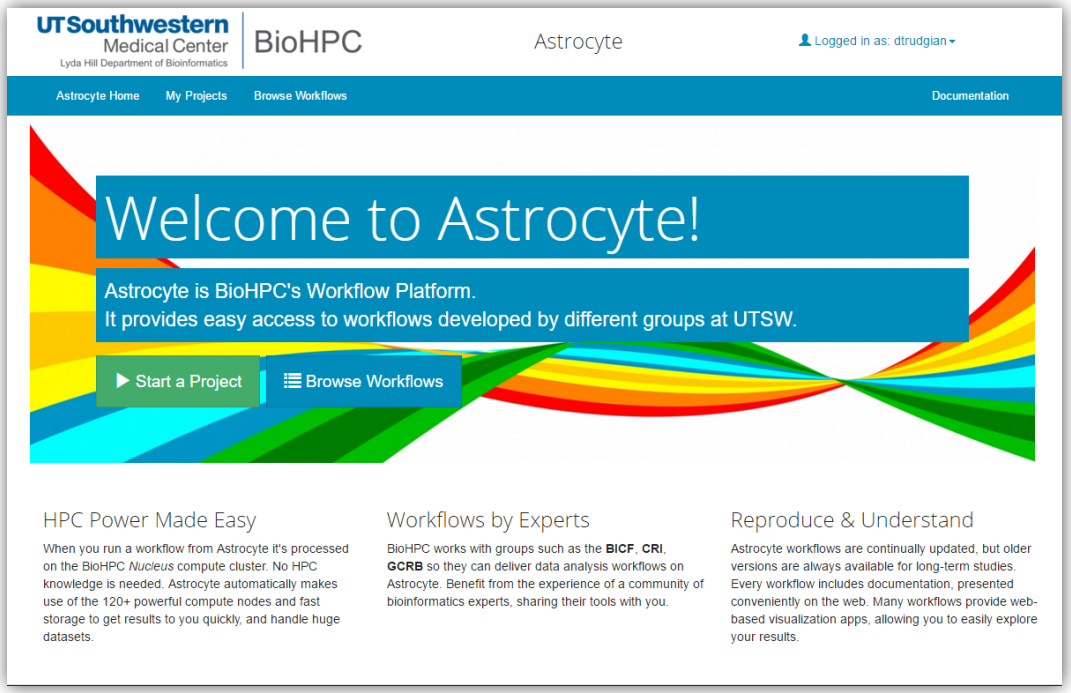
Astrocyte – BioHPC Workflow Platform

Allows groups to give easy-access to their analysis pipelines via the web



Standardized Workflows

Simple Web Forms

Online documentation & results visualization*

Workflows run on HPC cluster without developer or user needing cluster knowledge

astrocyte.biohpc.swmed.edu

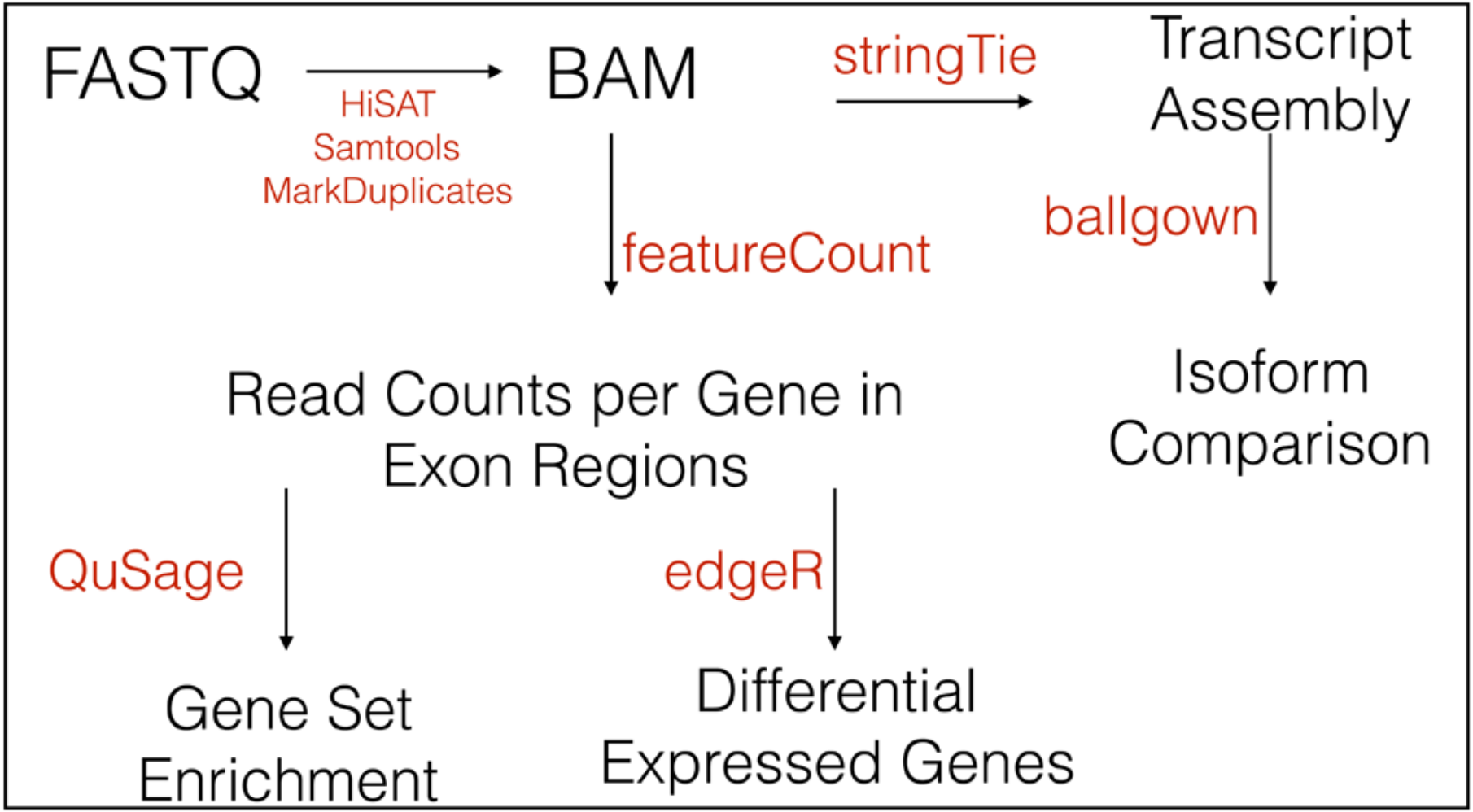
Slide contribution: David Trudgian@BioHPC

