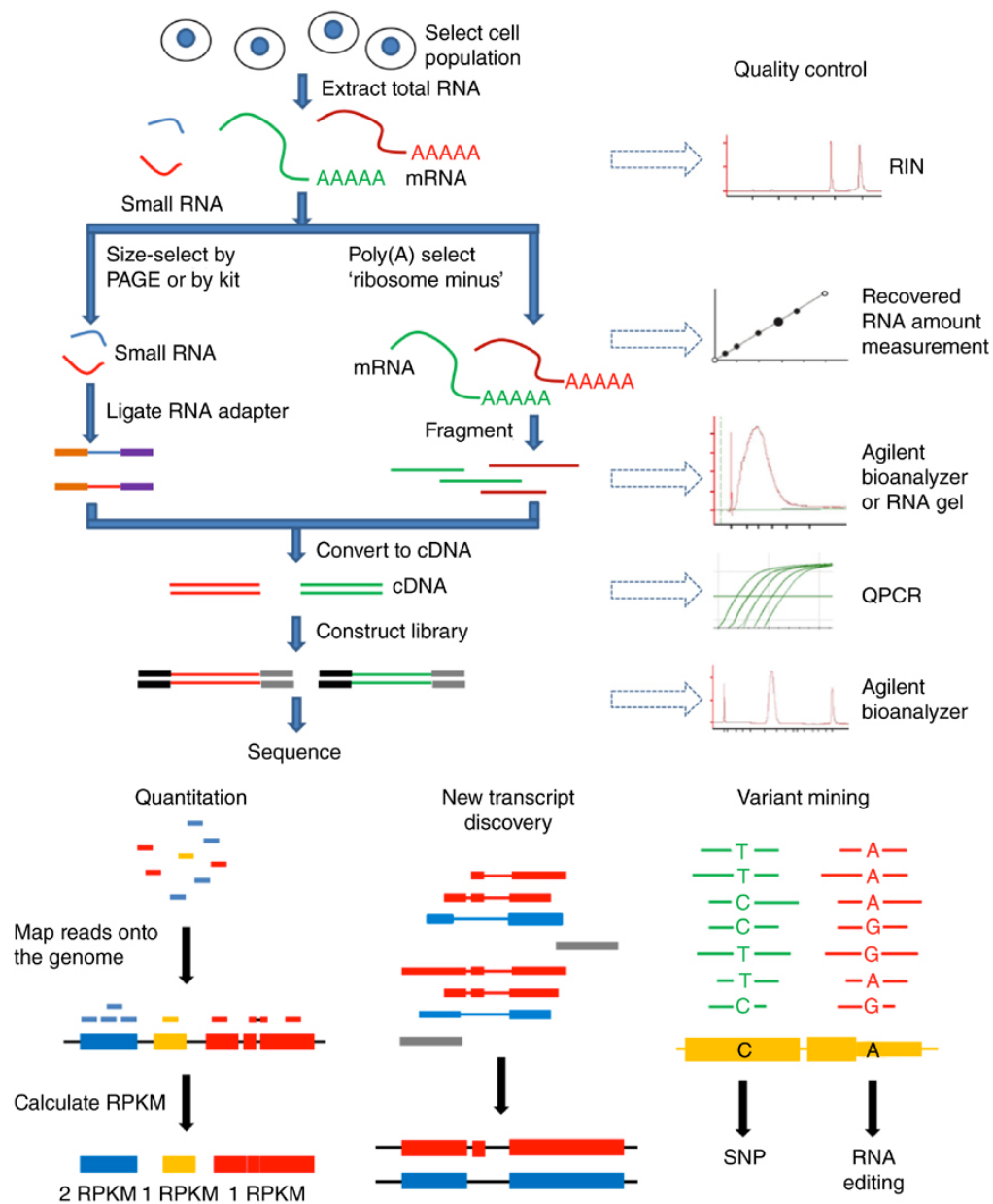


Introduction to RNASeq Analysis with BICF's Astrocyte Workflow

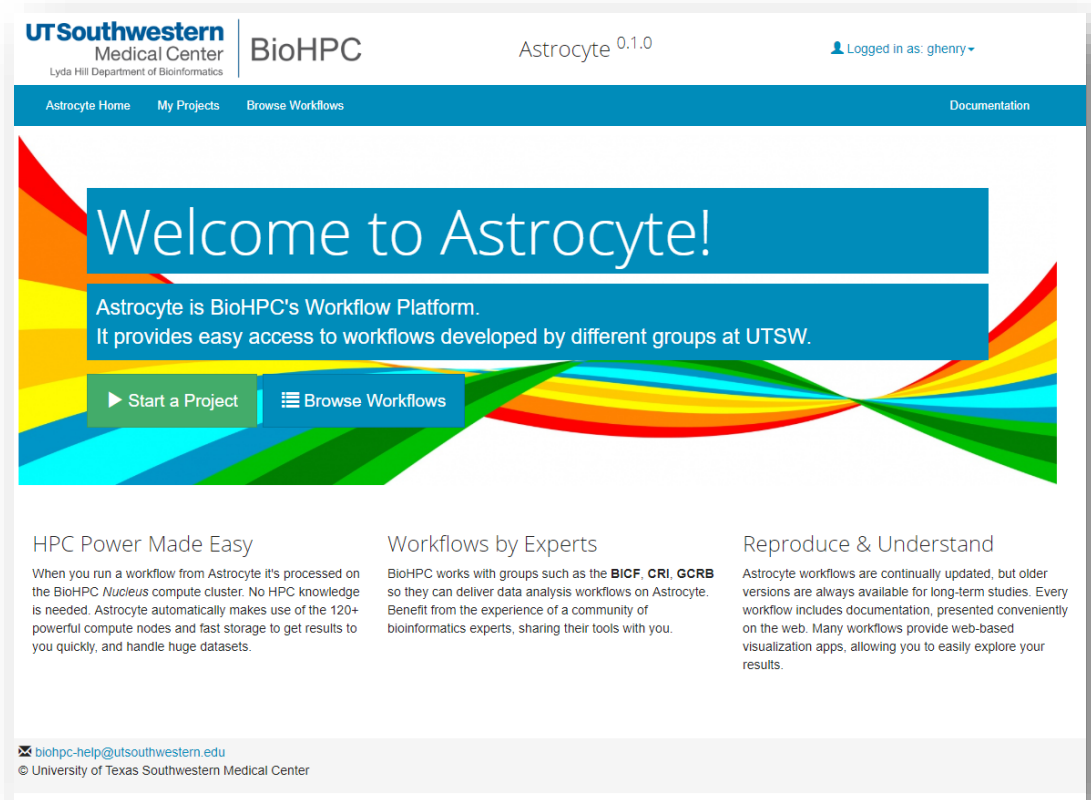
Gervaise H. Henry
Department of Urology
(BICF Fellow)



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



The screenshot shows the Astrocyte web interface. At the top, the header includes the UT Southwestern Medical Center logo, BioHPC, Astrocyte 0.1.0, and a user login status 'Logged in as: ghenry'. Below the header is a navigation bar with links: Astrocyte Home, My Projects, Browse Workflows, and Documentation. The main content area features a large blue banner with the text 'Welcome to Astrocyte!' and 'Astrocyte is BioHPC's Workflow Platform. It provides easy access to workflows developed by different groups at UTSW.' Below the banner are two buttons: 'Start a Project' and 'Browse Workflows'. The page is decorated with a colorful, abstract graphic of overlapping triangles in red, orange, yellow, green, and blue. Below the banner, there are three columns of text: 'HPC Power Made Easy', 'Workflows by Experts', and 'Reproduce & Understand'. At the bottom, there is a footer with the email 'biohpc-help@utsouthwestern.edu' and the copyright notice '© University of Texas Southwestern Medical Center'.

