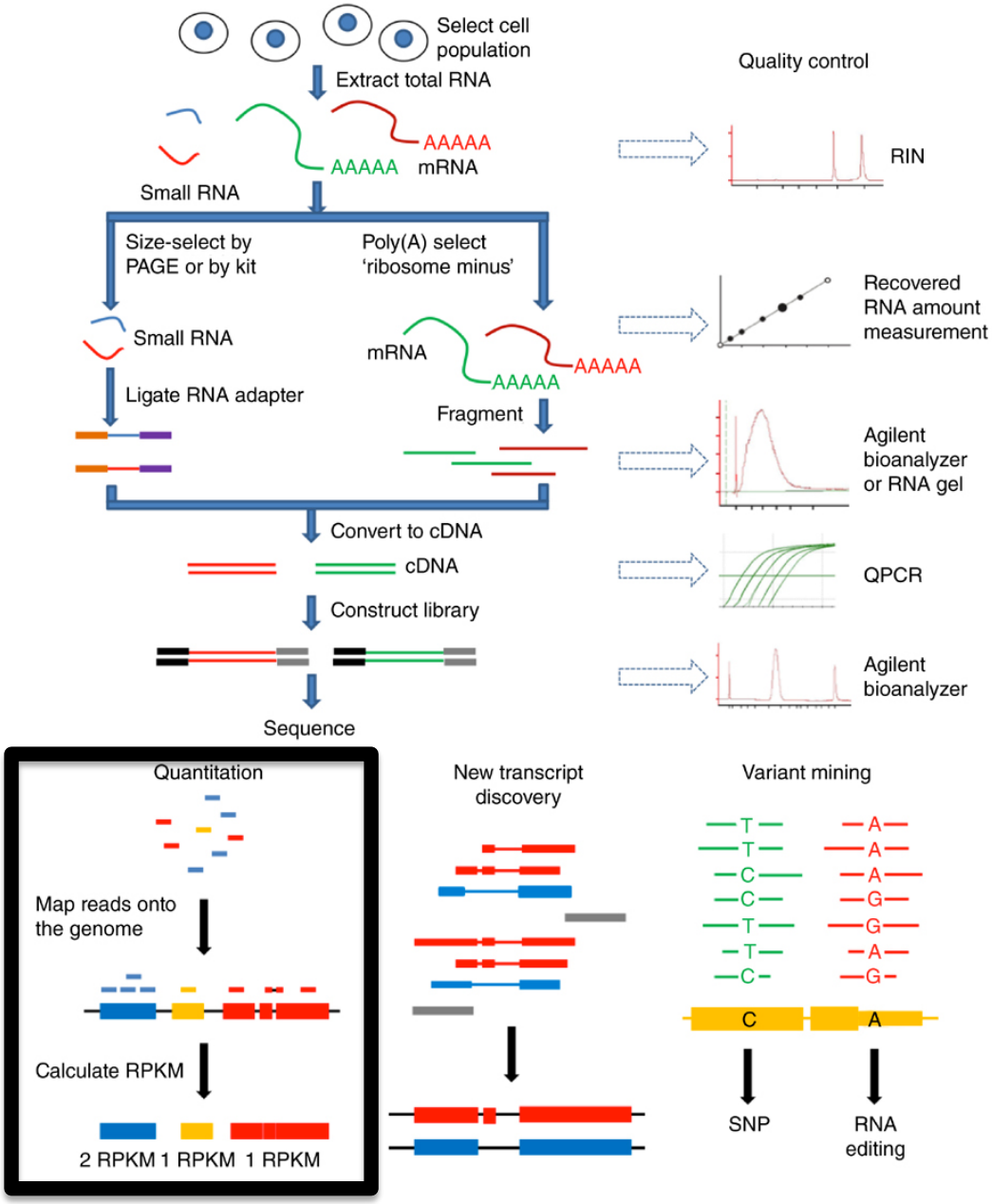
```
png(file="samples_heatmap.png",bg ="transparent",height=768,width=1024)
heatmap.2(as.matrix(sampleDists), col = bluered(100),RowSideColors = col.blocks,srtRow=45,srtCol=45,trace="none", margins=c(5, 5))
dev.off()

#Compare Samples using PCA
png(file="pca.png",bg ="transparent",height=768,width=1024)
print(plotPCA(rld, intgroup="SampleGroup"),col.hab=col.blocks)
dev.off()
```