Browse workflows

Available Workflows

CHILDREN'S MEDICAL CENTER RESEARCH INSTITUTE AT UT SOUTHWESTERN	**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 **Conatact:** 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 occurences of words in a test 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 **Conatact:** biohpc-help@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 a simple RNASeq analysis workflow using TrimGalore, HiSAT,FeatureCounts, StringTie and statistical analysis using EdgeR and Ballgown, plus a simple R Shiny visualization application.	**Current Version:** rnaseq_bicf - 0.1.0 **Author:** Brandi Cantarel **Conatact:** 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 - 0.0.1 **Author:** Brandi Cantarel **Conatact:** 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 TrimGalore, BWA, Speedseq, GATK, Samtools and Platypus. SNPs and Indels are integrated using BAYSIC; then annotated using SNPEFF and SnpSift.	**Current Version:** germline_bicf - 0.0.7 **Author:** Brandi Cantarel **Conatact:** biohpc-help@utsouthwestern.edu	Run Workflow Documentation All Versions
UTSouthwestern Medical Center | BioHPC	**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 - 0.0.4 **Author:** GCRB **Conatact:** biohpc-help@utsouthwestern.edu	Run Workflow Documentation All Versions

RNASeq Analysis Pipeline



http://www.utsouthwestern.edu/labs/bioinformatics/services/data-analysis/rnaseq-pipeline.html

RNAseq Analysis Essence

- Preprocessing and normalization
- Differential gene expression analysis
- QC
- Visualization
- Pathway and gene sets enrichment analysis
- Different splicing isoforms
- Fusion and variants

Create a new project

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

Project Name | Create New Project

Existing Projects

ID	Name	Created	Workflows Run	Input Files	Size	Actions
PRJ21	RNAseq_test	Aug. 23, 2016, 3:03 p.m.	0	0	0 bytes	

Projects Shared with Me

ID	Name	Created	Workflows Run	Input Files	Size	Actions
PRJ10	test	June 1, 2016, 5:02 p.m. by Brandi Cantarel	4	10	218.5 GB	

Add data to your project

Project 21 - RNAseq_test

Owner: bchen4

Created: Aug. 23, 2016, 3:03 p.m. by bchen4

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

Share With User

Shared With

Add data to your project

Upload files from the web

You can upload any size of 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/bchen4** on BioHPC to import them into your project directly.

Import Selected Files | Finished importing files

For NGS experiment, this is recommended.

Search:

	File	Size
☐	KO3_R2.fastq	4.4 GB
☑	WT1_R1.fastq	4.0 GB
☑	WT2_R1.fastq	4.1 GB
☐	KO4_R2.fastq	4.5 GB
☐	KO2_R1.fastq	4.0 GB
☐	WT2_R2.fastq	4.1 GB
☐	KO2_R2.fastq	4.0 GB
☐	KO4_R1.fastq	4.5 GB
☐	WT1_R2.fastq	4.0 GB
☐	KO3_R1.fastq	4.4 GB

Showing 1 to 10 of 10 entries 2 rows selected

Previous 1 Next

Select all | Deselect all

Make your design file

SampleID	SampleGroup	SubjectID	SampleName	FullPathToFqR1	FullPathToFqR2
SRR1551069	monocytes	53	53_Monocytes	SRR1551069_1.fastq.gz	SRR1551069_2.fastq.gz
SRR1551068	neutrophils	53	53_Neutrophils	SRR1551068_1.fastq.gz	SRR1551068_2.fastq.gz
SRR1551055	monocytes	21	21_Monocytes	SRR1551055_1.fastq.gz	SRR1551055_2.fastq.gz
SRR1551054	neutrophils	21	21_Neutrophils	SRR1551054_1.fastq.gz	SRR1551054_2.fastq.gz
SRR1551048	monocytes	20	20_Monocytes	SRR1551048_1.fastq.gz	SRR1551048_2.fastq.gz
SRR1551047	neutrophils	20	20_Neutrophils	SRR1551047_1.fastq.gz	SRR1551047_2.fastq.gz
SRR1550987	monocytes	44	44_Monocytes	SRR1550987_1.fastq.gz	SRR1550987_2.fastq.gz
SRR1550986	neutrophils	44	44_Neutrophils	SRR1550986_1.fastq.gz	SRR1550986_2.fastq.gz

```
SampleID
    This ID should match the name in the fastq file ie S0001.R1.fastq.gz the sample ID is S0001
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
FullPathToFqR1
Name of the fastq file R1
FullPathToFqR2
Name of the fastq file R2
```

Make your design file

- 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

Select your data files and set up workflow and submit

Project

Project 28: RNASeqTest

Name for this run

test_0.1.1

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

SRR1550987_1.fastq.gz
SRR1550986_2.fastq.gz
SRR1550986_1.fastq.gz
SRR1551069_2.fastq.gz
SRR1551069_1.fastq.gz

SELECT YOUR FILES

In the case that the sequence libraries where generated using a stranded specific protocol.

Unstranded

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.

Paired End

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

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

design.pe.txt

Reference genome for alignment

Human GRCh38

Gene Set Definitions used for QuSAGE Analysis -- see http://software.broadinstitute.org/gsea/msigdb/ for geneset descriptions

Hallmark Gene Sets

Run Workflow

http://software.broadinstitute.org/gsea/msigdb/index.jsp

Project is running

Run 'test0906_3' in Project 'test'

