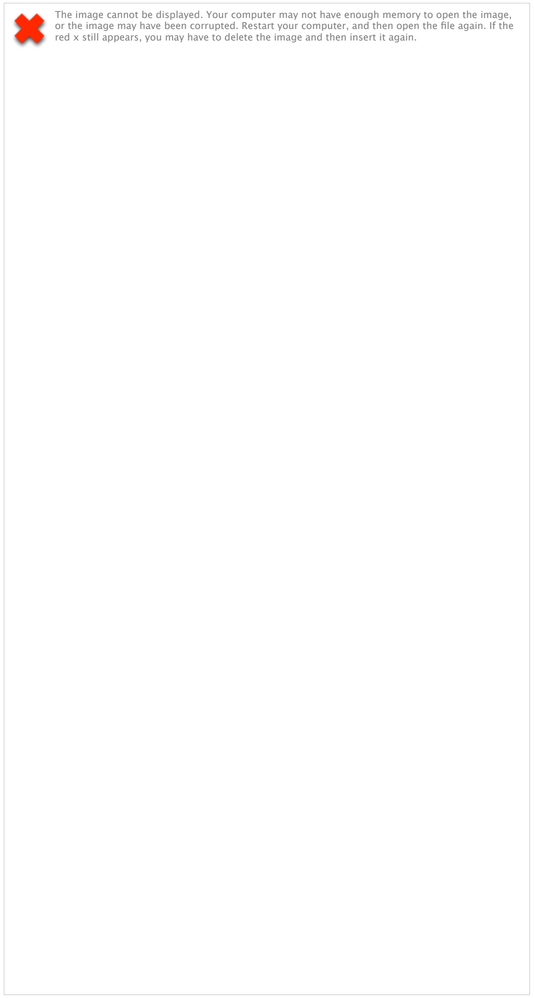
Visualizing RNA-Seq Differential Expression Results with CummeRbund

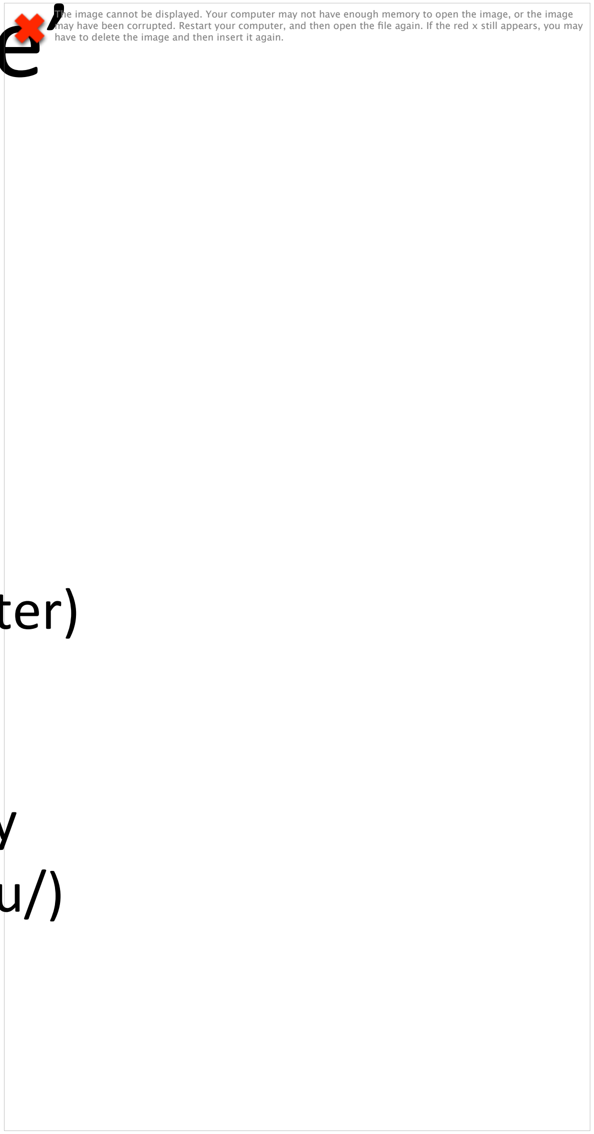
RNA-Seq Pipeline
‘The Tuxedo Suite’



The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

- Software is all free and downloadable from the internet!

- Run locally (on your computer) using a linux platform or
- through the web based bioinformatics site Galaxy (https://main.g2.bx.psu.edu/)



The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Trapnell et al. (2012) Nature Protocols 7 (3) 562-578.

Files you will need to analyze RNA-seq data using Tuxedo Suite

- RNA-Seq files-FASTQ (Sanger) format
  - FASTQ is a form of FASTA (sequence) file which includes quality scores
- Your genome file (FASTA file)
- Genome annotation file (either GFF3 or GTF file)

R Programming Language

- R is a programming language traditionally used for statistical and graphical analysis
- While all other Tuxedo Suite programs are run in Linux, the final 'visualization' step-CummeRbund-is run in R