UT Southwestern Medical Center
Lyda Hill Department of Bioinformatics

BioHPC

Astrocyte 0.1.0

Logged in as: ghenry

Astrocyte Home My Projects Browse Workflows Documentation

Welcome to Astrocyte!

Astrocyte is BioHPC's Workflow Platform.
It provides easy access to workflows developed by different groups at UTSW.

▶ Start a Project

≡ Browse Workflows

HPC Power Made Easy

When you run a workflow from Astrocyte it's processed on the BioHPC *Nucleus* compute cluster. No HPC knowledge is needed. Astrocyte automatically makes use of the 120+ powerful compute nodes and fast storage to get results to you quickly, and handle huge datasets.

Workflows by Experts

BioHPC works with groups such as the **BICF**, **CR1**, **GCRB** so they can deliver data analysis workflows on Astrocyte. Benefit from the experience of a community of bioinformatics experts, sharing their tools with you.

Reproduce & Understand

Astrocyte workflows are continually updated, but older versions are always available for long-term studies. Every workflow includes documentation, presented conveniently on the web. Many workflows provide web-based visualization apps, allowing you to easily explore your results.















✉ biohpc-help@utsouthwestern.edu
© University of Texas Southwestern Medical Center

Available Workflows

UTSouthwestern Medical Center BioHPC Astrocyte 0.1.0 [Login as gherry](#)

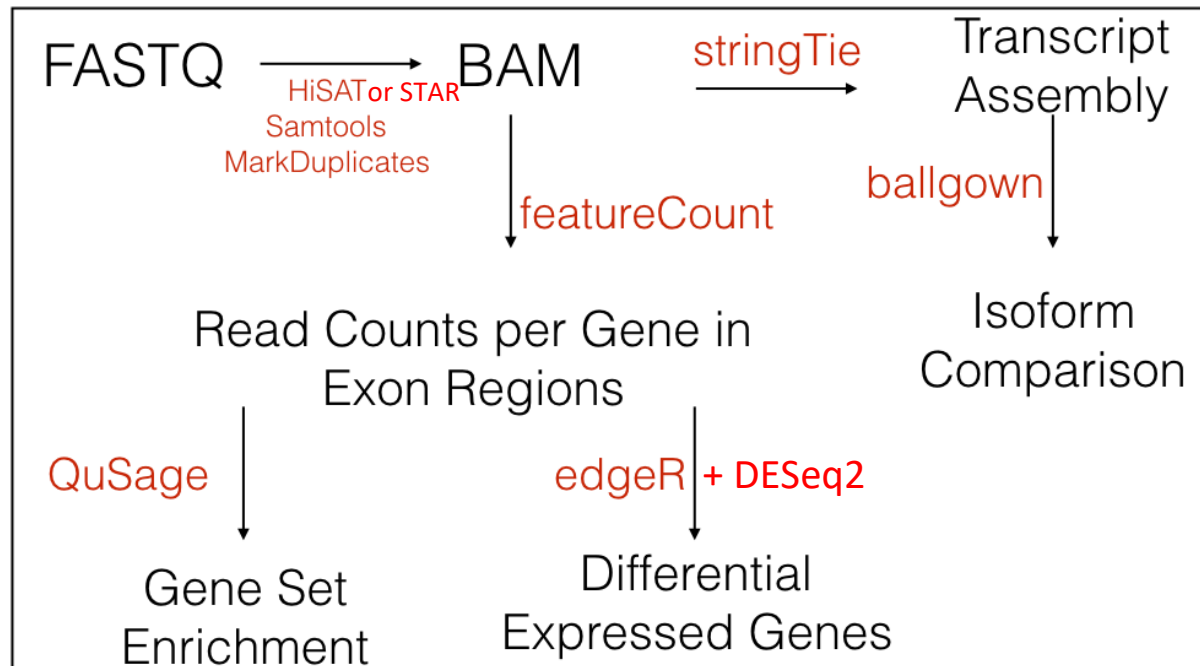
[Astrocyte Home](#) [My Projects](#) [Browse Workflows](#) [Documentation](#)

Available Workflows

 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: Sarah Chen Contact: bicfp-help@utsouthwestern.edu	Run Workflow Documentation All Versions
 BICF	BICF 16S rRNA Microbiome Analysis Workflow This is the BioHPC-BICF 16S rRNA Sequence Analysis for Microbiome. It implements the mothur mdseq SOP see the documentation.	Current Version: rna_16sAnalysis - 0.1.0 Author: Brand Cantarel Contact: bicf@utsouthwestern.edu	Run Workflow Documentation All Versions
	Astrocyte Example ChIP-seq Workflow This is an example workflow package for the BioHPC astrocyte workflow system. It implements a simple ChIP-seq analysis workflow using BEDTools and MACS, plus a simple R Shiny visualization application.	Current Version: astrocyte_example - 0.0.5 Author: David Trugnan Contact: bicfp-help@utsouthwestern.edu	Run Workflow Documentation All Versions
 BioHPC	Example Wordcount Workflow This is a minimal test workflow package that counts the occurrences 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 Trugnan Contact: bicfp-help@utsouthwestern.edu	Run Workflow Documentation All Versions
 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 Trugnan Contact: bicfp-help@utsouthwestern.edu	Run Workflow Documentation All Versions
	CRISPR flow cytometry workflow This workflow implements a package workflow pipeline to analyze CRISPR data generated by flow cytometry.	Current Version: Flowcytometry Workflow - 0.1.0 Author: Ling Li Contact: ling.li@utsouthwestern.edu	Run Workflow Documentation All Versions
 BICF	BICF RNA-seq Analysis Workflow This is a workflow package for the BioHPC-BICF RNA-seq workflow system. It implements differential expression analysis, gene set enrichment analysis, gene fusion analysis and variant identification using RNA-seq data.	Current Version: rnaseq_bicf - 0.4.2 Author: Brand Cantarel Contact: bicfp-help@utsouthwestern.edu	Run Workflow Documentation All Versions
 BICF	BICF RNA-seq Variant Analysis Workflow This workflow is OBSOLETE! The Max-BICF workflow includes variant analysis and differential expression analysis as one easy to use workflow.	Current Version: rnaseq_variant_bicf - 0.0.11 Author: Brand Cantarel Contact: bicfp-help@utsouthwestern.edu	Run Workflow Documentation All Versions
 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: Brand Cantarel Contact: bicfp-help@utsouthwestern.edu	Run Workflow Documentation All Versions
 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 TruSomatic, ENTP, SpeedSeq, GATK, Samtools and Platypus. SNPs and indels are integrated using BAYESIC, then annotated using SnpEff and SnpSift.	Current Version: germline_bicf - 0.2.2 Author: Brand Cantarel Contact: bicfp-help@utsouthwestern.edu	Run Workflow Documentation All Versions
	Astrocyte CRIS ChIP-seq Workflow This is an CRIS ChIP-seq workflow package for the BioHPC astrocyte workflow system. It implements a simple ChIP-seq analysis workflow.	Current Version: CRIS_chipseq - 0.1.3 Author: CRIS Contact: bicfp-help@utsouthwestern.edu	Run Workflow Documentation All Versions
	Astrocyte GC-RB ChIP-seq Workflow This is an GC-RB ChIP-seq workflow package for the BioHPC astrocyte workflow system. It implements a simple ChIP-seq analysis workflow.	Current Version: gcrb_chipseq - 0.1.3 Author: GC-RB Contact: bicfp-help@utsouthwestern.edu	Run Workflow Documentation All Versions
	Normalized data profiling and visualization This workflow uses R and Cytoscape 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 - 0.0.1 Author: Ling Shi Contact: ling.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 a DICOM image files from histocycle microscope. The workflow performs the task of converting DICOM files and then mapping with the reference atlas to perform image registration and segmentation. The output files are converted back to DICOM stack.	Current Version: aMAP_Workflow - 0.0.5 Author: Apoorva Ajay Contact: Apoorva.Ajay@utsouthwestern.edu	Run Workflow Documentation All Versions

bicfp-help@utsouthwestern.edu
© University of Texas Southwestern Medical Center

BICF RNA-Seq Analysis Workflow



Using BICF RNA-Seq Analysis Workflow on Astrocyte

UT Southwestern Medical Center | BioHPC | Astrocyte 0.4.2

Documentation for BICF RNASeq Analysis Workflow (0.4.2)