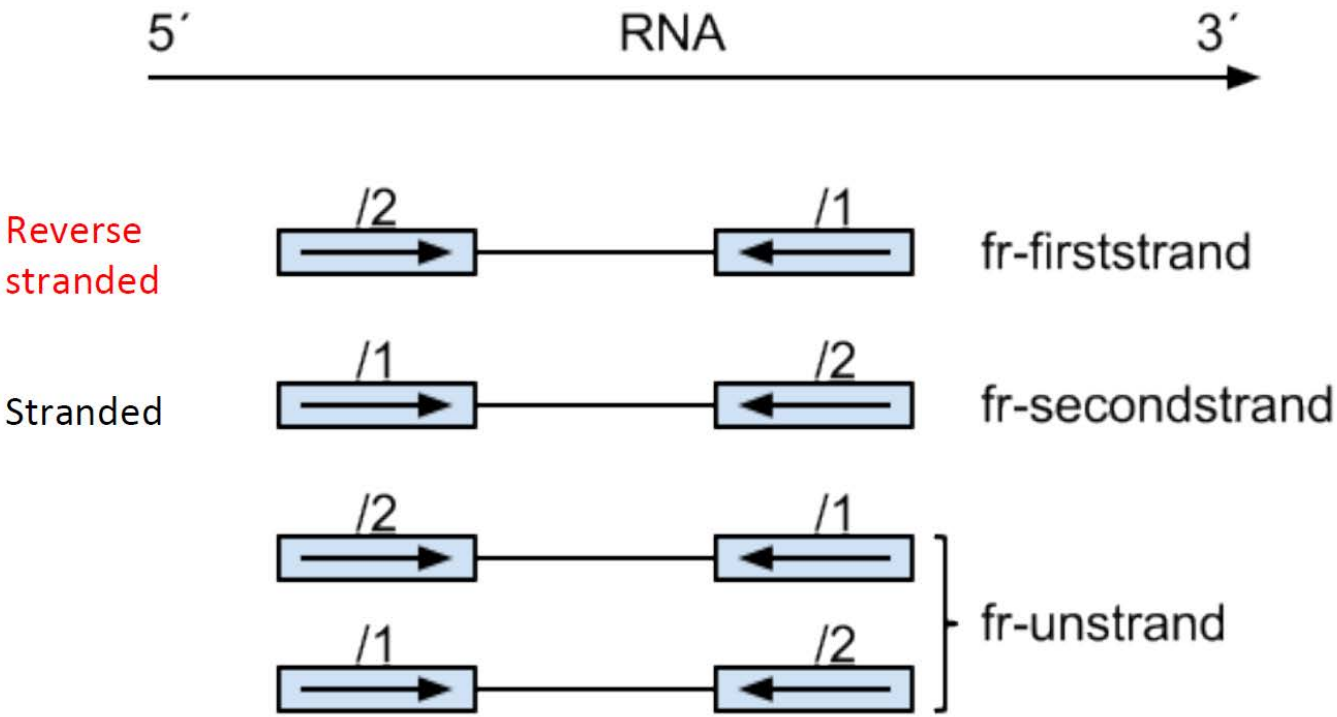
ABOUT RNA-SEQ


Select cell population
Extract total RNA
AAAAA
AAAAA
mRNA
Small RNA
Size-select by PAGE or by kit
Poly(A) select 'ribosome minus'
Small RNA
mRNA
AAAAA
AAAAA
Ligate RNA adapter
Fragment
Convert to cDNA
cDNA
Construct library
Sequence
Quality control
RIN
Recovered RNA amount measurement
Agilent bioanalyzer or RNA gel
QPCR
Agilent bioanalyzer
Quantitation
Map reads onto the genome
Calculate RPKM
2 RPKM 1 RPKM 1 RPKM
New transcript discovery
Variant mining
T
T
C
C
T
T
C
A
A
A
G
G
A
G
C
A
SNP
RNA editing

Experimental Design Affecting Your Analysis

- Whole transcriptome vs mRNA
- Single-end vs paired-end
  - Paired-end produces more accurate alignments
  - Paired-end allows for transcript level analysis
  - Single-end is cheaper
- Number of Reads
  - 10-50M is a good range
  - Aim for at least 20M
- Read Length
  - Longer reads produce better alignments, min 50bp paired-end or 100bp single-end for gene quantification

Experimental Design Affecting Your Analysis



Experimental Design Affecting Your Analysis

- Number of Samples
  - Your power to detect an effect depends on
    - Effect size (difference between groups)
    - Within group variance
    - Sample size
  - More samples the better, min 3 per group
    - 5 samples sequenced to 20M read each offer more power than 2 samples sequenced to 50M reads each
- Stranded
  - Can distinguish expression of overlapping genes

Useful Tools

- **Gene Set Enrichment Analysis (GSEA)**
  https://software.broadinstitute.org/gsea/index.jsp
- **Molecular Signatures Database (MSigDB)**
  http://software.broadinstitute.org/gsea/msigdb/index.jsp
- **Gene Pattern**
  https://genepattern.broadinstitute.org/gp/pages/login.jsf
  Use countable.logCPM.txt to generate .gct file or edgeR.results.txt to generate .rnk file in excel as inputs
- **Morpheus (user-defined specific heatmaps)**
  https://software.broadinstitute.org/morpheus/
- **Alternative/complex designs**
  https://www.bioconductor.org/packages/release/bioc/html/edgeR.html
  https://www.bioconductor.org/packages/release/bioc/html/DESeq2.html
  Use counTable.txt as input
- **Homer Motif Analysis**
  http://homer.ucsd.edu/homer/motif/
  Use edgeR.results.txt as input