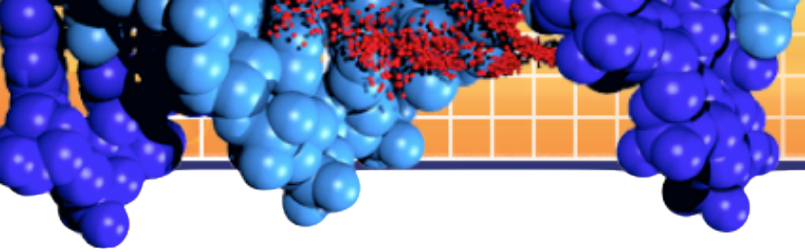


Bioinformatics Short Course: RNA-Seq Data Analysis

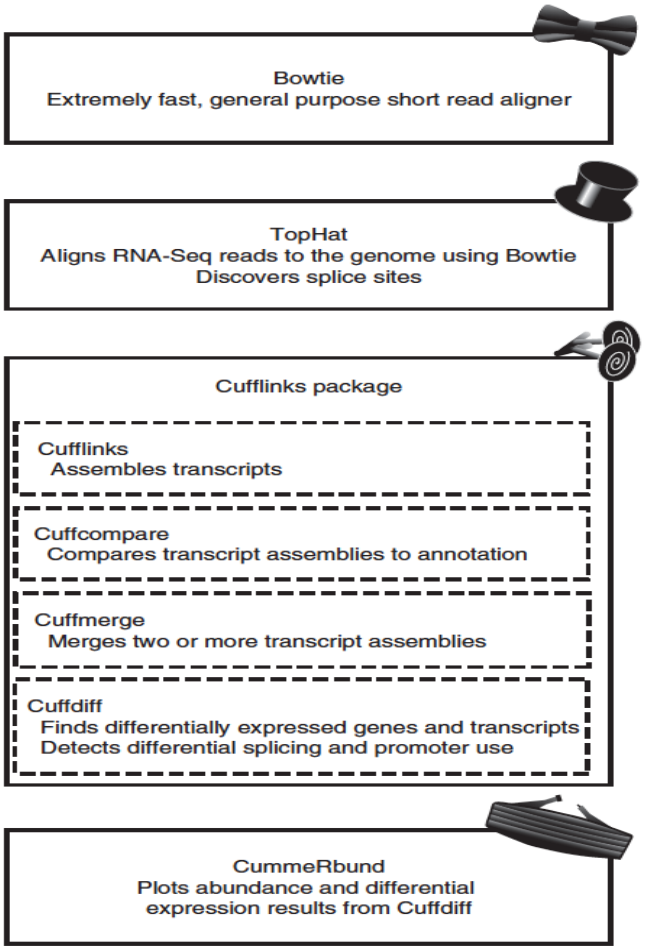
Part VI: Expression Analysis (Exercises)

Chuming Chen Ph. D.
University of Delaware
May 22-23, 2012

Summary (Lecture)

- Transcriptome assembly strategies
- Short read aligners
- Alignment format and SAMtools
- Alignment visualization

Software Components of Tuxedo Suite Tools



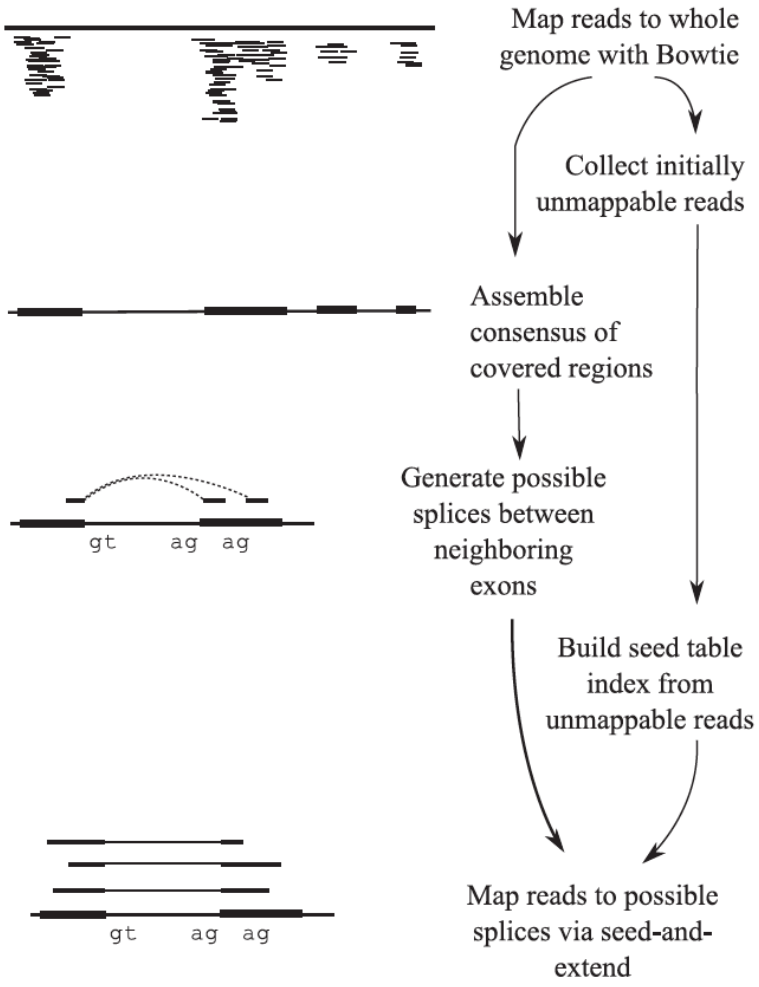
- **Bowtie** forms the algorithmic core of TopHat, which align reads to the reference genome.
- **TopHat**'s read alignments are assembled by **Cufflinks** and its associated utility program (**Cuffmerge**, **Cuffcompare**) can produce a transcriptome annotation of the genome.
- **Cuffdiff** quantifies this transcriptome across multiple conditions using the TopHat read alignments.
- **CummeRbund** explores and visualizes the differential expression data (Genes and Transcripts) generated by Cuffdiff.

(Trapnell et al., **Nat Protoc**. 2012 Mar 1;7(3):562-78)

Bowtie

- An ultrafast, memory-efficient short read aligner.
- It uses an extremely economical data structure called the Burrows-Wheeler index to store the reference genome sequence and allows it to be searched rapidly at a rate of tens of millions reads per CPU hour.
- It makes a number of compromises to achieve its high speed:
  - If one or more exact matches exist for a read, it is guaranteed to find one.
  - If the best match is not exact match, then it is not guaranteed in all cases to find the highest quality alignment.
  - It may fail to align reads with multiple mismatches.
- Furthermore, Bowtie does not allow alignments between a read and the genome to contain large gaps; hence, it cannot align reads that span introns. TopHat was created to address this limitation.
- Web Site: http://bowtie-bio.sourceforge.net/index.shtml

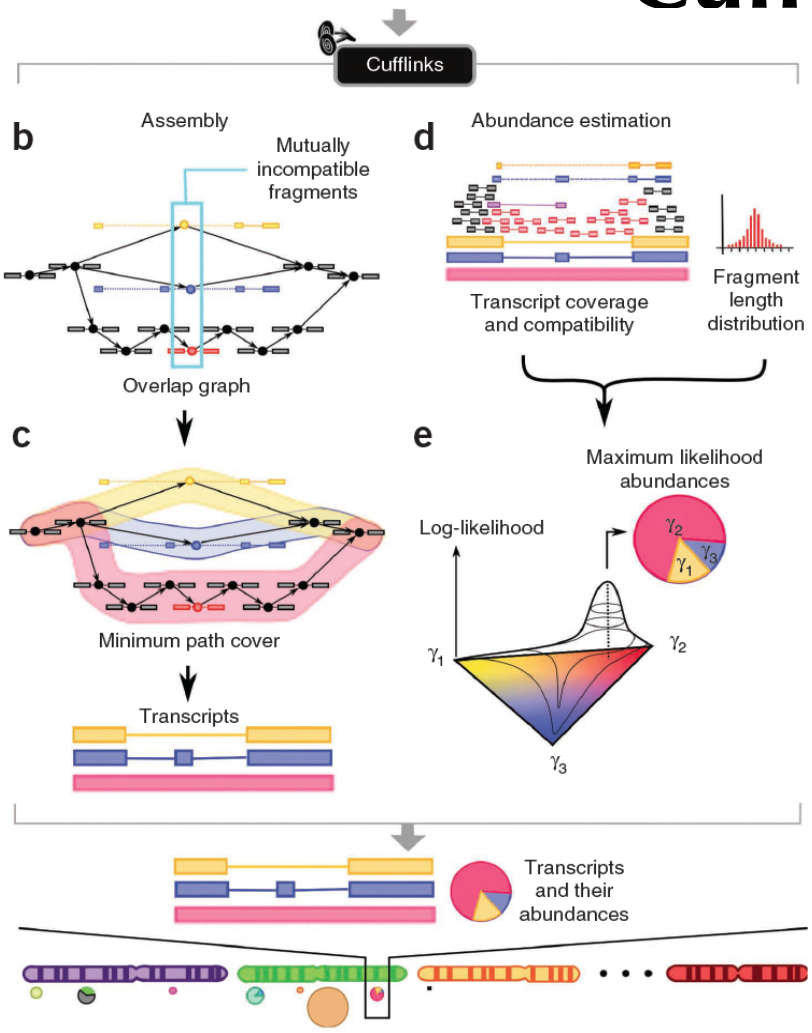
TopHat



- Use Bowtie as alignment engine.
- Break up reads Bowtie cannot align into segments then align them independently.
- When several of a read's segments aligned to the genome far apart, TopHat infers that the read spans a splice junction and estimate the splice site.
- By using the 'initially unmapped' reads, TopHat can build an index of splice sites in the transcriptome on the fly without a prior gene or splice site annotations.
- Web site: http://tophat.cbcb.umd.edu/

(Trapnell et al. **Bioinformatics**. 2009 May 1;25(9):1105-11)

Cufflinks



- Assembles individual transcripts from RNA-Seq reads that have been aligned to the genome.
- Reports as few full-length transcript fragments or 'transfrags' as are needed to 'explain' all the splicing events in the input data.
- Quantifies the expression level of each transfrag in the sample using a rigorous statistical model of RNA-Seq to filter out background or artifactual transfrags such as immature primary transcripts.
- Quantifies transcript abundance using a reference annotation.
- Web site: http://cufflinks.cbcb.umd.edu/

(Trapnell et al. **Nat Biotechnol**. 2010 May;28(5):511-5)

Cuffcompare

- In addition to differential expression analysis, people are often interested in discovering new genes and transcripts.
- Gaps in sequencing coverage will cause breaks in transcript reconstruction and make it difficult to distinguish full-length novel transcripts from partial fragments.
- Cuffcompare can compare the Cufflinks assemblies to reference annotation files and help sort out new genes from known ones.
- Web site: http://cufflinks.cbcb.umd.edu/manual.html#cuffcompare

Print Summary Reports

```
$ ls -tlr cufflinks_out_*/*map
-rw-r--r-- 1 chenc cwu 481209 May 17 14:35 cufflinks_out_FL1/cuffcmp.transcripts.gtf.tmap
-rw-r--r-- 1 chenc cwu 196073 May 17 14:35 cufflinks_out_FL1/cuffcmp.transcripts.gtf.refmap
-rw-r--r-- 1 chenc cwu 479849 May 17 14:35 cufflinks_out_FL2/cuffcmp.transcripts.gtf.tmap
-rw-r--r-- 1 chenc cwu   6144 May 17 14:35 cufflinks_out_FL2/cuffcmp.transcripts.gtf.refmap
-rw-r--r-- 1 chenc cwu 482416 May 17 14:35 cufflinks_out_LL1/cuffcmp.transcripts.gtf.tmap
-rw-r--r-- 1 chenc cwu   4466 May 17 14:35 cufflinks_out_LL1/cuffcmp.transcripts.gtf.refmap
-rw-r--r-- 1 chenc cwu 476989 May 17 14:35 cufflinks_out_LL2/cuffcmp.transcripts.gtf.tmap
-rw-r--r-- 1 chenc cwu   3880 May 17 14:35 cufflinks_out_LL2/cuffcmp.transcripts.gtf.refmap
```

.tmap

This tab delimited file lists the most closely matching reference transcript for each Cufflinks transcript. There is one row per Cufflinks transcript.

.refmap

This tab delimited file lists, for each reference transcript, which cufflinks transcripts either fully or partially match it. There is one row per reference transcript.

The following bash script prints a simple table for each assembly that lists how many transcripts in each assembly are complete matches to the know transcripts, how many are partial matches etc.

```
$ find . -name *.tmap | while read file; do echo $file;  awk 'NR > 1 { s[$3]++ } END { for (j in s) { print
j, s[j] }} ' $file; done
```

Cuffcompare Summary Reports

```
./cufflinks_out_FL1/cuffcmp.transcripts.gtf.tmap
u 116
i 20
j 471
x 1
o 6
c 42
p 87
= 2973
e 25
s 2
./cufflinks_out_FL2/cuffcmp.transcripts.gtf.tmap
u 110
i 21
j 479
x 1
o 6
c 42
p 76
= 2971
e 24
s 1
./cufflinks_out_LL1/cuffcmp.transcripts.gtf.tmap
u 129
i 23
j 481
x 1
o 6
c 43
p 89
= 2963
e 24
s 1
./cufflinks_out_LL2/cuffcmp.transcripts.gtf.tmap
u 111
i 17
j 468
x 1
o 9
c 42
p 71
= 2968
e 21
```

Code	Description
=	Complete match of intron chain
c	Contained
j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment
i	A transfrag falling entirely within a reference transcript
o	Generic exonic overlap with a reference transcript
P	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
u	Unknown, intergenic transcript
x	Exonic overlap with reference on the opposite strand
s	An intron of the transfrag overlaps a reference intro on the opposite strand (likely due to read mapping errors)
-	.tracking file only, indicates multiple classification

(http://cufflinks.cbcb.umd.edu/manual.html#cuffcompare)