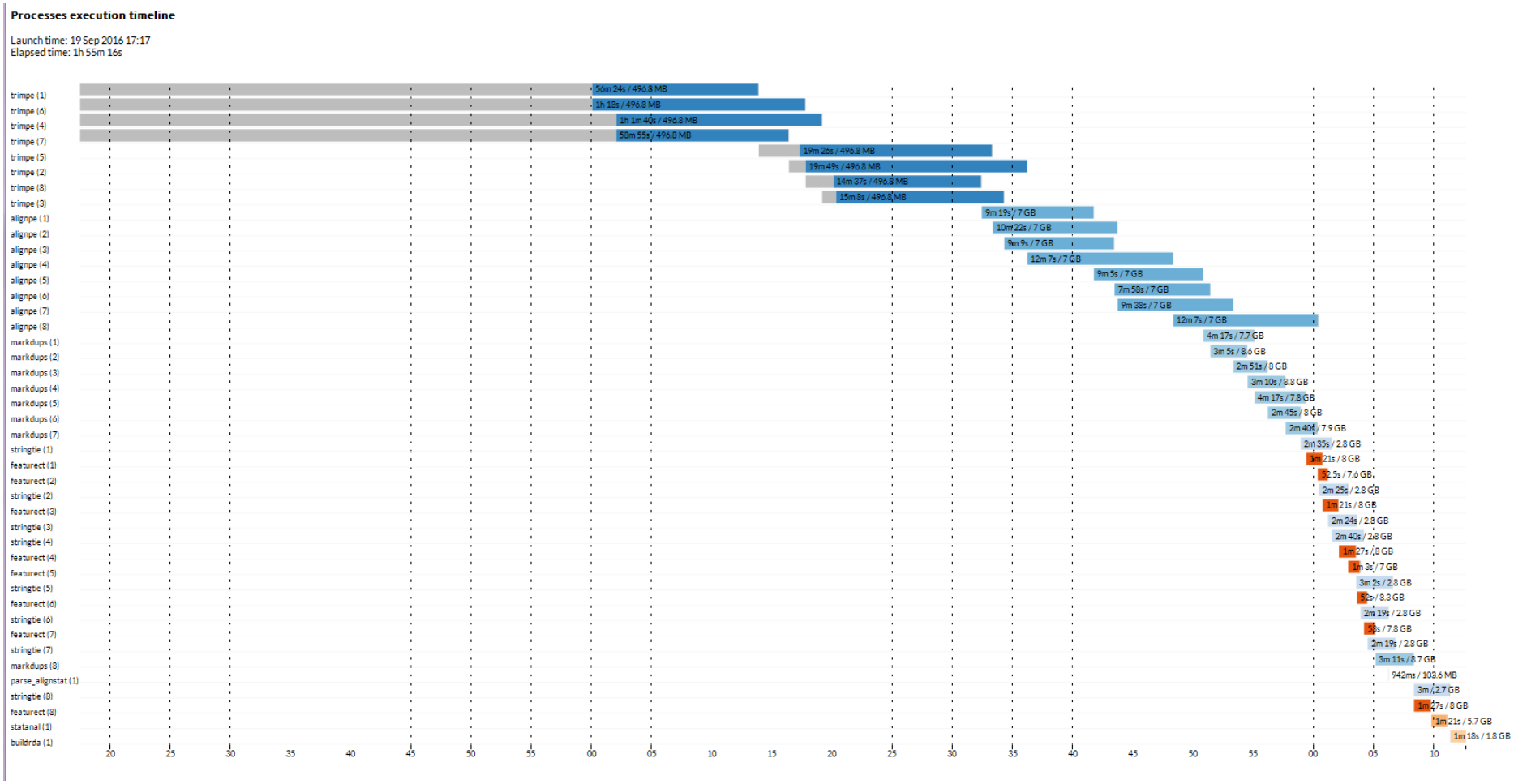
Run Information

Running Workflow	BICF RNASeq Analysis Workflow brandi.cantarel/rnaseq_nextflow.git / 0.0.13
Status	RUNNING
Created	Sept. 6, 2016, 9:28 p.m. by bchen4
Size	1.8 MB

Parameters

Parameter	Value
design	design.pe.txt
genome	/project/shared/bicf_workflow_ref/GRCh38
pairs	pe
fastqs	SRR1551054_1.fastq.gz
fastqs	SRR1551054_2.fastq.gz
fastqs	SRR1551055_1.fastq.gz
fastqs	SRR1551055_2.fastq.gz
fastqs	SRR1551068_1.fastq.gz
fastqs	SRR1551068_2.fastq.gz
fastqs	SRR1551069_1.fastq.gz
fastqs	SRR1551069_2.fastq.gz
markdups	mark

Timeline of the whole run



Download/visualize your 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/bchen4

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

Vizapp need about 30s to start if there is no queue. You need to refresh the page.

Output Browser

- geneset.shiny.gmt (46.4 KB)
- SRR1551054.bam (1.6 GB)
- SRR1551048.bam (1.4 GB)
- SRR1551054.flagstat.txt (444 bytes)
- SRR1551055.cts (9.8 MB)
- SRR1550987_fastqc.html (322.6 KB)
- SRR1551054.hisatout.txt (832 bytes)
- SRR1551069.flagstat.txt (443 bytes)
- SRR1551069.cts (9.8 MB)
- countTable.stats.txt (15.9 KB)
- pca.png (13.9 KB)
- SRR1551054_fastqc.zip (425.0 KB)

You can also choose individual files to download to your local computer

Comparisons

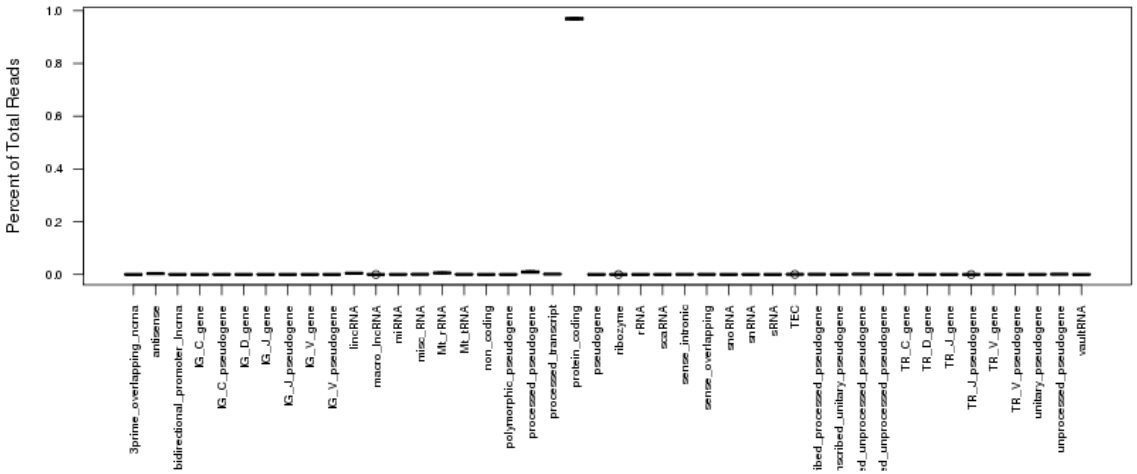
- Comparisons are based on SampleGroup
  - All pair-wise comparisons
  - Could be identified by file name
    - A_B.edgeR.txt
    - Log fold change will be A/B
    - If you want B/A, -1*logFC

