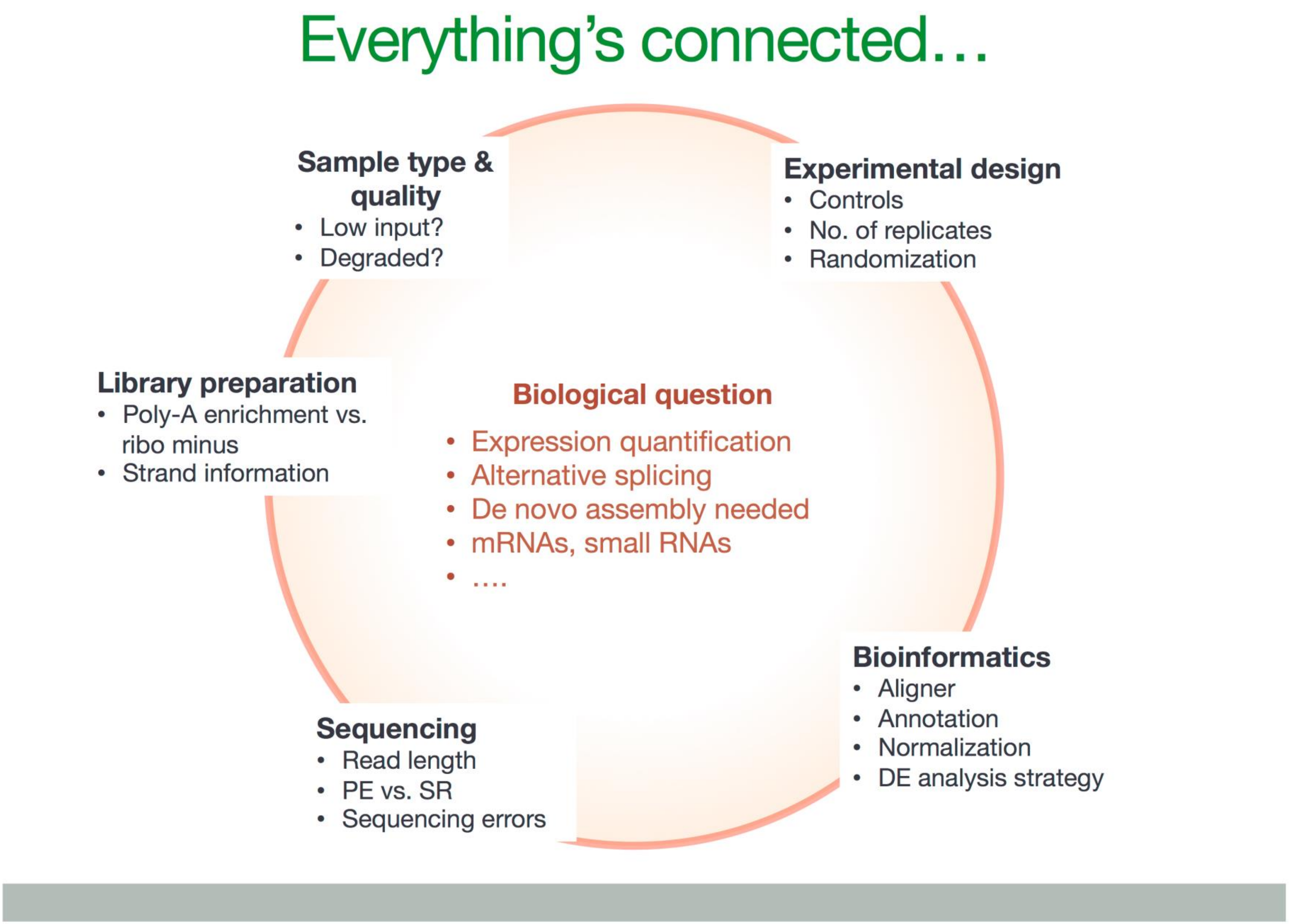
RNA-Seq analysis using R: Differential expression and transcriptome assembly

Beibei Chen Ph.D
BICF
2/1/2017

Agenda

- Brief about RNA-seq and experiment design
- Gene oriented analysis
  - Gene quantification
  - Gene differential analysis
  - Comparison model
- Astrocyte introduction
- Transcript oriented analysis
  - Transcripts assembly and quantification
  - Transcripts differential expression



Everything's connected…

Sample type & quality

- Low input?
- Degraded?

Experimental design

- Controls
- No. of replicates
- Randomization

Library preparation

- Poly-A enrichment vs. ribo minus
- Strand information

Biological question

- Expression quantification
- Alternative splicing
- De novo assembly needed
- mRNAs, small RNAs
- ….

Bioinformatics

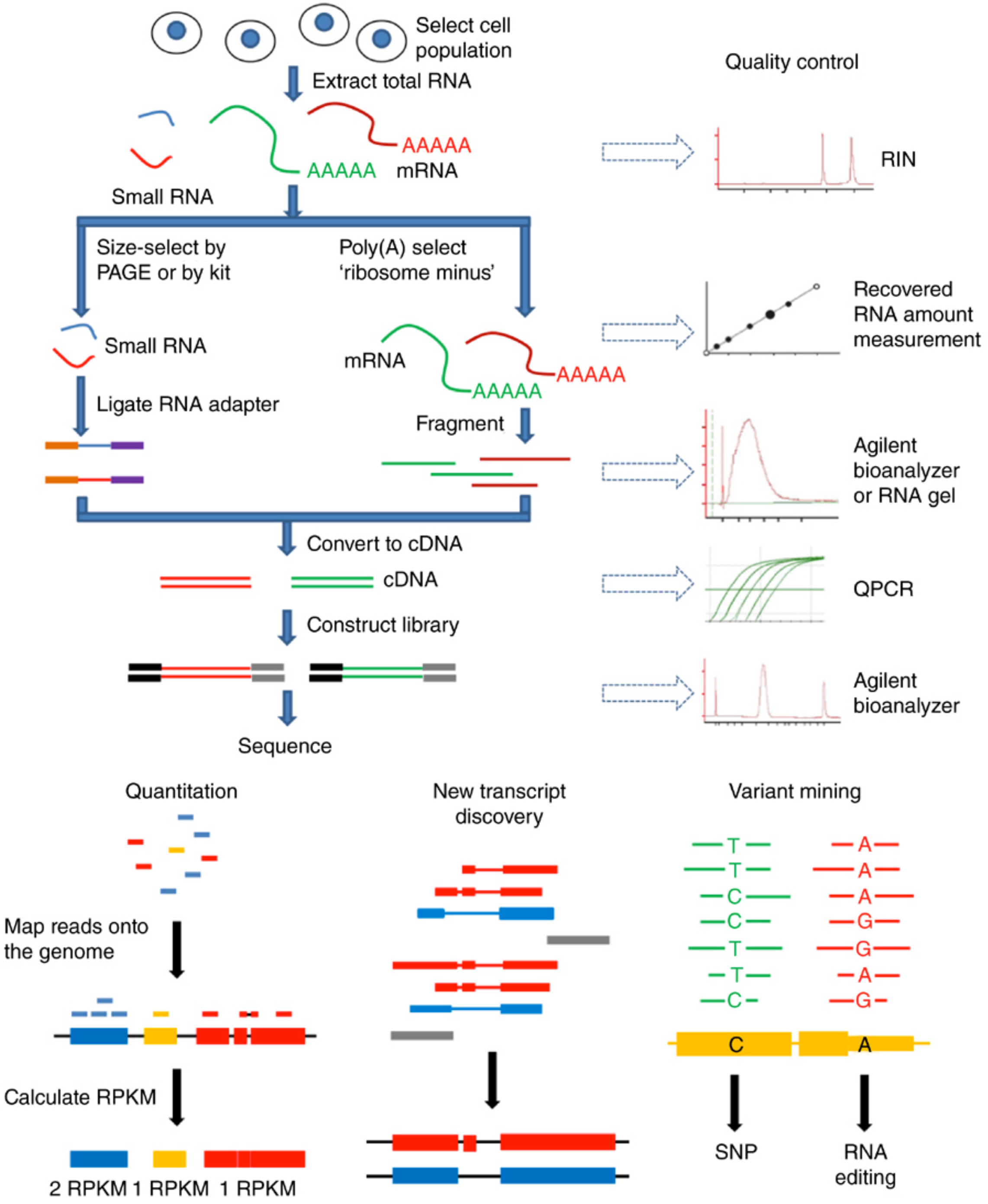
- Aligner
- Annotation
- Normalization
- DE analysis strategy

Sequencing

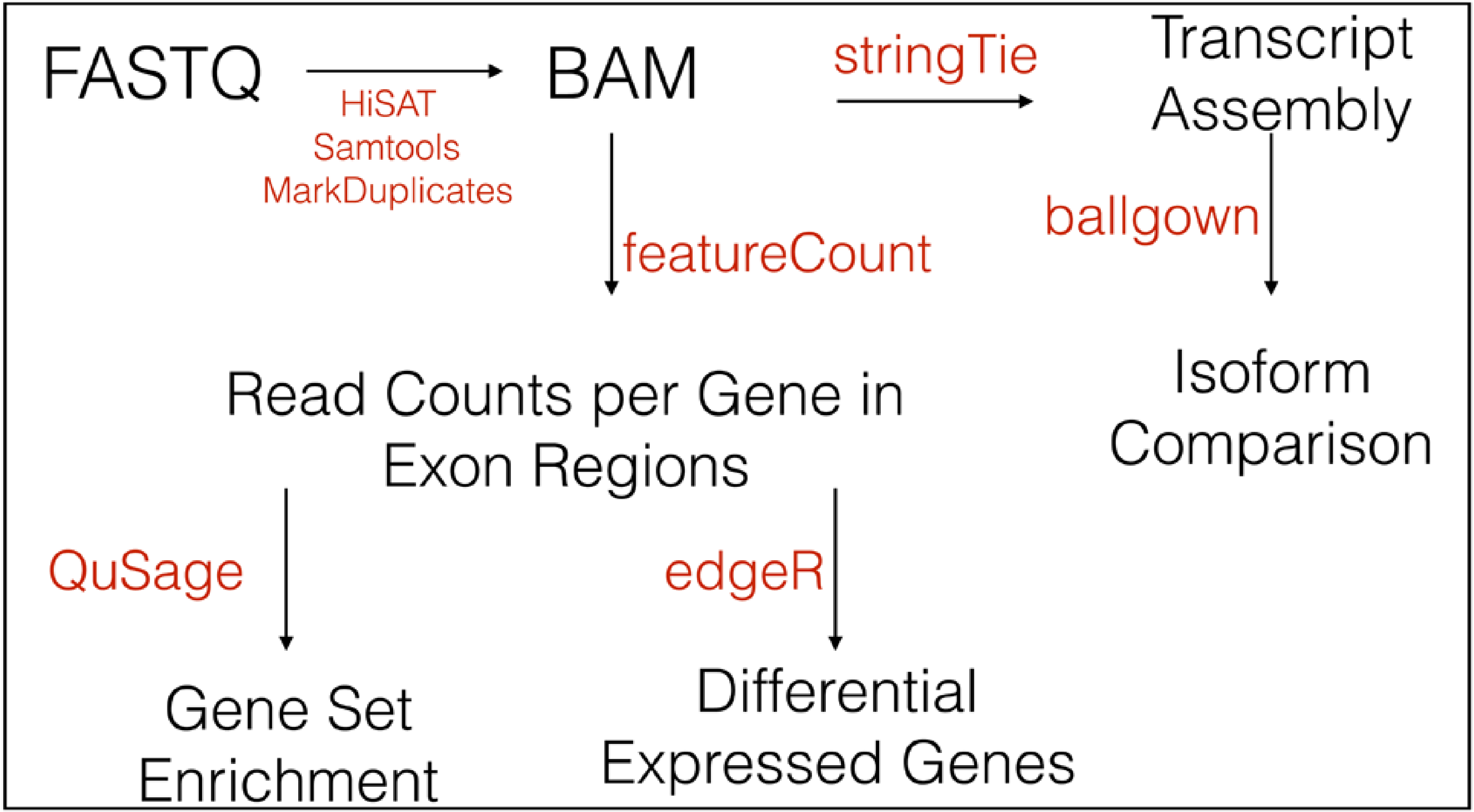
- Read length
- PE vs. SR
- Sequencing errors

Everything's connected slide by Dündar et al. (2015)

General RNA-seq Workflow



RNASeq Analysis Pipeline


FASTQ

HiSAT
Samtools
MarkDuplicates

BAM

stringTie

Transcript Assembly

featureCount

ballgown

Read Counts per Gene in Exon Regions

Isoform Comparison

QuSage

edgeR

Gene Set Enrichment

Differential Expressed Genes

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

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 at least 20M
- Read Length
  - Longer reads produce better alignments, min 50 bp paired or 100bp single for gene quantification
  - ChIP-seq, smallRNA-seq, RIP-seq, CLIP-seq: 50nt single-end