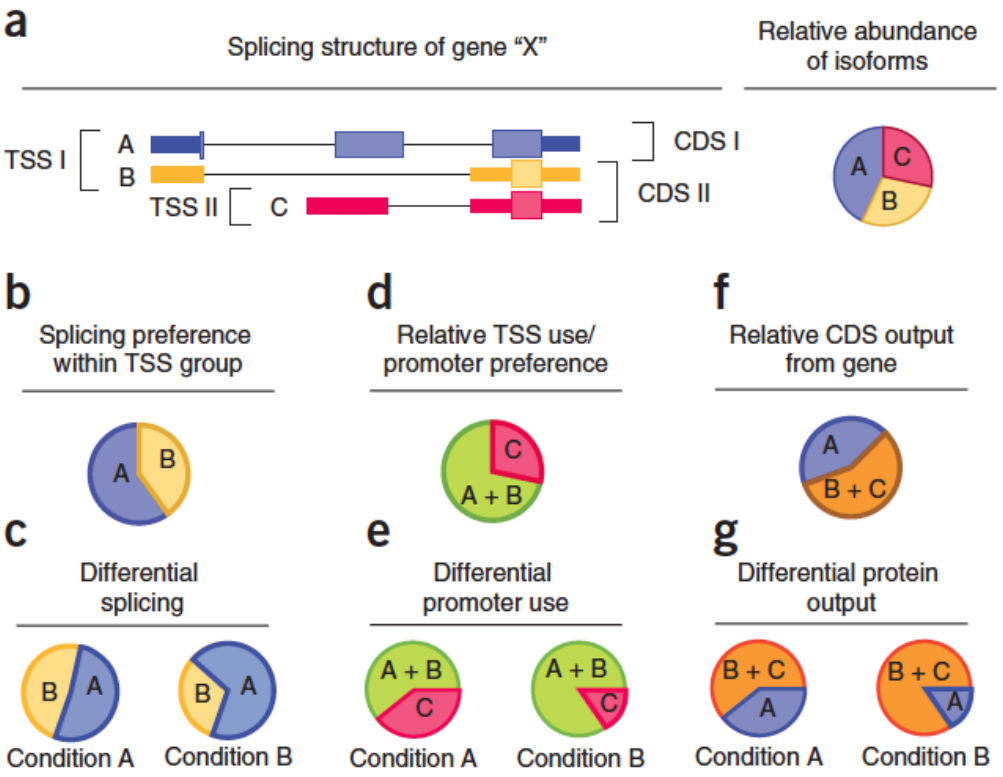
Cuffmerge

- In multi-sample RNA-Seq experiment, sometime it is necessary to pool the data and assemble them into a comprehensive set of transcripts before differential analysis.
- Pool aligned reads from all samples and run Cufflinks once on them is not recommended:
  - Assembly becomes more computationally expensive as read depth increases.
  - Complex mixture of splice isoforms for many genes may lead to the incorrectly assembled transcripts.
- As a 'meta-assembler', Cuffmerge parsimoniously merges the individually assemblies by Cufflinks by treating the assembled transfrags the way Cufflinks treats the reads.
- It can also performs a reference annotation-based transcript (RABT) (Roberts et al. 2011) assembly to merge reference transcripts with assembled sample transfrags to produce a single annotation file for downstream differential analysis.
- Web site: http://cufflinks.cbcb.umd.edu/manual.html#cuffmerge

Cuffdiff

- Calculates expression in two or more samples and tests the statistical significance of each observed change in expression between them.
- The statistical model assumptions:
  - The number of reads produced by each transcript is proportional to its abundance.
  - It fluctuates due to technical variability during library preparation and sequencing, and the biological variability between replicates of the same experiment.
- Allows multiple technical or biological replicate sequencing libraries per condition.
- Reports gene and transcript expression level changes in tabular format, which includes fold change (in log2), P-value (both raw and corrected for multiple hypotheses testing), gene and transcript related information such as name and location in the genome.
- Web site: http://cufflinks.cbcb.umd.edu/manual.html#cuffdiff

Cuffdiff (additional differential analysis)



- Identify genes that are differentially spliced or regulated via promoter switching
- Group isoforms of a gene that have the same TSS (derived from the same pre-mRNA, changes in abundance reflect the differential splicing of common pre-mRNA).
- Total expression levels of a TSS group is the sum of expression levels of the isoforms within it.
- Relative abundance between multiple TSSs reflect the changes in TSS (promoter) preferences between condition.

(Trapnell et al., **Nat Protoc**. 2012 Mar 1;7(3):562-78)

GTF format

- GTF stands for **G**ene **T**ransfer **F**ormat.
- The tab-delimited file includes fields below:
  - <seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes] [comments]

(http://mblab.wustl.edu/GTF22.html)

chr1	Cufflinks	transcript	84015	84983	1	-	.	gene_id "ENSGALG00000009775"; transcript_id "ENSGALT00000015896";
chr1	Cufflinks	exon	84015	84983	1	-	.	gene_id "ENSGALG00000009775"; transcript_id "ENSGALT00000015896"; exon_number "1";
chr1	Cufflinks	transcript	6268	21192	1000	+	.	gene_id "CUFF.1"; transcript_id "ENSGALT00000015891"; FPKM "26.6821513228";
chr1	Cufflinks	exon	6268	6477	1000	+	.	gene_id "CUFF.1"; transcript_id "ENSGALT00000015891"; exon_number "1"; FPKM "26.6821513228";
chr1	Cufflinks	exon	16287	16386	1000	+	.	gene_id "CUFF.1"; transcript_id "ENSGALT00000015891"; exon_number "2"; FPKM "26.6821513228";
chr1	Cufflinks	exon	18353	18470	1000	+	.	gene_id "CUFF.1"; transcript_id "ENSGALT00000015891"; exon_number "3"; FPKM "26.6821513228";
chr1	Cufflinks	exon	19705	19806	1000	+	.	gene_id "CUFF.1"; transcript_id "ENSGALT00000015891"; exon_number "4"; FPKM "26.6821513228";
chr1	Cufflinks	exon	20015	20196	1000	+	.	gene_id "CUFF.1"; transcript_id "ENSGALT00000015891"; exon_number "5"; FPKM "26.6821513228”;
chr1	Cufflinks	exon	20399	20505	1000	+	.	gene_id "CUFF.1"; transcript_id "ENSGALT00000015891"; exon_number "6"; FPKM "26.6821513228";
chr1	Cufflinks	exon	20595	21192	1000	+	.	gene_id "CUFF.1"; transcript_id "ENSGALT00000015891"; exon_number "7"; FPKM "26.6821513228";

GTF Field Definitions

<seqname> - The name of the sequence. Commonly, this is the chromosome ID or contig ID. Note that the coordinates used must be unique within each sequence name in all GTFs for an annotation set.

<source> - The source column should be a unique label indicating where the annotations came from --- typically the name of either a prediction program or a public database.

<feature> - The following feature types are required: "CDS", "start_codon", "stop_codon". The features "5UTR", "3UTR", "inter", "inter_CNS", "intron_CNS" and "exon" are optional. All other features will be ignored. The types must have the correct capitalization shown here.

<start> <end> - Integer start and end coordinates of the feature relative to the beginning of the sequence named in <seqname>. <start> must be less than or equal to <end>. Sequence numbering starts at 1. Values of <start> and <end> that extend outside the reference sequence are technically acceptable, but they are discouraged.

<score> - The score field indicates a degree of confidence in the feature's existence and coordinates. The value of this field has no global scale but may have relative significance when the <source> field indicates the prediction program used to create this annotation. It may be a floating point number or integer, and not necessary and may be replaced with a dot.

<frame> - 0 indicates that the feature begins with a whole codon at the 5' most base. 1 means that there is one extra base (the third base of a codon) before the first whole codon and 2 means that there are two extra bases (the second and third bases of the codon) before the first codon. Note that for reverse strand features, the 5' most base is the <end> coordinate.

(http://mblab.wustl.edu/GTF22.html)

BED format

- BED format provides a flexible way to define the data lines that are displayed in an annotation track BED lines have 12 fields.
  - Required fields:
    - chrom, chromStart, chromEnd
  - Additional optional fields:
    - name, score, strand, thickStart, thickEnd, itemRgb, blockCount, blockSizes, blockStarts.

(http://genome.ucsc.edu/FAQ/FAQformat#format1)

```
track name=junctions description="TopHat junctions"
chr1    13983  19738  JUNC00000001    1    +    13983  19738  255,0,0    2    46,34   0,5721
chr1    20459  20649  JUNC00000002    1    +    20459  20649  255,0,0    2    46,55   0,135
chr1    33994  35271  JUNC00000003    1    +    33994  35271  255,0,0    2    28,69   0,1208
chr1    41809  42420  JUNC00000004    2    +    41809  42420  255,0,0    2    97,61   0,550
chr1    42392  45268  JUNC00000005    3    +    42392  45268  255,0,0    2    81,74   0,2802
chr1    46828  48793  JUNC00000006    1    +    46828  48793  255,0,0    2    46,38   0,1927
```

Exercise 4

View Alignment, Coverage, and Isoforms

(SAMtools, IGV)

Index Alignment Files for IGV

```
$ cat ~/rnaseq-shared/pbs_scripts/samtools_index.qs
#PBS -N SamtoolsIndex
#PBS -S /bin/bash
#PBS -V
#PBS -l ncpus=1,walltime=16:00:00,cput=10:00:00,mem=2000mb,nodes=1:ppn=4
#PBS -q rnaseq

cd $PBS_O_WORKDIR
samtools faidx index/gallus_chr1.fa
ln -s accepted_hits.bam tophat_out_FL1/FL1.bam
ln -s accepted_hits.bam tophat_out_FL2/FL2.bam
ln -s accepted_hits.bam tophat_out_LL1/LL1.bam
ln -s accepted_hits.bam tophat_out_LL2/LL2.bam

samtools index tophat_out_FL1/FL1.bam
samtools index tophat_out_FL2/FL2.bam
samtools index tophat_out_LL1/LL1.bam
samtools index tophat_out_LL2/LL2.bam

ln -s transcripts.gtf cufflinks_out_FL1/FL1_transcripts.gtf
ln -s transcripts.gtf cufflinks_out_FL2/FL2_transcripts.gtf
ln -s transcripts.gtf cufflinks_out_LL1/LL1_transcripts.gtf
ln -s transcripts.gtf cufflinks_out_LL2/LL2_transcripts.gtf

$ qsub ~/rnaseq-shared/pbs_scripts/samtools_index.qs
90297.biohen.dbi.local
$ ls –ltr tophat_out_*/
$ ls -tlr index/
total 412968
-rw-r--r-- 1 chenc cwu  48797843 May 17 14:03 gallus_chr1.4.ebwt
-rw-r--r-- 1 chenc cwu     89909 May 17 14:03 gallus_chr1.3.ebwt
-rw-r--r-- 1 chenc cwu  60083414 May 17 14:06 gallus_chr1.1.ebwt
-rw-r--r-- 1 chenc cwu  24398928 May 17 14:06 gallus_chr1.2.ebwt
-rw-r--r-- 1 chenc cwu  60083414 May 17 14:08 gallus_chr1.rev.1.ebwt
-rw-r--r-- 1 chenc cwu  24398928 May 17 14:08 gallus_chr1.rev.2.ebwt
-rw-r--r-- 1 chenc cwu 205013902 May 17 15:07 gallus_chr1.fa
-rw-r--r-- 1 chenc cwu        23 May 17 15:07 gallus_chr1.fa.fai
```

