Visualizing RNA-Seq Differential Expression Results with CummeRbund

RNA-Seq Pipeline 'The Tuxedo Suite'

e image cannot be displayed. Your computer may not have enough memory to open the image, or the image ay have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may ve 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/)

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

(<u>http://www.r-project.org/</u>)-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)

	2 • 🔒 🔒	🖌 🖌 Go to fil	e/function								🖲 wDi RNA-seq analys
O	ntitled1 × gene.di	iff ×						Workspace History			
										es 🚭 🔒 🖙 Import Dataset 🗸 🎸	(
	gene_id	sample_1	sample_2	status	value_1	value_2	log2_fold_change	test_stat	P		
1	ENSMUSG0000000001	q1	q2	ОК	2.67284e+01	2.04593e+01	-3.85618e-01	3.62551e-01	7		
2	ENSMUSG0000000003	q1	q2	NOTEST	0.00000e+00	0.00000e+00	0.00000e+00	0.00000e+00	1		
3	ENSMUSG0000000028	q1	q2	ОК	6.95414e-01	3.65471e+01	5.71574e+00	-5.18622e+00	2		
4	ENSMUSG0000000031	q1	q2	ОК	1.17518e+01	0.00000e+00	-1.79769e+308	-1.79769e+308	5		
5	ENSMUSG0000000037	q1	q2	NOTEST	0.00000e+00	5.59712e-01	1.79769e+308	1.79769e+308	1		
6	ENSMUSG0000000049	q1	q2	ОК	2.55276e+03	2.46989e+00	-1.00134e+01	5.52937e+00	3		
7	ENSMUSG0000000056	q1	q2	ОК	1.03316e+01	4.08957e+01	1.98488e+00	-1.78927e+00	7		
8	ENSMUSG0000000058	q1	q2	ок	2.23735e+00	2.25028e+00	8.30915e-03	-8.09860e-03	9		
isplay	red 1000 rows of 37,611 (36,611 omitted)								
	ala (Daalataa (uDi d	1166		alumin (a)	=					Files Plots Parkages Help	
.ons	ole ~/Desktop/wDI_d	lift_out/wDI	KNA-seq and	aiysis/ 🟳	/			C		rites rious rackages neip	
lati is ou c /pe Nat is ype citc	orm: x86_64-apple- free software and ine welcome to redi 'license()' or 'li unal language supp a collaborative pr 'contributors()' f tion()' on how to 'demo()' for some ustart()' for some	-darwin9.8. comes with istribute i icence()' f port but ru roject with for more in cite R or demos, 'he HTML browse	0/x86_64 (ABSOLUTEL t under cc for distrik unning in c many cont formation R packages elp()' for printerfac	(64-bit) Y NO WAR Partain co- pution do an Englis tributors and s in pub ³ on-line to he ³	RRANTY. onditions. etails. sh locale s. lications. help, or lo.						
ype	'q()' to quit R.					R					