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

Astrocyte RNASeq Workflow Package

Workflow SOP

This SOP describes the analysis pipeline of RNA sequencing data. This pipeline includes: (1) quality control, (2) variant calling analysis, (3) identification of fusion genes, and (4) differential analysis of gene expression and isoform expression. The result R data of the statistical analysis can be visualized using R shiny service.

For each file this workflow:

- 1) Trim the ends of sequences with remaining adapter or quality scores < 25. Remove any sequence less than 30bp after trimming, then gene rate a file for capturing information about how many sequences were trimmed.
- 2) Trimmed Fastq files are aligned to the selected reference genome using HISAT2 (Kim et al. 2015) or STAR (Dobin et al. 2013).
- 3) RNA-Seq counts using SAMtools.
- 4) FeatureCounts (Liao, Smyth, and Shi 2014) and featureCounts (Liao et al. 2014) and StringTie (Pertea et al. 2015) using the Genomic Feature Table (GTF) file (Pertea et al. 2015).
- 5) Basic pairwise differential expression analysis is performed using edgeR (Robinson et al. 2010) and DESeq2.
- 6) Abundance of transcripts are calculated using ballgown (Frazee et al. 2014).
- 7) Identify gene fusions or fused transcripts using STAR-fusion.

Workflow Parameters

fastq - Choose one or more Illumina read files to process.
 genome - Choose a genomic reference (genome).
 stranded - If a stranded library is created, please select the appropriate protocol to ensure strand specific analysis workflow. The default is to remove duplicates, you can skip this function.
 pairs - Indicate if this is paired-end or single-end sequencing data.
 dex - Perform Differential Expression analysis, please skip if there are < 3 sample groups present.
 fusion - Select a set of fusion generators for pathway analysis.
 fusion - Perform gene fusion identification.
 design - This file matches the fastq files to data about the sample.

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

There are some optional columns that might help with the analysis: Tissue, CellType, Culture, Date, Sequencing, Organism, CellType, Culture, Date, Age, Race, Ethnicity, Age.

Test Data

Credits

This example workflow is derived from original scripts kindly contributed by the Bioinformatics Core Facility (BCF), Department of Bioinformatics.

Workflow SOP

References

- 1. FastQC: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- 2. STAR: <https://github.com/alexdobin/STAR>
- 3. HISAT2: <https://github.com/DaehwanKim/hisat2>
- 4. SAMtools: <http://www.htslib.org/>
- 5. DESeq2: <https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>
- 6. edgeR: <https://bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/edgeR.html>
- 7. StringTie: <https://github.com/corbett/stringtie>
- 8. ballgown: <https://github.com/alexdobin/ballgown>
- 9. STAR-fusion: <https://github.com/alexdobin/star-fusion>

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

UTSouthwestern
Medical Center
Lyda Hill Department of Bioinformatics

BioHPC

Astrocyte ^{0.1.0}

Logged in as: ghenry

[Astrocyte Home](#) [My Projects](#) [Browse Workflows](#) [Documentation](#)

My Projects

In Astrocyte **projects** are used to organize your work. You upload **input data** into a project, and can then run **workflows** against this input data. Try to separate your work into natural projects, so that you can easily share them with other users if required.

+ Start a New Project

Create New Project

Existing Projects

ID	Name	Created	Workflows Run	Input Files	Size	Actions
PRJ1187	DPrF	April 26, 2018, 2:26 p.m.	1	17	286.9 GB	Delete
PRJ1039	D_SLOB	June 21, 2017, 2:47 p.m.	2	31	225.1 GB	Delete
PRJ97	BCH	Feb. 18, 2017, 6:05 p.m.	1	15	96.7 GB	Delete

Projects Shared with Me

ID	Name	Created	Workflows Run	Input Files	Size	Actions
PRJ1080	Ganesh	Sept. 20, 2017, 9:29 a.m. by douglas strand	16	26	1.1 TB	

biohpc-help@utsouthwestern.edu

© University of Texas Southwestern Medical Center

Using BICF RNA-Seq Analysis Workflow on Astrocyte

UTSouthwestern
Medical Center
Lyda Hill Department of Bioinformatics

BioHPC

Astrocyte 0.1.0

Logged in as: ghenry

[Astrocyte Home](#) [My Projects](#) [Browse Workflows](#) [Documentation](#)

Project 1226 - BCH_New

Owner: Gervaise Henry (ghenry)

Created: June 26, 2018, 12:13 p.m. by Gervaise Henry (ghenry)

Input data in this project

To run a workflow against input data you need to upload it into this project. Click the button below to add new files from your web browser or the BioHPC cluster. You can also download or delete existing files from the project in the list below.

Add Data To This Project

No input data has been added to this project. Please upload files to use them with a workflow.

Workflows run in this project

Astrocyte provides many workflow created by different groups at UTSW for you to run against your data. To begin, make sure you have added input data into your project and then click the 'Run a workflow' button to choose a workflow to run.

Run a workflow in this project

You haven't run any workflows in this project. Upload some input data, and then click the 'Run Workflow' button above to begin.

Sharing

You can share access to this project with another user. Anyone you share the project with will be able to upload/remove data, run and remove workflows, but cannot delete the project itself. You can only share with a user that has logged into Astrocyte at least once.

Share With User

biohpc-help@utsouthwestern.edu

© University of Texas Southwestern Medical Center

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 1(▼) entries

Search:

	File	Size
<input type="checkbox"/>	BCH/S085B.fastq.gz	690.9 MB
<input type="checkbox"/>	BCH/S132INB.fastq.gz	1.3 GB
<input type="checkbox"/>	BCH/S114B.fastq.gz	846.1 MB
<input type="checkbox"/>	BCH/S110B.fastq.gz	1.5 GB
<input type="checkbox"/>	BCH/S117B.fastq.gz	854.8 MB
<input type="checkbox"/>	BCH/design.txt	345 bytes
<input type="checkbox"/>	BCH/S118B.fastq.gz	863.2 MB
<input type="checkbox"/>	BCH/S089B.fastq.gz	849.7 MB
<input type="checkbox"/>	20170615_CPL127_4055/S228F_OE.fastq.gz	801.7 MB

Using BICF RNA-Seq Analysis Workflow on Astrocyte

The screenshot displays the Astrocyte 0.1.0 web interface. At the top, the header includes the UT Southwestern Medical Center logo, BioHPC branding, the version number 0.1.0, and a user login status 'Logged in as: ghenry'. A navigation bar contains links for 'Astrocyte Home', 'My Projects', 'Browse Workflows', and 'Documentation'. The main content area is titled 'Project 1226 - BCH_New' and shows the owner as 'Gervaise Henry (ghenry)' and the creation date as 'June 26, 2018, 12:13 p.m. by Gervaise Henry (ghenry)'. There are three main sections: 1. 'Input data in this project' with a button 'Add Data To This Project' and a message stating 'No input data has been added to this project. Please upload files to use them with a workflow.' 2. 'Workflows run in this project' with a button 'Run a workflow in this project' (highlighted by a black arrow) and a message stating 'You haven't run any workflows in this project. Upload some input data, and then click the 'Run Workflow' button above to begin.' 3. 'Sharing' with a text box for sharing access and a 'Share With User' button. The footer contains contact information: 'biohpc-help@utsouthwestern.edu' and '© University of Texas Southwestern Medical Center'.