- Your login ends with odd number, i. e. rna7

  ssh –X rna7@glycine.dbi.udel.edu

- Your login ends with even number, i. e. rna8

  ssh –X rna8@biohen.dbi.udel.edu

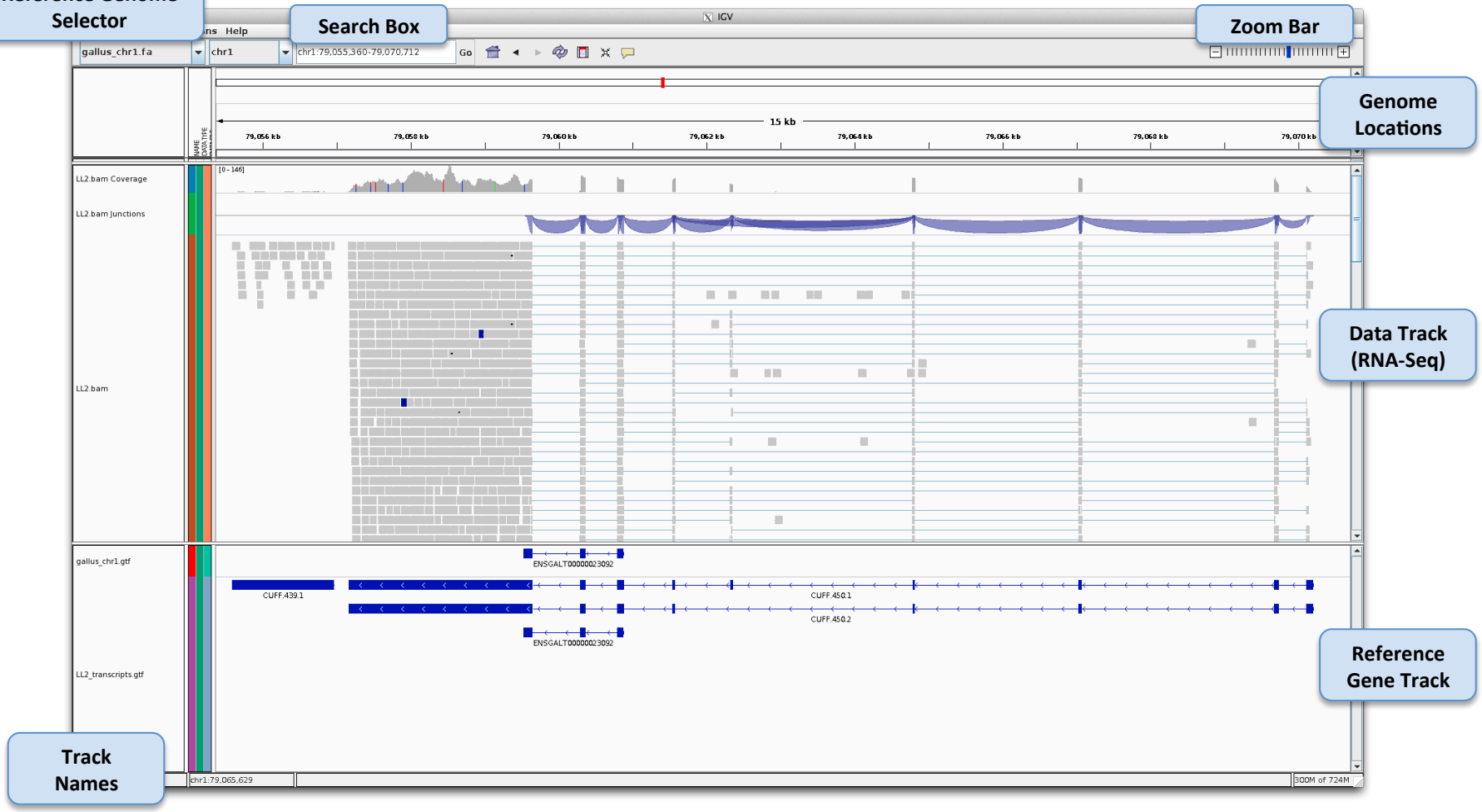
  Qsub –IXV –q rnaseq

Launch IGV

```
$ qsub -I -X -V -q rnaseq

$ cd ~/rnaseq-work

$ igv.sh
```

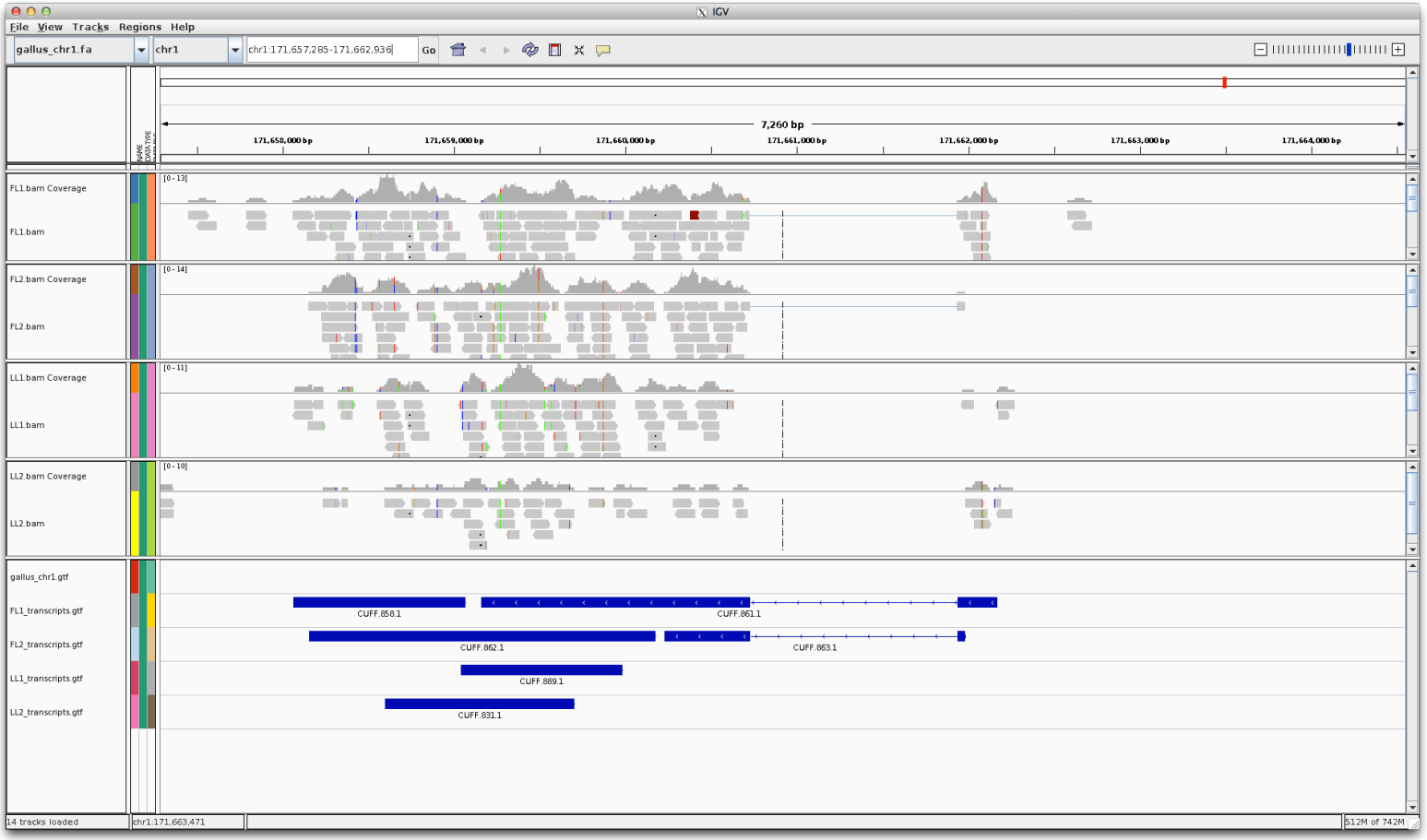
- Click “File”, then click “Load Genome from File …”, then select **gallus_chr1.fa** from the directory called **reference**.
- Click “File”, then click “Load from File …”, then select **gallus_chr1.gtf** from the directory called **reference**.
- Click “File”, then click “Load from File …”, then select **LL2.bam** from the directory called **tophat_out_LL2**.
- Click “File”, then click “Load from File …”, then select **LL2_transcripts.gtf** from the directory called **cufflinks_out_LL2**.
- Type in **chr1:79055360-79070712** in the search box and click “Go”.
- Right click **LL2_transcripts.gtf** track and select “Expanded”.

View Splice Junctions and Isoforms of MFAP5



View Differentially Expressed Novel Gene

- Type in **chr1:171657285-171662936** in the search box and click “Go”.



Exercise 6

Explore Differential Analysis Results
(CummeRbund)

Tabular view of differentially expressed genes

```
$ cd ~/rnaseq-work
$ head -3 cuffdiff_out/gene_exp.diff
```

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant	
XLOC_000001	XLOC_000001	LOC425783	chr1:6267-21192	FL	LL	NOTEST	37.6264	96.7431	1.36241	-1.4275	0.153436	1	no	
XLOC_000002	XLOC_000002	GOLGB1	chr1:33922-67653	FL	LL	OK	135.819	155.192	0.192365	-0.820333	0.412026	0.832533	no	

```
$ grep yes cuffdiff_out/gene_exp.diff  | cut –f2-
```

XLOC_000151	DOCK4	chr1:28911240-29077867	FL	LL	OK	315.97	190.759	-0.728074	3.96029	7.48577e-05	0.00764796	yes		
XLOC_000206	LGR5	chr1:38117567-38203984	FL	LL	OK	101.952	91.104	1.51355	-4.5102	6.47663e-06	0.000794035	yes		
XLOC_000509	B3TZB5_CHICK	chr1:71254727-71270462	FL	LL	OK	3619.21	4752.44	0.392991	-3.55812	0.000373513	0.0190803	yes		
XLOC_000581	Q5F3N3_CHICK	chr1:80541032-80550319	FL	LL	OK	291.884	712.317	1.28712	-4.82633	1.39072e-06	0.000284171	yes		
XLOC_000640	CCD80_CHICK	chr1:86865454-86889448	FL	LL	OK	1500.05	2234.23	0.574764	-4.70744	2.50853e-06	0.000384432	yes		
XLOC_000711	PTGFRN	chr1:95337409-95402844	FL	LL	OK	209.043	85.674	0.883563	-3.84727	0.000119441	0.00915213	yes		
XLOC_000818	SRPX	chr1:116557728-116600189	FL	LL	OK	259.464	592.13	1.19038	-3.65185	0.000260356	0.0159599	yes		
XLOC_000875	MXRA5	chr1:132390250-132407787	FL	LL	OK	378.182	786.069	1.05558	-8.89053	0	0	yes		
XLOC_001059	Q6DMS3_CHICK	chr1:176287499-176321543	FL	LL	OK	3713.88	5652.7	0.606012	-3.86014	0.000113324	0.00915213	yes		
XLOC_001315	LAMB1_CHICK	chr1:15859496-15899985	FL	LL	OK	487.081	772.948	0.666211	-3.56962	0.000357498	0.0190803	yes		
XLOC_001494	FAM109B	chr1:51307286-51312438	FL	LL	OK	214.384	28.553	0.999238	-3.24371	0.00117986	0.0482168	yes		
XLOC_001708	MFAP5	chr1:79057160-79080614	FL	LL	OK	4509.86	781.72	0.588585	-3.36647	0.000761379	0.035902	yes		
XLOC_001826	PROS1	chr1:92892486-92925118	FL	LL	OK	1121.44	96.313	-1.17604	6.85379	7.19202e-12	2.20436e-09	yes		
XLOC_001875	ADAMTS1	chr1:106484837-106492792	FL	LL	OK	369.098	197.532	-0.901915		3.33526	0.000852187	0.0373136	yes	
XLOC_002164	-	chr1:171658063-171662169	FL	LL	OK	366.494	156.924	-1.22372	3.8001	0.000144639	0.00985155	yes		

```
$ head -1 cuffdiff_out/gene_exp.diff > sig_diff_genes.txt
$ grep yes cuffdiff_out/gene_exp.diff | cut -f2- >> sig_diff_genes.txt
$ cat sig_diff_genes.txt
```

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant	
XLOC_000151	DOCK4	chr1:28911240-29077867	FL	LL	OK	315.979	190.759	-0.728074	3.96029	7.48577e-05	0.00764796	yes		
XLOC_000206	LGR5	chr1:38117567-38203984	FL	LL	OK	101.959	291.104	1.51355	-4.5102	6.47663e-06	0.000794035	yes		
XLOC_000509	B3TZB5_CHICK	chr1:71254727-71270462	FL	LL	OK	3619.21	4752.44	0.392991	-3.55812	0.000373513	0.0190803	yes		
XLOC_000581	Q5F3N3_CHICK	chr1:80541032-80550319	FL	LL	OK	291.884	712.317	1.28712	-4.82633	1.39072e-06	0.000284171	yes		
XLOC_000640	CCD80_CHICK	chr1:86865454-86889448	FL	LL	OK	1500.05	2234.23	0.574764	-4.70744	2.50853e-06	0.000384432	yes		
XLOC_000711	PTGFRN	chr1:95337409-95402844	FL	LL	OK	209.046	385.674	0.883563	-3.84727	0.000119441	0.00915213	yes		
XLOC_000818	SRPX	chr1:116557728-116600189	FL	LL	OK	259.464	592.13	1.19038	-3.65185	0.000260356	0.0159599	yes		
XLOC_000875	MXRA5	chr1:132390250-132407787	FL	LL	OK	378.182	786.069	1.05558	-8.89053	0	0	yes		
XLOC_001059	Q6DMS3_CHICK	chr1:176287499-176321543	FL	LL	OK	3713.88	5652.7	0.606012	-3.86014	0.000113324	0.00915213	yes		
XLOC_001315	LAMB1_CHICK	chr1:15859496-15899985	FL	LL	OK	487.081	772.948	0.666211	-3.56962	0.000357498	0.0190803	yes		
XLOC_001494	FAM109B	chr1:51307286-51312438	FL	LL	OK	214.389	428.553	0.999238	-3.24371	0.00117986	0.0482168	yes		
XLOC_001708	MFAP5	chr1:79057160-79080614	FL	LL	OK	4509.81	6781.72	0.588585	-3.36647	0.000761379	0.035902	yes		
XLOC_001826	PROS1	chr1:92892486-92925118	FL	LL	OK	1121.45	496.313	-1.17604	6.85379	7.19202e-12	2.20436e-09	yes		
XLOC_001875	ADAMTS1	chr1:106484837-106492792	FL	LL	OK	369.098	197.532	-0.901915		3.33526	0.000852187	0.0373136	yes	
XLOC_002164	-	chr1:171658063-171662169	FL	LL	OK	366.494	156.924	-1.22372	3.8001	0.000144639	0.00985155	yes		

What is R?

- R is a data analysis software for statistical analysis, data visualization and predictive modeling.
- R is a complete, interactive, object-oriented programming language.
- R is an environment for statistical analysis, providing functions for virtually every data manipulation, statistical modeling.
- R is an open-source software project.
- R is a community of leading statisticians and computer scientists and thousands of contributors.
- http://www.r-project.org/
- http://www.bioconductor.org/

CummeRbund

- A user-friendly R package to help manage, visualize and integrate all the data generated by Cuffdiff analysis.
- Simplify the data exploration task such as plotting and cluster analysis of expression data.
- Scripted plotting automates the plot generation and reuse analyses from previous experiments.
- Transform Cuffdiff data into R statistical computing environment enables other advanced statistical analysis and plotting packages.
- Takes the various output files from a Cuffdiff run and creates a SQLite database of the results describing appropriate relationships between genes, transcripts, transcription start sites, and CDS regions to allow efficiently retrieval and exploration.
- Web site: http://compbio.mit.edu/cummeRbund/

Start R interactive shell

```
$ qsub –I –X –V –q rnaseq

$ cd ~/rnaseq-work

$ R

R version 2.15.0 (2012-03-30)
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-unknown-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Basic Syntax of R language

```
> x <- c(1,2,3,4,5,6)   # Create ordered collection (vector)
> y <- x^2              # Square the elements of x
> print(y)              # print (vector) y
[1]  1  4  9 16 25 36
> mean(y)               # Calculate average (arithmetic mean) of (vector) y; result is scalar
[1] 15.16667
> var(y)                # Calculate sample variance
[1] 178.9667
> lm_1 <- lm(y ~ x)     # Fit a linear regression model "y = f(x)" or "y = B0 + (B1 * x)" store the results as lm_1
> print(lm_1)           # Print the model from the (linear model object) lm_1

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     -9.333        7.000

> summary(lm_1)         # Compute and print statistics for the fit of the (linear model object) lm_1

Call:
lm(formula = y ~ x)

Residuals:
1       2       3       4       5       6
3.3333 -0.6667 -2.6667 -2.6667 -0.6667  3.3333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.3333     2.8441  -3.282 0.030453 *
x             7.0000     0.7303   9.585 0.000662 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.055 on 4 degrees of freedom
Multiple R-squared: 0.9583,     Adjusted R-squared: 0.9478
F-statistic: 91.88 on 1 and 4 DF,  p-value: 0.000662

> library(cairoDevice)  # external package provides Cairo() function.
> Cairo()               # Open an R graphics device based on the Cairo vector graphics
> par(mfrow=c(2, 2))    # Request 2x2 plot layout
> plot(lm_1)       # Diagnostic plot of regression model
> pdf("lm_1.pdf")  # starts the graphics device driver for producing PDF file "lm_1.pdf"
> par(mfrow=c(2,2))
> plot(lm_1)
> dev.off()        # shuts down the specified (by default the current) device
> list.files(pattern="pdf")
[1] "lm_1.pdf
```

Create a CummeRbund Database from Cuffdiff Output

```
> library(cummeRbund)
Loading required package: RSQLite
Loading required package: DBI
Loading required package: ggplot2
Loading required package: reshape2
> cuff_data <- readCufflinks('cuffdiff_out')
Creating database cuffdiff_out/cuffData.db
Reading cuffdiff_out/genes.fpkm_tracking
Checking samples table...
Populating samples table...
Writing genes table
Reshaping geneData table
Recasting
Writing geneData table
Reading cuffdiff_out/gene_exp.diff
Writing geneExpDiffData table
Reading cuffdiff_out/promoters.diff
Writing promoterDiffData table
Reading cuffdiff_out/isoforms.fpkm_tracking
Checking samples table...
OK!
Writing isoforms table
Reshaping isoformData table
Recasting
Writing isoformData table
Reading cuffdiff_out/isoform_exp.diff
Writing isoformExpDiffData table
Reading cuffdiff_out/tss_groups.fpkm_tracking
Checking samples table...
OK!
Writing TSS table
Reshaping TSSData table
Recasting
Writing TSSData table
Reading cuffdiff_out/tss_group_exp.diff
Writing TSSExpDiffData table
Reading cuffdiff_out/splicing.diff
Writing splicingDiffData table
Reading cuffdiff_out/cds.fpkm_tracking
Checking samples table...
OK!
Writing CDS table
Reshaping CDSData table
Recasting
Writing CDSData table
Reading cuffdiff_out/cds_exp.diff
Writing CDSExpDiffData table
Reading cuffdiff_out/cds.diff
Writing CDSDiffData table
Indexing Tables...
>
```

Accessing Data