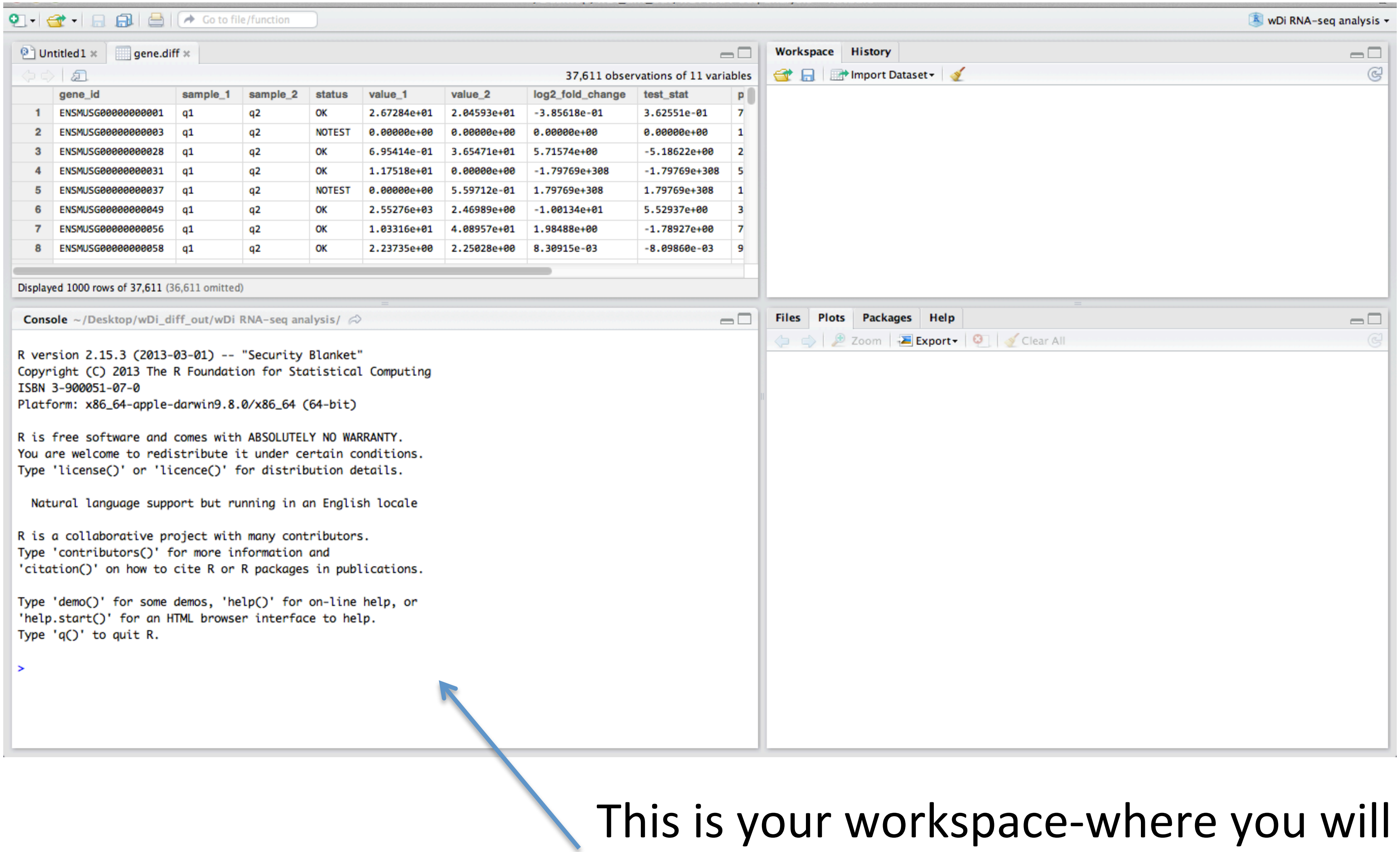
- Download R

(http://www.r-project.org/)-you can use this to run CummeRbund, however it is a bit more primitive than Rstudio (I find RStudio is easier to use)

- Download RStudio-

(http://www.rstudio.com/ide/download/desktop)

RStudio



37,611 observations of 11 variables

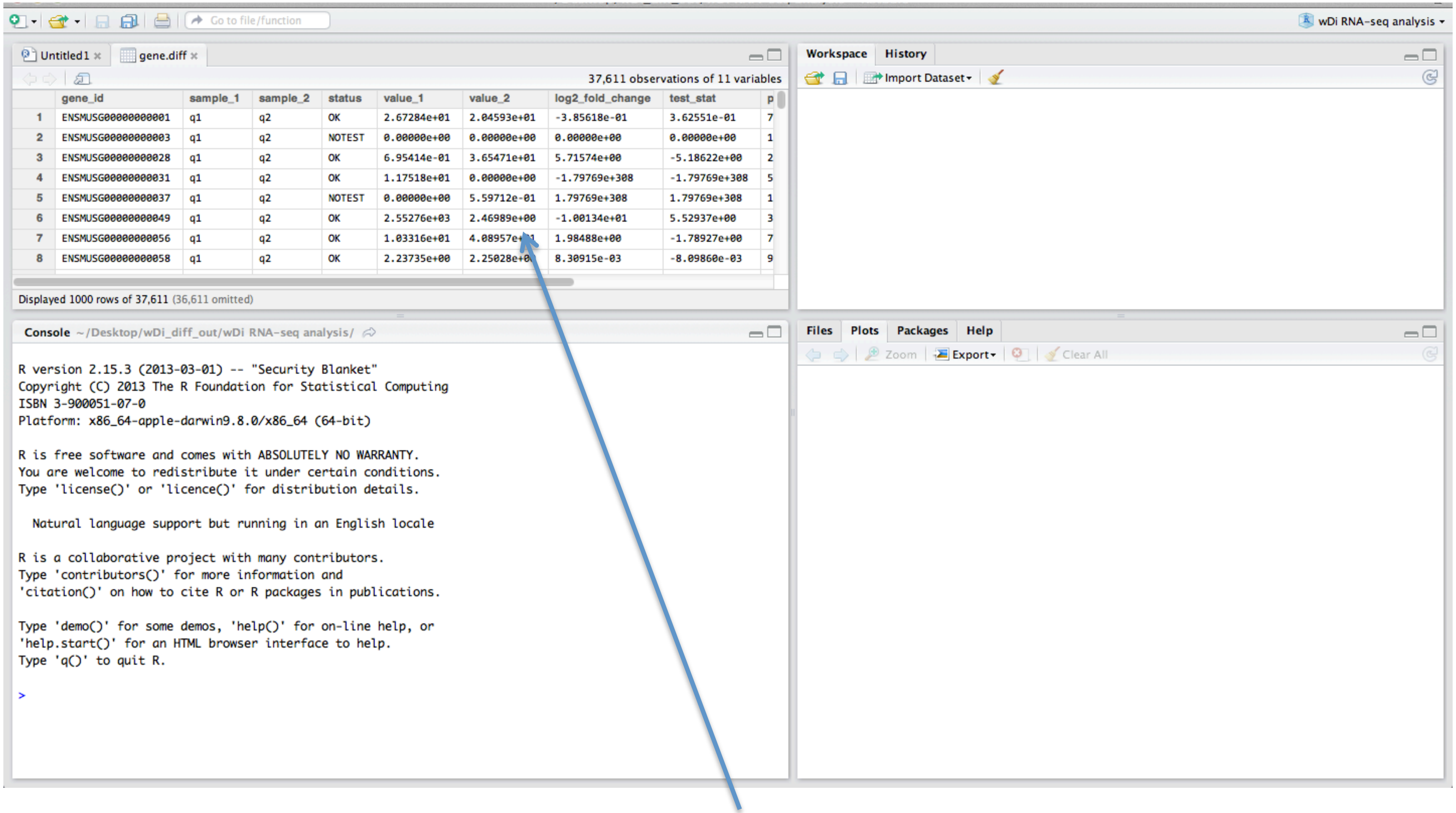
	gene_id	sample_1	sample_2	status	value_1	value_2	log2_fold_change	test_stat	p
1	ENSMUSG00000000001	q1	q2	OK	2.67284e+01	2.04593e+01	-3.85618e-01	3.62551e-01	7
2	ENSMUSG00000000003	q1	q2	NOTEST	0.00000e+00	0.00000e+00	0.00000e+00	0.00000e+00	1
3	ENSMUSG00000000028	q1	q2	OK	6.95414e-01	3.65471e+01	5.71574e+00	-5.18622e+00	2
4	ENSMUSG00000000031	q1	q2	OK	1.17518e+01	0.00000e+00	-1.79769e+308	-1.79769e+308	5
5	ENSMUSG00000000037	q1	q2	NOTEST	0.00000e+00	5.59712e-01	1.79769e+308	1.79769e+308	1
6	ENSMUSG00000000049	q1	q2	OK	2.55276e+03	2.46989e+00	-1.00134e+01	5.52937e+00	3
7	ENSMUSG00000000056	q1	q2	OK	1.03316e+01	4.08957e+01	1.98488e+00	-1.78927e+00	7
8	ENSMUSG00000000058	q1	q2	OK	2.23735e+00	2.25028e+00	8.30915e-03	-8.09860e-03	9

Displayed 1000 rows of 37,611 (36,611 omitted)

```
R version 2.15.3 (2013-03-01) -- "Security Blanket"
Copyright (C) 2013 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin9.8.0/x86_64 (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.

  Natural language support but running in an English locale

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.

>
```

This is your workspace-where you will type all commands!