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

• •	2 - 🔒 🔒	🔿 🖌 Go to fi	le/function								🛝 wDi RNA-seq analysis 🕇
🖭 Ur	ntitled1 × gene.di	iff ×					Workspace History				
	1						37,611 obser	ables	🕣 🕞 🖙 Import Dataset - 🔏	G	
	gene_id	sample_1	sample_2	status	value_1	value_2	log2_fold_change	test_stat	p		
1	ENSMUSG0000000001	q1	q2	ОК	2.67284e+01	2.04593e+01	-3.85618e-01	3.62551e-01	7		
2	ENSMUSG0000000003	q1	q2	NOTEST	0.00000e+00	0.00000e+00	0.00000e+00	0.00000e+00	1		
3	ENSMUSG000000028	q1	q2	ОК	6.95414e-01	3.65471e+01	5.71574e+00	-5.18622e+00	2		
4	ENSMUSG0000000031	q1	q2	ОК	1.17518e+01	0.00000e+00	-1.79769e+308	-1.79769e+308	5		
5	ENSMUSG0000000037	q1	q2	NOTEST	0.00000e+00	5.59712e-01	1.79769e+308	1.79769e+308	1		
6	ENSMUSG0000000049	q1	q2	ОК	2.55276e+03	2.46989e+00	-1.00134e+01	5.52937e+00	3		
7	ENSMUSG0000000056	q1	q2	ОК	1.03316e+01	4.08957e+1	1.98488e+00	-1.78927e+00	7		
8	ENSMUSG0000000058	q1	q2	ОК	2.23735e+00	2.25028e+0	8.30915e-03	-8.09860e-03	9		
Display	red 1000 rows of 37 611 (36.611 omitter	Ð				-				
Dispia		50,011 onnittet	2)		-						-
Cons	ole ~/Desktop/wDi_d	liff_out/wDi	RNA-seq and	alysis/ 🛱	>			-		Files Plots Packages Help	
Platf Platf R is You a Type Nat R is Type 'cita Type 'help Type >	free software and ire welcome to redi 'license()' or 'li ural language supp a collaborative pr 'contributors()' f tition()' on how to 'demo()' for some .start()' for an H 'q()' to quit R.	-darwin9.8. comes with istribute i icence()' f port but ru roject with for more in cite R or demos, 'he iTML browse	.0/x86_64 (h ABSOLUTEL it under ce for distrib unning in c h many cont nformation R packages elp()' for er interfac	(64-bit) Y NO WAI ertain coution do an Engli: and in pub on-line re to he	RRANTY. onditions. etails. sh locale s. lications. help, or lp.						
_										1	

This is where any data tables you create will appear!

9 (🛫 - 1 🗄 🖽 1 🚍	Go to file/function								wDi RNA-seq analysis 🕶
Untitled1 × gene.di	ff×				Workspace History	-0			
00 D					iables	🞯 🔒 🖙 Import Dataset 🗸 🎸	C		
gene_id	sample_1 sample	2 status	value_1	value_2	log2_fold_change	test_stat	p		
1 ENSMUSG0000000001	q1 q2	ОК	2.67284e+01	2.04593e+01	-3.85618e-01	3.62551e-01	7		
2 ENSMUSG0000000003	q1 q2	NOTEST	0.00000e+00	0.00000e+00	0.00000e+00	0.00000e+00	1		
3 ENSMUSG000000028	q1 q2	ОК	6.95414e-01	3.65471e+01	5.71574e+00	-5.18622e+00	2		
4 ENSMUSG0000000031	q1 q2	ОК	1.17518e+01	0.00000e+00	-1.79769e+308	-1.79769e+308	5		
5 ENSMUSG0000000037	q1 q2	NOTEST	0.00000e+00	5.59712e-01	1.79769e+308	1.79769e+308	1		
6 ENSMUSG0000000049	q1 q2	ОК	2.55276e+03	2.46989e+00	-1.00134e+01	5.52937e+00	3		
7 ENSMUSG0000000056	q1 q2	ОК	1.03316e+01	4.08957e+01	1.98488e+00	-1.78927e+00	7		
8 ENSMUSG0000000058	q1 q2	OK	2.23735e+00	2.25028e+00	8.30915e-03	-8.09860e-03	9		
								r	
Displayed 1000 rows of 37,611 (36,611 omitted)								
Console ~/Desktop/wDi_d	iff_out/wDi RNA-seq	analysis/	>					Files Plots Packages Help	
R version 2.15.3 (2013- Copyright (C) 2013 The ISBN 3-900051-07-0 Platform: x86_64-apple- R is free software and You are welcome to redi Type 'license()' or 'li Natural language supp R is a collaborative pr Type 'contributors()' f 'citation()' on how to Type 'demo()' for some 'help.start()' for an H Type 'q()' to quit R.	03-01) "Securi R Foundation for darwin9.8.0/x86_6 comes with ABSOLU stribute it under cence()' for dist port but running i roject with many c for more informati cite R or R packa demos, 'help()' f ITML browser inter	ty Blanket Statistica 4 (64-bit) TELY NO WA certain c ribution d n an Engli ontributor on and ges in pub face to he	RRANTY. conditions. letails. sh locale rs. clications. e help, or clp.		Clear All				

