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

- **Sequence**
  - Read length
  - PE vs. SR
  - Sequencing errors

- **Biological question**
  - Expression quantification
  - Alternative splicing
  - De novo assembly needed
  - mRNAs, small RNAs
  - ....

- **Bioinformatics**
  - Aligner
  - Annotation
  - Normalization
  - DE analysis strategy

*Everything's connected* slide by Dündar et al. (2015)
General RNA-seq Workflow
RNASeq Analysis Pipeline

- FASTQ → BAM (HiSAT, Samtools, MarkDuplicates)
- BAM → featureCount
- stringTie → Transcript Assembly (ballgown)
- Read Counts per Gene in Exon Regions
- QuSage → Gene Set Enrichment
- edgeR → Differential Expressed Genes
- Isoform Comparison

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
How to decide strand

Reverse stranded

Stranded

fr-firststrand

fr-secondstrand

fr-unstrand
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

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

- 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:

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

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

```r
> 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_5  6.919684  7.071573  7.328437  7.606885
gene_6  8.493105  8.696510  8.923458  9.210440
```

```r
> 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:

```r
> 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

Hierarchical clustering dendrogram

Distance calculation

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

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 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.
Real example
Test for differential expressed genes

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

\begin{verbatim}
> glmFit()
> design <- model.matrix(~ group)
> fit <- glmFit(y, design)
> res <- glmLRT(fit, coef=2)
\end{verbatim}
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="monocytes")

• Use coefficient to extract results
  – res <- glmLRT(fit, coef=2)
Make a model and extract results

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

```r
> 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
  – results

```r
> d <- model.matrix(~SubjectID + SampleGroup)
> d

  (Intercept) SubjectID21 SubjectID44 SubjectID53 SampleGroupneutrophils
  1       1         0         0         1         1
  2       1         0         0         1         1
  3       1         1         0         0         0
  4       1         1         0         0         0
  5       1         0         0         0         0
  6       1         0         0         0         0
  7       1         0         1         0         0
  8       1         0         1         0         0
```

More complicated comparison models

- Time series: treatment and control, 5 time points