Vizapp: QC general stat

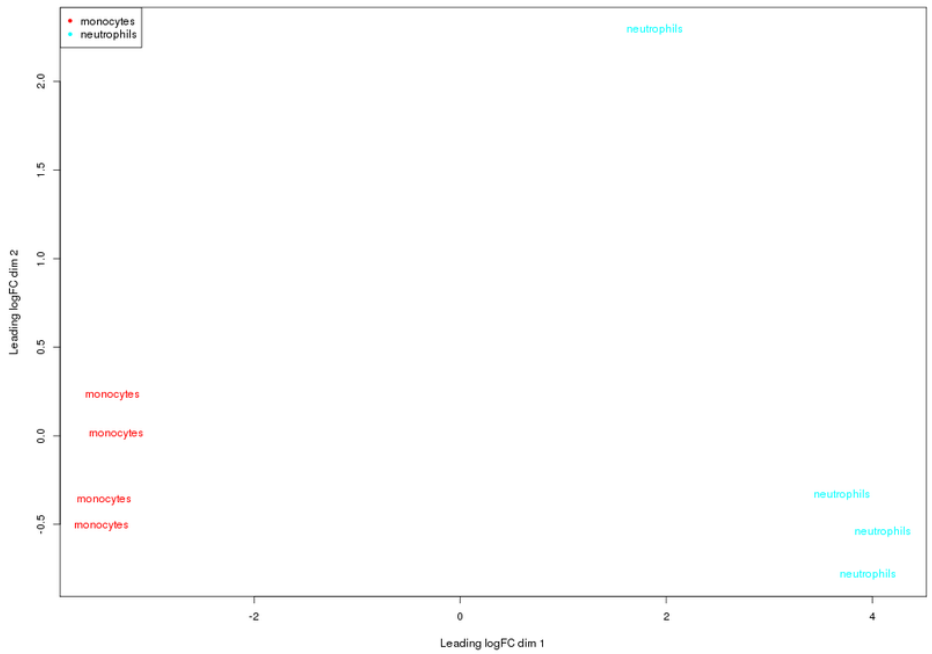
	Sample	Type	ReadPerc	ReadCt	TotalReads
	All	["protein_coding"]	All	All	All
28	SRR1550986	protein_coding	0.972	22464205	23104512
72	SRR1550987	protein_coding	0.965	25377897	26289950
116	SRR1551047	protein_coding	0.974	28651979	29406536
160	SRR1551048	protein_coding	0.967	25645837	26512740
204	SRR1551054	protein_coding	0.974	29351633	30149319
248	SRR1551055	protein_coding	0.966	24706269	25587382
292	SRR1551068	protein_coding	0.972	29958958	30820979
336	SRR1551069	protein_coding	0.966	22278607	23074264

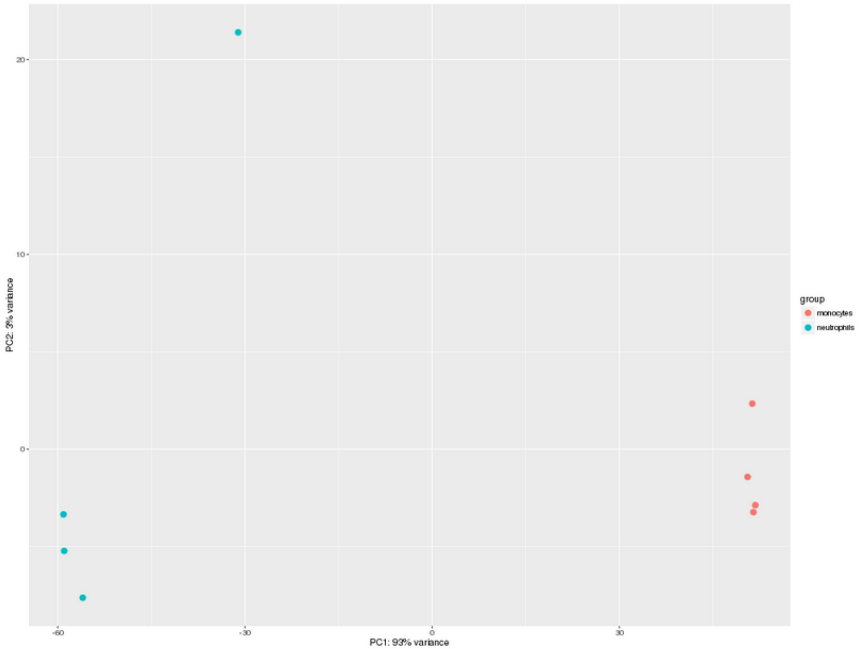
Showing 1 to 8 of 8 entries (filtered from 352 total entries)