Experimental Design Affecting Your Analysis

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

Strand-specific RNA-seq

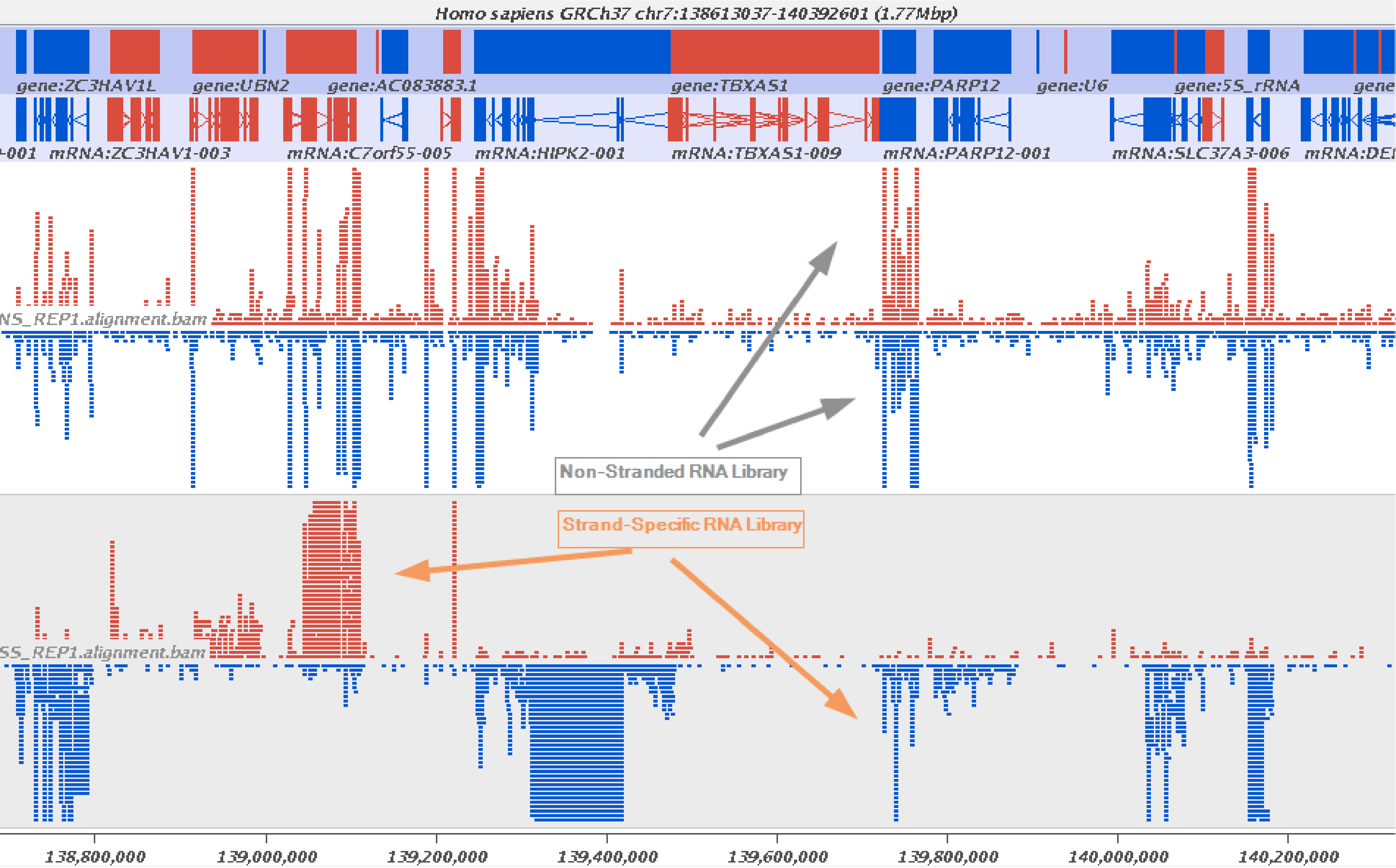
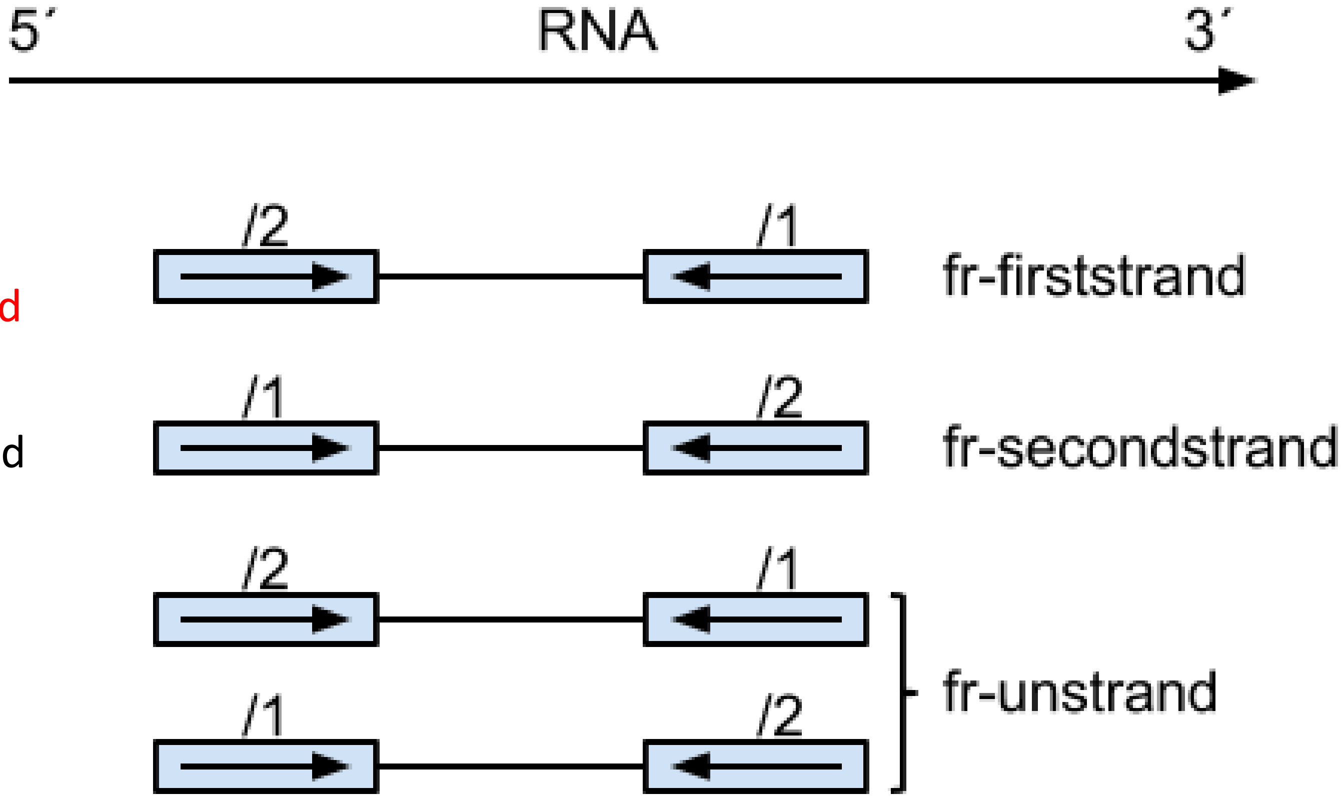


image from GATC Biotech

How to decide strand



Reverse stranded

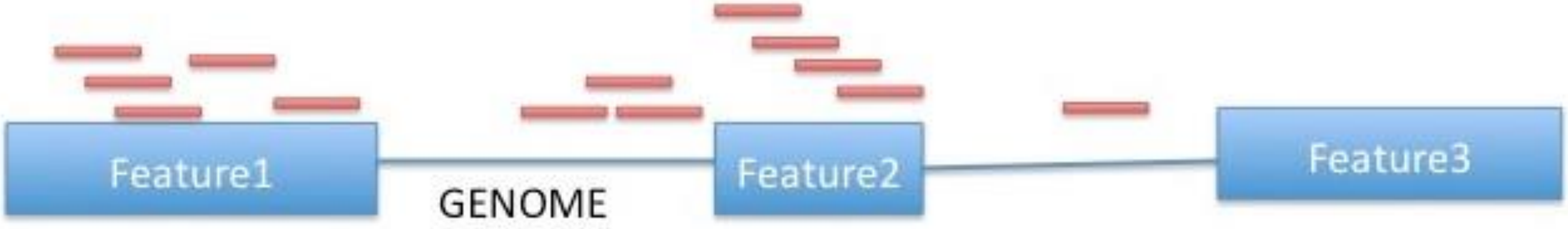
Stranded

Agenda

- Brief about RNA-seq and experiment design
- Gene oriented analysis
  - Gene quantification
  - Gene differential analysis
  - Comparison model
- Astrocyte introduction
- Transcript oriented analysis
  - Transcripts assembly and quantification
  - Transcripts differential expression

Gene quantification

- In RNA-Seq, the abundance level of a gene is measured by the number of reads that map to that gene/exon.



Tool to use: featureCounts

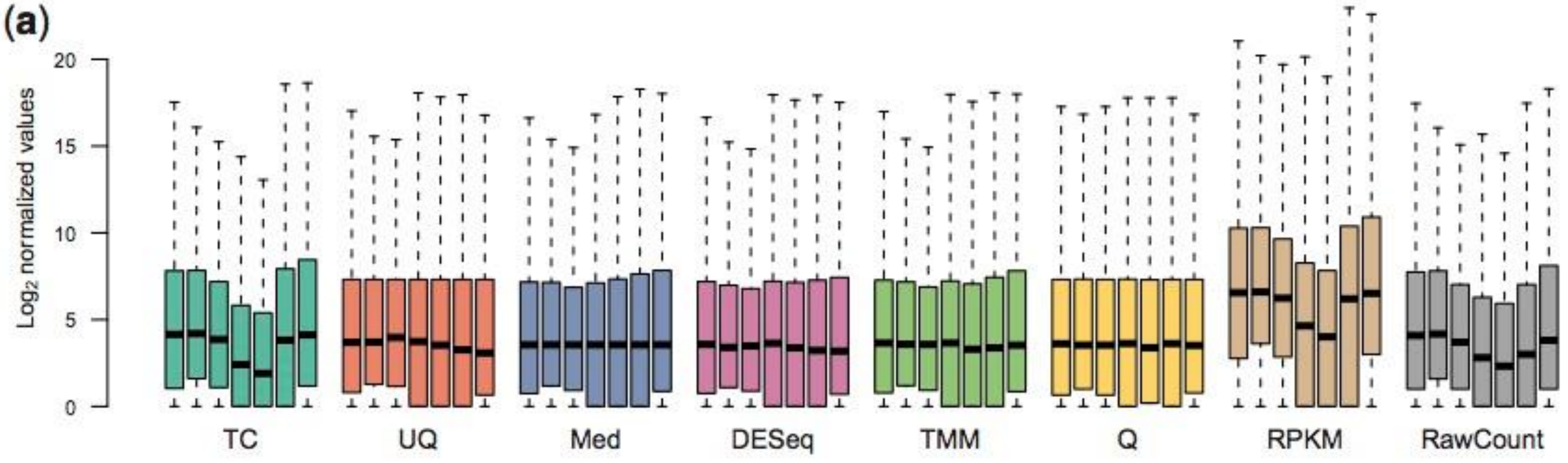
Differential expressed gene detection

- Normalization
- Explore your data

Why normalize

- To smooth out technical variations among the samples
  - Sequencing depth: genes have more reads in a deeper sequenced library
  - Gene length: longer genes are likely to have more reads than the shorter genes

Effects of different normalization methods



Assuming reads count distribution should be the same

- Total count (TC): Gene counts are divided by the total number of mapped reads
- Upper Quartile (UQ):Gene counts are divided by the upper quartile of counts
- Median (Med): Gene counts are divided by the median counts
- Quantile (Q): Matching distributions of gene counts across samples (limma)
- Reads Per Kilobase per Million mapped reads (RPKM): Re-scales gene counts
- to correct for differences in both library sizes and gene length

Assuming most gene are not differentially expressed

- DESeq
- Trimmed Mean of *M*-values (TMM): edgeR

doi:10.1093/bib/bbs046

Trimmed Mean M values (TMM)

- Applied in edgeR package
- Rationale:
  - TMM is the weighted mean of log ratios between this test and the reference.
  - TMM should be close to 0 according to the hypothesis of low DE. If it is not, its value provides an estimate of the correction factor that must be applied to the library sizes (and not the raw counts) in order to fulfill the hypothesis
- Reference sample can be assigned or the sample whose upper quartile is closest to the mean upper quartile is used

Trimmed Mean M values (TMM)

Gene-wise log-fold-changes $M_g = \log_2 \frac{Y_{gk}/N_k}{Y_{gk'}/N_{k'}}$

Absolute expression levels

$$A_g = \frac{1}{2}\log_2\left(Y_{gk}/N_k \bullet Y_{gk'}/N_{k'}\right) \text{ for } Y_{g\bullet} \neq 0$$

- By default, trim the M_g values by 30% and the A_g values by 5% (can be tailored in program)
- Weights are from the delta method on Binomial data
- Normalization factor for sample *k* using reference sample *r* is calculated as:

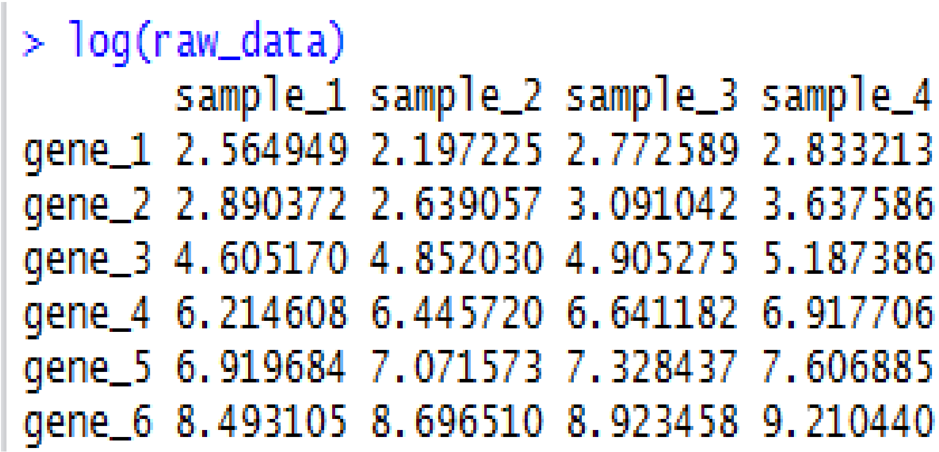
$$\log_2(TMM_k^{(r)}) = \frac{\sum_{g \in G^*} w_{gk}^r M_{gk}^r}{\sum_{g \in G^*} w_{gk}^r} \text{ where } M_{gk}^r = \frac{\log_2\left(\frac{Y_{gk}}{N_k}\right)}{\log_2\left(\frac{Y_{gr}}{N_r}\right)} \text{ and } w_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}};$$

$Y_{gk}, Y_{gr} > 0.$

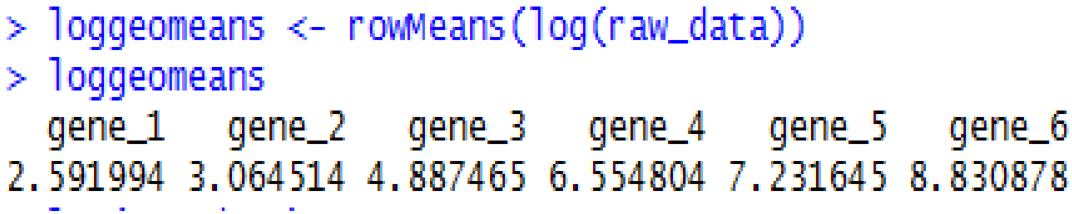
Median-of-ratios normalization

- Applied in DESeq and DESeq2
- Rationale:
  - Calculate the ratio of between a test and a pseudosample (For each gene, the geometric mean of all samples)
  - Non-DE genes should have similar read counts across samples, leading to a ratio of 1.
  - Assuming most genes are not DE, the median of this ratio for the lane provides an estimate of the correction factor that should be applied to all read counts of this lane to fulfill the hypothesis

Median-of-ratios normalization



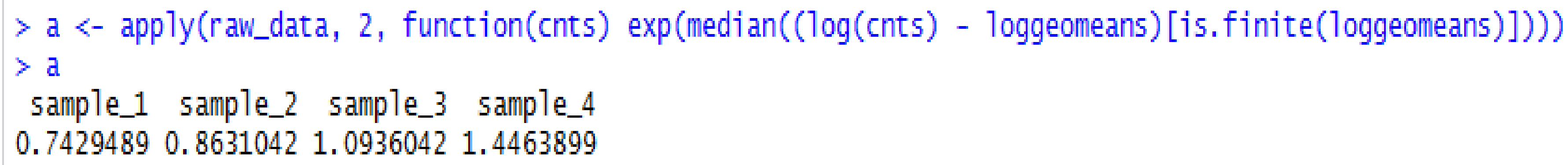
```
> log(raw_data)
       sample_1 sample_2 sample_3 sample_4
gene_1 2.564949 2.197225 2.772589 2.833213
gene_2 2.890372 2.639057 3.091042 3.637586
gene_3 4.605170 4.852030 4.905275 5.187386
gene_4 6.214608 6.445720 6.641182 6.917706
gene_5 6.919684 7.071573 7.328437 7.606885
gene_6 8.493105 8.696510 8.923458 9.210440
```



```
> loggeomeans <- rowMeans(log(raw_data))
> loggeomeans
  gene_1   gene_2   gene_3   gene_4   gene_5   gene_6 
2.591994 3.064514 4.887465 6.554804 7.231645 8.830878 
```

Pseudo sample

Get the median of log ratio of test comparing to pseudo sample:



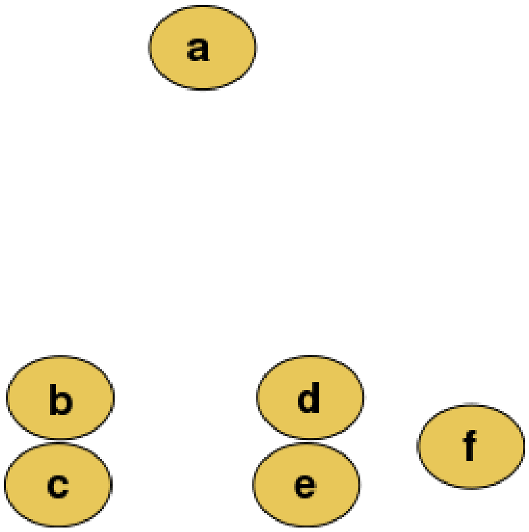
```
> a <- apply(raw_data, 2, function(cnts) exp(median((log(cnts) - loggeomeans)[is.finite(loggeomeans)])))
> a
 sample_1  sample_2  sample_3  sample_4 
0.7429489 0.8631042 1.0936042 1.4463899 
```

Data exploration

- Use log transformed normalized gene reads count
- Check if replicates from the same group are well concordance and grouped together
  - Hierarchy clustering
  - PCA plot

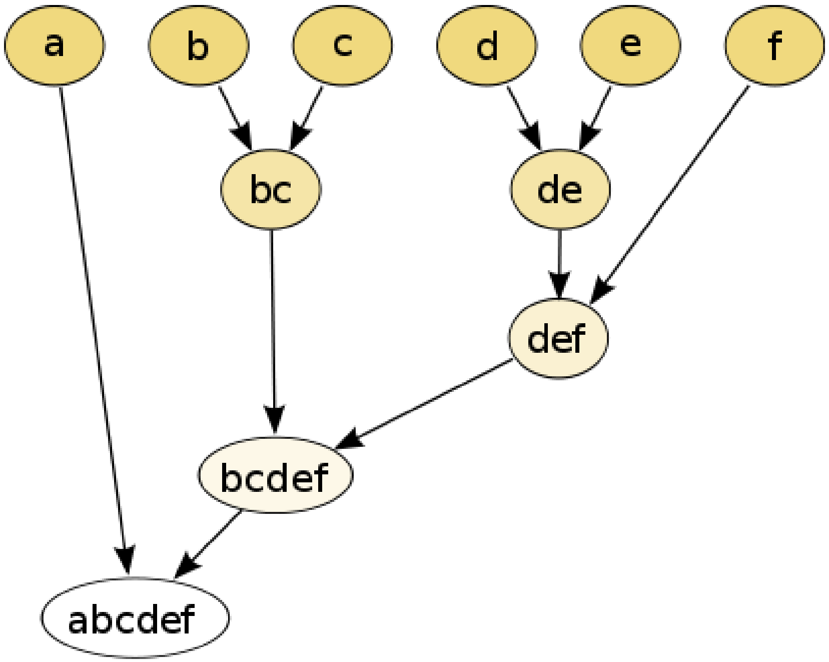
Hierarchy clustering

Raw data



Distance calculation

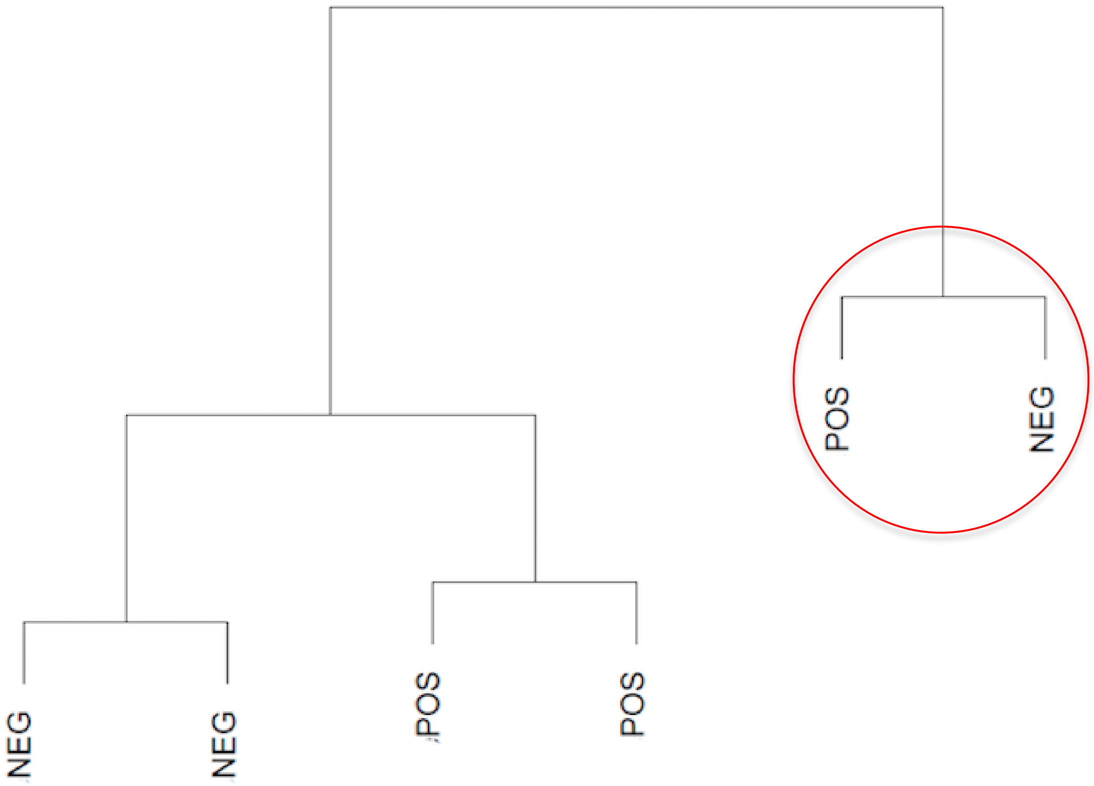
hierarchical clustering dendrogram



Euclidean distance: $\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$

https://en.wikipedia.org/wiki/Hierarchical_clustering

Hierarchy plot example

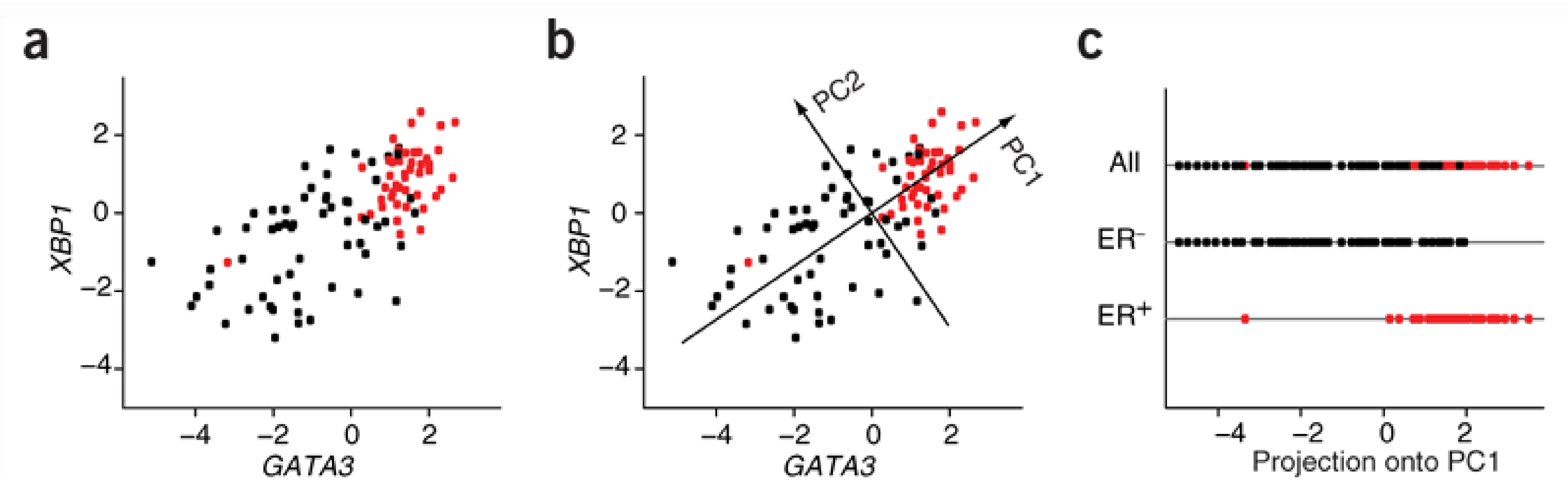


Samples prepared a year ago

Principal component analysis (PCA)

- A mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set
- It identifies directions, called principal components, along which the variation in the data is maximal
- By using a few components, each sample can be represented by relatively few numbers instead of by values for thousands of variables.