RStudio



37,611 observations of 11 variables

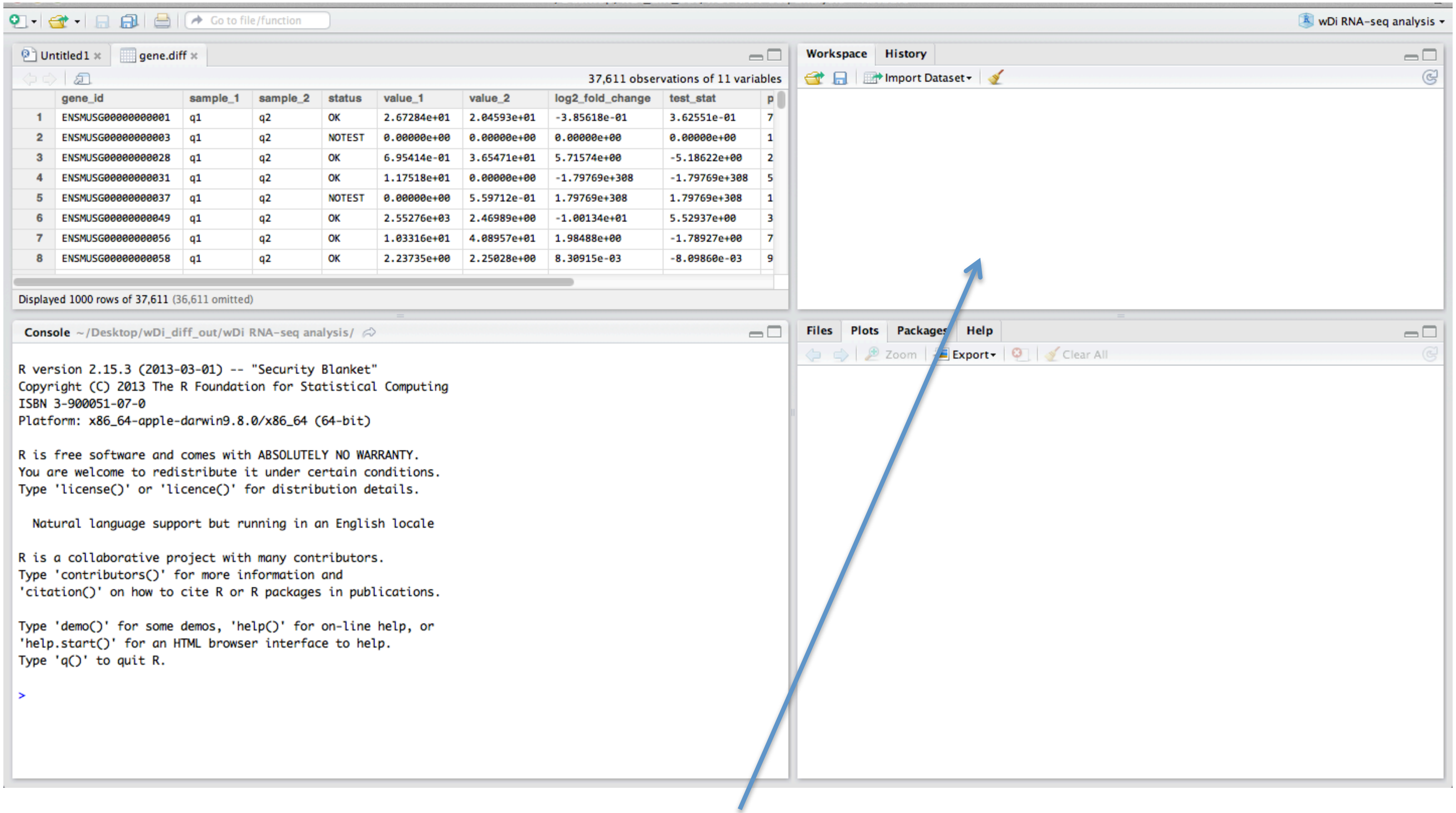
	gene_id	sample_1	sample_2	status	value_1	value_2	log2_fold_change	test_stat	p
1	ENSMUSG00000000001	q1	q2	OK	2.67284e+01	2.04593e+01	-3.85618e-01	3.62551e-01	7
2	ENSMUSG00000000003	q1	q2	NOTEST	0.00000e+00	0.00000e+00	0.00000e+00	0.00000e+00	1
3	ENSMUSG00000000028	q1	q2	OK	6.95414e-01	3.65471e+01	5.71574e+00	-5.18622e+00	2
4	ENSMUSG00000000031	q1	q2	OK	1.17518e+01	0.00000e+00	-1.79769e+308	-1.79769e+308	5
5	ENSMUSG00000000037	q1	q2	NOTEST	0.00000e+00	5.59712e-01	1.79769e+308	1.79769e+308	1
6	ENSMUSG00000000049	q1	q2	OK	2.55276e+03	2.46989e+00	-1.00134e+01	5.52937e+00	3
7	ENSMUSG00000000056	q1	q2	OK	1.03316e+01	4.08957e+01	1.98488e+00	-1.78927e+00	7
8	ENSMUSG00000000058	q1	q2	OK	2.23735e+00	2.25028e+00	8.30915e-03	-8.09860e-03	9

Displayed 1000 rows of 37,611 (36,611 omitted)

```
R version 2.15.3 (2013-03-01) -- "Security Blanket"
Copyright (C) 2013 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin9.8.0/x86_64 (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.

  Natural language support but running in an English locale

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.

>
```

This is where any data tables you create will appear!

RStudio



37,611 observations of 11 variables

	gene_id	sample_1	sample_2	status	value_1	value_2	log2_fold_change	test_stat	p
1	ENSMUSG00000000001	q1	q2	OK	2.67284e+01	2.04593e+01	-3.85618e-01	3.62551e-01	7
2	ENSMUSG00000000003	q1	q2	NOTEST	0.00000e+00	0.00000e+00	0.00000e+00	0.00000e+00	1
3	ENSMUSG00000000028	q1	q2	OK	6.95414e-01	3.65471e+01	5.71574e+00	-5.18622e+00	2
4	ENSMUSG00000000031	q1	q2	OK	1.17518e+01	0.00000e+00	-1.79769e+308	-1.79769e+308	5
5	ENSMUSG00000000037	q1	q2	NOTEST	0.00000e+00	5.59712e-01	1.79769e+308	1.79769e+308	1
6	ENSMUSG00000000049	q1	q2	OK	2.55276e+03	2.46989e+00	-1.00134e+01	5.52937e+00	3
7	ENSMUSG00000000056	q1	q2	OK	1.03316e+01	4.08957e+01	1.98488e+00	-1.78927e+00	7
8	ENSMUSG00000000058	q1	q2	OK	2.23735e+00	2.25028e+00	8.30915e-03	-8.09860e-03	9