UT Southwestern Medical Center
Lyda Hill Department of Bioinformatics

BioHPC

Astrocyte 0.1.0

Logged in as: ghenry

Astrocyte Home My Projects Browse Workflows Documentation

Project 1226 - BCH_New

Owner: Gervaise Henry (ghenry)

Created: June 26, 2018, 12:13 p.m. by Gervaise Henry (ghenry)

Input data in this project

To run a workflow against input data you need to upload it into this project. Click the button below to add new files from your web browser or the BioHPC cluster. You can also download or delete existing files from the project in the list below.

Add Data To This Project

No input data has been added to this project. Please upload files to use them with a workflow.

Workflows run in this project

Astrocyte provides many workflow created by different groups at UTSW for you to run against your data. To begin, make sure you have added input data into your project and then click the 'Run a workflow' button to choose a workflow to run.

Run a workflow in this project

You haven't run any workflows in this project. Upload some input data, and then click the 'Run Workflow' button above to begin.

Sharing

You can share access to this project with another user. Anyone you share the project with will be able to upload/remove data, run and remove workflows, but cannot delete the project itself. You can only share with a user that has logged into Astrocyte at least once.

Share With User

biohpc-help@utsouthwestern.edu
© University of Texas Southwestern Medical Center

Available Workflows

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.

The screenshot shows the 'Available Workflows' section of the BioHPC Astrocyte v1.0 interface. The table lists the following workflows:

Workflow Name	Description	Current Version	Contact	Try Workflow	Documentation	No Version
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.	0.1.0	Author: Subal Chak	Try Workflow	Documentation	No Version
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.	0.1.0	Author: Subal Chak	Try Workflow	Documentation	No Version
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.	0.1.0	Author: Subal Chak	Try Workflow	Documentation	No Version
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.	0.1.0	Author: Subal Chak	Try Workflow	Documentation	No Version
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.	0.1.0	Author: Subal Chak	Try Workflow	Documentation	No Version
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.	0.1.0	Author: Subal Chak	Try Workflow	Documentation	No Version
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.	0.1.0	Author: Subal Chak	Try Workflow	Documentation	No Version
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.	0.1.0	Author: Subal Chak	Try Workflow	Documentation	No Version
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.	0.1.0	Author: Subal Chak	Try Workflow	Documentation	No Version
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.	0.1.0	Author: Subal Chak	Try Workflow	Documentation	No Version

Using BICF RNA-Seq Analysis Workflow on Astrocyte

UTSouthwestern
Medical Center
Linda Yee Department of Bioinformatics

BioHPC

Astrocyte 0.1.0

Logged in as: ghenny

[Astrocyte Home](#) [My Projects](#) [Browse Workflows](#) [Documentation](#)

Running Workflow brandi.cantarel/rnaseq_nextflow.git (0.4.2)

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.

UTSouthwestern
Medical Center

BICF

This workflow provided by
UTSW Bioinformatics Core Facility

Parameters

Project

Project 1226: BCH_New

Name for this run

BCH

One or more input paired-end FASTQ files from a RNASeq experiment and a design file with the link between the same name and the sample group

S1108.fastq.gz
S0856.fastq.gz
S0858.fastq.gz
design.txt

In the case that the sequence libraries were generated using a stranded specific protocol:

Stranded

In single-end sequencing, the sequencer reads a fragment from only one end to the other, generating the sequence of base pairs. In paired-end reading it starts at one read, finishes this direction at the specified read length, and then starts another round of reading from the opposite end of the fragment.

Single End

Alignment tool

HiSAT2

Run Star Fusion

No Fusion Detection

Duplicate reads are defined as originating from the same original fragment of DNA. Duplicates are identified as read pairs having identical 5-prime positions (coordinate and strand) for both reads in a mate pair and optionally, matching unique molecular identifier reads.

Remove Duplicates

Runs deSeq2 and EdgeR

Run Statistical Analysis

A design file listing pairs of sample name and sample group. Columns must include: SampleID, SampleName, SampleGroup, FullPathToFqR1, FullPathToFqR2

design.txt

Reference genome for alignment

Human GRCh38

Gene Set Definitions used for GSEA Analysis -- see <http://software.broadinstitute.org/gsea/msigdb/> for gene set descriptions

Hallmark Gene Sets

Run Workflow

Using BICF RNA-Seq Analysis Workflow on Astrocyte

UT Southwestern Medical Center
Lyda Hill Department of Bioinformatics

BioHPC

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 RNA-Seq Analysis Workflow brandi.cantarelli/maiseq_newflow.gtf / 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	S132N8.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	default
stranded	1
markdup	picard
geneset	h.at.v5.1.symbol.gtf

Input Files

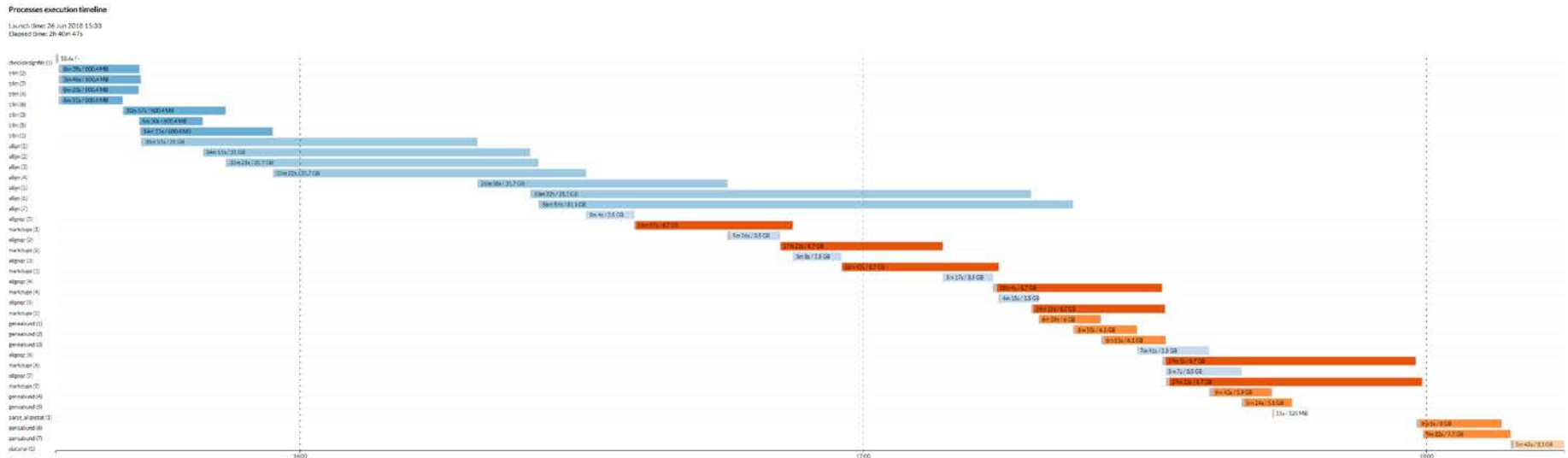
Filename	Size
design.txt	345 bytes
S132N8.fastq.gz	1.3 GB
S118B.fastq.gz	863.2 MB
S117B.fastq.gz	854.8 MB
S114B.fastq.gz	846.1 MB
S110B.fastq.gz	1.5 GB
S089B.fastq.gz	849.7 MB
S085B.fastq.gz	690.9 MB

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.