Simple example of PCA



Separate breast cancer ER+ from ER- : get profiles with only two genes

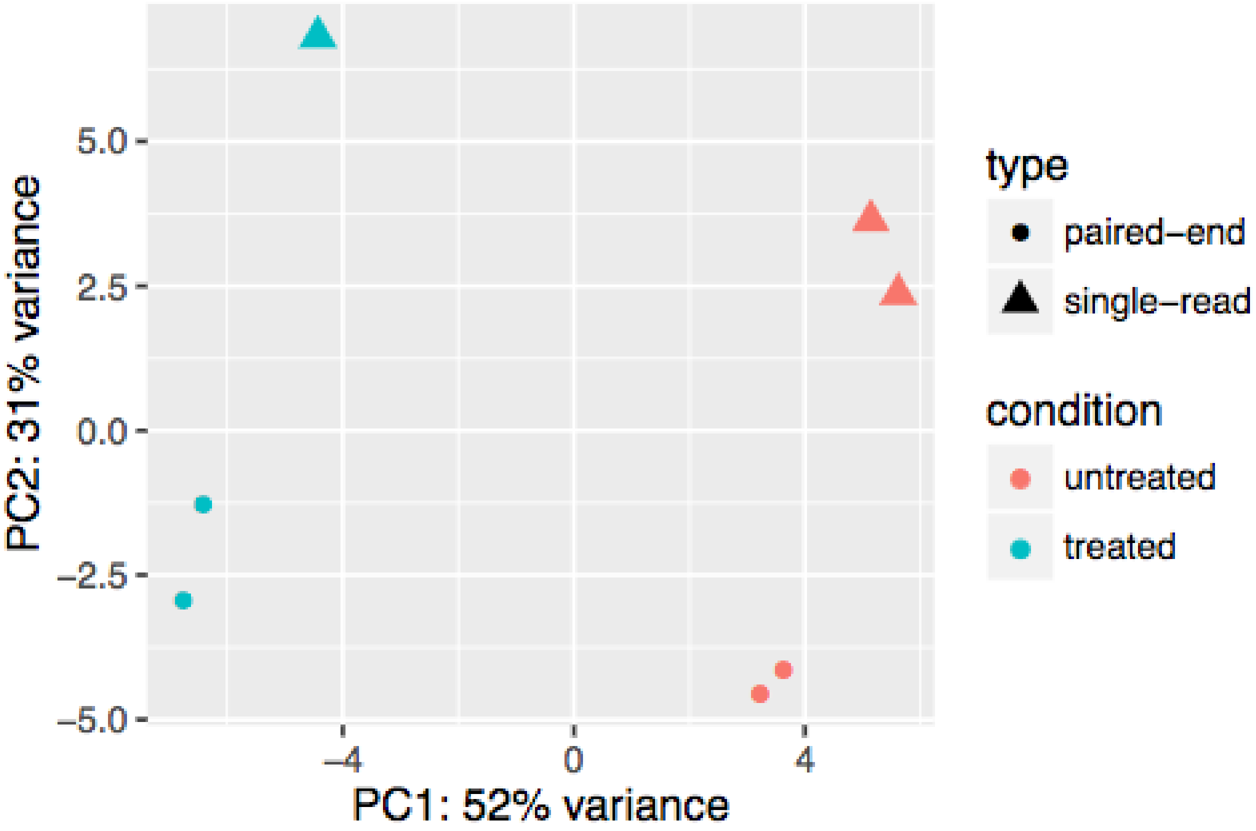
(a) Each dot represents a breast cancer sample plotted against its expression levels for two genes. (ER$^+$, red; ER$^-$, black).

(b) (**b**) PCA identifies the two directions (PC1 and PC2) along which the data have the largest spread.

(c) (**c**) Samples plotted in one dimension using their projections onto the first principal component (PC1) for ER$^+$, ER$^-$ and all samples separately.

Nature Biotechnology 26, 303 - 304 (2008)

Real example



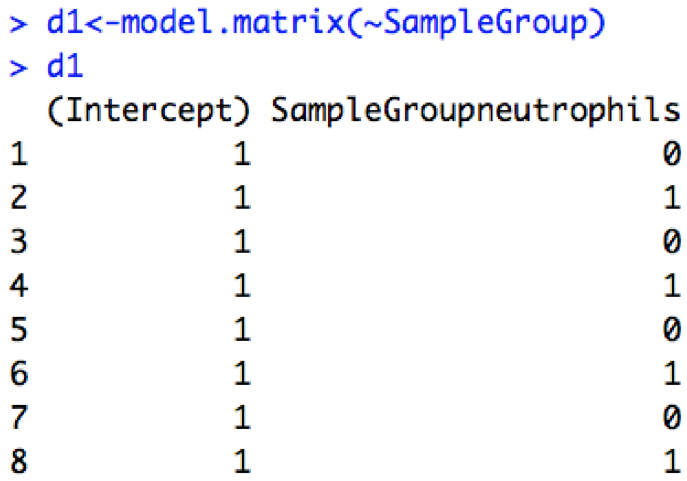
Test for differential expressed genes

- General liner model: negative binomial distribution
- For each gene,
- Typical commands

```
>glmFit()
>design <- model.matrix(~group)
>fit <- glmFit(y, design)
>res<-glmLRT(fit, coef=2)
```

Make a model and extract results

- Basic: treatment vs control
- Question: what’s the difference between monocytes and neutrophils
  - design <- model.matrix(~SampleGroup)



```
> d1<-model.matrix(~SampleGroup)
> d1
  (Intercept) SampleGroupneutrophils
1           1                      0
2           1                      1
3           1                      0
4           1                      1
5           1                      0
6           1                      1
7           1                      0
8           1                      1
```

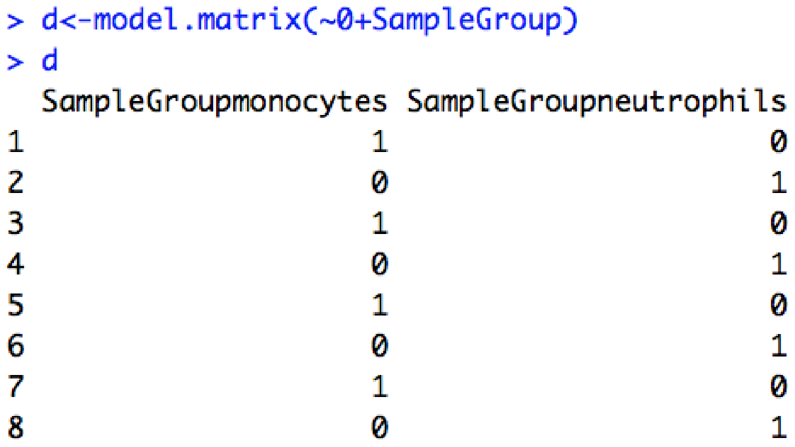
Set control as baseline (intercept) and compare treatment to it

relevel(factor(SampleGroup),ref=“monocyotes”)

- Use coefficient to extract results
  - res<-glmLRT(fit, coef=2)

Make a model and extract results

- Model without interception:
  - model.matrix(~0+SampleGroup)

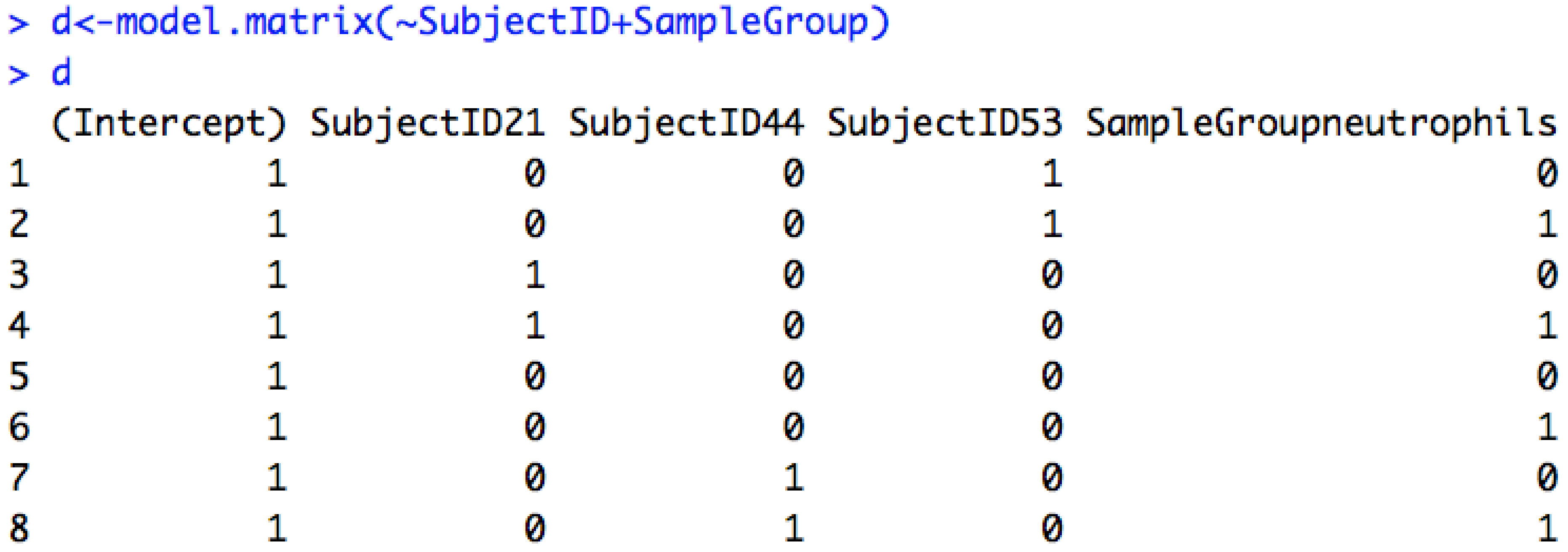


```
> d<-model.matrix(~0+SampleGroup)
> d
  SampleGroupmonocytes SampleGroupneutrophils
1                    1                      0
2                    0                      1
3                    1                      0
4                    0                      1
5                    1                      0
6                    0                      1
7                    1                      0
8                    0                      1
```

- Use contrast vector
  - Res<-glmLRT(fit, contrast=c(-1,1))
- Use a contrast function
  - My.contrast <- makeContrasts(SampleGroupneutrophils-SampleGroupmonocytes, level=design)

Comparison model

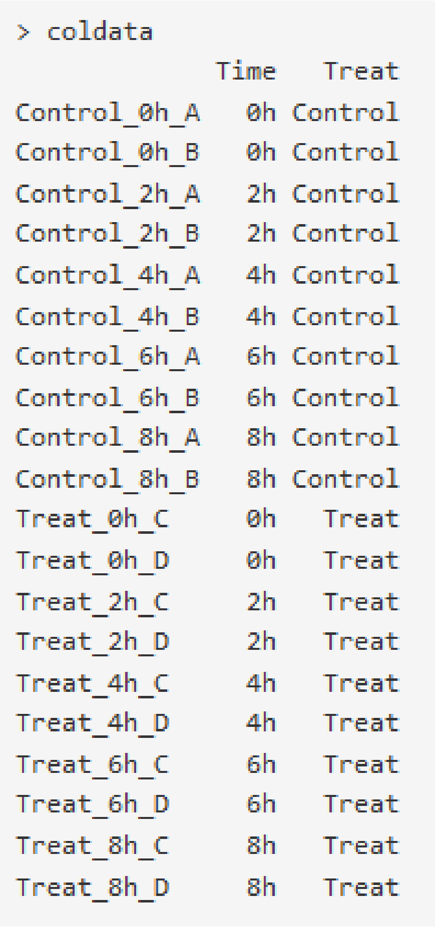
- Batch effect (additive model)
- Question: I want to account for the individual since I think individual difference will affect
  - resultsmodel.matrix(~SampleGroup+SubjectID)



```
> d<-model.matrix(~SubjectID+SampleGroup)
> d
  (Intercept) SubjectID21 SubjectID44 SubjectID53 SampleGroupneutrophils
1           1           0           0           1                      0
2           1           0           0           1                      1
3           1           1           0           0                      0
4           1           1           0           0                      1
5           1           0           0           0                      0
6           1           0           0           0                      1
7           1           0           1           0                      0
8           1           0           1           0                      1
```