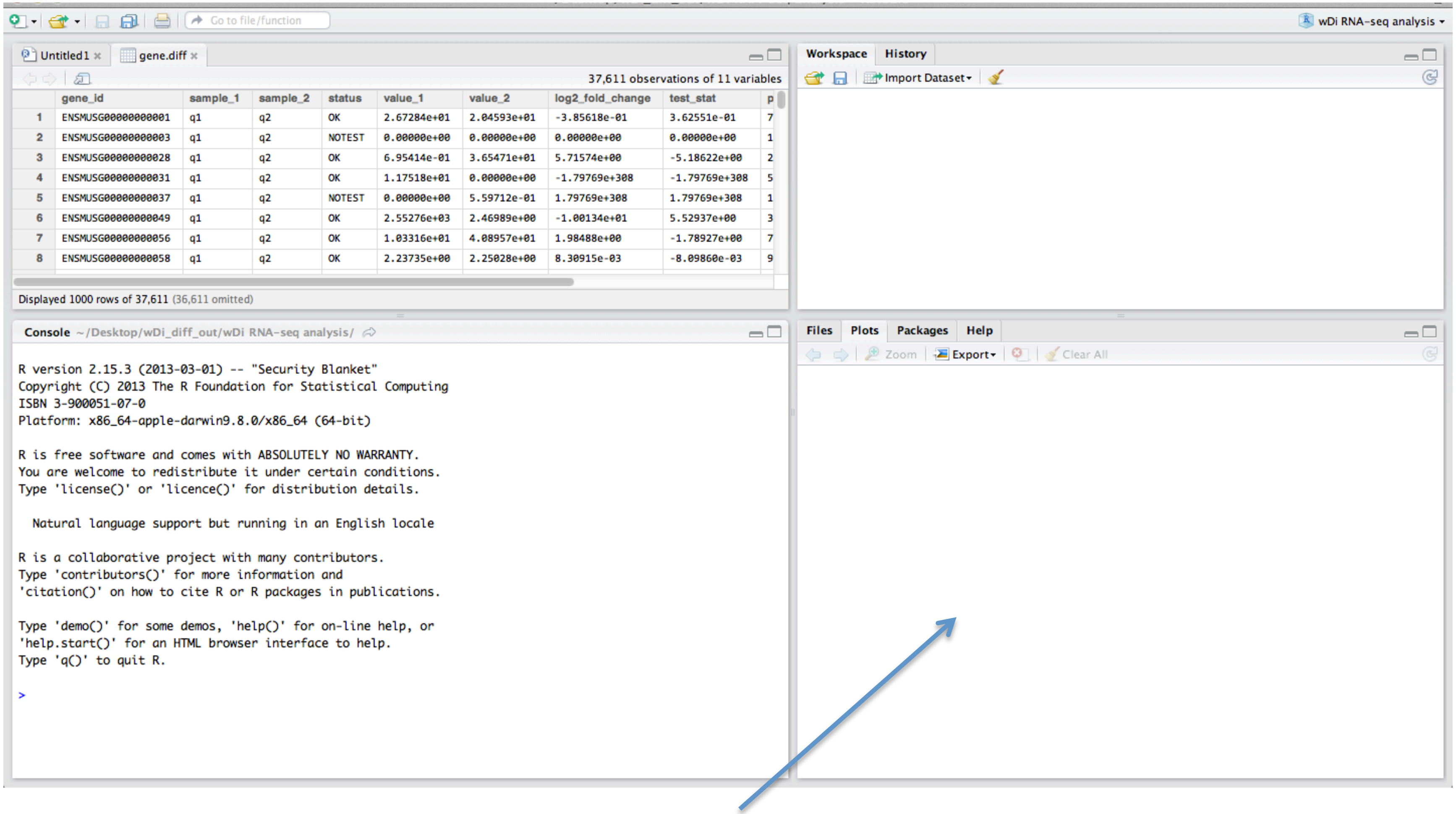
Displayed 1000 rows of 37,611 (36,611 omitted)

Console ~/Desktop/wDi_diff_out/wDi RNA-seq analysis/

```
R version 2.15.3 (2013-03-01) -- "Security Blanket"
Copyright (C) 2013 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin9.8.0/x86_64 (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.

  Natural language support but running in an English locale

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.

>
```

This is where any ‘objects’ or gene sets you create will appear!

RStudio



37,611 observations of 11 variables

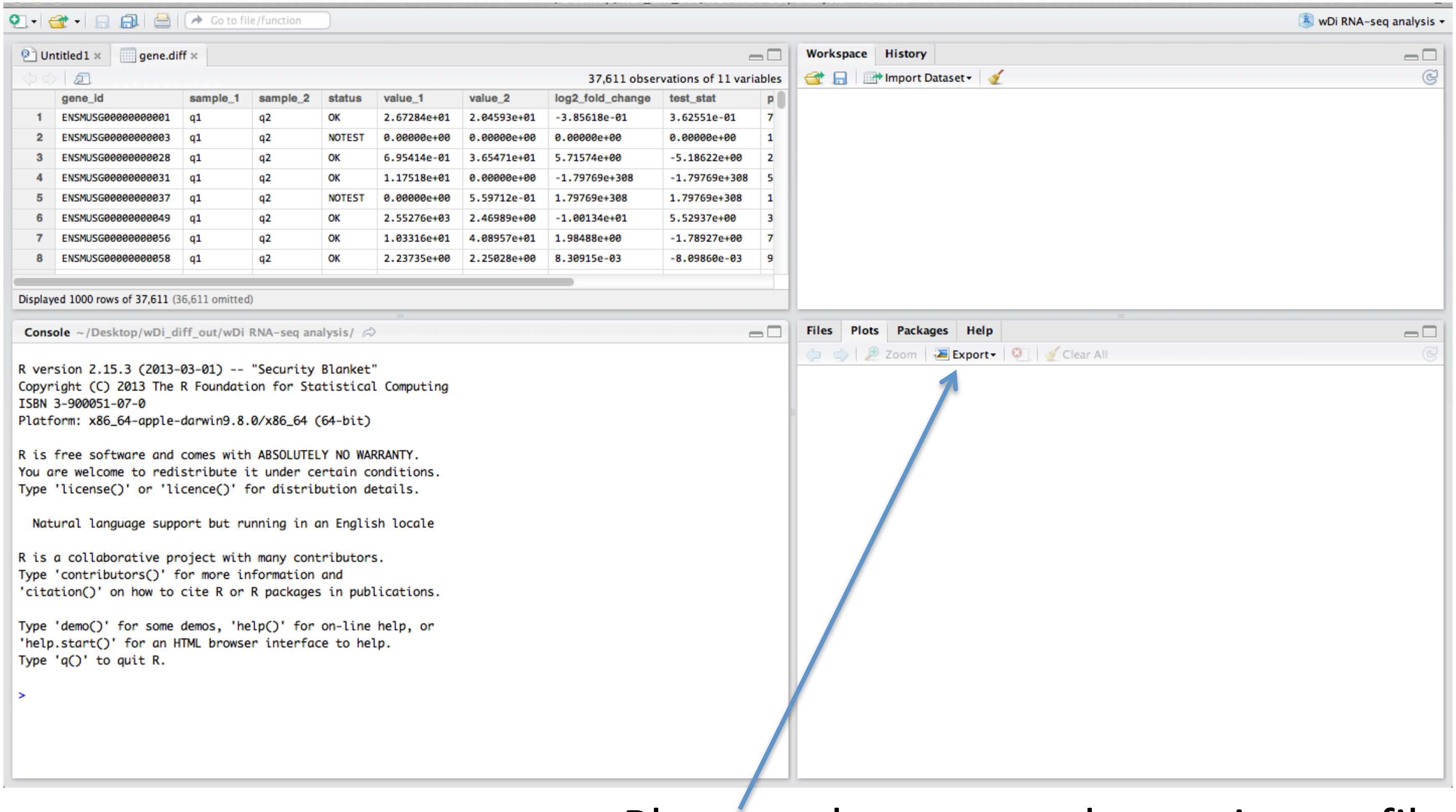
	gene_id	sample_1	sample_2	status	value_1	value_2	log2_fold_change	test_stat	p
1	ENSMUSG00000000001	q1	q2	OK	2.67284e+01	2.04593e+01	-3.85618e-01	3.62551e-01	7
2	ENSMUSG00000000003	q1	q2	NOTEST	0.00000e+00	0.00000e+00	0.00000e+00	0.00000e+00	1
3	ENSMUSG00000000028	q1	q2	OK	6.95414e-01	3.65471e+01	5.71574e+00	-5.18622e+00	2
4	ENSMUSG00000000031	q1	q2	OK	1.17518e+01	0.00000e+00	-1.79769e+308	-1.79769e+308	5
5	ENSMUSG00000000037	q1	q2	NOTEST	0.00000e+00	5.59712e-01	1.79769e+308	1.79769e+308	1
6	ENSMUSG00000000049	q1	q2	OK	2.55276e+03	2.46989e+00	-1.00134e+01	5.52937e+00	3
7	ENSMUSG00000000056	q1	q2	OK	1.03316e+01	4.08957e+01	1.98488e+00	-1.78927e+00	7
8	ENSMUSG00000000058	q1	q2	OK	2.23735e+00	2.25028e+00	8.30915e-03	-8.09860e-03	9