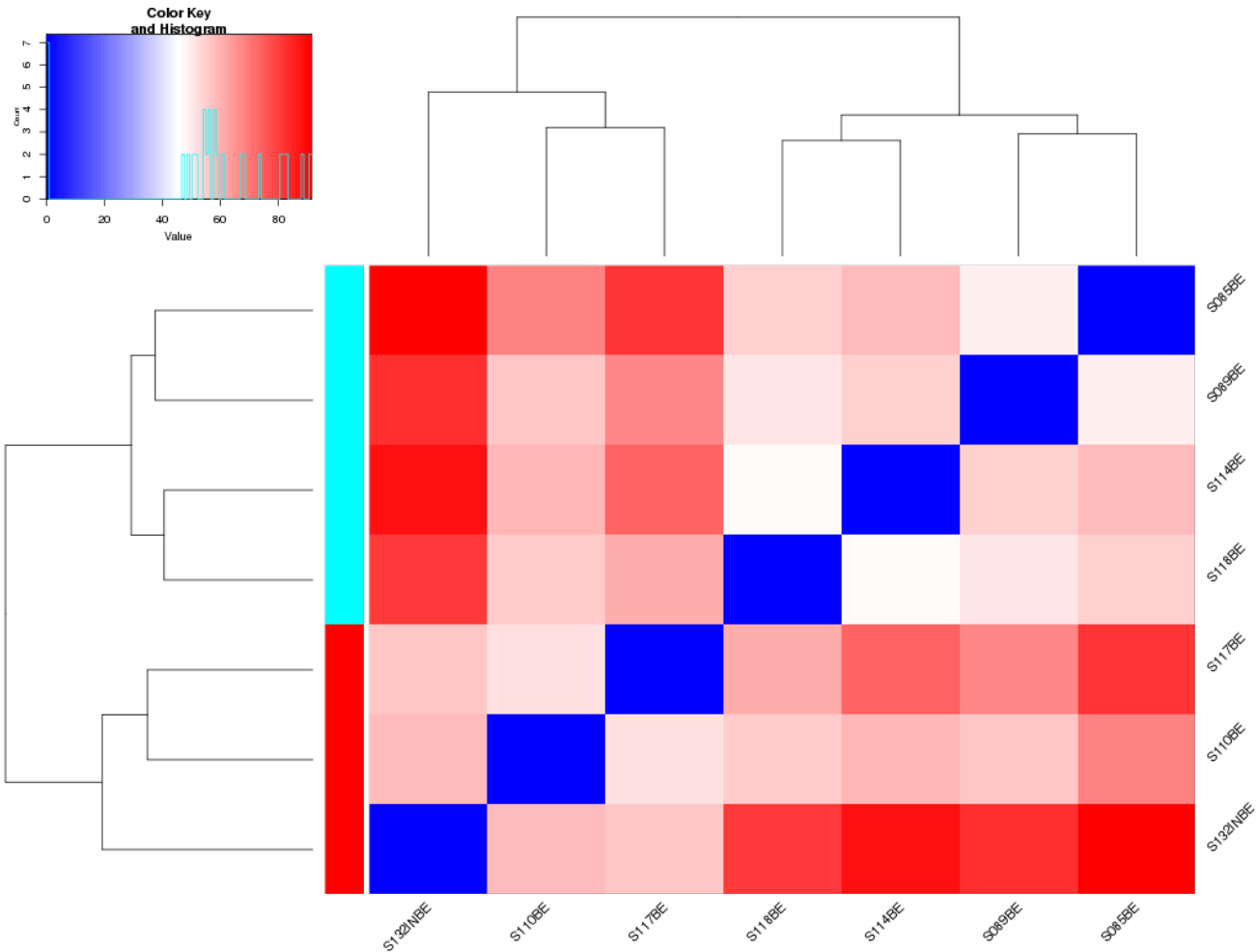
Current Directory: (/)

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

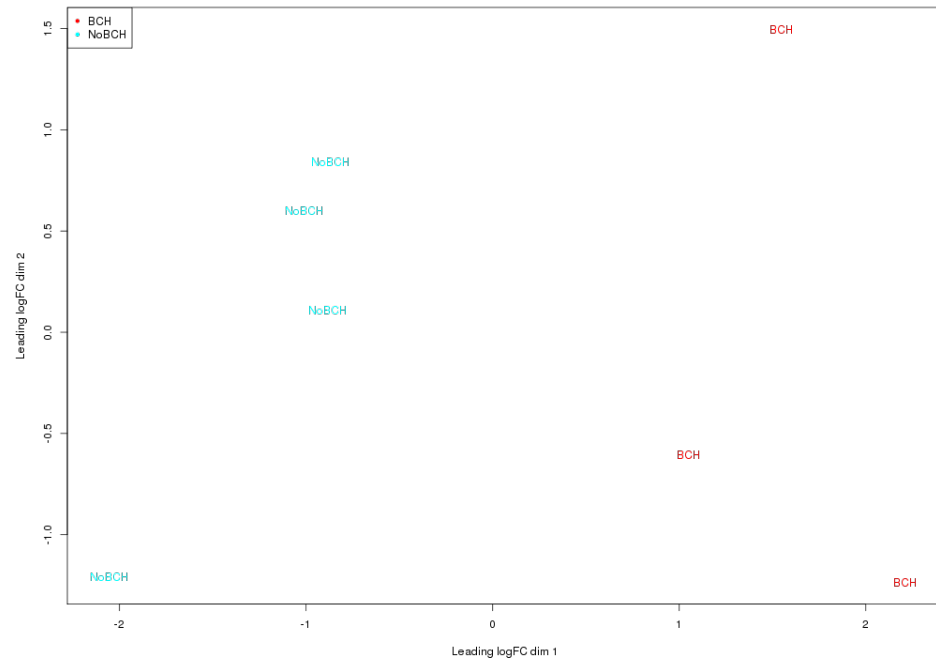
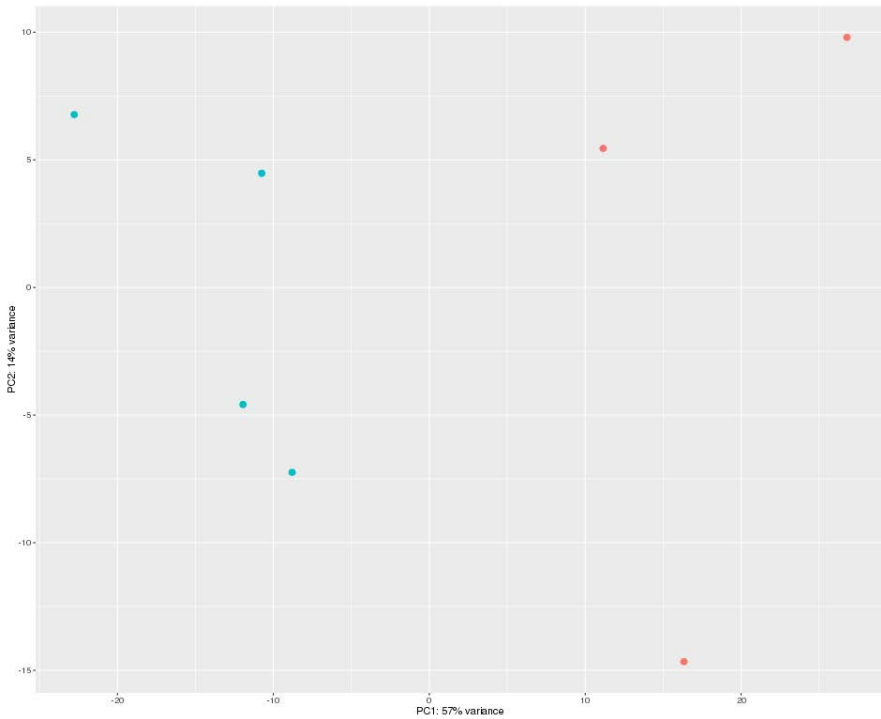
Results