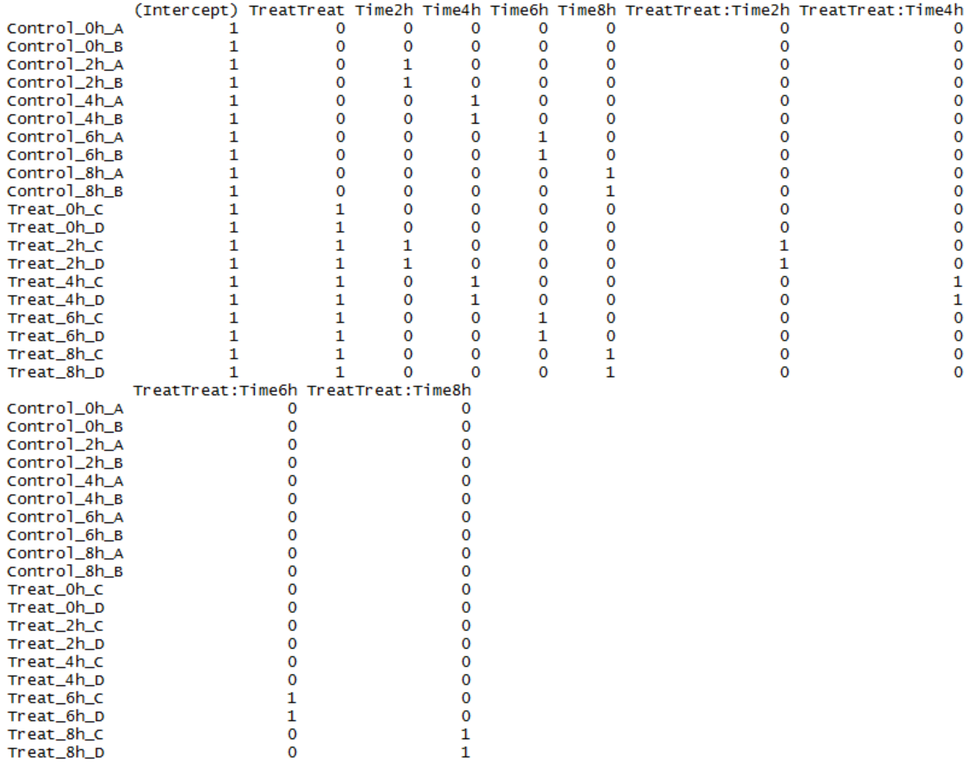
More complicated comparison models

- Time series: treatment and control, 5 time points



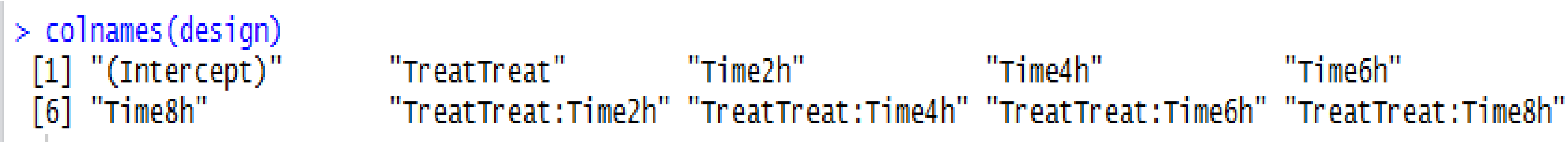
```
> coldata
             Time   Treat
Control_0h_A   0h Control
Control_0h_B   0h Control
Control_2h_A   2h Control
Control_2h_B   2h Control
Control_4h_A   4h Control
Control_4h_B   4h Control
Control_6h_A   6h Control
Control_6h_B   6h Control
Control_8h_A   8h Control
Control_8h_B   8h Control
Treat_0h_C     0h   Treat
Treat_0h_D     0h   Treat
Treat_2h_C     2h   Treat
Treat_2h_D     2h   Treat
Treat_4h_C     4h   Treat
Treat_4h_D     4h   Treat
Treat_6h_C     6h   Treat
Treat_6h_D     6h   Treat
Treat_8h_C     8h   Treat
Treat_8h_D     8h   Treat
```

```
design<- model.matrix
(~Treat+Time+Treat:Time, data=coldata)
```



```
             (Intercept) TreatTreat Time2h Time4h Time6h Time8h TreatTreat:Time2h TreatTreat:Time4h
Control_0h_A           1          0      0      0      0      0                 0                 0
Control_0h_B           1          0      0      0      0      0                 0                 0
Control_2h_A           1          0      1      0      0      0                 0                 0
Control_2h_B           1          0      1      0      0      0                 0                 0
Control_4h_A           1          0      0      1      0      0                 0                 0
Control_4h_B           1          0      0      1      0      0                 0                 0
Control_6h_A           1          0      0      0      1      0                 0                 0
Control_6h_B           1          0      0      0      1      0                 0                 0
Control_8h_A           1          0      0      0      0      1                 0                 0
Control_8h_B           1          0      0      0      0      1                 0                 0
Treat_0h_C             1          1      0      0      0      0                 0                 0
Treat_0h_D             1          1      0      0      0      0                 0                 0
Treat_2h_C             1          1      1      0      0      0                 1                 0
Treat_2h_D             1          1      1      0      0      0                 1                 0
Treat_4h_C             1          1      0      1      0      0                 0                 1
Treat_4h_D             1          1      0      1      0      0                 0                 1
Treat_6h_C             1          1      0      0      1      0                 0                 0
Treat_6h_D             1          1      0      0      1      0                 0                 0
Treat_8h_C             1          1      0      0      0      1                 0                 0
Treat_8h_D             1          1      0      0      0      1                 0                 0
             TreatTreat:Time6h TreatTreat:Time8h
Control_0h_A                 0                 0
Control_0h_B                 0                 0
Control_2h_A                 0                 0
Control_2h_B                 0                 0
Control_4h_A                 0                 0
Control_4h_B                 0                 0
Control_6h_A                 0                 0
Control_6h_B                 0                 0
Control_8h_A                 0                 0
Control_8h_B                 0                 0
Treat_0h_C                   0                 0
Treat_0h_D                   0                 0
Treat_2h_C                   0                 0
Treat_2h_D                   0                 0
Treat_4h_C                   0                 0
Treat_4h_D                   0                 0
Treat_6h_C                   1                 0
Treat_6h_D                   1                 0
Treat_8h_C                   0                 1
Treat_8h_D                   0                 1
```

More complicated comparison models



```
> colnames(design)
 [1] "(Intercept)"      "TreatTreat"       "Time2h"           "Time4h"           "Time6h"
 [6] "Time8h"           "TreatTreat:Time2h" "TreatTreat:Time4h" "TreatTreat:Time6h" "TreatTreat:Time8h"
```

- Intercept: (control at time 0)
- Coef=2
  - baseline comparison of treat and control
- Coef=7
  - difference between treat and control at 2h
- Coef=3:6
  - difference at any time of control comparing to control baseline
- Coef=7:10
  - difference at any time of treat comparing to control at that time

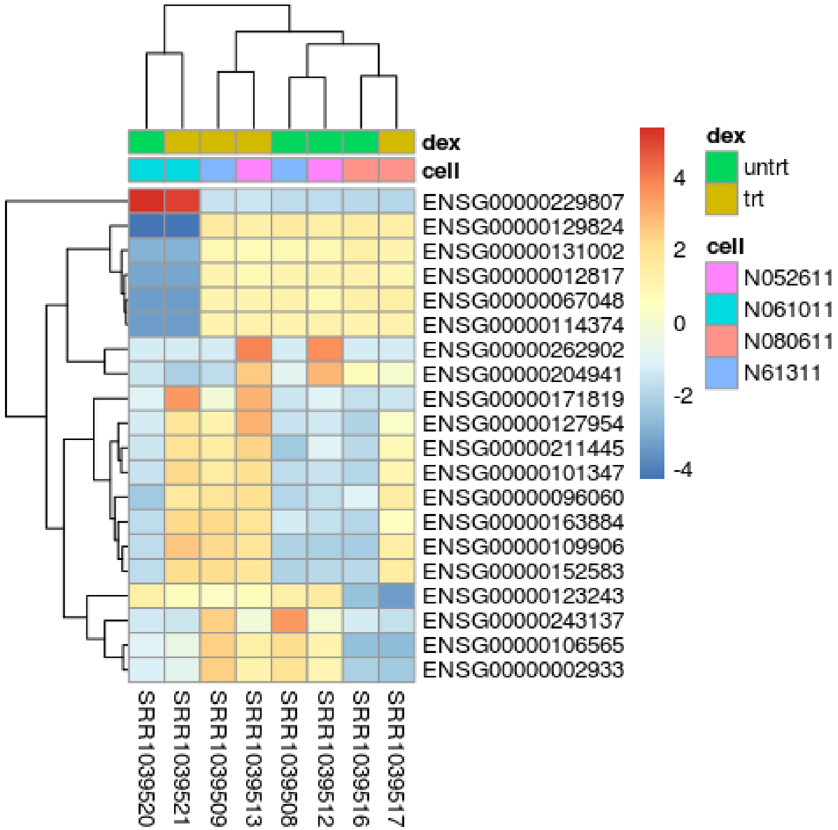
Test for differential expressed genes

- After GLMs are fit for each gene
- Wald test: whether each model treatment coefficient differs significantly from zero
- Multiple testing adjust
  - For a genome with 10,000 gene, using p<=0.05 as cutoff, there are 500 genes are significant by chance
  - BH method

https://bioconductor.org/help/course-materials/2010/BioC2010/2010-07-29_DESeq_Bioc-Meeting_Seattle.pdf

Define differential expressed genes

FDR and/or logFC cutoff



GSEA
Gene Set Enrichment Analysis

PANTHER
Classification System

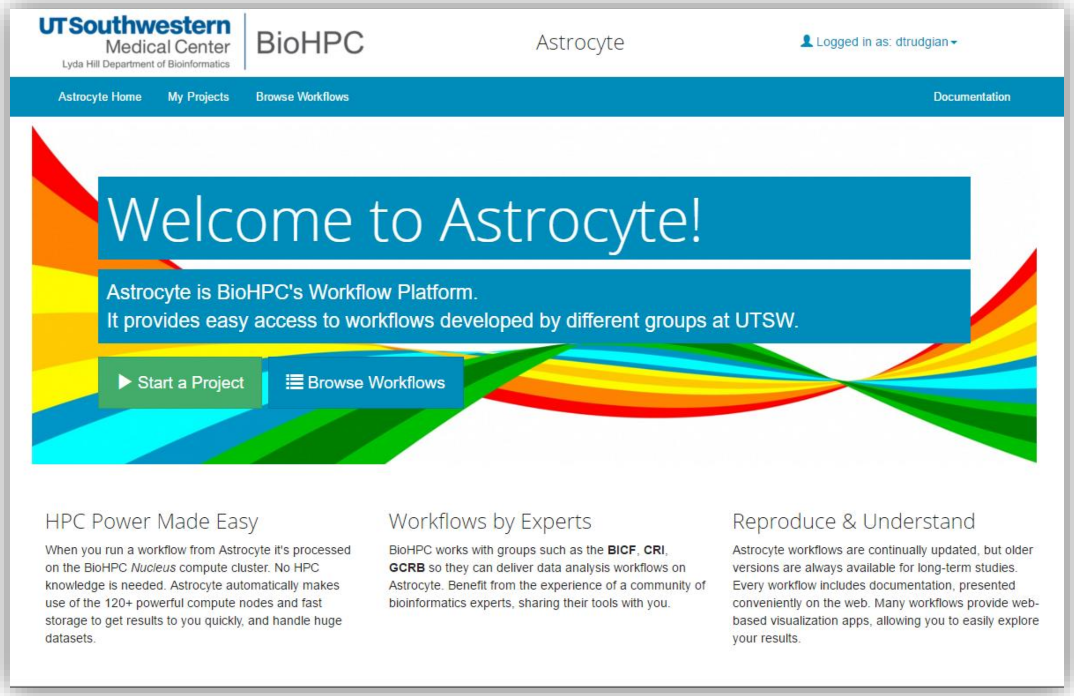
DAVID
BIOINFORMATICS
DATABASE

Agenda

- Brief about RNA-seq and experiment design
- Gene oriented analysis
  - Gene quantification
  - Gene differential analysis
  - Comparison model
- Astrocyte introduction
- Transcript oriented analysis
  - Transcripts assembly and quantification
  - Transcripts differential expression

Astrocyte – BioHPC Workflow Platform

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



Welcome to Astrocyte!

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

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**, **CRI**, **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.

Standardized Workflows

Simple Web Forms

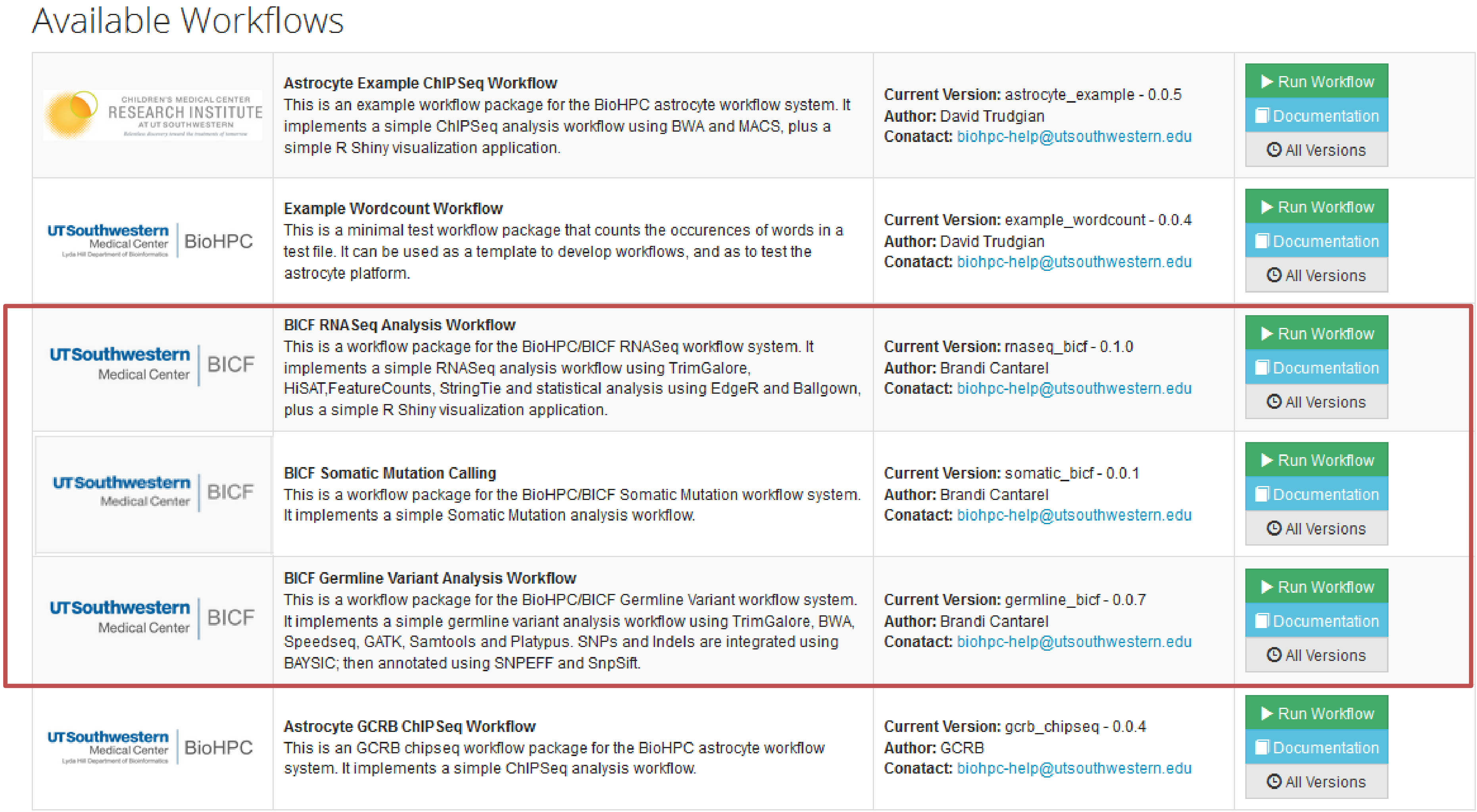
Online documentation & results visualization*