Previous 1 Next



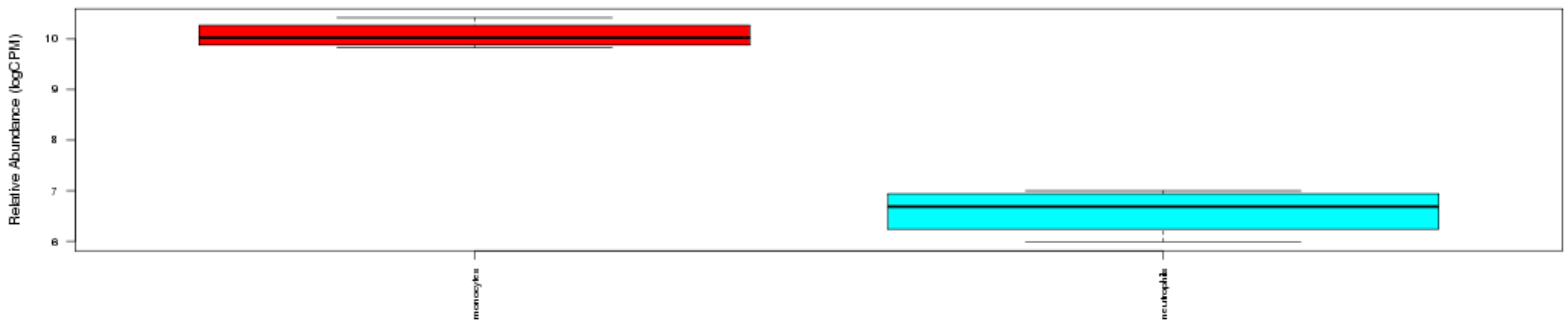
Vizapp: QC MSD and PCA



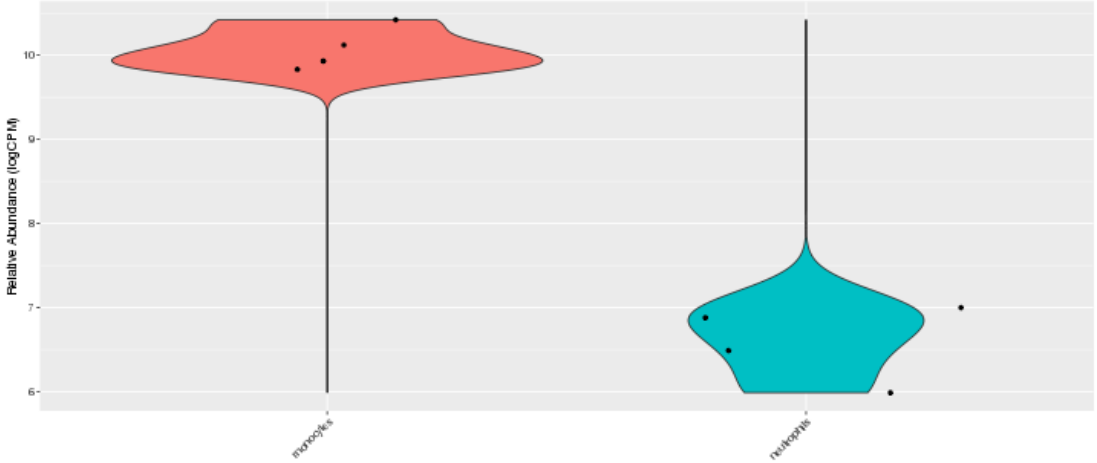


Vizapp: Gene Compare

Gene Compare



Relative Abudance of IL1B 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 Abudance of IL1B 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 indicating the interquartile range, as in boxplot above.

Vizapp: DEA

Show 10 entries Search:

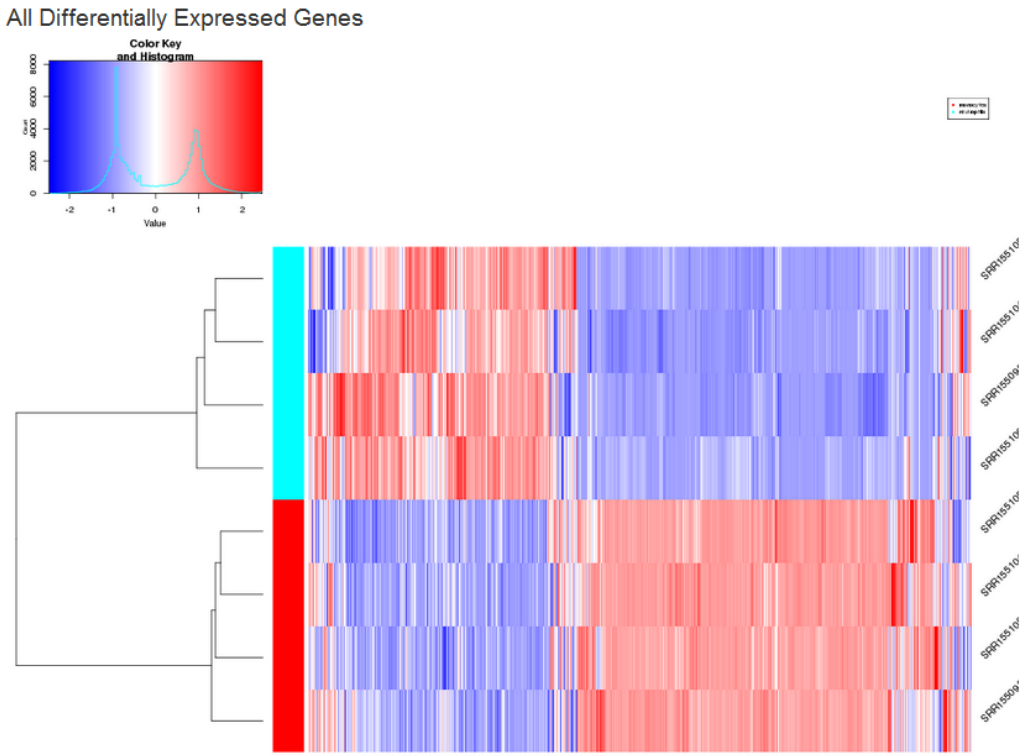
	symbol	ensembl	chrom	start	end	type	logFC	logCPM	PValue	monocytes	neutrophils	rawP	fdr	bonf
		All				All	All	All	All	All	All	All	All	All
1	A1BG-AS1	ENSG00000268895	chr19	58347751	58355183	antisense	3.65721935316179	0.123039159086969	6.54535890771608e-14	1.19	0	6.54535890771608e-14	2.43018562147972e-13	8.63201932749597e-10
2	A3GALT2	ENSG00000184389	chr1	33306766	33321098	protein_coding	-3.43466819993849	0.271181524758484	2.72822682115086e-15	0	0.68	2.72822682115086e-15	1.13896344784228e-14	3.59798553173375e-11
3	AAAS	ENSG00000094914	chr12	53307456	53324864	protein_coding	1.05698637142307	3.79650933915081	0.00163235320859952	2.52	1.85	0.00163235320859952	0.00251166422996272	1
4	AACS	ENSG00000081760	chr12	125065379	125143333	protein_coding	1.48791521833416	2.45474643479604	0.0000198772329606549	2.16	0.97	0.0000198772329606549	0.0000369108628956796	0.262140948285116
5	AAED1	ENSG00000158122	chr9	96639577	96655303	protein_coding	5.45806434600363	3.22181280065653	4.28638853670029e-37	2.5	0	4.28638853670029e-37	1.24239323125282e-35	5.65288920220034e-33
8	AAMP	ENSG00000127837	chr2	218264123	218270257	protein_coding	1.95818108290593	5.19726071074492	8.78068291199973e-9	2.88	2.13	8.78068291199973e-9	2.17504970404682e-8	0.000115799646243452
9	AANAT	ENSG00000129673	chr17	76453351	76470117	protein_coding	-1.12932073254575	0.970550145754216	0.00191886678302253	0.79	0.86	0.00191886678302253	0.00292724293053802	1
10	AAR2	ENSG00000131043	chr20	36236459	36270918	protein_coding	2.8817930888318	3.52693374975034	1.28882384659221e-15	2.54	0.89	1.28882384659221e-15	5.53108001589914e-15	1.69970088888581e-11
11	AARS	ENSG00000090861	chr16	70252295	70289543	protein_coding	3.55049814703604	2.74145701251323	2.2831883116971e-20	2.34	0.23	2.2831883116971e-20	1.49210542391781e-19	3.01106874546614e-16
12	AARS2	ENSG00000124608	chr6	44299654	44313326	protein_coding	3.46727438389513	1.97481858117748	1.467279578421e-18	2.1	0	1.467279578421e-18	8.16820729430821e-18	1.93504830802162e-14