```
> options(width=200)
> gene.feature <- features(genes(cuff_data))
> head(gene.feature)
      gene_id class_code nearest_ref_id    gene_short_name              locus length coverage gene_id
1 XLOC_000001       <NA>           <NA>          LOC425783    chr1:6267-21806     NA       NA    <NA>
2 XLOC_000002       <NA>           <NA>             GOLGB1   chr1:33922-67653     NA       NA    <NA>
3 XLOC_000003       <NA>           <NA>       Q5ZMV0_CHICK   chr1:69103-83497     NA       NA    <NA>
4 XLOC_000004       <NA>           <NA>             RABL2B  chr1:99078-102964     NA       NA    <NA>
5 XLOC_000005       <NA>           <NA> ENSGALG00000023985 chr1:108633-114770     NA       NA    <NA>
6 XLOC_000006       <NA>           <NA>             SHANK3 chr1:171646-199289     NA       NA    <NA>
> gene.fpkm <- fpkm(genes(cuff_data))
> head(gene.fpkm)
      gene_id sample_name     fpkm  conf_hi  conf_lo quant_status
1 XLOC_000001          FL  37.6272  80.6297   0.0000           OK
2 XLOC_000001          LL  96.7472 161.2450  32.2491           OK
3 XLOC_000002          FL 135.8230 167.5380 104.1070           OK
4 XLOC_000002          LL 155.1970 190.3030 120.0920           OK
5 XLOC_000003          FL 132.7550 190.9640  74.5450           OK
6 XLOC_000003          LL 140.7280 200.0020  81.4548           OK
> isoform.fpkm <- fpkm(isoforms(cuff_data))
> head(isoform.fpkm)
      isoform_id sample_name        fpkm   conf_hi  conf_lo quant_status
1 TCONS_00000001          FL 37.62720000  80.62970  0.00000           OK
2 TCONS_00000001          LL 96.74720000 161.24500 32.24910           OK
3 TCONS_00000002          FL  0.00355677   3.43438  0.00000           OK
4 TCONS_00000002          LL  0.00000000   0.00000  0.00000           OK
5 TCONS_00000003          FL 37.31330000  59.59370 15.03290           OK
6 TCONS_00000003          LL 22.01600000  40.60530  3.42675           OK
> gene.diff <- diffData(genes(cuff_data))
> head(gene.diff)
     gene_id sample_1 sample_2 status  value_1  value_2 log2_fold_change     test_stat   p_value  q_value significant
1 XLOC_000001       FL       LL NOTEST  37.6272  96.7472      1.36245e+00 -1.42753e+00 0.1534270 1.000000          no
2 XLOC_000002       FL       LL     OK 135.8230 155.1970      1.92379e-01 -8.20334e-01 0.4120260 0.829944          no
3 XLOC_000003       FL       LL NOTEST 132.7550 140.7280      8.41509e-02 -1.91872e-01 0.8478430 1.000000          no
4 XLOC_000004       FL       LL NOTEST   0.0000  30.2349     1.79769e+308  1.79769e+308 0.0786496 1.000000          no
5 XLOC_000005       FL       LL NOTEST   0.0000  22.0152     1.79769e+308  1.79769e+308 0.2970490 1.000000          no
6 XLOC_000006       FL       LL NOTEST  71.7574  44.4741     -6.90162e-01   5.59906e-01 0.5755430 1.000000          no
```

Inspect the Differentially Expressed Genes

```
> cuff_data
CuffSet instance with:
        2 samples
        2445 genes
        4914 isoforms
        3047 TSS
        1578 CDS
        2445 promoters
        3047 splicing
        1351 relCDS
> gene_diff_data <- diffData(genes(cuff_data))
> sig_gene_data <- subset(gene_diff_data, (significant == 'yes'))
> nrow(sig_gene_data)
[1] 15
> sig_gene_data
         gene_id sample_1 sample_2 status  value_1  value_2 log2_fold_change test_stat     p_value     q_value significant
151  XLOC_000151       FL       LL     OK  315.986  190.767        -0.728054   3.96393 7.37261e-05 7.53235e-03         yes
206  XLOC_000206       FL       LL     OK  101.960  291.114         1.513580  -4.51017 6.47750e-06 7.94142e-04         yes
509  XLOC_000509       FL       LL     OK 3619.320 4752.600         0.392999  -3.55611 3.76389e-04 1.92272e-02         yes
581  XLOC_000581       FL       LL     OK  291.897  712.350         1.287130  -4.82627 1.39117e-06 2.84262e-04         yes
640  XLOC_000640       FL       LL     OK 1500.080 2234.310         0.574785  -4.70454 2.54439e-06 3.89928e-04         yes
711  XLOC_000711       FL       LL     OK  209.051  385.690         0.883583  -3.84734 1.19406e-04 9.14950e-03         yes
818  XLOC_000818       FL       LL     OK  259.468  592.153         1.190420  -3.65196 2.60251e-04 1.59534e-02         yes
875  XLOC_000875       FL       LL     OK  378.188  786.100         1.055610  -8.88649 0.00000e+00 0.00000e+00         yes
1059 XLOC_001059       FL       LL     OK 3713.950 5652.920         0.606044  -3.85972 1.13519e-04 9.14950e-03         yes
1315 XLOC_001315       FL       LL     OK  487.095  772.977         0.666222  -3.56953 3.57626e-04 1.92272e-02         yes
1494 XLOC_001494       FL       LL     OK  214.394  428.568         0.999259  -3.24377 1.17959e-03 4.82059e-02         yes
1708 XLOC_001708       FL       LL     OK 4509.900 6781.980         0.588611  -3.36642 7.61506e-04 3.59079e-02         yes
1826 XLOC_001826       FL       LL     OK 1121.490  496.331        -1.176040   6.86799 6.51124e-12 1.99569e-09         yes
1875 XLOC_001875       FL       LL     OK  369.099  197.538        -0.901878   3.33496 8.53125e-04 3.73547e-02         yes
2164 XLOC_002164       FL       LL     OK  366.506  156.929        -1.223730   3.80781 1.40204e-04 9.54943e-03         yes
>
```

Inspect the Differentially Expressed Transcripts

```
> isoform_diff_data <- diffData(isoforms(cuff_data), 'FL', 'LL')

> sig_isoform_data <- subset(isoform_diff_data, (significant == 'yes'))

> nrow(sig_isoform_data)
[1] 72

> head(sig_isoform_data, 20)
```

	isoform_id	isoform_id.1	sample_1	sample_2	status	value_1	value_2	log2_fold_change	test_stat	p_value	q_value	significant
36	TCONS_00000036	TCONS_00000036	FL	LL	OK	184.9200	0.0000	-1.79769e+308	-1.79769e+308	1.13350e-05	3.50250e-04	yes
58	TCONS_00000058	TCONS_00000058	FL	LL	OK	156.3640	0.0000	-1.79769e+308	-1.79769e+308	2.00928e-03	2.34290e-02	yes
186	TCONS_00000186	TCONS_00000186	FL	LL	OK	55.8696	239.2600	2.09844e+00	-4.11086e+00	3.94184e-05	9.36945e-04	yes
328	TCONS_00000328	TCONS_00000328	FL	LL	OK	190.1450	46.9916	-2.01663e+00	4.14806e+00	3.35304e-05	8.28871e-04	yes
458	TCONS_00000458	TCONS_00000458	FL	LL	OK	90.6414	191.2480	1.07720e+00	-2.98451e+00	2.84032e-03	3.02641e-02	yes
603	TCONS_00000603	TCONS_00000603	FL	LL	OK	1216.0800	2034.8300	7.42673e-01	-2.76010e+00	5.77836e-03	4.95976e-02	yes
669	TCONS_00000669	TCONS_00000669	FL	LL	OK	0.0000	195.1090	1.79769e+308	1.79769e+308	4.89356e-07	3.02422e-05	yes
670	TCONS_00000670	TCONS_00000670	FL	LL	OK	268.3530	76.1287	-1.81762e+00	3.05952e+00	2.21695e-03	2.49104e-02	yes
717	TCONS_00000717	TCONS_00000717	FL	LL	OK	258.5230	124.5110	-1.05402e+00	2.92780e+00	3.41367e-03	3.56297e-02	yes
760	TCONS_00000760	TCONS_00000760	FL	LL	OK	0.0000	251.2270	1.79769e+308	1.79769e+308	2.51525e-08	1.94303e-06	yes
786	TCONS_00000786	TCONS_00000786	FL	LL	OK	355.3370	755.0020	1.08729e+00	-3.57399e+00	3.51589e-04	4.93823e-03	yes
799	TCONS_00000799	TCONS_00000799	FL	LL	OK	0.0000	211.7210	1.79769e+308	1.79769e+308	1.98595e-06	8.18210e-05	yes
800	TCONS_00000800	TCONS_00000800	FL	LL	OK	165.6600	0.0000	-1.79769e+308	-1.79769e+308	7.13557e-05	1.52633e-03	yes
855	TCONS_00000855	TCONS_00000855	FL	LL	OK	374.3220	0.0000	-1.79769e+308	-1.79769e+308	1.48892e-10	2.30038e-08	yes
936	TCONS_00000936	TCONS_00000936	FL	LL	OK	28.8865	196.6750	2.76735e+00	-4.15966e+00	3.18728e-05	8.20726e-04	yes
976	TCONS_00000976	TCONS_00000976	FL	LL	OK	173.9980	0.0000	-1.79769e+308	-1.79769e+308	2.50821e-05	7.04578e-04	yes
984	TCONS_00000984	TCONS_00000984	FL	LL	OK	524.1740	259.3030	-1.01540e+00	2.79985e+00	5.11268e-03	4.71587e-02	yes
986	TCONS_00000986	TCONS_00000986	FL	LL	OK	98.2900	292.0360	1.57103e+00	-2.78568e+00	5.34153e-03	4.78415e-02	yes
1122	TCONS_00001122	TCONS_00001122	FL	LL	OK	161.4060	0.0000	-1.79769e+308	-1.79769e+308	2.66003e-04	3.91404e-03	yes
1201	TCONS_00001201	TCONS_00001201	FL	LL	OK	0.0000	178.4750	1.79769e+308	1.79769e+308	5.23773e-03	4.76018e-02	yes>

Inspect the Differentially Expressed TSS Groups (optional)

```
> tss_diff_data <- diffData(TSS(cuff_data), 'FL', 'LL')

> sig_tss_data <- subset(tss_diff_data, (significant == 'yes'))

> nrow(sig_tss_data)
[1] 26

> sig_tss_data
```

	TSS_group_id	TSS_group_id.1	sample_1	sample_2	status	value_1	value_2	log2_fold_change	test_stat	p_value	q_value	significant
34	TSS1028	TSS1028	FL	LL	OK	149.2180	0.0000	-1.79769e+308	-1.79769e+308	7.16239e-05	4.24930e-03	yes
118	TSS1103	TSS1103	FL	LL	OK	378.1880	786.1000	1.05561e+00	-8.88649e+00	0.00000e+00	0.00000e+00	yes
157	TSS1139	TSS1139	FL	LL	OK	518.8100	1450.4400	1.48321e+00	-3.40323e+00	6.65948e-04	2.22216e-02	yes
368	TSS1329	TSS1329	FL	LL	OK	1692.8200	3086.8200	8.66697e-01	-5.53756e+00	3.06716e-08	4.86145e-06	yes
653	TSS1586	TSS1586	FL	LL	OK	2214.9800	1726.3500	-3.59568e-01	3.10359e+00	1.91186e-03	4.66201e-02	yes
685	TSS1614	TSS1614	FL	LL	OK	320.7450	121.4850	-1.40065e+00	3.82435e+00	1.31117e-04	6.34922e-03	yes
697	TSS1625	TSS1625	FL	LL	OK	232.5470	64.4287	-1.85175e+00	3.28215e+00	1.03019e-03	2.90433e-02	yes
882	TSS1792	TSS1792	FL	LL	OK	4292.0800	5177.5000	2.70579e-01	-3.17760e+00	1.48498e-03	3.92282e-02	yes
956	TSS1859	TSS1859	FL	LL	OK	214.0030	415.5230	9.57295e-01	-3.60364e+00	3.13790e-04	1.17025e-02	yes
1035	TSS193	TSS193	FL	LL	OK	315.9860	190.7670	-7.28054e-01	3.96393e+00	7.37261e-05	4.24930e-03	yes
1172	TSS2052	TSS2052	FL	LL	OK	0.0000	154.9110	1.79769e+308	1.79769e+308	2.13577e-04	8.46300e-03	yes
1242	TSS2115	TSS2115	FL	LL	OK	2850.3400	4538.4600	6.71070e-01	-4.09094e+00	4.29626e-05	3.02648e-03	yes
1243	TSS2116	TSS2116	FL	LL	OK	429.4060	104.6850	-2.03629e+00	3.34006e+00	8.37602e-04	2.57562e-02	yes
1331	TSS2196	TSS2196	FL	LL	OK	327.8470	1349.8600	2.04172e+00	-5.68208e+00	1.33070e-08	2.81220e-06	yes
1381	TSS2240	TSS2240	FL	LL	OK	111.7450	294.3440	1.39729e+00	-3.43213e+00	5.98853e-04	2.10929e-02	yes
1417	TSS2273	TSS2273	FL	LL	OK	1121.4900	496.3310	-1.17604e+00	6.86799e+00	6.51124e-12	2.06406e-09	yes
1489	TSS2338	TSS2338	FL	LL	OK	369.0990	197.5380	-9.01878e-01	3.33496e+00	8.53125e-04	2.57562e-02	yes
1518	TSS2364	TSS2364	FL	LL	OK	349.9230	177.6840	-9.77723e-01	3.15932e+00	1.58137e-03	4.01035e-02	yes
1780	TSS260	TSS260	FL	LL	OK	90.6414	220.4320	1.28209e+00	-3.79027e+00	1.50481e-04	6.36032e-03	yes
1894	TSS2702	TSS2702	FL	LL	OK	366.5060	156.9290	-1.22373e+00	3.80781e+00	1.40204e-04	6.34922e-03	yes
2413	TSS427	TSS427	FL	LL	OK	295.8820	139.4110	-1.08568e+00	3.27580e+00	1.05362e-03	2.90433e-02	yes
2503	TSS508	TSS508	FL	LL	OK	817.5910	300.0690	-1.44609e+00	5.33737e+00	9.43041e-08	1.19578e-05	yes
2576	TSS574	TSS574	FL	LL	OK	173.9980	0.0000	-1.79769e+308	-1.79769e+308	2.50821e-05	2.18723e-03	yes
2590	TSS587	TSS587	FL	LL	OK	1364.3900	313.3070	-2.12261e+00	3.87503e+00	1.06612e-04	5.63264e-03	yes
2747	TSS728	TSS728	FL	LL	OK	291.8970	659.7650	1.17649e+00	-4.19242e+00	2.75992e-05	2.18723e-03	yes
2830	TSS802	TSS802	FL	LL	OK	1500.0800	2234.3100	5.74785e-01	-4.70454e+00	2.54439e-06	2.68857e-04	yes