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

astrocyte.biohpc.swmed.edu

Slide contribution: David Trudgian@BioHPC

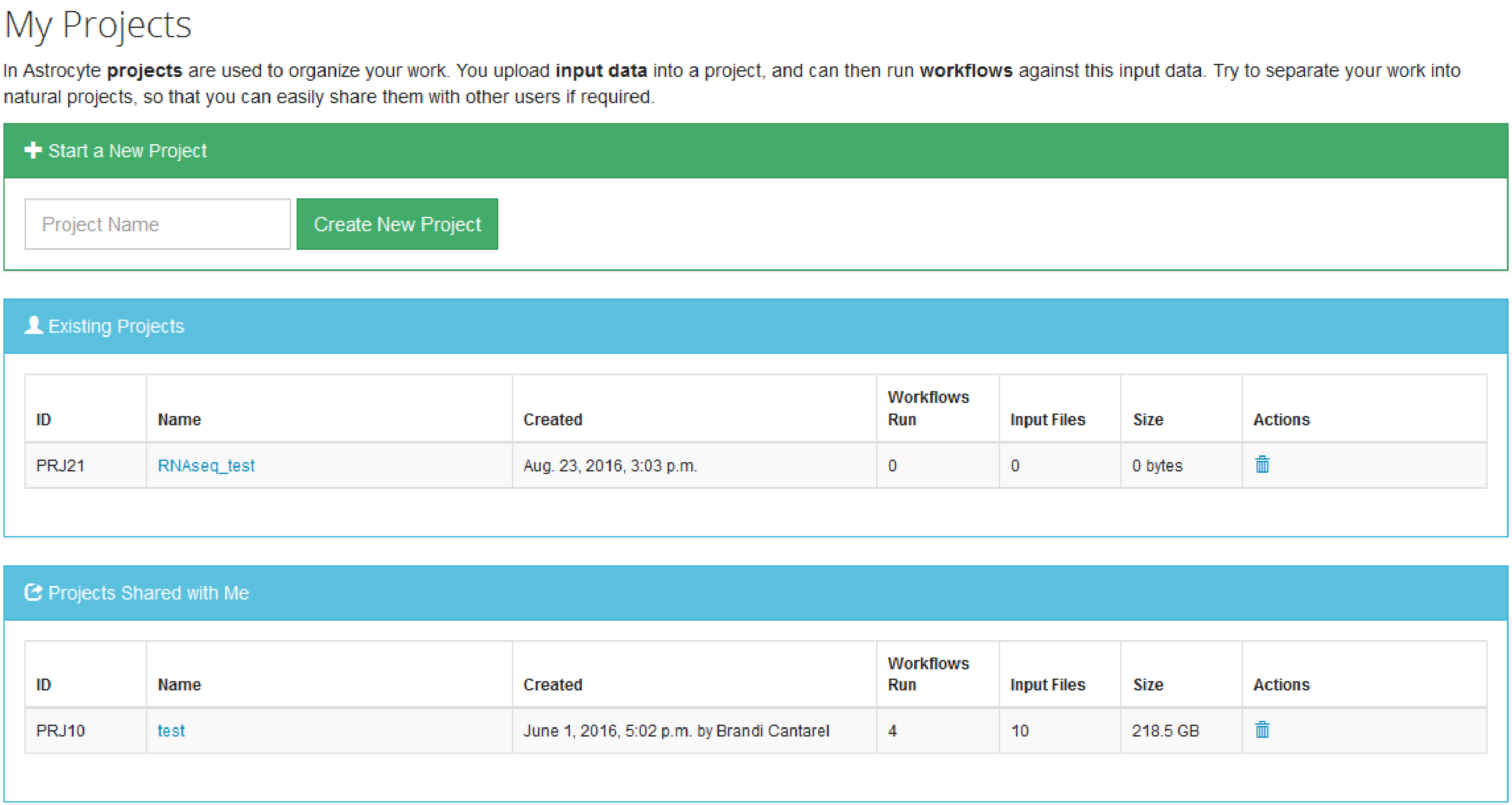
Browse workflows



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 Lyda Hill Department of Bioinformatics 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 Lyda Hill Department of Bioinformatics 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

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.

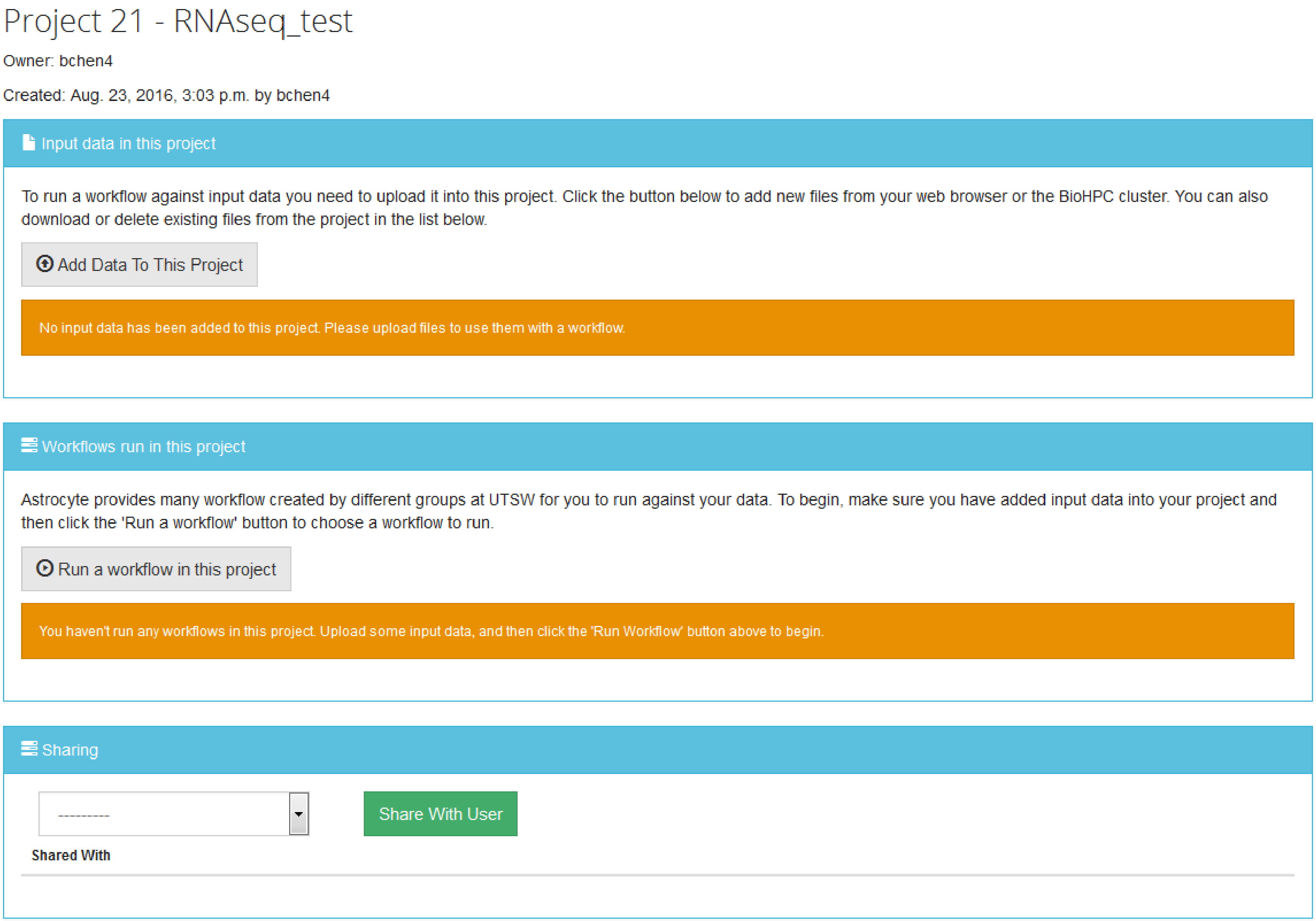
Existing Projects

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.

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

Workflows run in this project

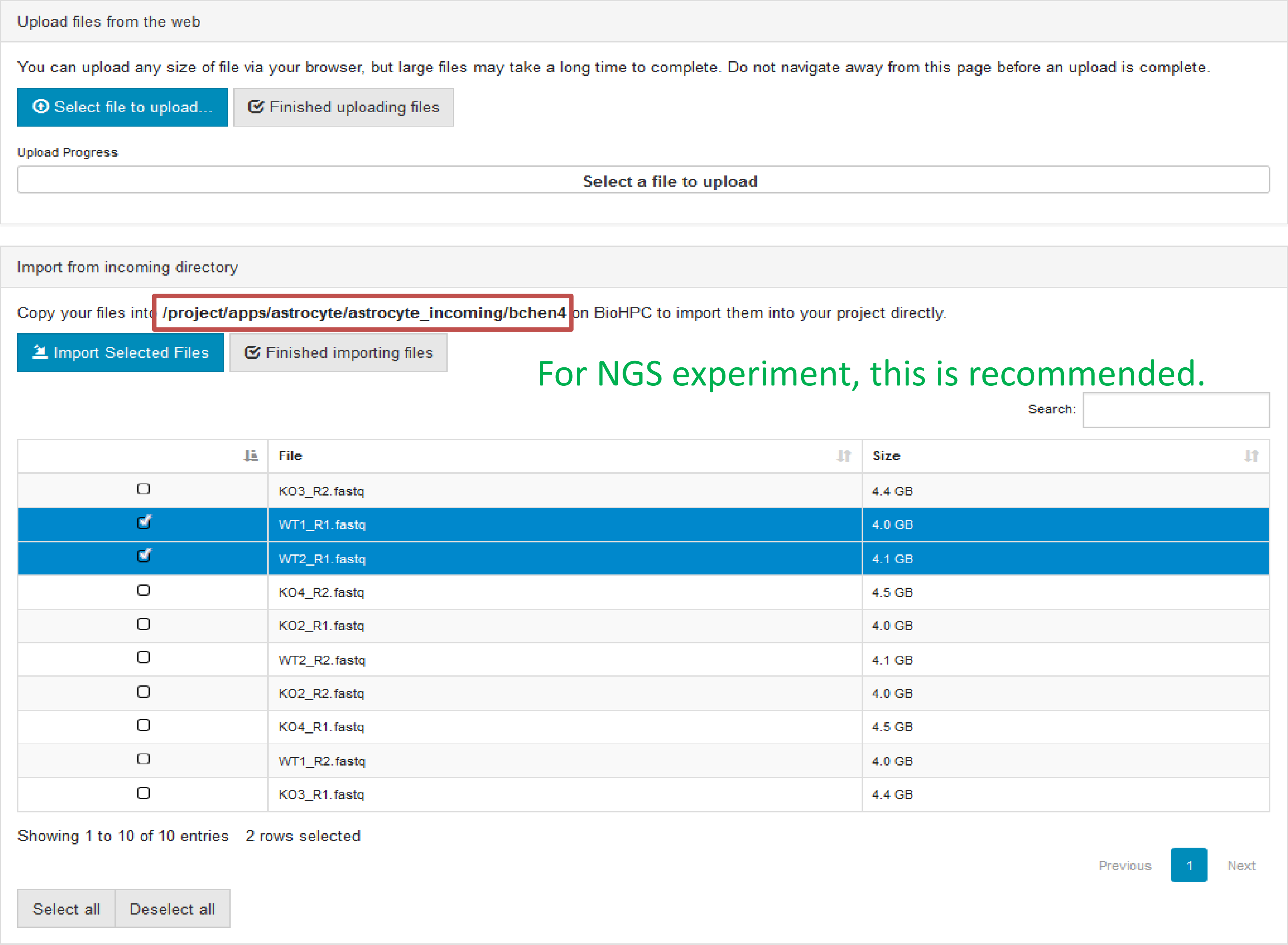
Astrocyte provides many workflow created by different groups at UTSW for you to run against your data. To begin, make sure you have added input data into your project and then click the 'Run a workflow' button to choose a workflow to run.

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

Sharing

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.

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.