											
S085BE_str ngtie	S089BE_str ngtie	S110BE_str ngtie	S114BE_str ngtie	S117BE_str ngtie	S118BE_str ngtie	S132INBE_s tringtie	alignment. summary.tx t	BCH_NoBC H.deseq2.t xt	BCH_NoBC H.edgeR.txt	BCH_NoBC H.heatmap. deseq2.pn g	BCH_NoBC H.heatmap. edgeR.png
											
BCH_NoBC H.qusage.r da	bg.rda	countTable. fpm.txt	countTable. logCPM.txt	countTable. stats.txt	countTable. txt	design.shin y.txt	design.vali d.txt	edgeR.resu lts.txt	geneset.shi ny.gmt	mds.png	pca.png
											
S085BE.alig nerout.txt	S085BE.ba m	S085BE.cts	S085BE.de dup.bam	S085BE fla gstat.txt	S085BE.fpk m.txt	S085BE_fas tqc.html	S085BE_fas tqc.zip	S089BE.alig nerout.txt	S089BE.ba m	S089BE.cts	S089BE.de dup.bam
											
S089BE fla gstat.txt	S089BE.fpk m.txt	S089BE_fas tqc.html	S089BE_fas tqc.zip	S110BE.alig nerout.txt	S110BE.ba m	S110BE.cts	S110BE.de dup.bam	S110BE fla gstat.txt	S110BE.fpk m.txt	S110BE_fas tqc.html	S110BE_fas tqc.zip
											
S114BE.alig nerout.txt	S114BE.ba m	S114BE.cts	S114BE.de dup.bam	S114BE fla gstat.txt	S114BE.fpk m.txt	S114BE_fas tqc.html	S114BE_fas tqc.zip	S117BE.alig nerout.txt	S117BE.ba m	S117BE.cts	S117BE.de dup.bam
											
S117BE fla gstat.txt	S117BE.fpk m.txt	S117BE_fas tqc.html	S117BE_fas tqc.zip	S118BE.alig nerout.txt	S118BE.ba m	S118BE.cts	S118BE.de dup.bam	S118BE fla gstat.txt	S118BE.fpk m.txt	S118BE_fas tqc.html	S118BE_fas tqc.zip
											
S132INBE.a lignerout.tx t	S132INBE.b am	S132INBE.c ts	S132INBE.d edup.bam	S132INBE.fl agstat.txt	S132INBE.f pkm.txt	S132INBE_f astqc.html	S132INBE_f astqc.zip	samples_h eatmap.pn g			