Showing 1 to 10 of 8,899 entries Previous 1 2 3 4 5 ... 890 Next

GetCSV-Comp

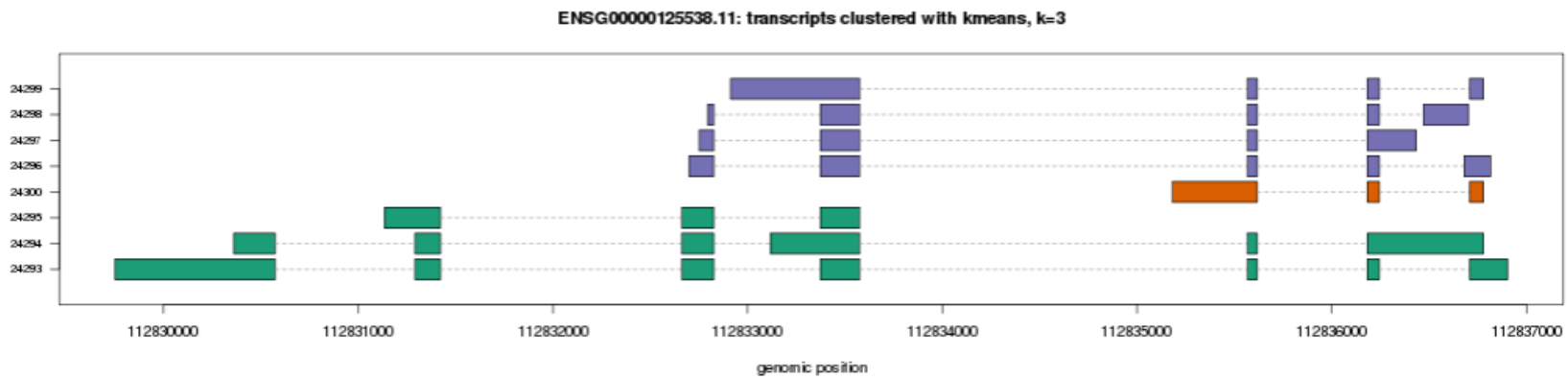
- Uses edgeR results
- Filter gene list by different parameters
- Sort by different columns
- Data table downloading

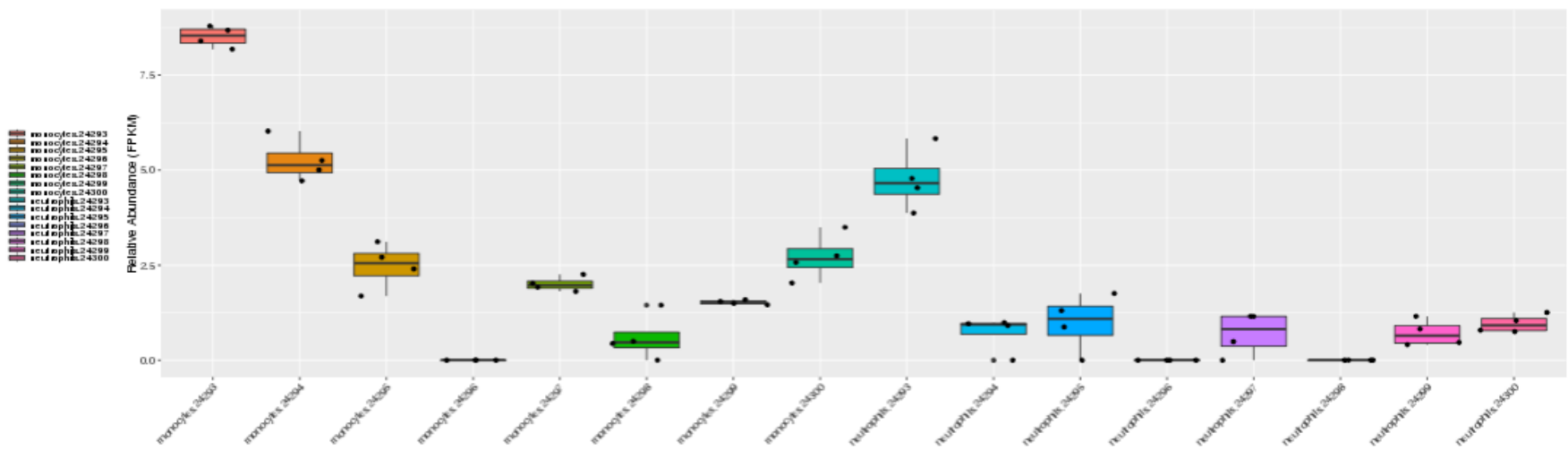
Vizapp: DEA heatmap



- Filter gene list by different parameters
- Choose different comparisons
- Support user define gene list (gene official symbol)
- Support pathway