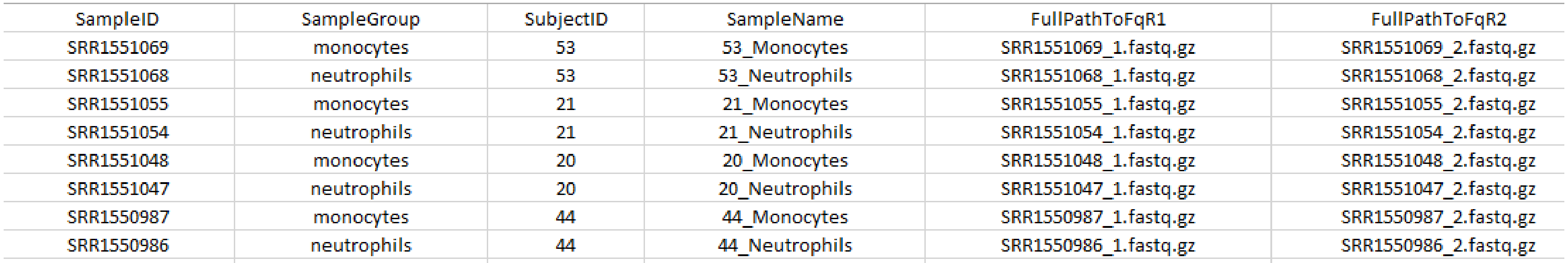
For NGS experiment, this is recommended.

Search:

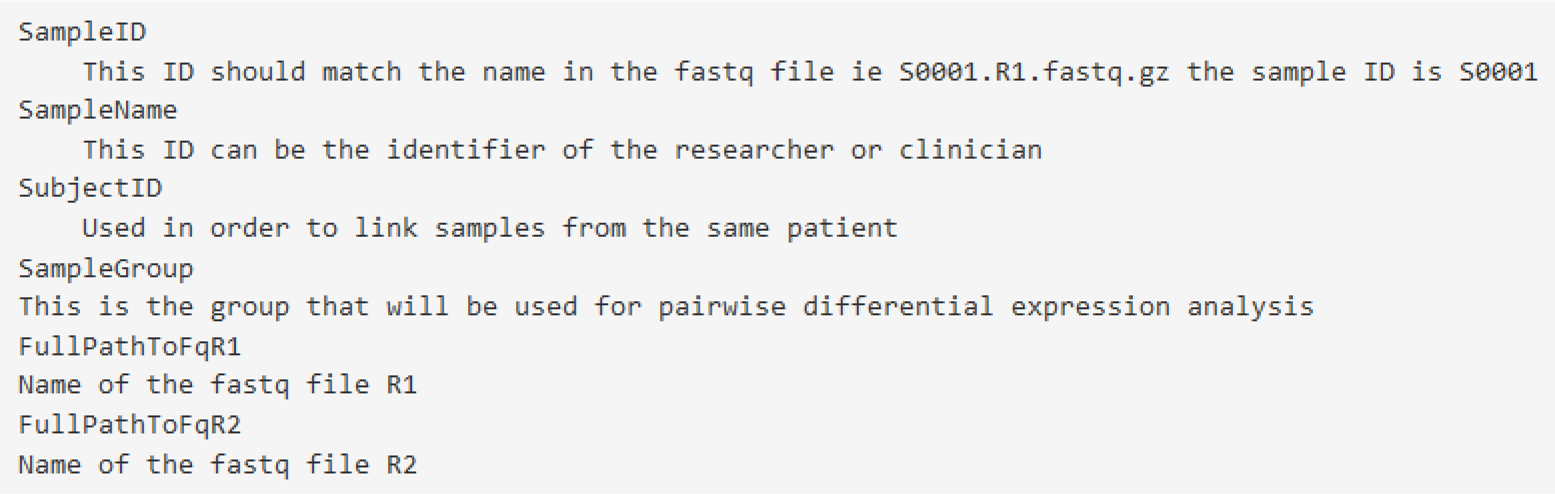
	File	Size
□	KO3_R2.fastq	4.4 GB
☑	WT1_R1.fastq	4.0 GB
☑	WT2_R1.fastq	4.1 GB
□	KO4_R2.fastq	4.5 GB
□	KO2_R1.fastq	4.0 GB
□	WT2_R2.fastq	4.1 GB
□	KO2_R2.fastq	4.0 GB
□	KO4_R1.fastq	4.5 GB
□	WT1_R2.fastq	4.0 GB
□	KO3_R1.fastq	4.4 GB