Displayed 1000 rows of 37,611 (36,611 omitted)

```
R version 2.15.3 (2013-03-01) -- "Security Blanket"
Copyright (C) 2013 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin9.8.0/x86_64 (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.

  Natural language support but running in an English locale

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.

>
```

This is where any plots you make will appear!

RStudio



37,611 observations of 11 variables

	gene_id	sample_1	sample_2	status	value_1	value_2	log2_fold_change	test_stat	p
1	ENSMUSG00000000001	q1	q2	OK	2.67284e+01	2.04593e+01	-3.85618e-01	3.62551e-01	7
2	ENSMUSG00000000003	q1	q2	NOTEST	0.00000e+00	0.00000e+00	0.00000e+00	0.00000e+00	1
3	ENSMUSG00000000028	q1	q2	OK	6.95414e-01	3.65471e+01	5.71574e+00	-5.18622e+00	2
4	ENSMUSG00000000031	q1	q2	OK	1.17518e+01	0.00000e+00	-1.79769e+308	-1.79769e+308	5
5	ENSMUSG00000000037	q1	q2	NOTEST	0.00000e+00	5.59712e-01	1.79769e+308	1.79769e+308	1
6	ENSMUSG00000000049	q1	q2	OK	2.55276e+03	2.46989e+00	-1.00134e+01	5.52937e+00	3
7	ENSMUSG00000000056	q1	q2	OK	1.03316e+01	4.08957e+01	1.98488e+00	-1.78927e+00	7
8	ENSMUSG00000000058	q1	q2	OK	2.23735e+00	2.25028e+00	8.30915e-03	-8.09860e-03	9

Displayed 1000 rows of 37,611 (36,611 omitted)

Console ~/Desktop/wDi_diff_out/wDi RNA-seq analysis/

```
R version 2.15.3 (2013-03-01) -- "Security Blanket"
Copyright (C) 2013 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin9.8.0/x86_64 (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.

  Natural language support but running in an English locale

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.

>
```

Plots can be exported as an image file (png, jpeg, tiff, bmp, svg or evs) or as a pdf

R basics

- In R when you type a command and add your open parenthesis (R automatically closes it for you
  - You type (and () appears
- Get working directory
  - getwd()
- Set working directory
  - setwd()
- This is pretty much all the R language you need to know to run CummeRbund-the rest of the language is specific to CummeRbund

CummeRbund

- Download CummeRbund-

(http://compbio.mit.edu/cummeRbund)

- -on the right hand side of the page (under Releases) select the version you need (Mac OS or Windows).
- This will download a compressed file into your downloads.
- Unzip this file.

Download Cuffdiff Files from Galaxy

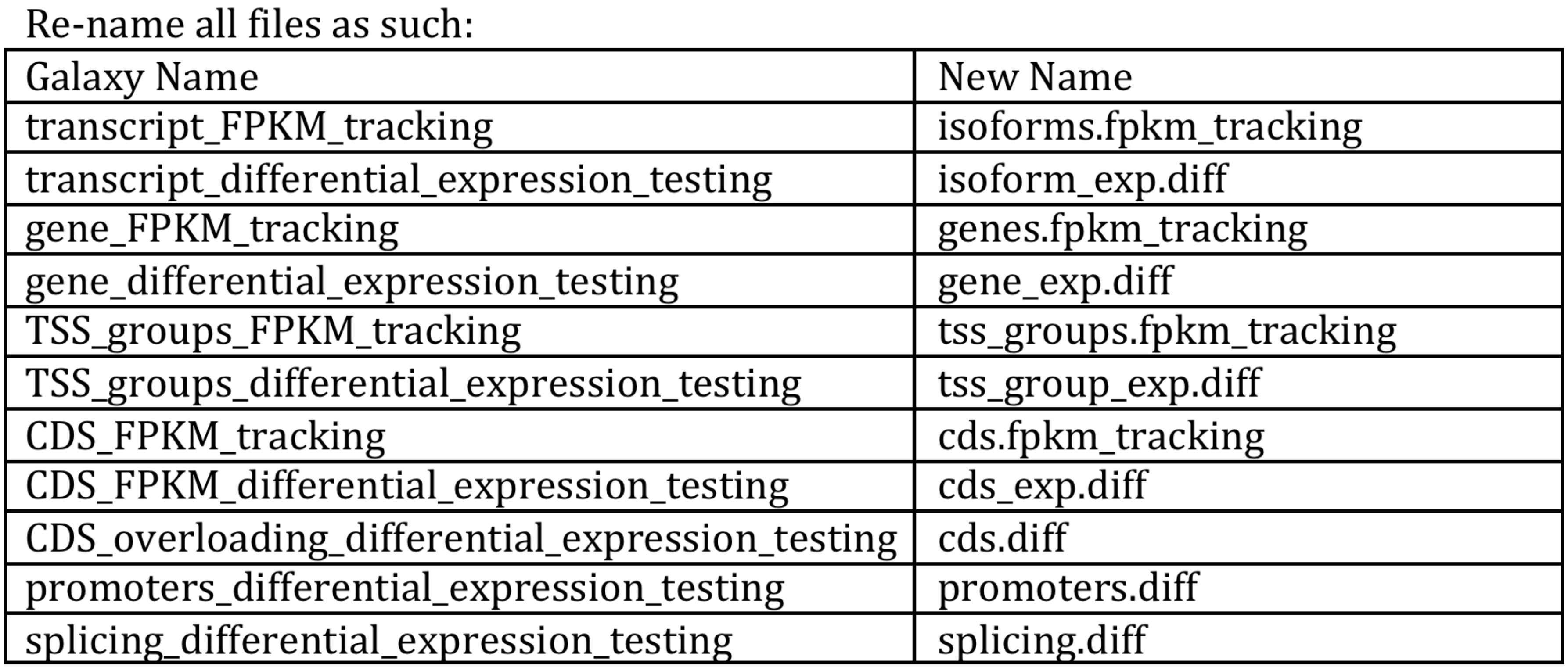
- Create a new folder on your Desktop called diff_out
- From Galaxy history: Download all 11 Cuffdiff output files.
- Once they are all downloaded, move all 11 files from your downloads folder (or wherever your downloads go) into the newly created diff_out folder on your Desktop.