```
>
```

Inspect the Differentially Expressed Coding Sequences (optional)

```
> cds_diff_data <- diffData(CDS(cuff_data), 'FL', 'LL')

> sig_cds_data <- subset(cds_diff_data, (significant=='yes'))

> nrow(sig_cds_data)
[1] 0

> options(width=300)
> sig_cds_data
 [1] CDS_id          CDS_id.1        sample_1        sample_2        status          value_1         value_2
log2_fold_change test_stat       p_value         q_value         significant
<0 rows> (or 0-length row.names)
>
```

Inspect the Differentially Spliced TSS Groups

```
> splicing_diff_data <- distValues(splicing(cuff_data))
> sig_splicing_data <- subset(splicing_diff_data, (significant =='yes'))
> nrow(sig_splicing_data)
[1] 47
> sig_splicing_data
```

	TSS_group_id	gene_id	sample_1	sample_2	status	value_1	value_2	JS_dist	test_stat	p_value	q_value	significant
75	TSS1065	XLOC_000843	FL	LL	OK	0	0	0.7943550	0.00000e+00	0.000010	4.84211e-05	yes
193	TSS1171	XLOC_000928	FL	LL	OK	0	0	0.8282090	0.00000e+00	0.000010	4.84211e-05	yes
376	TSS1336	XLOC_001063	FL	LL	OK	0	0	0.6959130	4.81453e-11	0.000010	4.84211e-05	yes
445	TSS1399	XLOC_001115	FL	LL	OK	0	0	0.3841220	4.81171e-13	0.000965	2.86387e-03	yes
473	TSS1423	XLOC_001136	FL	LL	OK	0	0	0.3270890	3.51497e-13	0.003080	8.33412e-03	yes
519	TSS1465	XLOC_001173	FL	LL	OK	0	0	0.2438710	3.10355e-09	0.000105	3.71538e-04	yes
653	TSS1586	XLOC_001277	FL	LL	OK	0	0	0.2920950	3.68499e-09	0.000010	4.84211e-05	yes
678	TSS1608	XLOC_001294	FL	LL	OK	0	0	0.2829270	0.00000e+00	0.000010	4.84211e-05	yes
681	TSS1610	XLOC_001294	FL	LL	OK	0	0	0.3673720	5.55112e-15	0.000010	4.84211e-05	yes
779	TSS17	XLOC_000014	FL	LL	OK	0	0	0.5690520	4.61742e-12	0.000045	1.72500e-04	yes
857	TSS177	XLOC_000138	FL	LL	OK	0	0	0.2992650	6.29022e-03	0.011925	2.49341e-02	yes
882	TSS1792	XLOC_001437	FL	LL	OK	0	0	0.2138870	7.30370e-07	0.000030	1.31429e-04	yes
961	TSS1863	XLOC_001497	FL	LL	OK	0	0	0.3512650	6.39427e-05	0.015730	3.14600e-02	yes
1004	TSS1901	XLOC_001528	FL	LL	OK	0	0	0.3539090	3.45009e-03	0.006755	1.63542e-02	yes
1010	TSS1907	XLOC_001534	FL	LL	OK	0	0	0.4555710	7.74047e-13	0.000010	4.84211e-05	yes
1235	TSS2109	XLOC_001704	FL	LL	OK	0	0	0.3807510	6.42791e-09	0.000045	1.72500e-04	yes
1242	TSS2115	XLOC_001708	FL	LL	OK	0	0	0.1249580	9.50066e-03	0.014960	3.05849e-02	yes
1248	TSS2120	XLOC_001711	FL	LL	OK	0	0	0.2654530	2.86046e-03	0.007515	1.77277e-02	yes
1331	TSS2196	XLOC_001767	FL	LL	OK	0	0	0.6164660	2.22045e-16	0.000020	9.20000e-05	yes
1440	TSS2294	XLOC_001842	FL	LL	OK	0	0	0.3589460	5.02387e-04	0.000565	1.73267e-03	yes
1483	TSS2332	XLOC_001873	FL	LL	OK	0	0	0.2136550	4.51033e-03	0.009315	2.09020e-02	yes
1484	TSS2333	XLOC_001873	FL	LL	OK	0	0	0.2136550	4.51033e-03	0.009315	2.09020e-02	yes
1490	TSS2339	XLOC_001876	FL	LL	OK	0	0	0.3673620	9.50080e-06	0.000090	3.31200e-04	yes
1571	TSS2411	XLOC_001937	FL	LL	OK	0	0	0.4031330	0.00000e+00	0.000010	4.84211e-05	yes
1609	TSS2446	XLOC_001962	FL	LL	OK	0	0	0.2565560	9.06389e-09	0.002140	5.96606e-03	yes
1669	TSS250	XLOC_000199	FL	LL	OK	0	0	0.3456680	2.15999e-03	0.005810	1.44465e-02	yes
1819	TSS2635	XLOC_002109	FL	LL	OK	0	0	0.3889090	1.99418e-12	0.000010	4.84211e-05	yes
1900	TSS2708	XLOC_002170	FL	LL	OK	0	0	0.0919622	3.56889e-05	0.011225	2.40270e-02	yes
1916	TSS2722	XLOC_002181	FL	LL	OK	0	0	0.4894800	2.00524e-05	0.005455	1.39406e-02	yes
1920	TSS2726	XLOC_002185	FL	LL	OK	0	0	0.8325550	0.00000e+00	0.000010	4.84211e-05	yes
1950	TSS2753	XLOC_002204	FL	LL	OK	0	0	0.1949760	1.13762e-08	0.000045	1.72500e-04	yes
1952	TSS2755	XLOC_002206	FL	LL	OK	0	0	0.8323410	8.11910e-09	0.000010	4.84211e-05	yes
2039	TSS2833	XLOC_002266	FL	LL	OK	0	0	0.5716820	0.00000e+00	0.000010	4.84211e-05	yes
2294	TSS32	XLOC_000026	FL	LL	OK	0	0	0.2341990	1.60051e-07	0.001715	4.93062e-03	yes
2376	TSS394	XLOC_000320	FL	LL	OK	0	0	0.5970190	0.00000e+00	0.000010	4.84211e-05	yes
2459	TSS469	XLOC_000382	FL	LL	OK	0	0	0.2153690	1.70573e-04	0.000485	1.59357e-03	yes
2468	TSS477	XLOC_000389	FL	LL	OK	0	0	0.6365770	0.00000e+00	0.000010	4.84211e-05	yes
2503	TSS508	XLOC_000413	FL	LL	OK	0	0	0.4785260	0.00000e+00	0.000010	4.84211e-05	yes
2552	TSS552	XLOC_000451	FL	LL	OK	0	0	0.8316180	0.00000e+00	0.000010	4.84211e-05	yes
2580	TSS578	XLOC_000465	FL	LL	OK	0	0	0.3955950	1.76709e-08	0.000125	4.25926e-04	yes
2590	TSS587	XLOC_000470	FL	LL	OK	0	0	0.5639740	9.62258e-03	0.000010	4.84211e-05	yes
2635	TSS627	XLOC_000503	FL	LL	OK	0	0	0.1397040	1.31973e-02	0.021495	4.20753e-02	yes
2643	TSS634	XLOC_000507	FL	LL	OK	0	0	0.2848940	7.53992e-05	0.011230	2.40270e-02	yes
2647	TSS638	XLOC_000509	FL	LL	OK	0	0	0.2987790	0.00000e+00	0.000010	4.84211e-05	yes
2739	TSS720	XLOC_000574	FL	LL	OK	0	0	0.2952790	2.13325e-06	0.004390	1.15394e-02	yes
2830	TSS802	XLOC_000640	FL	LL	OK	0	0	0.2039240	7.56921e-05	0.000545	1.72897e-03	yes
2966	TSS925	XLOC_000741	FL	LL	OK	0	0	0.8317730	0.00000e+00	0.000010	4.84211e-05	yes

Inspect the Genes with Differential Promoter Usage (optional)

```
> promoter_diff_data <- distValues(promoters(cuff_data))

> sig_promoter_data <- subset(promoter_diff_data, (significant == 'yes'))

> nrow(sig_promoter_data)
[1] 22

> sig_promoter_data
          gene_id sample_1 sample_2 status value_1 value_2  JS_dist   test_stat  p_value     q_value significant
10    XLOC_000010       FL       LL     OK       0       0 0.350489 2.05866e-06 0.000100 0.000750000         yes
26    XLOC_000026       FL       LL     OK       0       0 0.163359 1.88976e-03 0.003890 0.017426500         yes
76    XLOC_000076       FL       LL     OK       0       0 0.341414 2.98360e-03 0.014295 0.048733000         yes
349   XLOC_000349       FL       LL     OK       0       0 0.101063 6.15428e-03 0.008725 0.034440800         yes
370   XLOC_000370       FL       LL     OK       0       0 0.314046 6.62848e-04 0.009325 0.034968700         yes
413   XLOC_000413       FL       LL     OK       0       0 0.495559 0.00000e+00 0.000010 0.000125000         yes
462   XLOC_000462       FL       LL     OK       0       0 0.528164 7.99361e-15 0.000015 0.000140625         yes
465   XLOC_000465       FL       LL     OK       0       0 0.181609 1.91158e-04 0.003285 0.017075000         yes
470   XLOC_000470       FL       LL     OK       0       0 0.565331 1.89777e-11 0.000010 0.000125000         yes
620   XLOC_000620       FL       LL     OK       0       0 0.347070 6.35768e-04 0.005325 0.022187500         yes
813   XLOC_000813       FL       LL     OK       0       0 0.426070 1.55431e-15 0.000065 0.000541667         yes
818   XLOC_000818       FL       LL     OK       0       0 0.595225 1.28807e-05 0.000010 0.000125000         yes
903   XLOC_000903       FL       LL     OK       0       0 0.494402 2.22045e-16 0.000015 0.000140625         yes
1055  XLOC_001055       FL       LL     OK       0       0 0.438843 1.09297e-03 0.002580 0.014884600         yes
1304  XLOC_001304       FL       LL     OK       0       0 0.431044 3.05717e-07 0.000010 0.000125000         yes
1335  XLOC_001335       FL       LL     OK       0       0 0.275093 9.07483e-03 0.013430 0.047964300         yes
1450  XLOC_001450       FL       LL     OK       0       0 0.301625 5.66491e-05 0.003950 0.017426500         yes
1656  XLOC_001656       FL       LL     OK       0       0 0.169095 1.31302e-10 0.000535 0.003647730         yes
1658  XLOC_001658       FL       LL     OK       0       0 0.829814 0.00000e+00 0.000010 0.000125000         yes
1767  XLOC_001767       FL       LL     OK       0       0 0.503301 0.00000e+00 0.000010 0.000125000         yes
2177  XLOC_002177       FL       LL     OK       0       0 0.304342 1.76425e-05 0.001160 0.007250000         yes
2265  XLOC_002265       FL       LL     OK       0       0 0.512821 1.04072e-05 0.003415 0.017075000         yes
>
```

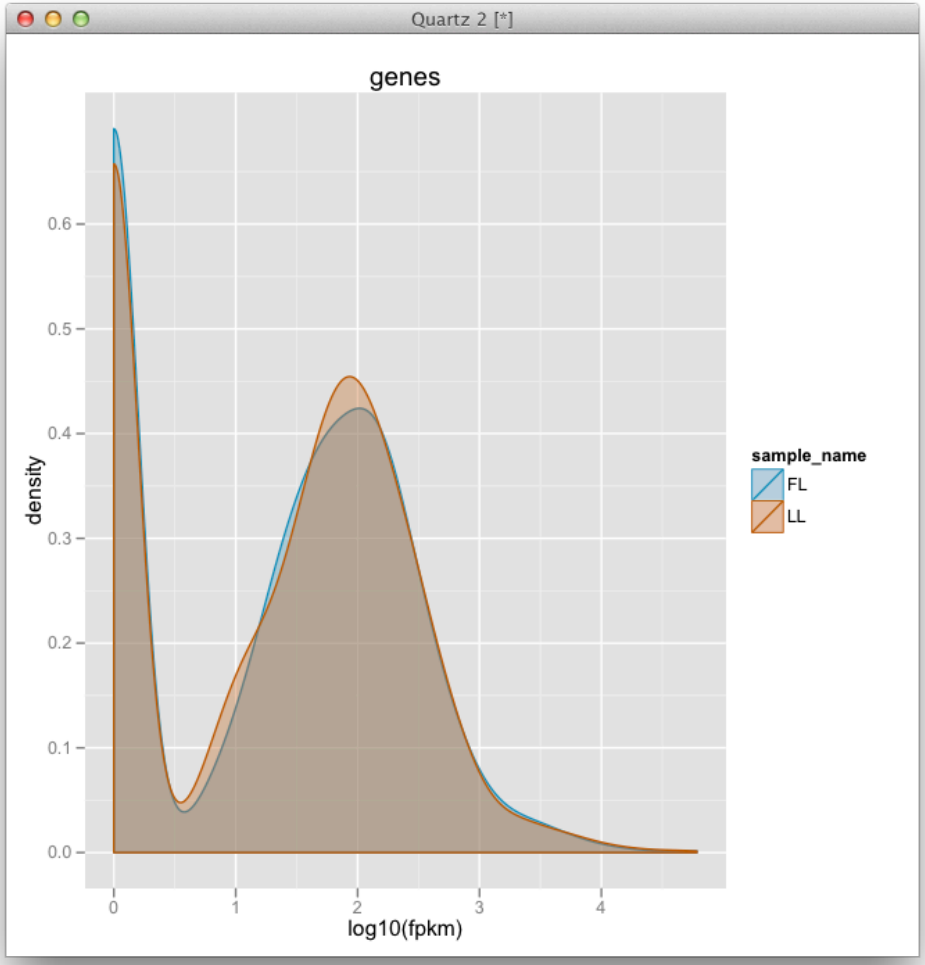
Inspect the Genes with Differential CDS Output (optional)

```
> relCDS_diff_data <- distValues(relCDS(cuff_data))

> sig_relCDS_data <- subset(relCDS_diff_data, (significant == 'yes'))