Showing 1 to 10 of 10 entries 2 rows selected

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
```

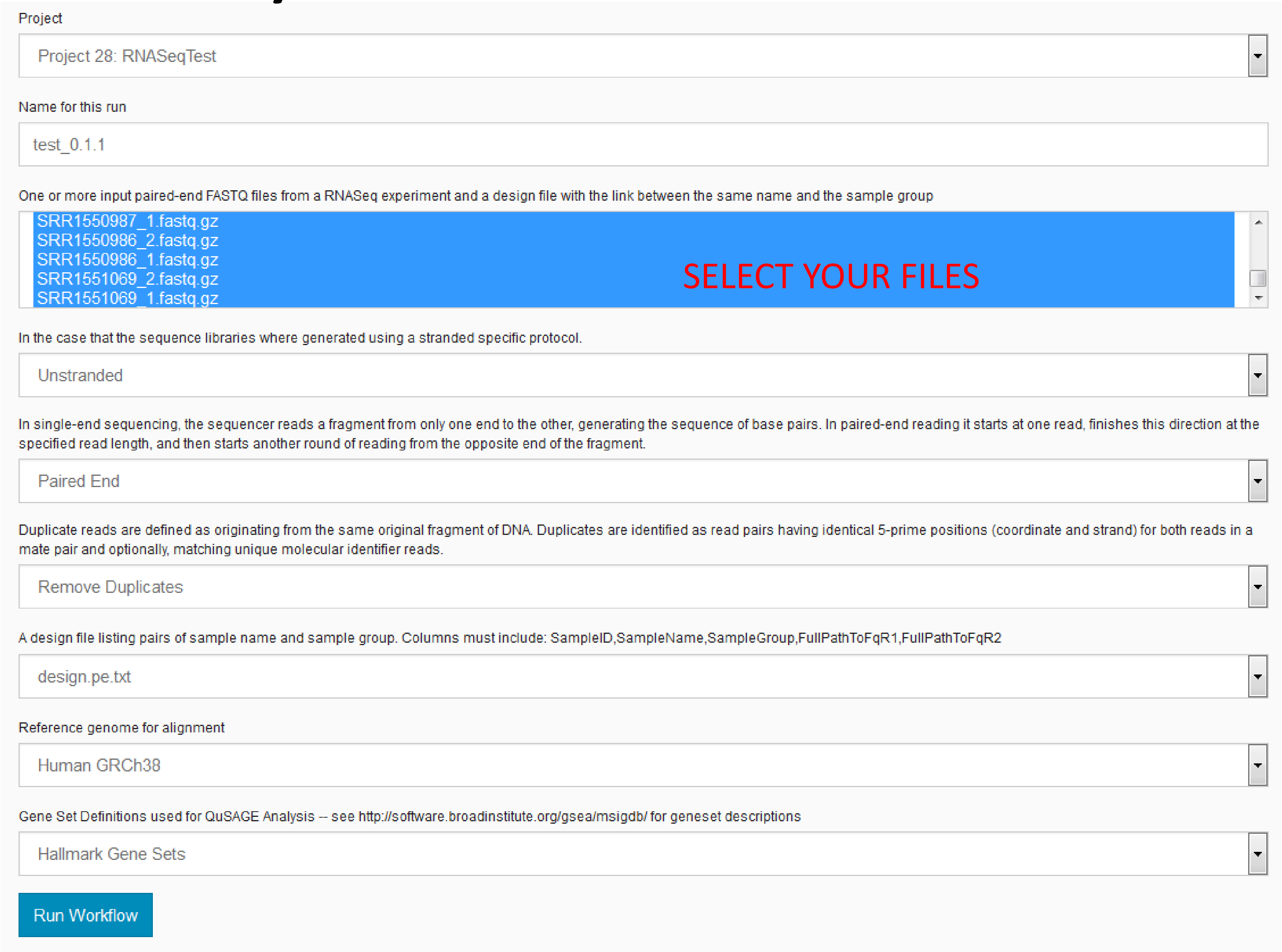
Make your design file

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

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

Select your data files 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

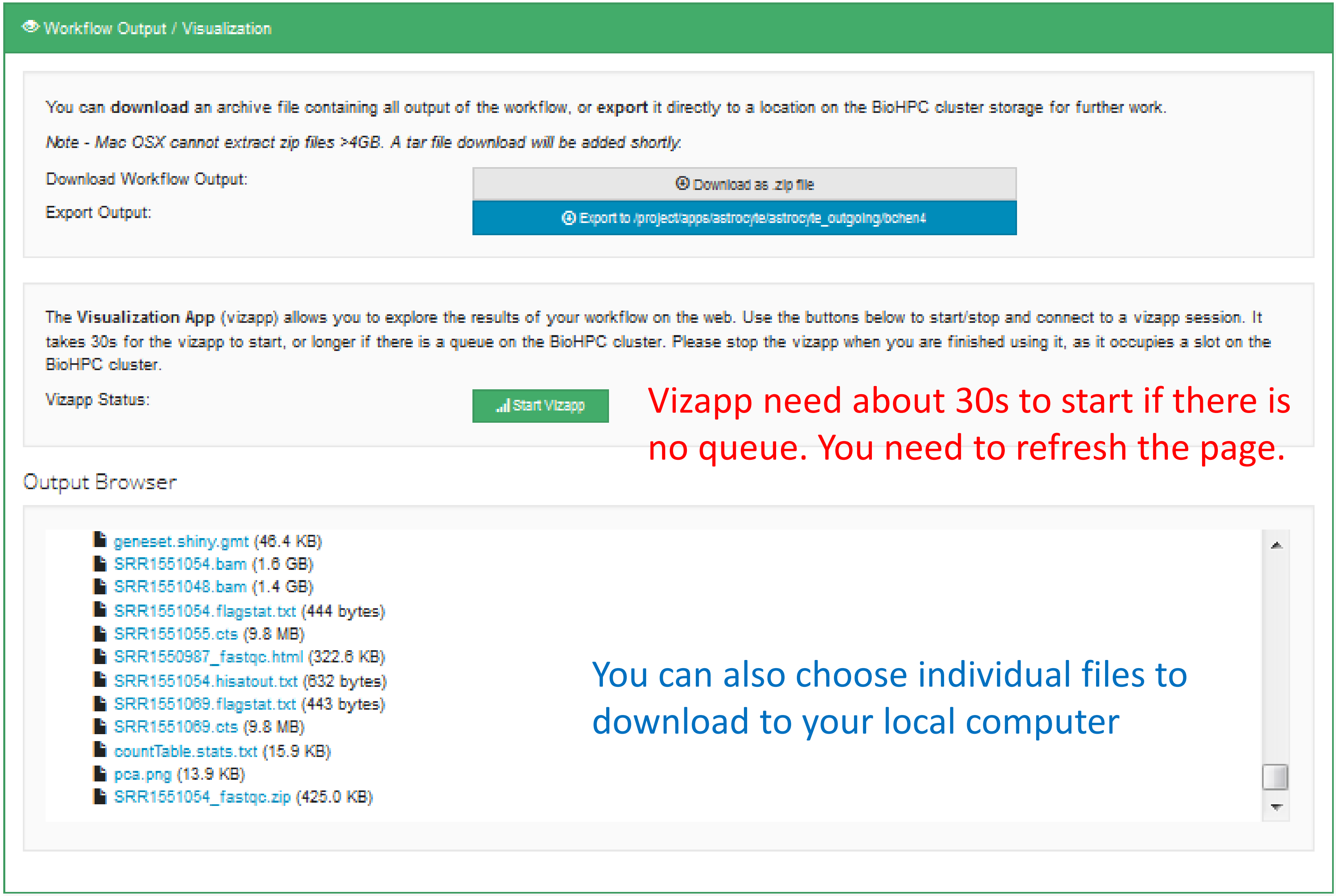
Reference genome for alignment

Human GRCh38

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

Hallmark Gene Sets

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:

Export Output:

Download as .zip file
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:

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 (632 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

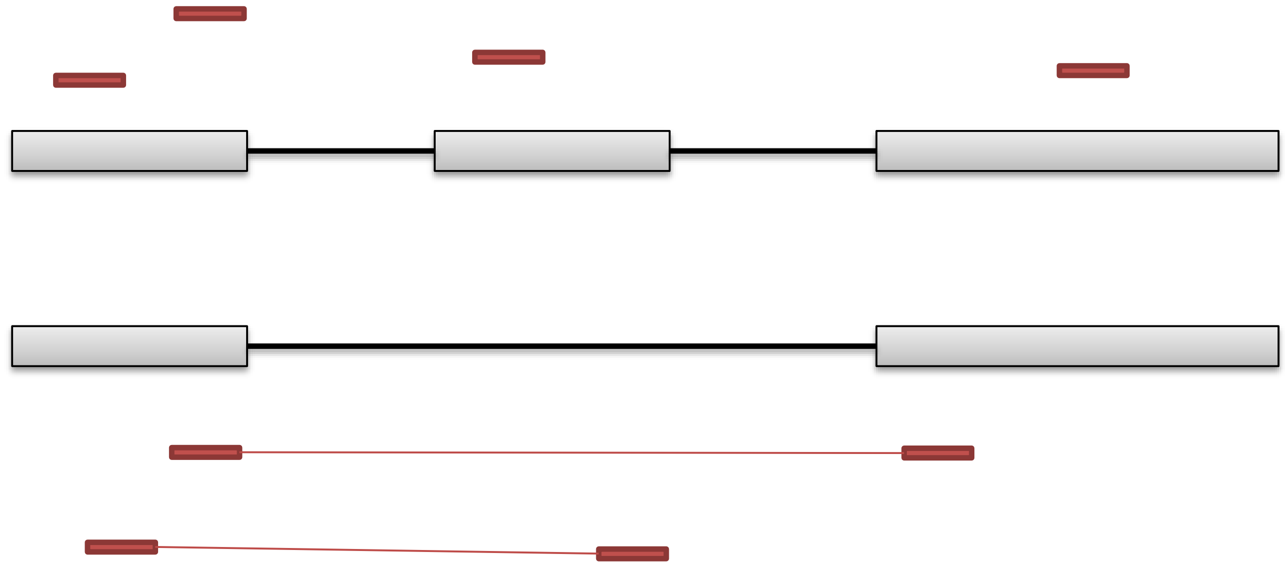
Agenda

- Brief about RNA-seq and experiment design
- Gene oriented analysis
  - Gene quantification
  - Gene differential analysis
  - Comparison model
- Astrocyte introduction
- Transcript oriented analysis
  - Transcripts assembly and quantification
  - Transcripts differential expression

Transcript oriented analysis

- Transcripts assembly and quantification
  - Stringtie
- Transcripts differential expression
  - Ballgown

Pair-end and single-endsequencing



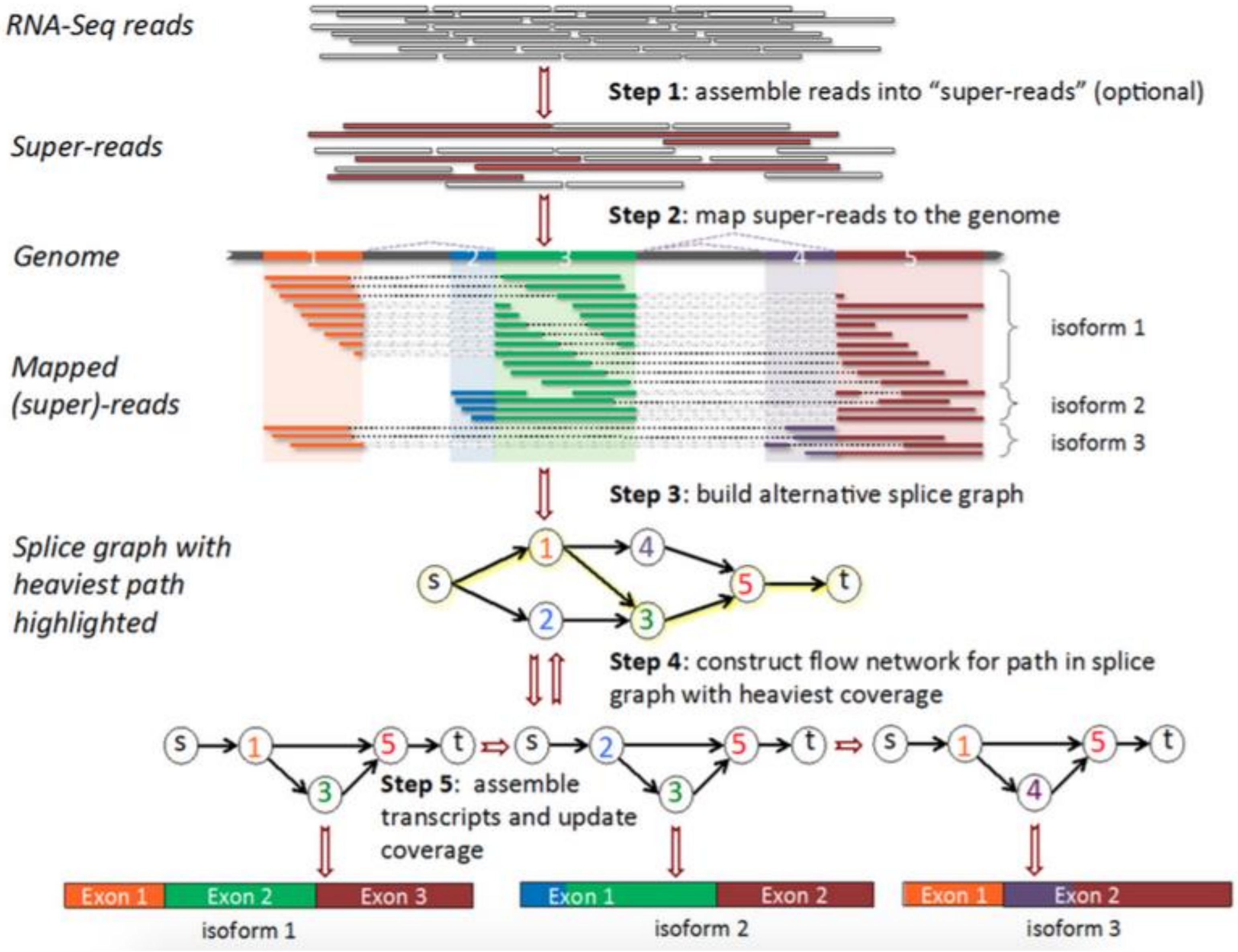
Single-end reads

Isoform 1

Isoform 2

Pair-end reads

StringTie workflow



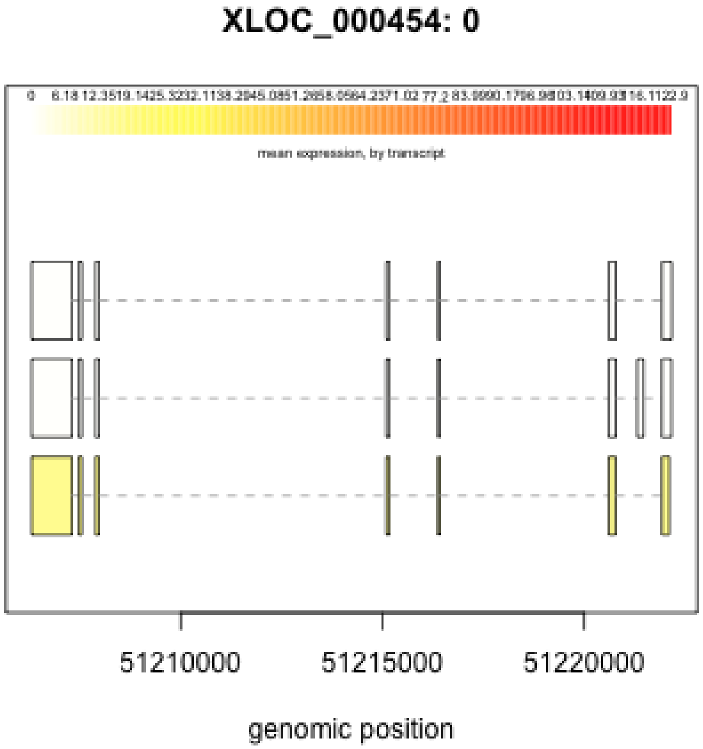
RNA-Seq reads
Step 1: assemble reads into "super-reads" (optional)
Super-reads
Step 2: map super-reads to the genome
Genome
1
2
3
4
5
Mapped (super)-reads
isoform 1
isoform 2
isoform 3
Step 3: build alternative splice graph
Splice graph with heaviest path highlighted
s
1
2
3
4
5
t
Step 4: construct flow network for path in splice graph with heaviest coverage
Step 5: assemble transcripts and update coverage
Exon 1
Exon 2
Exon 3
isoform 1
Exon 1
Exon 2
isoform 2
Exon 1
Exon 2
isoform 3

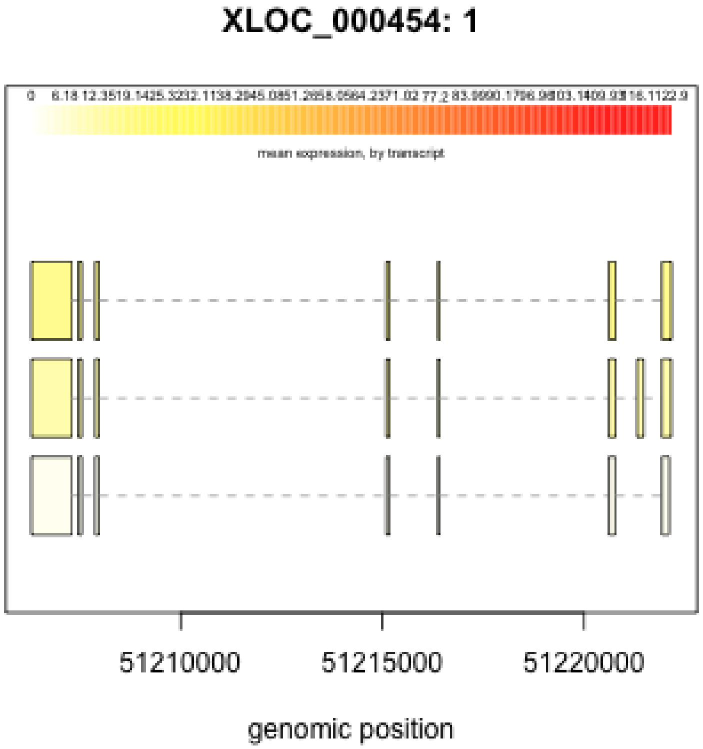
Nature Biotechnology 33, 290–295 (2015)

Ballgown

- Bridged the gap of transcripts assembly and differential expression analysis
    - – RSEM + edgeR
- Statistical methods are conceptual similar to limma
- Super fast

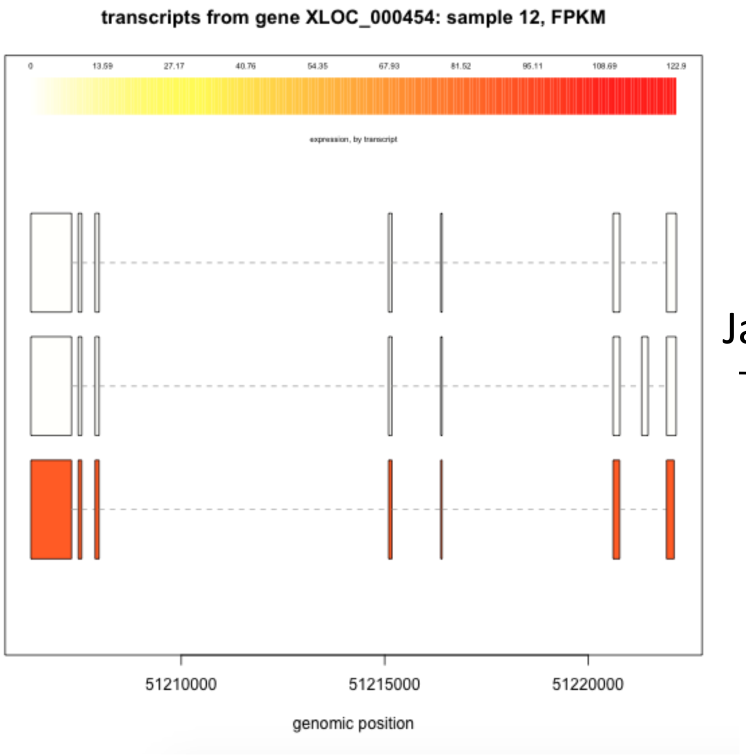
Ballgown visualization



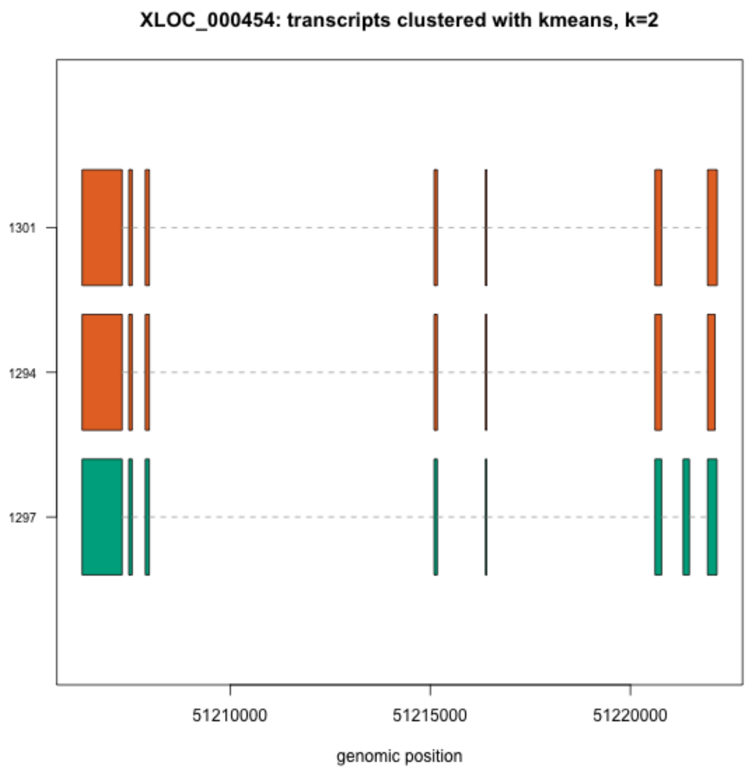


Ballgown: transcripts clustering

- Expression estimates are unreliable for very similar transcripts of a same gene



Jaccard distance →



Astrocyte Vizapp demo and workshop