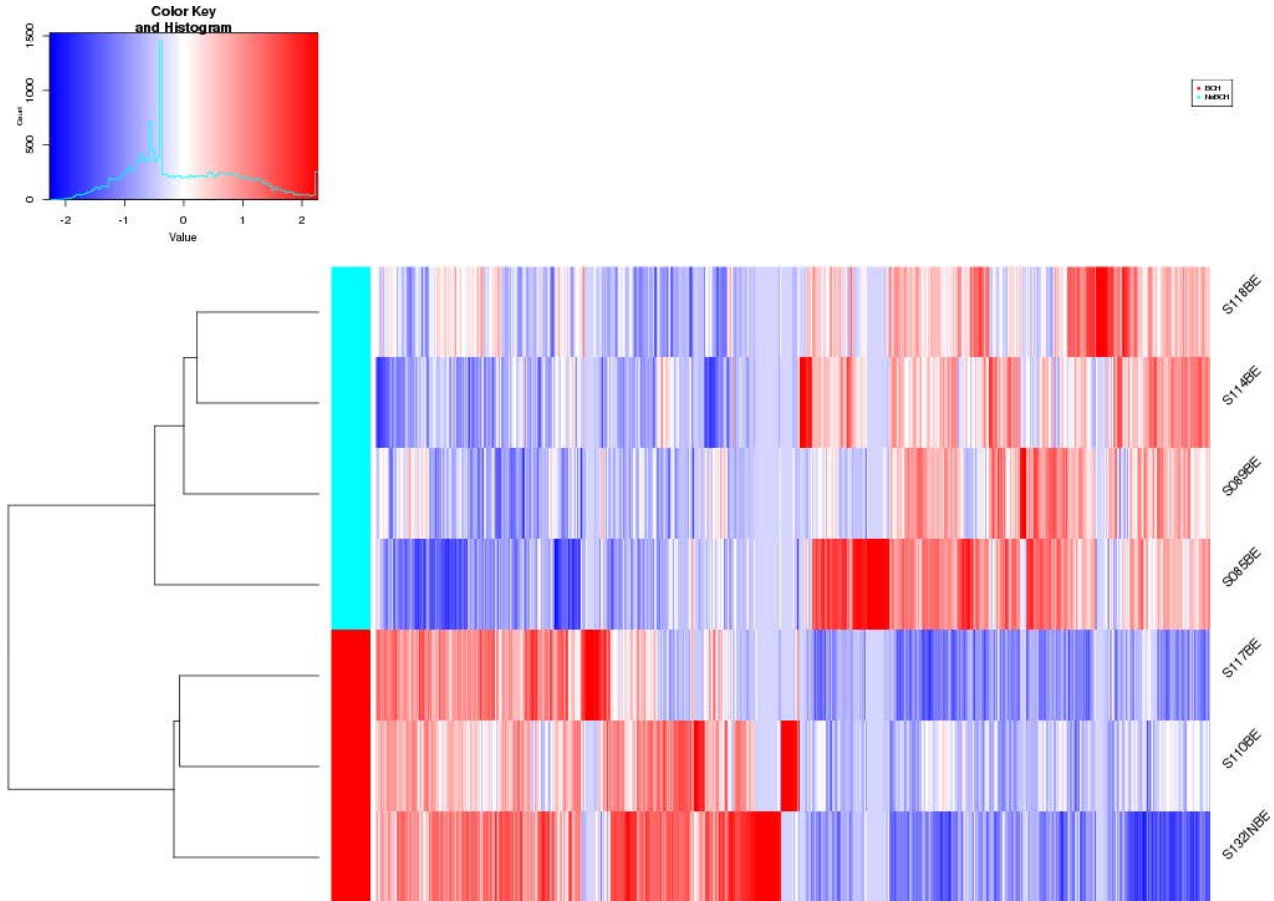
Results



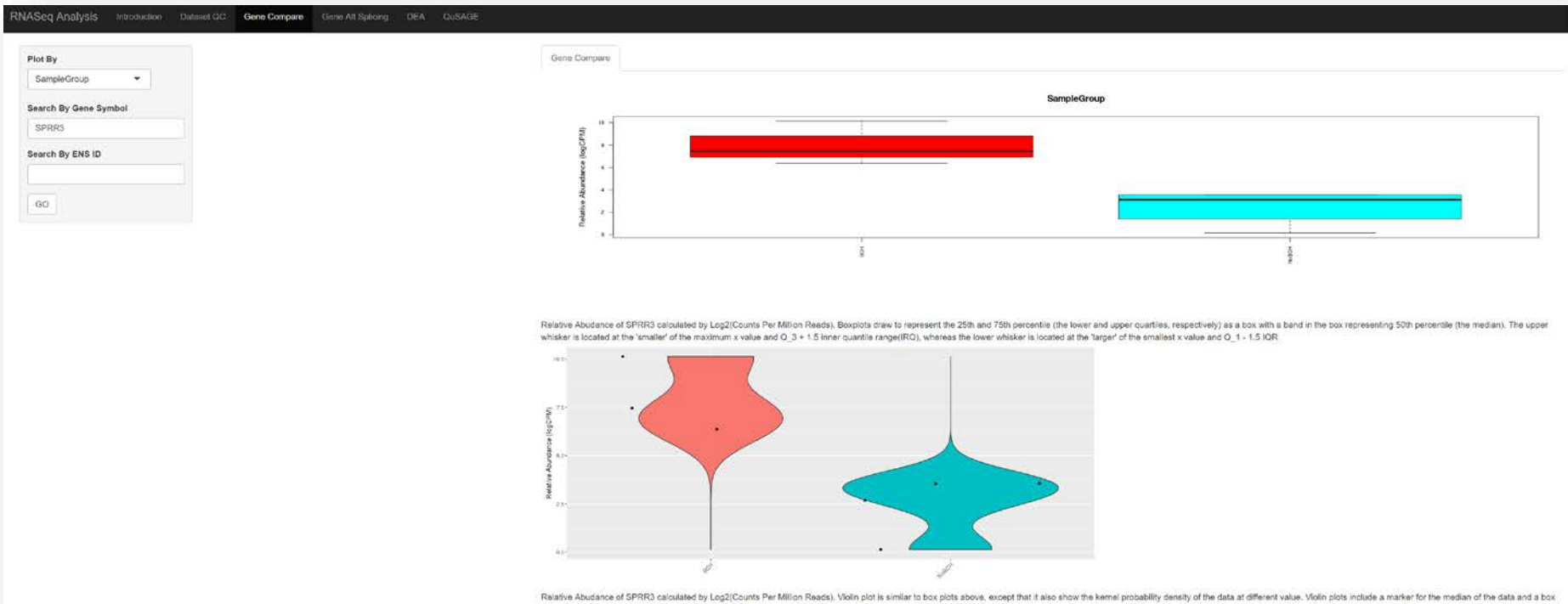
Results



Results



Results

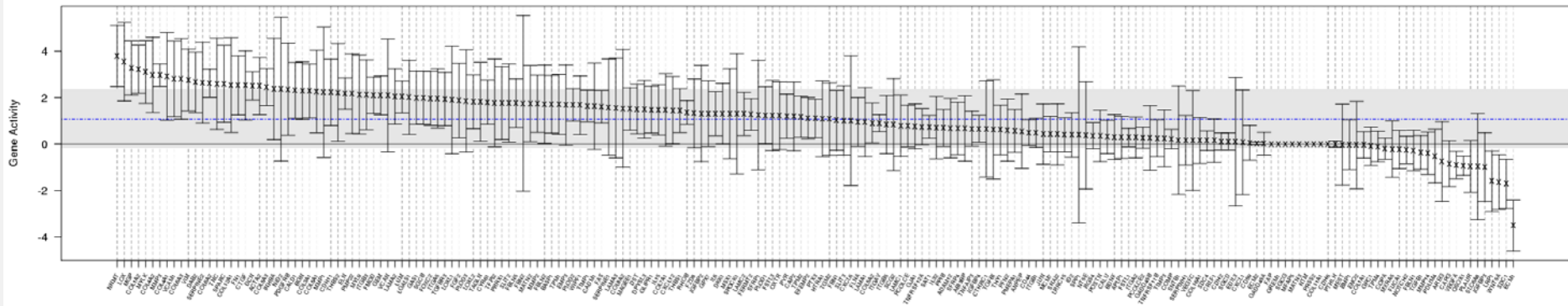


Results

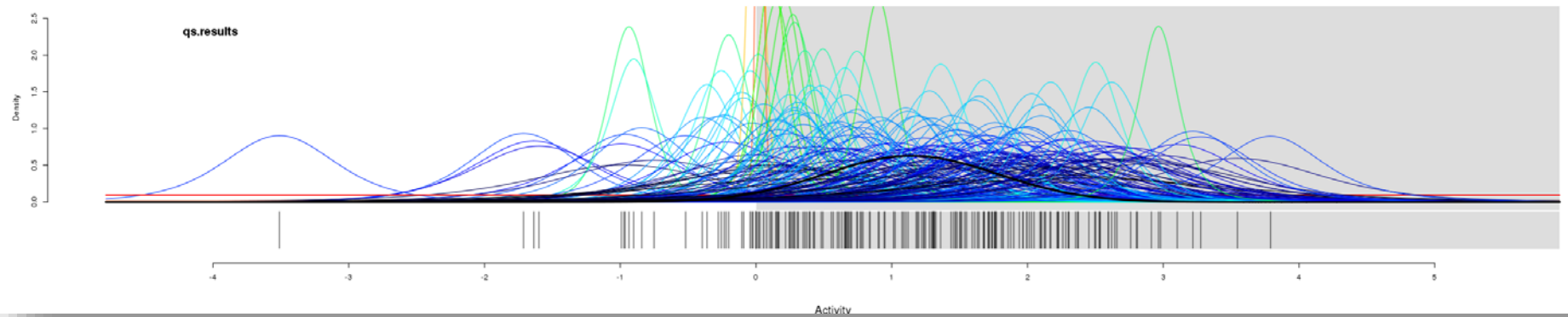
Gene Set Comparisons

Gene Set Comparison

EPITHELIAL_MESENCHYMAL_TRANSITION



qs.results



Under the hood

```
2 library(edgeR)
3 library(DESeq2)
4 library("RColorBrewer")
5 library("gplots")
6 library(qusage)
```

Under the hood

```
23 ##### Read in Data #####
24 genenames <- read.table(file="genenames.txt",header=TRUE,sep='\t')
25
26 tbl <- read.table('countTable.txt',header=TRUE,sep="\t")
27 tbl2 <- read.table('countTable.logCPM.txt',header=TRUE,sep="\t")
28 ct <- tbl[,4:length(tbl)]
29 row.names(ct) <- tbl$ENSEMBL
30
31 samples<- names(ct)
```