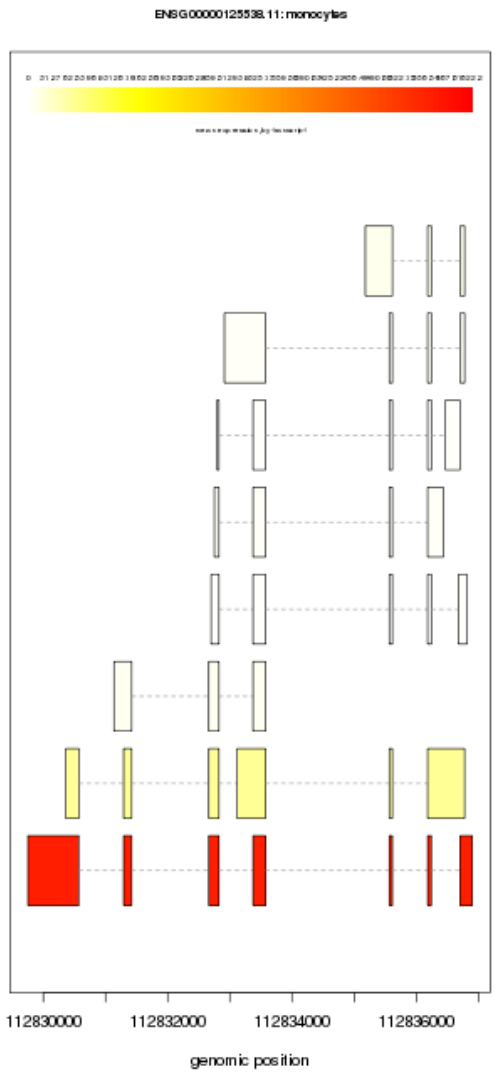
Vizapp: alternative splicing

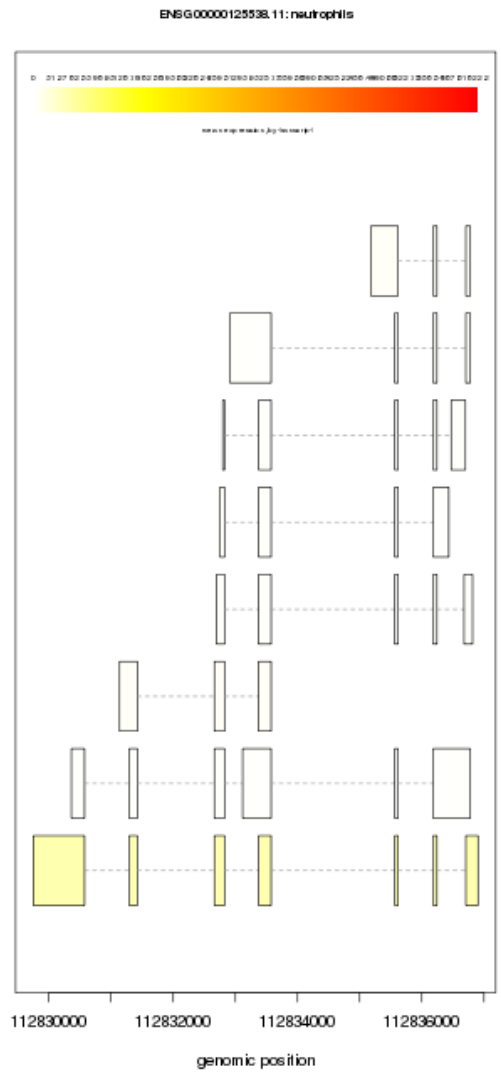




Different transcripts' expression in sample groups

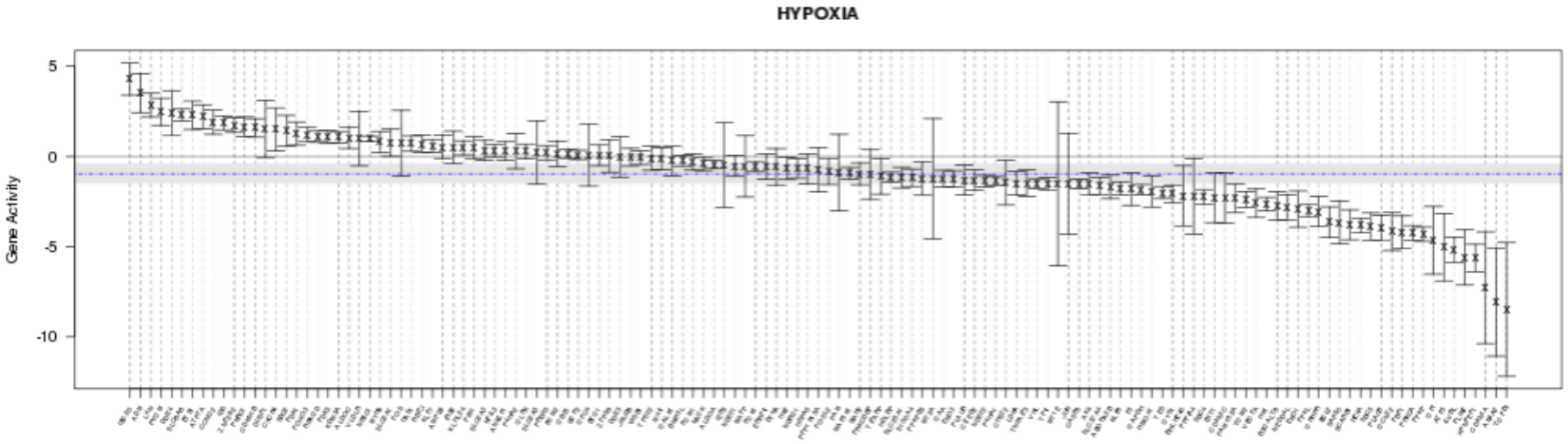
Vizapp: alternative splicing

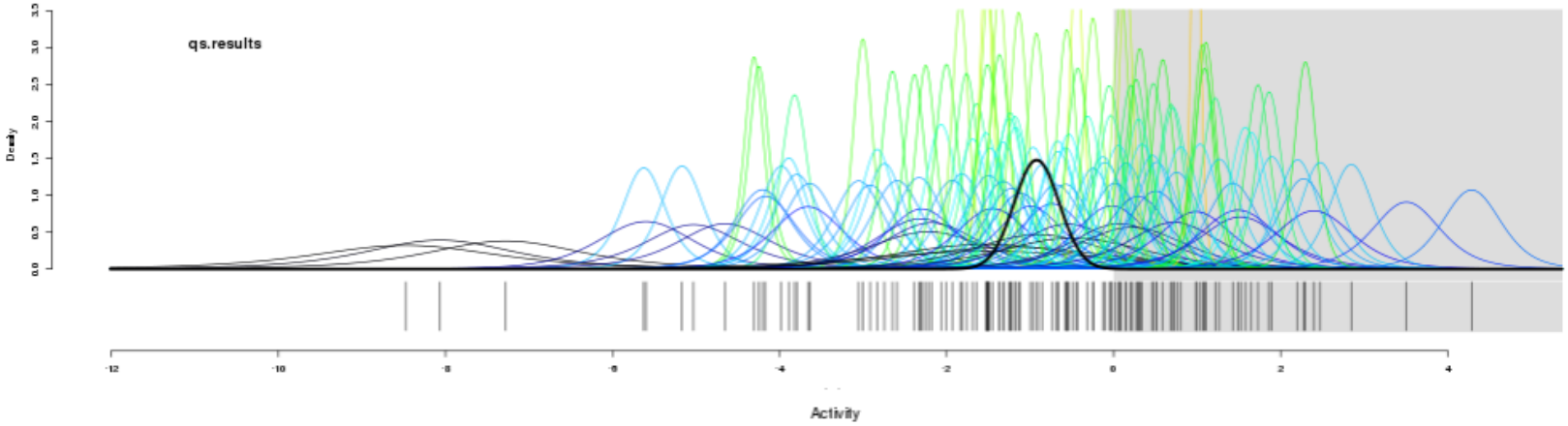




Vizapp: QuSAGE

Gene Set Comparisons | Gene Set Comparison





Common errors and solutions

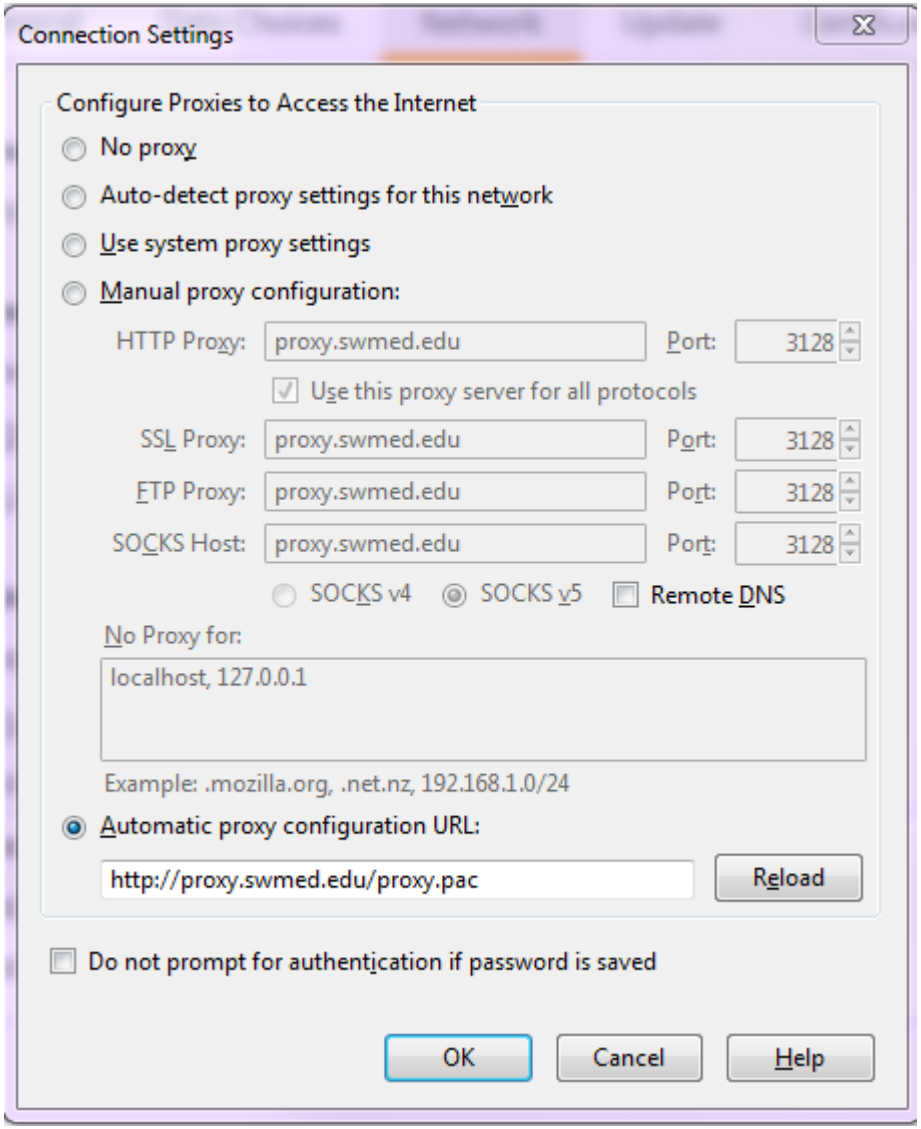
```
        Error running workflow. Diagnostic output

N E X T F L O W  ~  version 0.20.1
Launching main.nf
Didn't match any input files with entries in the design file

 -- Check script 'main.nf' at line: 49 or see '.nextflow.log' file for more details
```

- Make sure the delimiter is tab
- Make sure the column name are the same as mentioned in documentation
- Make sure the file names match

Common errors and solutions



- Not all files are uploaded
- It's about the proxy setting
- Use auto-detect proxy

Additional website for more options on data report

- Gene Set Enrichment Analysis (GSEA)

http://software.broadinstitute.org/gsea/index.jsp

MSigDB

http://software.broadinstitute.org/gsea/msigdb/index.jsp

Gene Pattern

http://software.broadinstitute.org/cancer/software/genepattern/

- User designed specific heatmaps by Morpheus

https://software.broadinstitute.org/morpheus/

- Complex designs

Factorial designs in edgeR or DEseq from countTable.csv

- Motif search/promoter analysis with Homer motif search

Different regulated gene list (edgeR.result)