> nrow(sig_relCDS_data)
[1] 2

> sig_relCDS_data
  gene_id sample_1 sample_2 status value_1 value_2  JS_dist   test_stat p_value q_value significant
1036 XLOC_001767       FL       LL     OK       0       0 0.503301 0.00000e+00   1e-05   4e-05         yes
1269 XLOC_002206       FL       LL     OK       0       0 0.728721 1.33311e-06   1e-05   4e-05         yes
>
```

Exercise 7

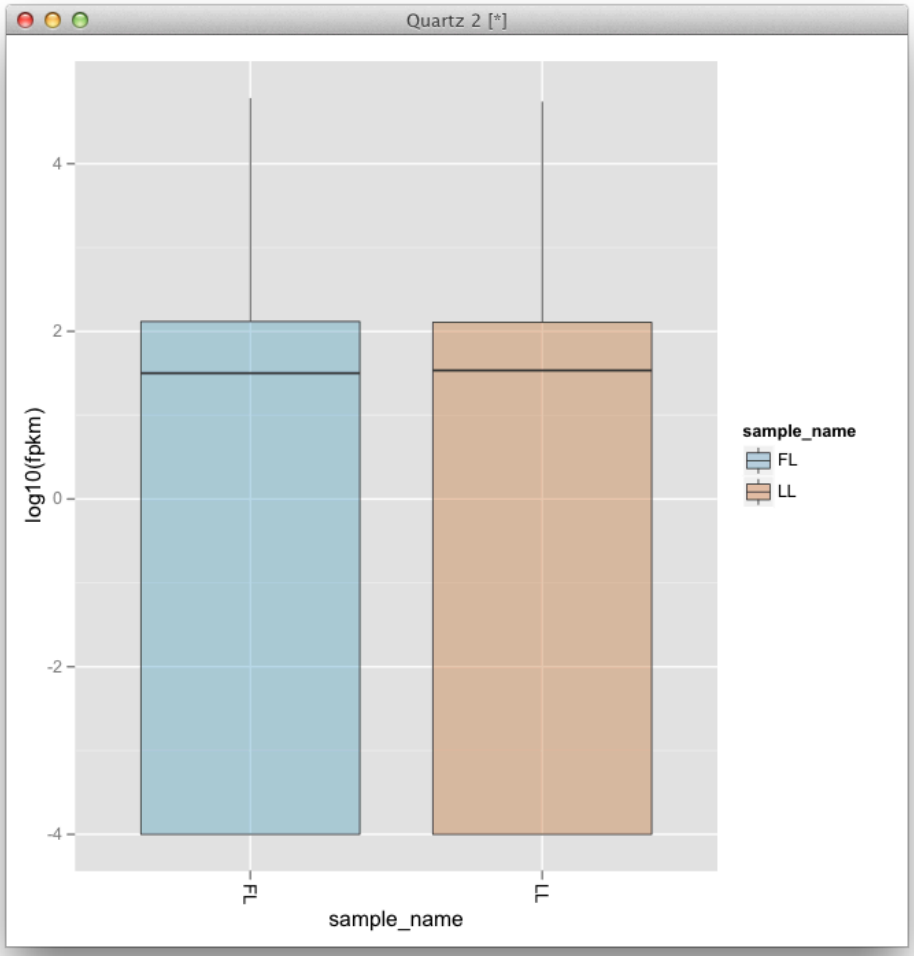
Visualizing the Differential Analysis Results
(CummeRbund)

Plot the Distribution of Expression Levels



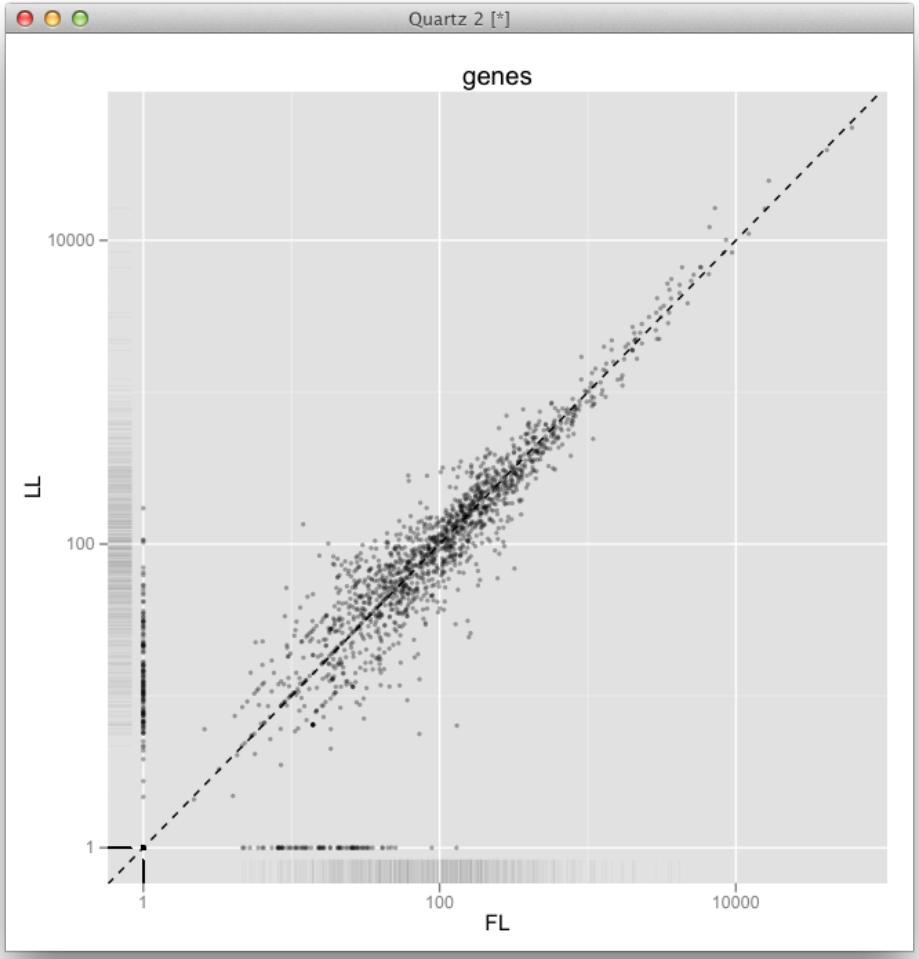

```
> library(cairoDevice)
> Cairo()
> csDensity(genes(cuff_data))
```

Boxplot View of Expression Levels



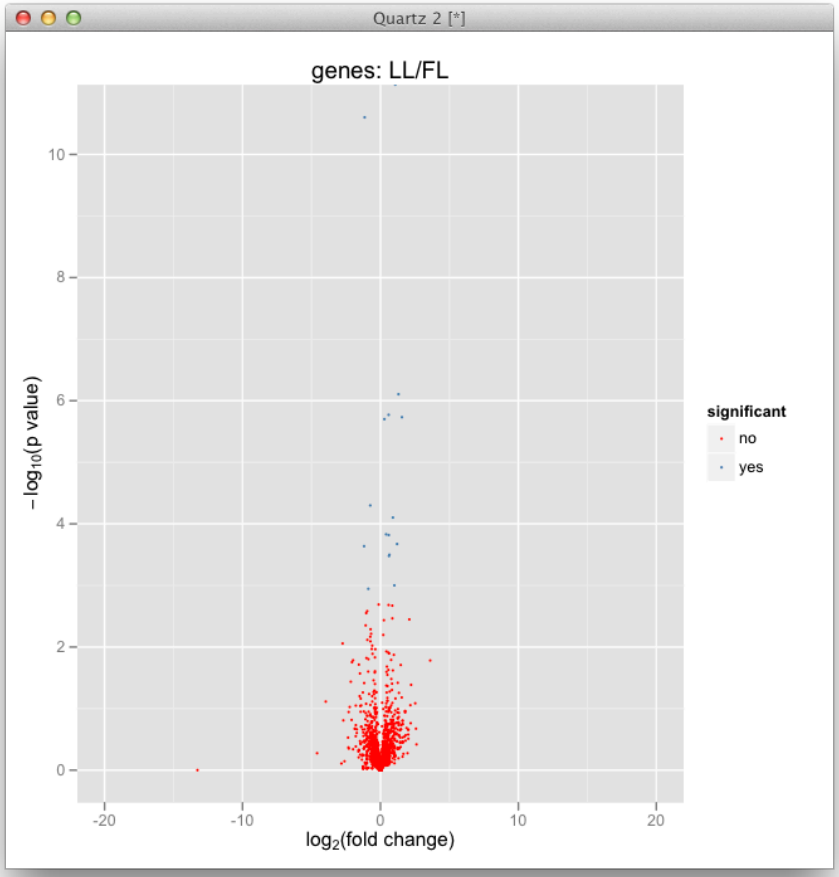

```
> csBoxplot(genes(cuff_data))
```

Compare the Expression Levels of Genes




```
> csScatter(genes(cuff_data), 'FL', 'LL')
```

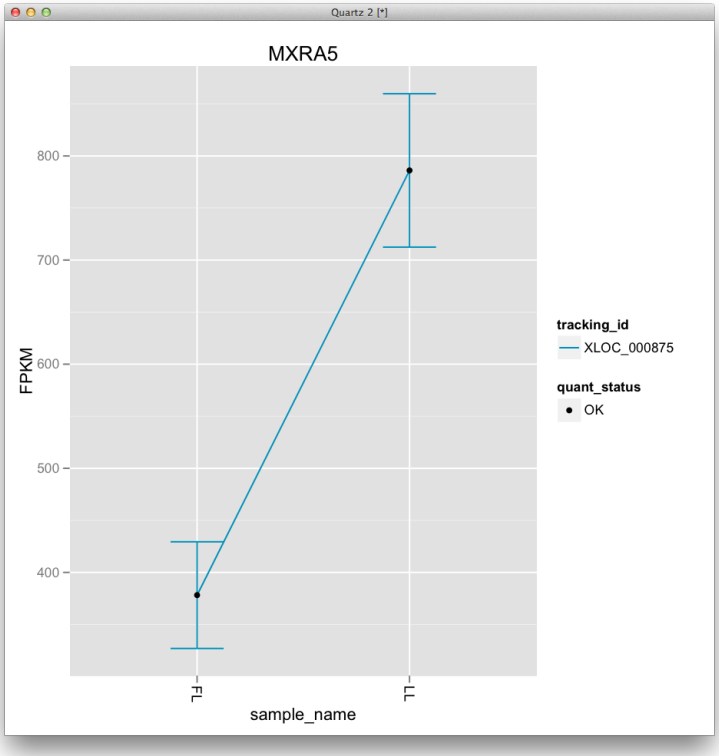
Create a Volcano Plot to Inspect DE Genes

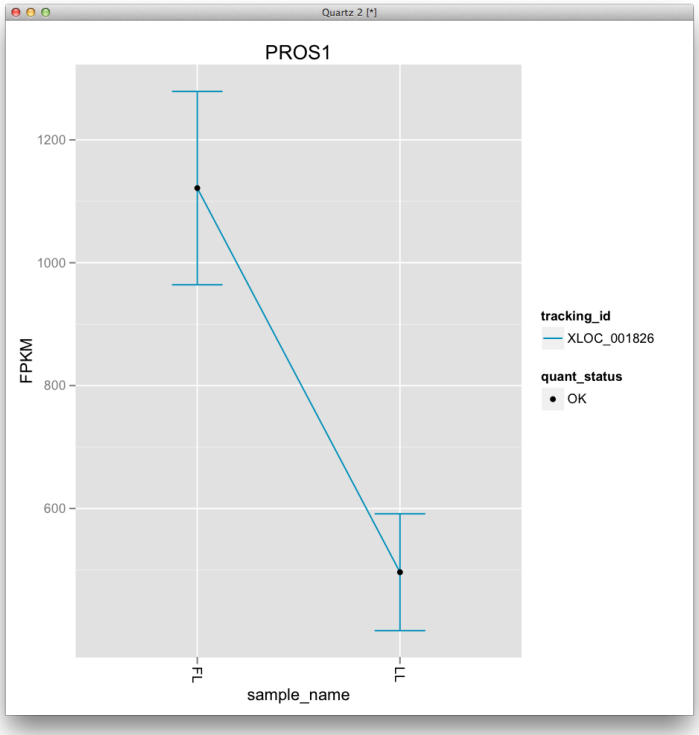