Re-Naming Cuffdiff Output Files

- All files must be re-named in order for CummeRbund to recognize them.

- All Galaxy downloaded file names will begin with something like: Galaxy56[Cuffdiff_on_data_45,_data_41,_and_data_3
- this should be fairly similar for all 11 files and we can ignore-what we care about is at the end of the Galaxy file name, *i.e.* transcript_FPKM_tracking. This is the part that tells you what the output is and how it must be re-named.

Renaming Galaxy Cuffdiff Files



Re-name all files as such:

Galaxy Name	New Name
transcript_FPKM_tracking	isoforms.fpkm_tracking
transcript_differential_expression_testing	isoform_exp.diff
gene_FPKM_tracking	genes.fpkm_tracking
gene_differential_expression_testing	gene_exp.diff
TSS_groups_FPKM_tracking	tss_groups.fpkm_tracking
TSS_groups_differential_expression_testing	tss_group_exp.diff
CDS_FPKM_tracking	cds.fpkm_tracking
CDS_FPKM_differential_expression_testing	cds_exp.diff
CDS_overloading_differential_expression_testing	cds.diff
promoters_differential_expression_testing	promoters.diff
splicing_differential_expression_testing	splicing.diff

- Once this is complete you can start analyzing data with CummeRbund!

Running R

- In the remaining slides text shown in BLACK are my explanations to you
- Text shown in BLUE are the commands you should input into RStudio
- Text shown in RED are lines of code output from RStudio if your command worked correctly

Visualize the Data with CummeRbund

- Open RStudio

```
R version 2.15.3 (2013-03-01) -- "Security Blanket"
Copyright (C) 2013 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin9.8.0/x86_64 (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.

  Natural language support but running in an English locale

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.
```

Install CummeRbund

- To install the CummeRbund package use the following commands:

```
> source('http://www.bioconductor.org/
biocLite.R')

> biocLite('cummeRbund')
```

Setting the Working Directory

- Get working directory

```
>getwd()
```

- This will tell you what your current working directory is.
- Set working directory-I usually set mine as my computer-note that this could be different on your computer but should be one level up from the Desktop

```
>setwd(“/Users/slatko”)
```

- I then usually check my working directory again-just to make sure it is set where I want it to be.

```
>getwd()
```

Load CummeRbund into R

- To load CummeRbund into R use the following command:

```
>library(cummeRbund)
Loading required package: BiocGenerics
Attaching package: ‘BiocGenerics’
The following object(s) are masked from ‘package:stats’:
   xtabs
The following object(s) are masked from ‘package:base’:
   anyDuplicated, cbind, colnames, duplicated, eval, Filter, Find, get,
   intersect, lapply, Map, mapply, mget, order, paste, pmax, pmax.int, pmin,
   pmin.int, Position, rbind, Reduce, rep.int, rownames, sapply, setdiff, table,
   tapply, union, unique
Loading required package: RSQLite
Loading required package: DBI
Loading required package: ggplot2
Loading required package: reshape2
Loading required package: fastcluster
Attaching package: ‘fastcluster’
The following object(s) are masked from ‘package:stats’:
   hclust
Loading required package: rtracklayer
Loading required package: GenomicRanges
Loading required package: IRanges
Loading required package: Gviz
Loading required package: grid
```

Creating a CummeRbund Database

- Now you must create a database out of your 11 cuffdiff output files.

```
> cuff_data<-readCufflinks('~/Desktop/diff_out’)
```

- Again-this will take a minute or two to run a number of lines of script (see next page) while creating a database file.
- Once this is complete you will notice your diff_out folder on your desktop now contains a file called cuff_data.db
  - This is your CummeRbund database!

```
Creating database ~/Desktop/mouse_diff_out/cuffData.db
Reading ~/Desktop/mouse_diff_out/genes.fpkm_tracking
Checking samples table...
Populating samples table...
Writing genes table
Reshaping geneData table
Recasting
Writing geneData table
Reading ~/Desktop/mouse_diff_out/gene_exp.diff
Writing geneExpDiffData table
Reading ~/Desktop/mouse_diff_out/promoters.diff
Writing promoterDiffData table
No records found in ~/Desktop/mouse_diff_out/promoters.diff
Reading ~/Desktop/mouse_diff_out/isoforms.fpkm_tracking
Checking samples table...
OK!
Writing isoforms table
Reshaping isoformData table
Recasting
Writing isoformData table
Reading ~/Desktop/mouse_diff_out/isoform_exp.diff
Writing isoformExpDiffData table
Reading ~/Desktop/mouse_diff_out/tss_groups.fpkm_tracking
Checking samples table...
OK!
Writing TSS table
No records found in ~/Desktop/mouse_diff_out/tss_groups.fpkm_tracking
TSS FPKM tracking file was empty.
Reading ~/Desktop/mouse_diff_out/tss_group_exp.diff
No records found in ~/Desktop/mouse_diff_out/tss_group_exp.diff
Reading ~/Desktop/mouse_diff_out/splicing.diff
No records found in ~/Desktop/mouse_diff_out/splicing.diff
Reading ~/Desktop/mouse_diff_out/cds.fpkm_tracking
Checking samples table...
OK!
Writing CDS table
No records found in ~/Desktop/mouse_diff_out/cds.fpkm_tracking
CDS FPKM tracking file was empty.
Reading ~/Desktop/mouse_diff_out/cds_exp.diff
No records found in ~/Desktop/mouse_diff_out/cds_exp.diff
Reading ~/Desktop/mouse_diff_out/cds.diff
No records found in ~/Desktop/mouse_diff_out/cds.diff
Indexing Tables...
```