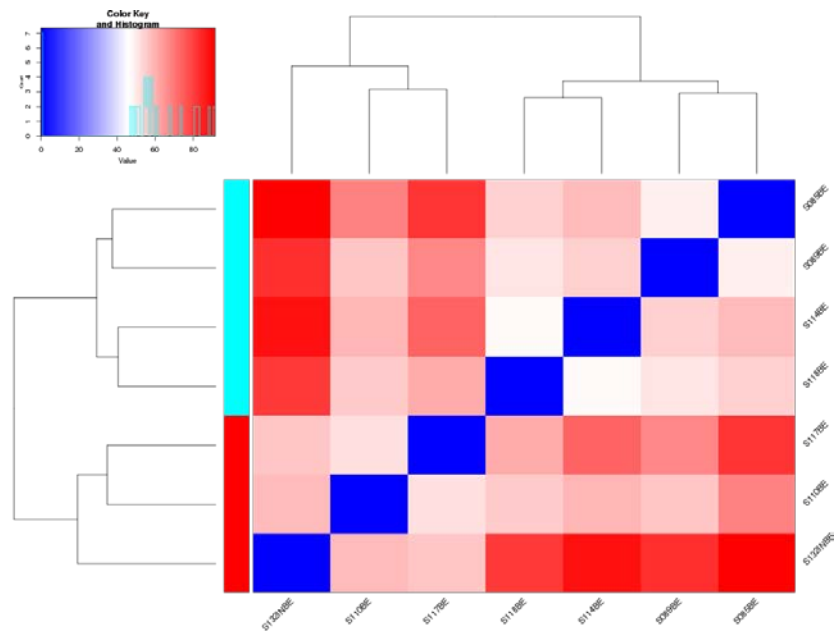
Under the hood

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


Under the hood

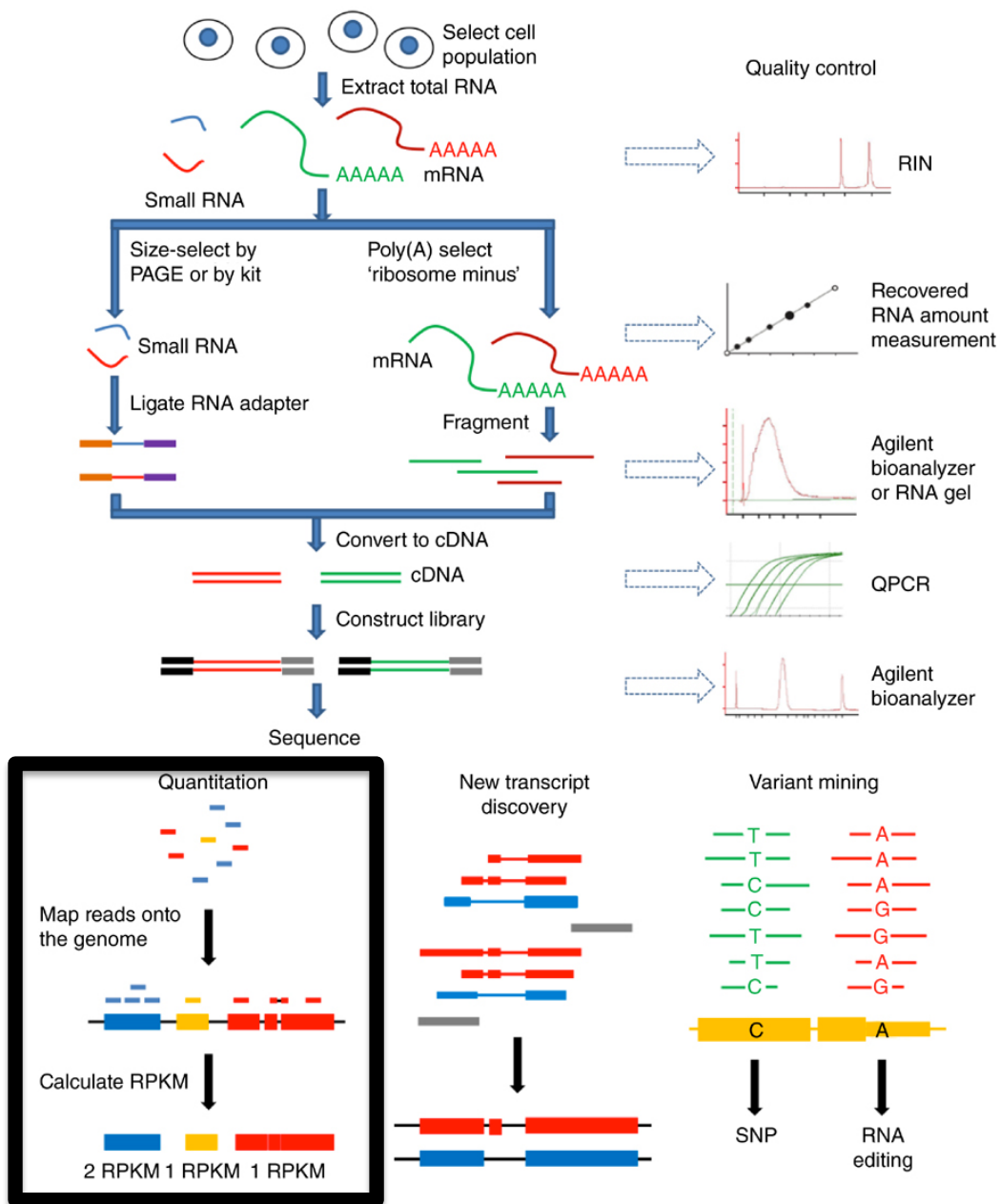
```
png(file="samples_heatmap.png",bg ="transparent",height=768,width=1024)
heatmap.2(as.matrix(sampleDists), col = bluered(100),RowSideColors = col.blocks,srtRow=45,srtCol=45,trace="none", margins=c(5, 5))
dev.off()

#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()
```





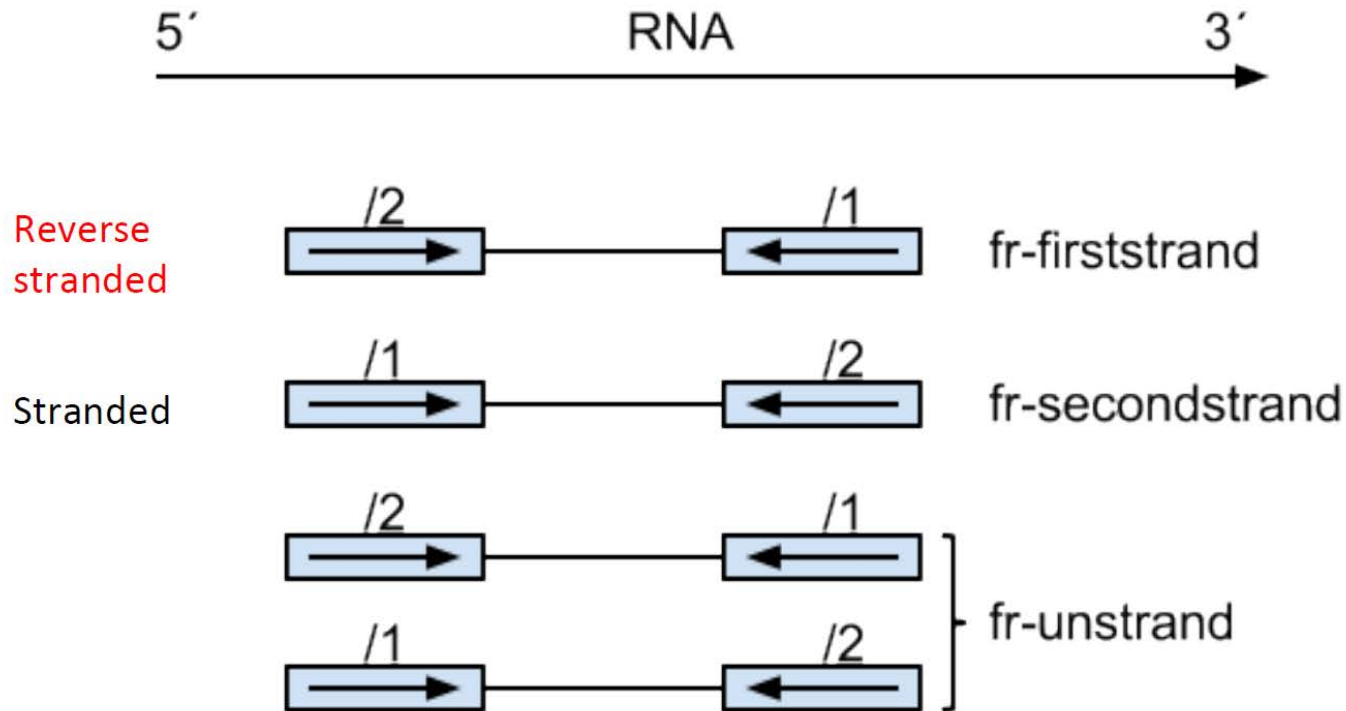
ABOUT RNA-SEQ



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