```
> csVolcano(genes(cuff_data), 'FL', 'LL')
> pdf("volcano.pdf")
> csVolcano(genes(cuff_data), 'FL', 'LL')
> dev.off()
```

Plot the Expression Levels for Genes of Interest

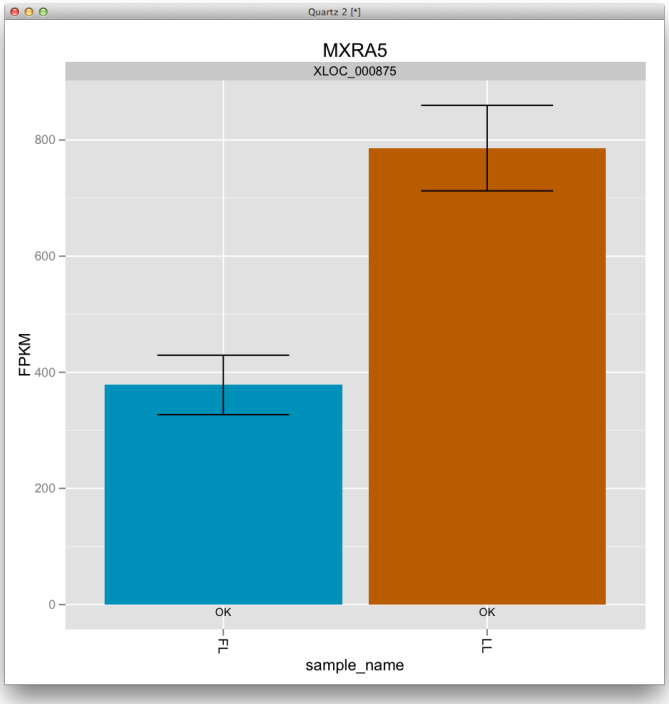
```
> mygene <- getGene(cuff_data, 'MXRA5')
> expressionPlot(mygene)
> mygene <- getGene(cuff_data, 'PROS1')
> expressionPlot(mygene)
```

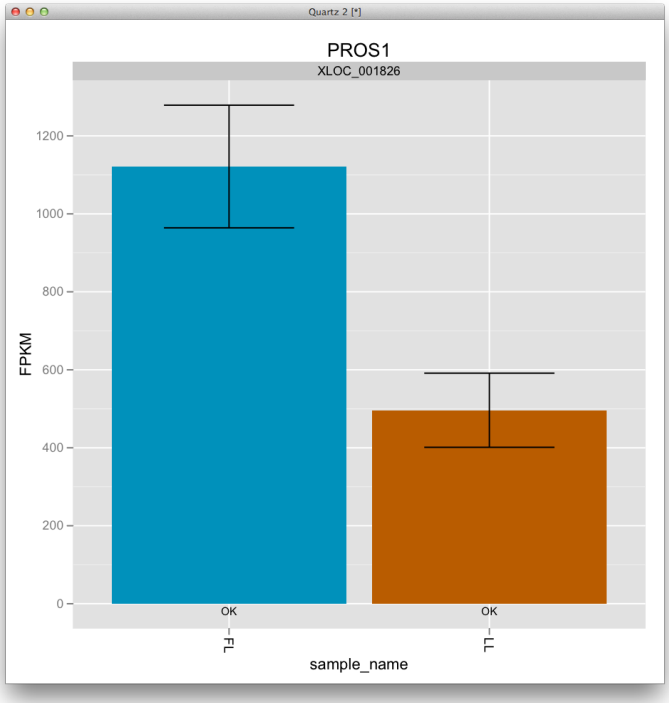




Barplot of the Expression Levels for Genes of Interest

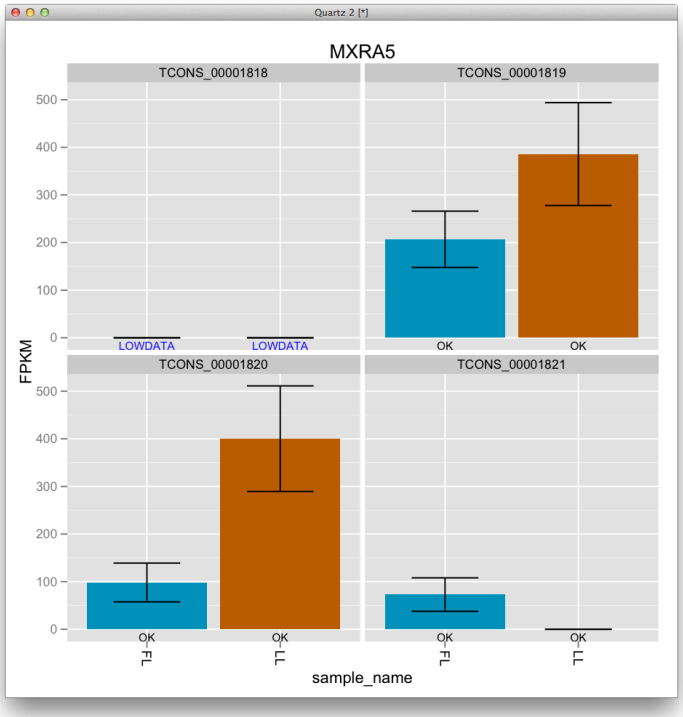
```
> mygene <- getGene(cuff_data, 'MXRA5')
> expressionBarplot(mygene)
> mygene <- getGene(cuff_data, 'PROS1')
> expressionBarplot(mygene)
```

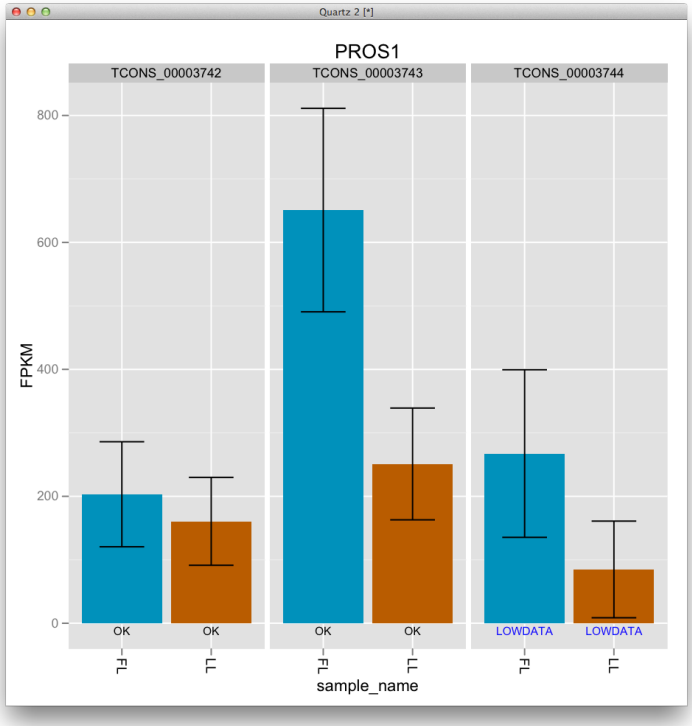




Barplot of Isoform Expression Levels for Genes of Interest (optional)

```
> mygene <- getGene(cuff_data, 'MXRA5')
> expressionBarplot(isoforms(mygene))
> mygene <- getGene(cuff_data, 'PROS1')
> expressionBarplot(isoforms(mygene))
```

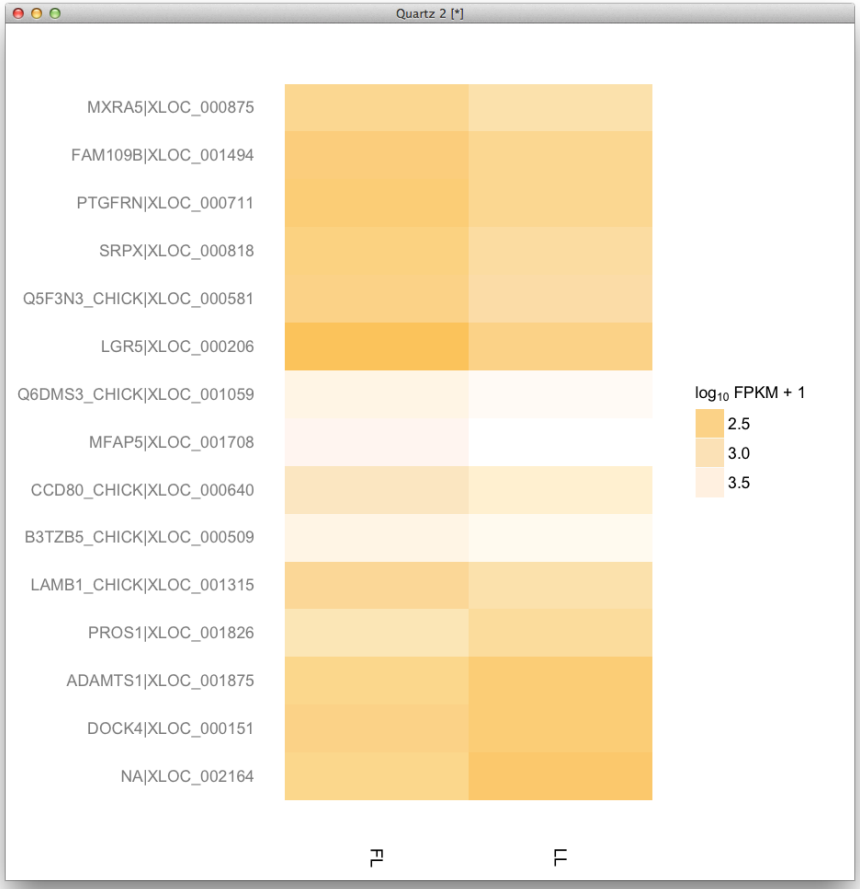




Create Gene Set from Significantly Regulated Genes

```
> mySigGeneIds <-getSig(cuff_data, alpha=0.05, level ="genes”)
> head(mySigGeneIds)
[1] "XLOC_000151" "XLOC_000206" "XLOC_000509" "XLOC_000581" "XLOC_000640" "XLOC_000711”
> length(mySigGeneIds)
[1] 15
> mySigGenes <- getGenes(cuff_data, mySigGeneIds)
Getting gene information:
        FPKM
        Differential Expression Data
        Annotation Data
Getting isoforms information:
        FPKM
        Differential Expression Data
        Annotation Data
Getting CDS information:
        FPKM
        Differential Expression Data
        Annotation Data
Getting TSS information:
        FPKM
        Differential Expression Data
        Annotation Data
Getting promoter information:
        distData
Getting splicing information:
        distData
Getting relCDS information:
        distData
> mySigGenes
CuffGeneSet instance for  15  genes

Slots:
         annotation
         fpkm
         diff
         isoforms         CuffFeatureSet instance of size 68
         TSS              CuffFeatureSet instance of size 27
         CDS              CuffFeatureSet instance of size 0
         promoters                CuffFeatureSet instance of size 15
         splicing                 CuffFeatureSet instance of size 27
         relCDS           CuffFeatureSet instance of size 15
```

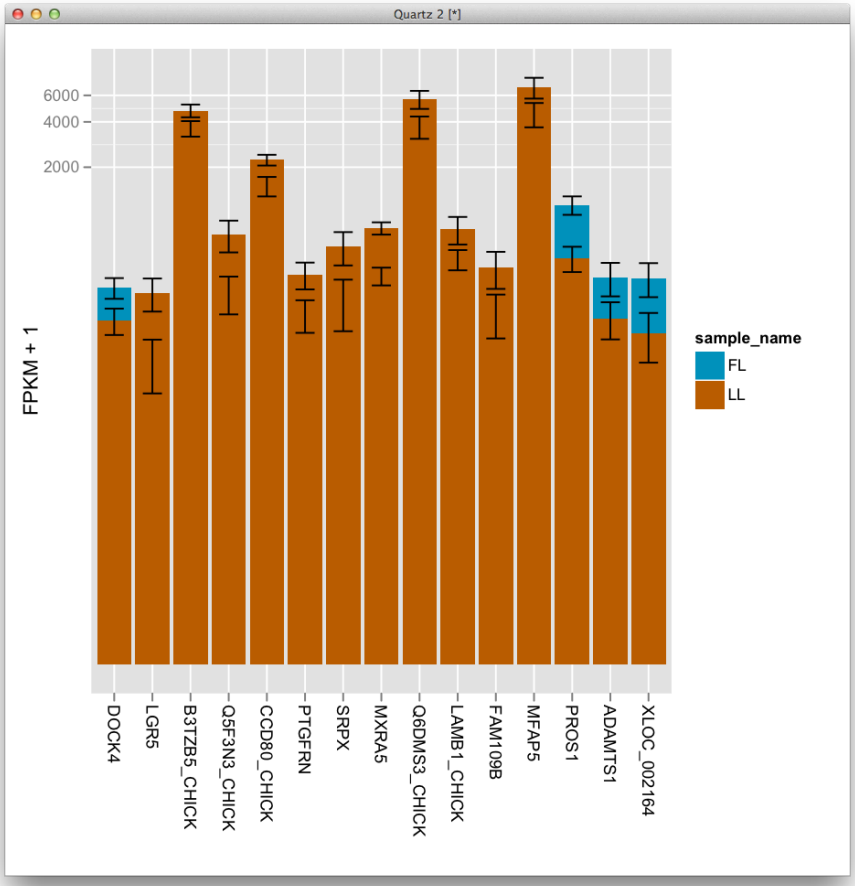
Heatmap of Significantly Regulated Genes




```
> csHeatmap(mySigGenes, cluster="both")
Using tracking_id, sample_name as id variables
Using  as id variables
>
```

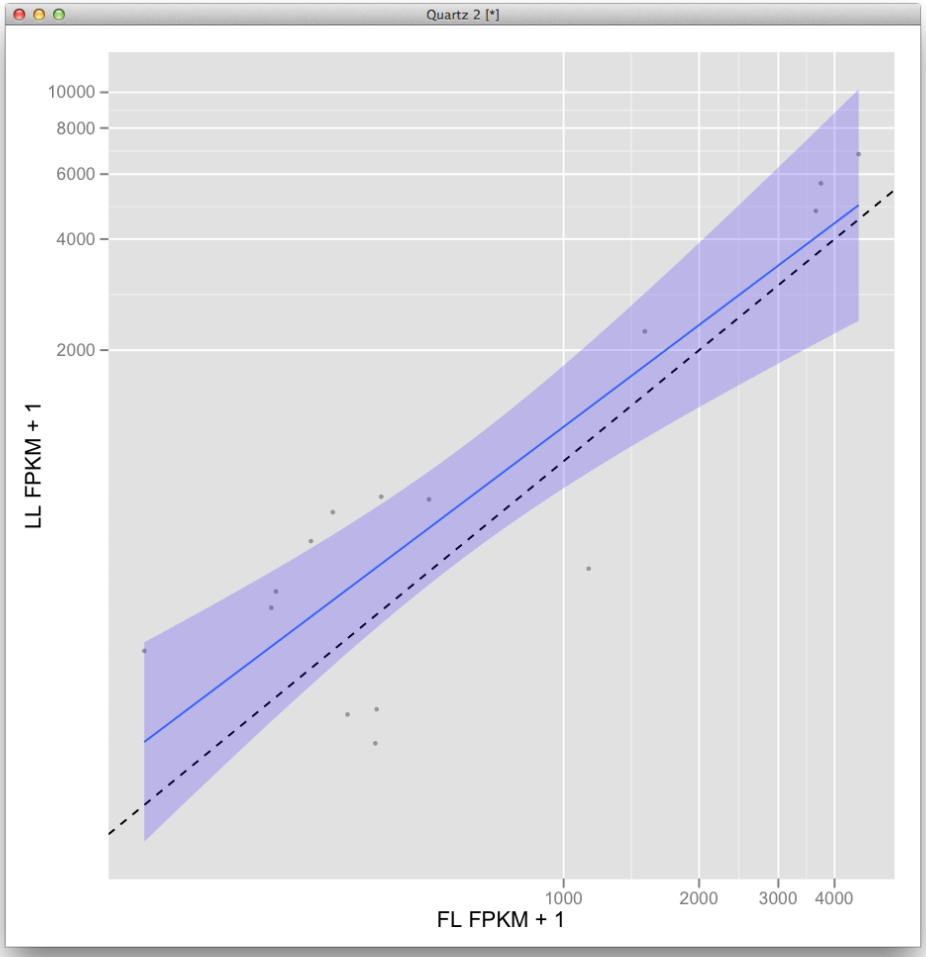
By default, the Jensen-Shannon distance is used as the clustering metric.

Barplot of Significantly Regulated Genes



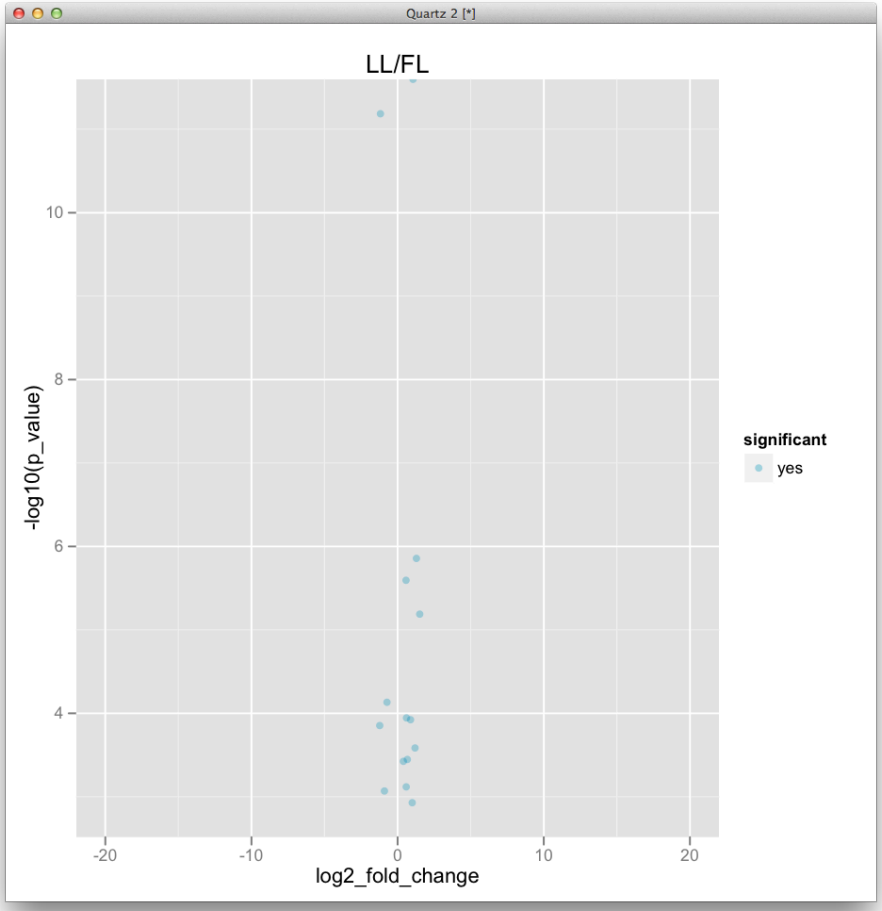