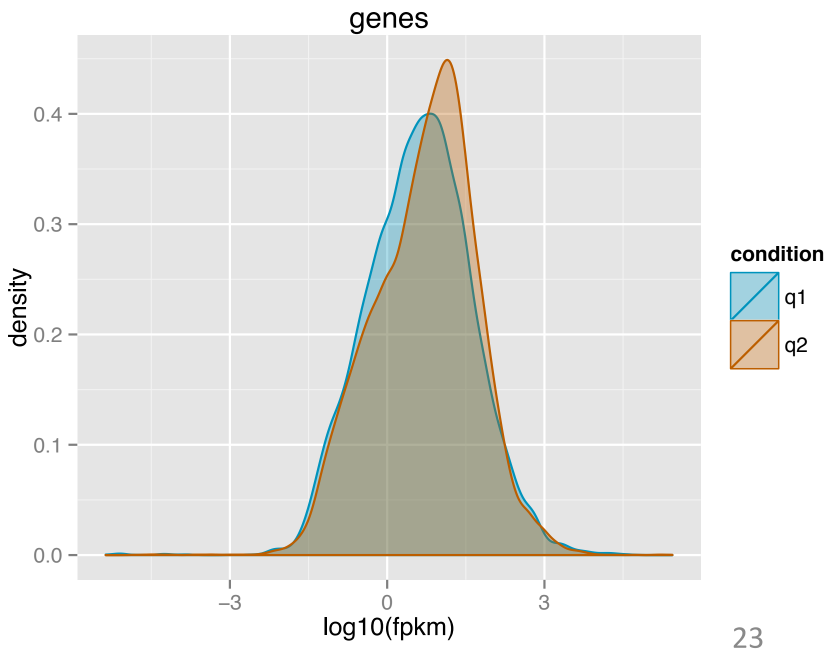
Now it is time to visualize your results!

Density Plot

- The density plot will show you the distribution of your RNA-seq read counts (fpkm)

```
> csDensity(genes(cuff_data))
```

This will plot data for genes. You can also do this with other data from Cuffdiff, *e.g.*, isoforms.



Volcano Plot

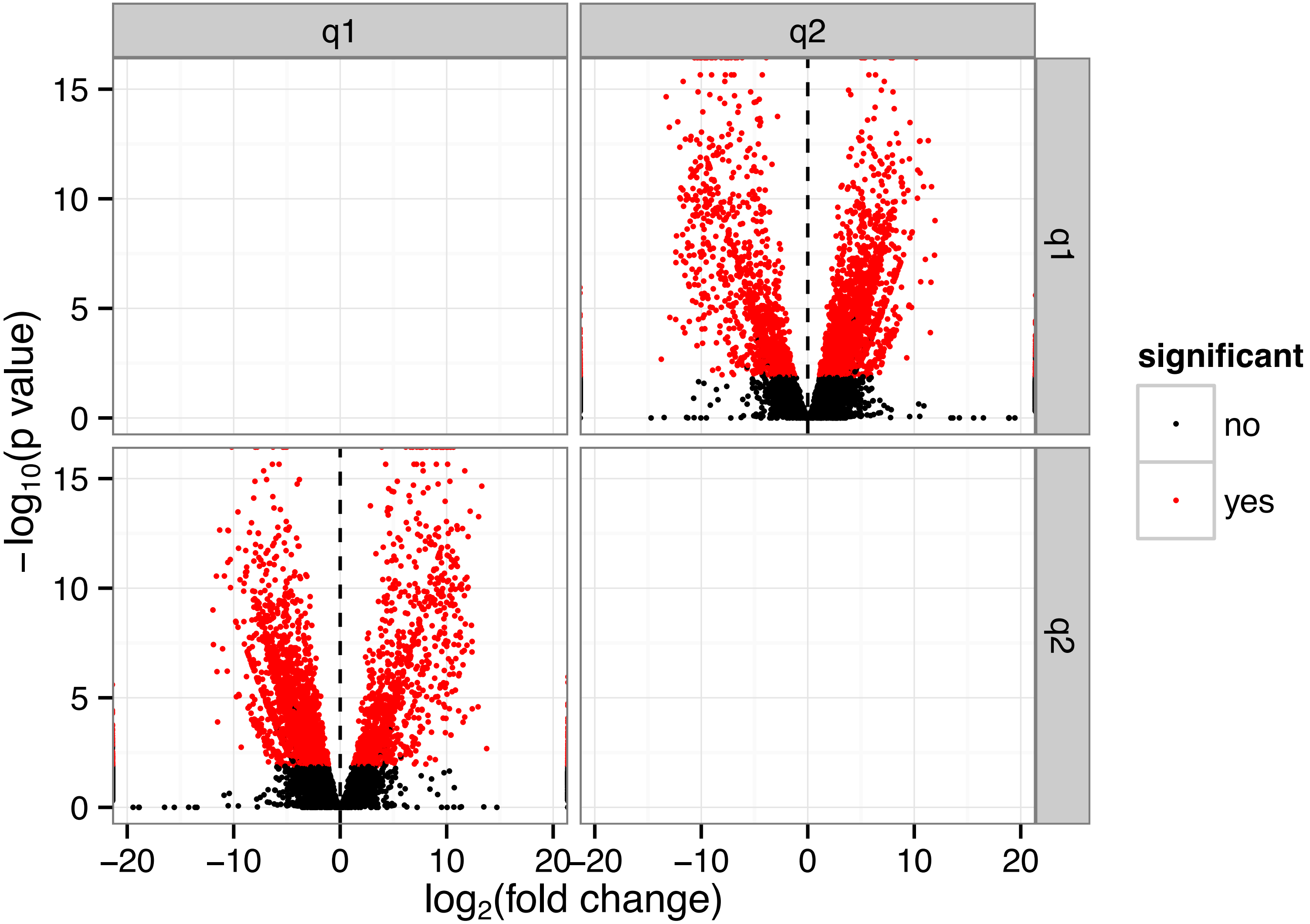
- A volcano plot is a scatter plot that also identifies differentially expressed genes (by color) between samples

```
>v<-csVolcanoMatrix(genes(cuff_data))
```

- This line creates a command (v)-to execute the command you must type the following line

```
>v
```

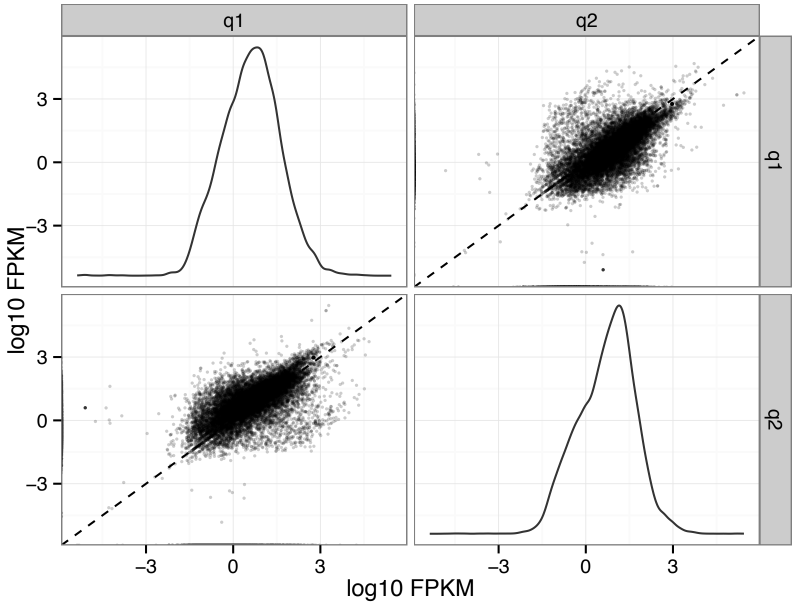
Volcano Matrix



Scatter Plot

- Shows differences in gene expression between two samples
  - If two samples were identical all dots (genes) would fall on the mid-line

>csScatter(genes(cuff_data))



Looking a Specific Genes of Interest

- 3 Genes
  - F9
  - Rdh7
  - Gapdh

Getting Gene Info

```
>myGeneId<-"F9"
> myGene<-getGene(cuff_data,myGeneId)
> myGene

CuffGene instance for gene ENSMUSG00000031138
Short name:     F9
Slots:
        annotation
        features
        fpkm
        repFpkm
        diff
        count
        isoforms  CuffFeature instance of size 1
        TSS       CuffFeature instance of size 0
        CDS       CuffFeature instance of size 0
```

This tells you how many isoforms of this gene there are.