```r
coldata
> 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
```

definition <- model.matrix
(~Treat+Time+Treat:Time, data=coldata)
More complicated comparison models

• 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
Define differential expressed genes

FDR and/or logFC cutoff
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

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
# Available Workflows

<table>
<thead>
<tr>
<th>Workflow Name</th>
<th>Description</th>
<th>Current Version</th>
<th>Author</th>
<th>Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Astrocyte Example ChiPSeq Workflow</strong></td>
<td>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.</td>
<td>0.0.5</td>
<td>David Trudgan</td>
<td><a href="mailto:biohpc-help@utsouthwestern.edu">biohpc-help@utsouthwestern.edu</a></td>
</tr>
<tr>
<td><strong>Example Wordcount Workflow</strong></td>
<td>This is a minimal test workflow package that counts the occurrences of words in a test file. It can be used as a template to develop workflows, and as to test the astrocyte platform.</td>
<td>0.0.4</td>
<td>David Trudgan</td>
<td><a href="mailto:biohpc-help@utsouthwestern.edu">biohpc-help@utsouthwestern.edu</a></td>
</tr>
<tr>
<td><strong>BICF RNA Seq Analysis Workflow</strong></td>
<td>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.</td>
<td>0.1.0</td>
<td>Brando Cantarel</td>
<td><a href="mailto:biohpc-help@utsouthwestern.edu">biohpc-help@utsouthwestern.edu</a></td>
</tr>
<tr>
<td><strong>BICF Somatic Mutation Calling</strong></td>
<td>This is a workflow package for the BioHPC/BICF Somatic Mutation workflow system. It implements a simple Somatic Mutation analysis workflow.</td>
<td>0.0.1</td>
<td>Brando Cantarel</td>
<td><a href="mailto:biohpc-help@utsouthwestern.edu">biohpc-help@utsouthwestern.edu</a></td>
</tr>
<tr>
<td><strong>BICF Germline Variant Analysis Workflow</strong></td>
<td>This is a workflow package for the BioHPC/BICF Germline Variant workflow system. It implements a simple germline variant analysis workflow using TrimGalore, EWA, Speedseq, GATK, Samtools and Platypus. SNPs and Indels are integrated using BAYSiC. Then annotated using SNPEFF and SnipSift.</td>
<td>0.0.7</td>
<td>Brando Cantarel</td>
<td><a href="mailto:biohpc-help@utsouthwestern.edu">biohpc-help@utsouthwestern.edu</a></td>
</tr>
<tr>
<td><strong>Astrocyte GCRB ChiPSeq Workflow</strong></td>
<td>This is an GCRB chipseq workflow package for the BioHPC astrocyte workflow system. It implements a simple ChiPSeq analysis workflow.</td>
<td>0.0.4</td>
<td>GCRB</td>
<td><a href="mailto:biohpc-help@utsouthwestern.edu">biohpc-help@utsouthwestern.edu</a></td>
</tr>
</tbody>
</table>
Create a new project

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.

### My Projects

#### Start a New Project

<table>
<thead>
<tr>
<th>Project Name</th>
<th>Create New Project</th>
</tr>
</thead>
</table>

#### Existing Projects

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Created</th>
<th>Workflows Run</th>
<th>Input Files</th>
<th>Size</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRJ21</td>
<td>RNAs_seq_test</td>
<td>Aug. 23, 2016, 3:03 p.m.</td>
<td>0</td>
<td>0</td>
<td>0 bytes</td>
<td></td>
</tr>
</tbody>
</table>

#### Projects Shared with Me

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Created</th>
<th>Workflows Run</th>
<th>Input Files</th>
<th>Size</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRJ10</td>
<td>test</td>
<td>June 1, 2016, 5:02 p.m. by Brandi Cantarel</td>
<td>4</td>
<td>10</td>
<td>218.5 GB</td>
<td></td>
</tr>
</tbody>
</table>
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
Add data to your project

For NGS experiment, this is recommended.

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT1_R1.fasq</td>
<td>4.0 GB</td>
</tr>
<tr>
<td>WT2_R1.fasq</td>
<td>4.1 GB</td>
</tr>
<tr>
<td>K03_R2.fasq</td>
<td>4.4 GB</td>
</tr>
<tr>
<td>K04_R2.fasq</td>
<td>4.5 GB</td>
</tr>
<tr>
<td>K02_R1.fasq</td>
<td>4.0 GB</td>
</tr>
<tr>
<td>WT2_R2.fasq</td>
<td>4.1 GB</td>
</tr>
<tr>
<td>K04_R1.fasq</td>
<td>4.5 GB</td>
</tr>
<tr>
<td>WT1_R2.fasq</td>
<td>4.0 GB</td>
</tr>
<tr>
<td>K03_R1.fasq</td>
<td>4.4 GB</td>
</tr>
</tbody>
</table>
Make your design file

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

SampleID

This ID should match the name in the fastq file i.e. 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

<table>
<thead>
<tr>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project 28: RNASeqTest</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name for this run</th>
</tr>
</thead>
<tbody>
<tr>
<td>test_0.1.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>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</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRR1550987_1.fastq.gz</td>
</tr>
<tr>
<td>SRR1550986_2.fastq.gz</td>
</tr>
<tr>
<td>SRR1550985_1.fastq.gz</td>
</tr>
<tr>
<td>SRR1551069_2.fastq.gz</td>
</tr>
<tr>
<td>SRR1551069_1.fastq.gz</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>In the case that the sequence libraries were generated using a stranded specific protocol:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstranded</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paired End</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remove Duplicates</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A design file listing pairs of sample name and sample group. Columns must include: SampleID,SampleName,SampleGroup,FullPathToFqR1,FullPathToFqR2</th>
</tr>
</thead>
<tbody>
<tr>
<td>design_pe.txt</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference genome for alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human GRCh38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene Set Definitions used for QUSAGE Analysis – see <a href="http://software.broadinstitute.org/gsea/msigdb">http://software.broadinstitute.org/gsea/msigdb</a> for gene set descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hallmark Gene Sets</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run Workflow</th>
</tr>
</thead>
</table>
Download/visualize your results

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

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-end sequencing

Single-end reads

Isoform 1

Isoform 2

Pair-end reads
StringTie workflow
Ballgown

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

XLOC_000454: 0

XLOC_000454: 1

mean expression, by transcript

genomic position
Ballgown: transcripts clustering

- Expression estimates are unreliable for very similar transcripts of a same gene
Astrocyte Vizapp demo and workshop