```
> expressionBarplot(mySigGenes)
ymax not defined: adjusting position using y instead
>
```

Scatter Plot of Significantly Regulated Genes



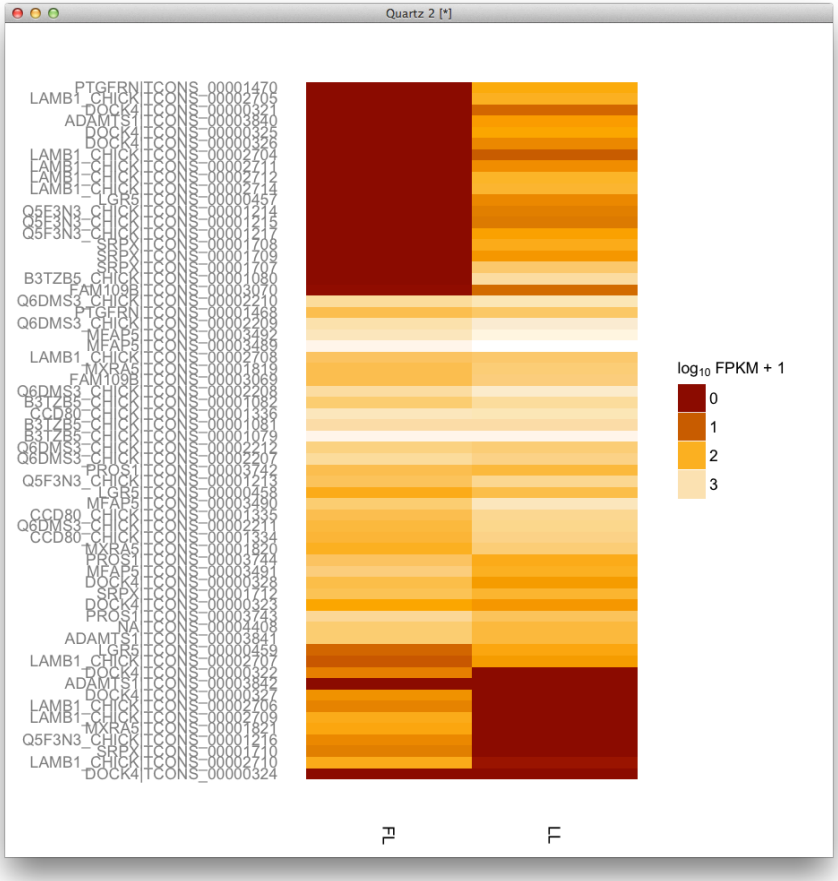

```
> csScatter(mySigGenes, 'FL', 'LL', smooth=T)
Using tracking_id, sample_name as id variables
>
```

Volcano Plot of Significantly Regulated Genes




```
> csVolcano(mySigGenes, 'FL', 'LL')
>
```

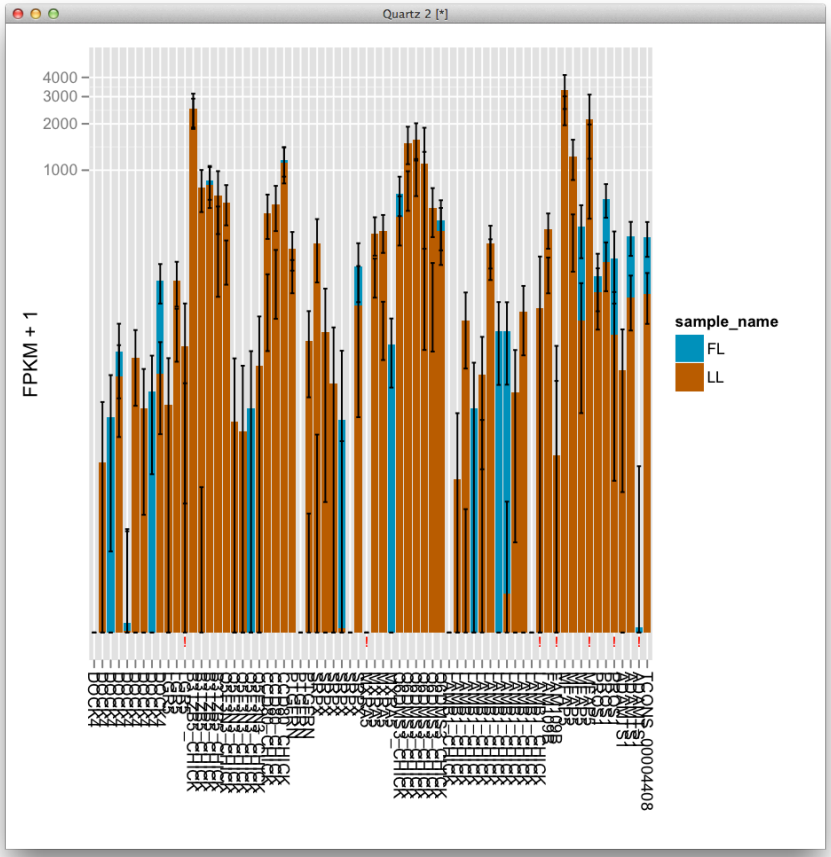
Heatmap of Significantly Regulated Isoforms (optional)




```
> csHeatmap(isoforms(mySigGenes), cluster="both")
Using tracking_id, sample_name as id variables
Using  as id variables
>
```

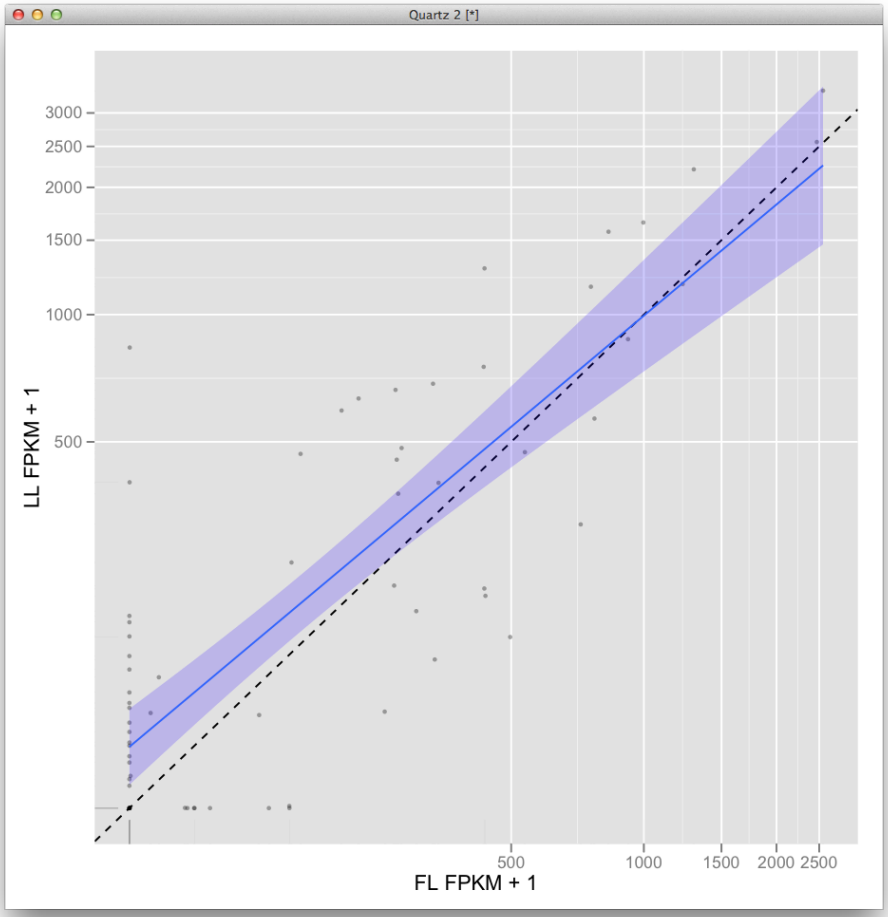
By default, the Jensen-Shannon distance is used as the clustering metric.

Barplot of Significantly Regulated Isoforms (optional)



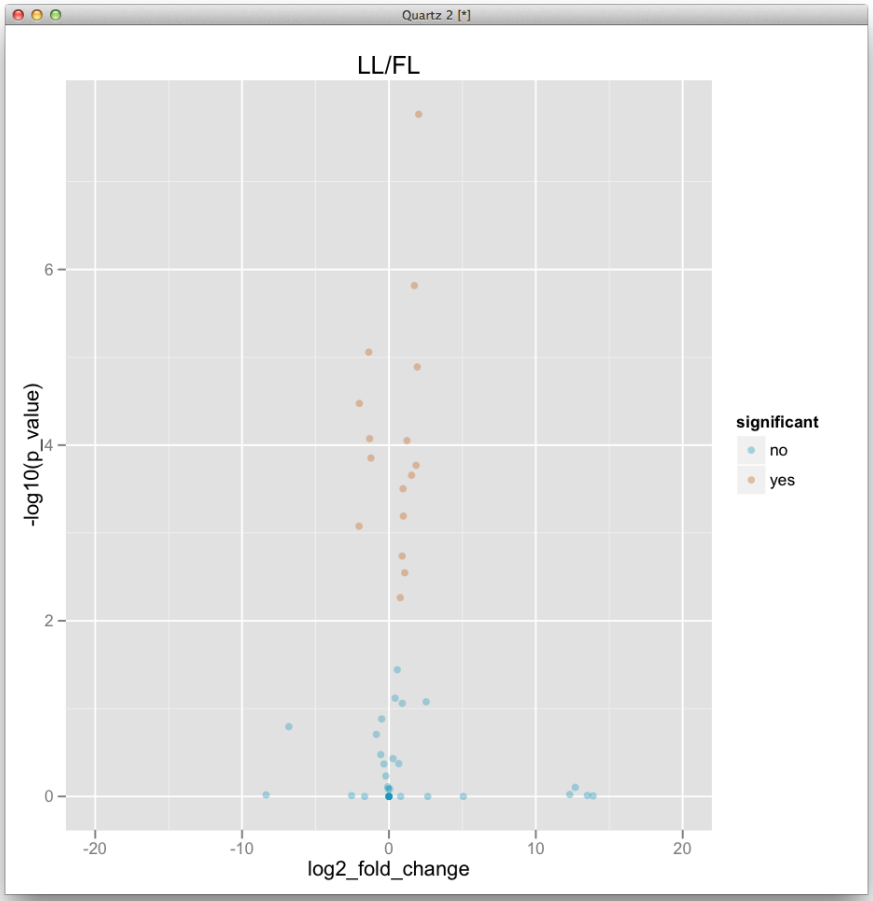

```
> expressionBarplot(isoforms(mySigGenes))
ymax not defined: adjusting position using y instead
>
```

Scatter Plot of Significantly Regulated Isoforms (optional)




```
> csScatter(isoforms(mySigGenes), 'FL', 'LL',
smooth=T)
Using tracking_id, sample_name as id variables
>
```

Volcano Plot of Significantly Regulated Isoforms (optional)




```
> csVolcano(isoforms(mySigGenes), 'FL', 'LL')
Warning message:
Removed 17 rows containing missing values
(geom_point).
>
```

Map Significantly Regulated Genes for iProXpress

```
$ mapSigGenes4iProXpress.pl reference/idmapping.txt cuffdiff_out/gene_exp.diff
F1NKB3 decrease     315.979(FL):190.759(LL):-0.728074(log2fc):decrease:7.48577e-05(pval):0.00764796(qval)
E1C2F8 increase     101.959(FL):291.104(LL):1.51355(log2fc):increase:6.47663e-06(pval):0.000794035(qval)
F1NXW6 increase     3619.21(FL):4752.44(LL):0.392991(log2fc):increase:0.000373513(pval):0.0190803(qval)
Q5F3N3 increase     291.884(FL):712.317(LL):1.28712(log2fc):increase:1.39072e-06(pval):0.000284171(qval)
F1NTI1 increase     1500.05(FL):2234.23(LL):0.574764(log2fc):increase:2.50853e-06(pval):0.000384432(qval)
F1NTD7 increase     209.046(FL):385.674(LL):0.883563(log2fc):increase:0.000119441(pval):0.00915213(qval)
F1NUG5 increase     259.464(FL):592.13(LL):1.19038(log2fc):increase:0.000260356(pval):0.0159599(qval)
E1BYQ7 increase     378.182(FL):786.069(LL):1.05558(log2fc):increase:0(pval):0(qval)
F1P4N9 increase     3713.88(FL):5652.7(LL):0.606012(log2fc):increase:0.000113324(pval):0.00915213(qval)
F1NUC1 increase     487.081(FL):772.948(LL):0.666211(log2fc):increase:0.000357498(pval):0.0190803(qval)
E1BXH5 increase     214.389(FL):428.553(LL):0.999238(log2fc):increase:0.00117986(pval):0.0482168(qval)
F1NIR2 increase     4509.81(FL):6781.72(LL):0.588585(log2fc):increase:0.000761379(pval):0.035902(qval)
E1C6L4 decrease     1121.45(FL):496.313(LL):-1.17604(log2fc):decrease:7.19202e-12(pval):2.20436e-09(qval)
F1P3T6 decrease     369.098(FL):197.532(LL):-0.901915(log2fc):decrease:0.000852187(pval):0.0373136(qval)
```

- Copy the output as shown in blue above, and paste it into the text box on the iProXpress web site at the URL below
  http://pir18.georgetown.edu/iproxpress2/
- Click submit button, you can now do GO Slim analysis and other analyses from there.
- Next section of this short course will cover more on this topic using full set of significantly regulated genes.

Summary

- Use Cuffdiff to identify differentially expressed genes/transcripts.
- Use CummeRbund to explore the Cufflinks RNA-Seq output.

References

- Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform first published online April 19, 2012 doi:10.1093/bib/bbs017.
- Langmead B et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
- Li H et al. The Sequence Alignment/Map format and SAMTools. Bioinformatics. 2009 Aug 15;25(16):2078-9.
- Roberts, A et al. Identification of novel transcripts in annotated genomes using RNA-seq. Bioinformatics 27, 2325–2329 (2011).
- Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009 May 1;25(9):1105-11.
- Trapnell C et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511-5.
- Trapnell C et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012 Mar 1;7(3):562-78.