Here you could also find out if your gene had more than one transcriptional start site (TSS)

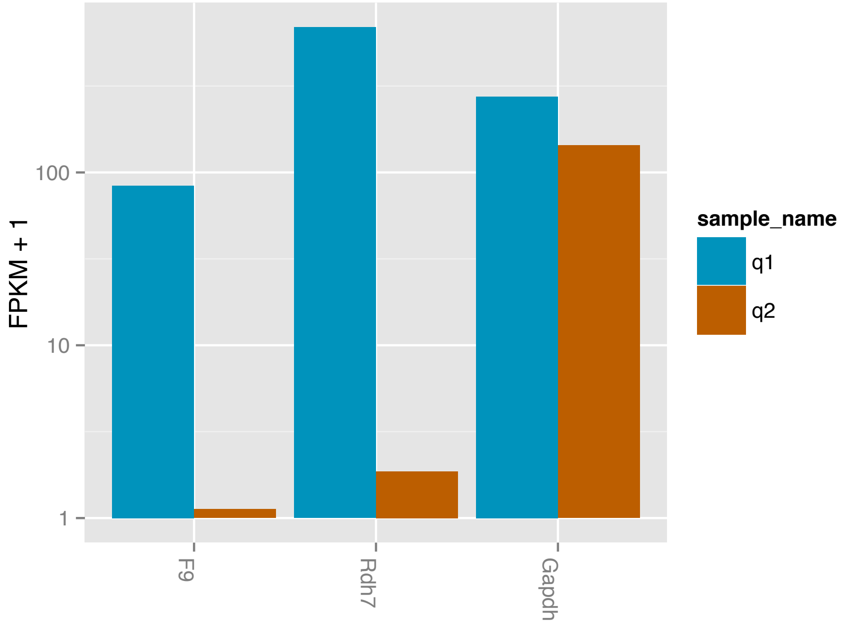
How many isoforms do Rdh7 and Gapdh have??

Looking at Groups of Genes

```
>myGeneIds<- c("F9","Rdh7", "Gapdh")
> myGenes <- getGenes(cuff_data,myGeneIds)
Getting gene information:
        FPKM
        Differential Expression Data
        Annotation Data
        Replicate FPKMs
        Counts
Getting isoforms information:
        FPKM
        Differential Expression Data
        Annotation Data
        Replicate FPKMs
        Counts
Getting CDS information:
        FPKM
        Differential Expression Data
        Annotation Data
        Replicate FPKMs
        Counts
Getting TSS information:
        FPKM
        Differential Expression Data
        Annotation Data
        Replicate FPKMs
        Counts
Getting promoter information:
        distData
Getting splicing information:
        distData
Getting relCDS information:
        distData
```

Plot Expression of 'Your Genes'

```
>gb<-expressionBarplot(myGenes,showErrorbars=FALSE)
Scale for 'colour' is already present. Adding another scale for 'colour', which will
replace the existing scale.
> gb
```

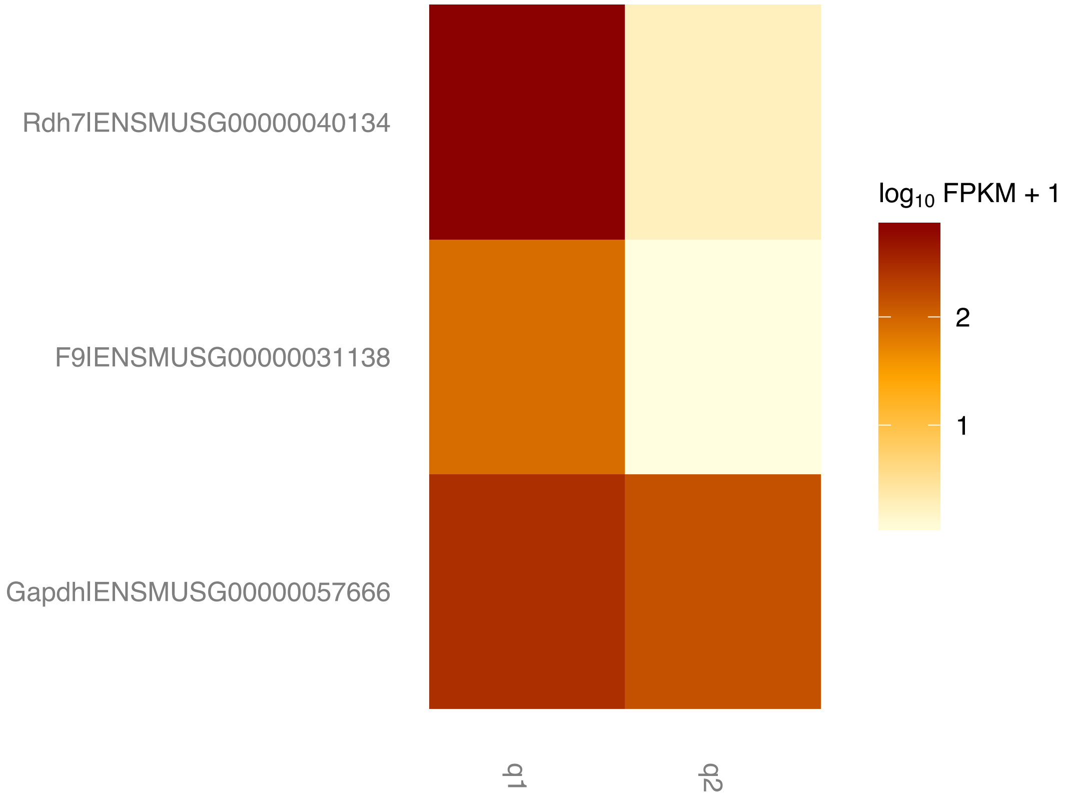


* The argument showErrobars=FALSE is necessary because of a lack of replicates. The default is showErrorbars=TRUE, but because there are no replicates there is no error to show!

Plot Expression of ‘Your Genes’- Heatmap

>h<-csHeatmap(myGenes)

> h



CummeRbund Conclusions

- Relatively easy to use
- Great way to visualize differential expression data from RNA-seq experiments
- This is just the beginning-CummeRbund can do much more!
- If interested, the complete CummeRbund manual can be found online

(http://compbio.mit.edu/cummeRbund/manual_2_0.html)