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

0 . • (🛫 • 🔒 🔒 🚔	🔿 Go to fi	le/function							(wDi RNA-seq analysis 🗸
🖭 Ur	ntitled1 × gene.di	ff×							Workspace History	-0	
	1.2						37,611 obser	bles	🕣 🕞 🖙 Import Dataset 🗸 🍯	Ç	
	gene_id	sample_1	sample_2	status	value_1	value_2	log2_fold_change	test_stat	p		
1	ENSMUSG0000000001	q1	q2	ОК	2.67284e+01	2.04593e+01	-3.85618e-01	3.62551e-01	7		
2	ENSMUSG0000000003	q1	q2	NOTEST	0.00000e+00	0.00000e+00	0.00000e+00	0.00000e+00	1		
3	ENSMUSG0000000028	q1	q2	ОК	6.95414e-01	3.65471e+01	5.71574e+00	-5.18622e+00	2		
4	ENSMUSG0000000031	q1	q2	ОК	1.17518e+01	0.00000e+00	-1.79769e+308	-1.79769e+308	5		
5	ENSMUSG0000000037	q1	q2	NOTEST	0.00000e+00	5.59712e-01	1.79769e+308	1.79769e+308	1		
6	ENSMUSG0000000049	q1	q2	ОК	2.55276e+03	2.46989e+00	-1.00134e+01	5.52937e+00	3		
7	ENSMUSG0000000056	q1	q2	ОК	1.03316e+01	4.08957e+01	1.98488e+00	-1.78927e+00	7		
8	ENSMUSG0000000058	q1	q2	ОК	2.23735e+00	2.25028e+00	8.30915e-03	-8.09860e-03	9		
		1	1								
Display	ed 1000 rows of 37,611 (3	6,611 omitted	i)								
	ala inclusion (contrat				=					Ellos Blots Dackages Help	
Cons	ole ~/Desktop/wDi_d	iff_out/wDi	RNA-seq an	alysis/ 🔗	>			_		Files Plots Packages help	
Burn	cion 2 15 2 (2012	02 013	"Cocuritu	Pl onlyat						🖕 🚽 🖉 Zoom 🛛 🔁 Export - 🧕 🥑 Clear All	G
Conve	sion 2.15.5 (2013-	03-01) P Foundati	Security	BLanket	1 Computing						
ISBN	3-900051-07-0	K Foundati		uciscicu	tt compacting						
Platf	orm: x86_64-apple-	darwin9.8	.0/x86_64	(64-bit)					1		
Ris	free software and	comes with	n ABSOLUTEI	LY NO WA	RRANTY.						
You a	re welcome to redi	stribute i	it under ce	ertain c	onditions.						
Type	'license()' or 'li	cence()'	for distrib	bution d	etails.						
Nat	ural language supp	ort but ru	unning in d	an Engli	sh locale						
R is	a collaborative pr	oject with	n many cont	tributor	s.						
Туре	'contributors()' f	or more in	nformation	and							
'cita	tion()' on how to	cite R or	R packages	s in pub	lications.						
Type	'demo()' for some	demos 'he	aln()' for	on-line	heln or						
'heln	start()' for an H	TML browse	er interfa	ce to he	ln.						
Type	'q()' to quit R.										
	-										
>											
									1		
							-				I
								nisis	M	vnere anv plots voll ma	Ke WIII

appear!

0 . • (2 • 🔒 🔒	🔿 Go to fi	le/function							🛞 wDi RNA-seq analysi
O Ur	ntitled1 × gene.dit	f×						Workspace History		
♦ ⇒ 27,611 observations of 11 variables									ables	🞯 🔒 📑 Import Dataset + 🥑
	gene_id	sample_1	sample_2	status	value_1	value_2	log2_fold_change	test_stat	p	
1	ENSMUSG0000000001	q1	q2	ОК	2.67284e+01	2.04593e+01	-3.85618e-01	3.62551e-01	7	
2	ENSMUSG0000000003	q1	q2	NOTEST	0.00000e+00	0.00000e+00	0.00000e+00	0.00000e+00	1	
3	ENSMUSG0000000028	q1	q2	ОК	6.95414e-01	3.65471e+01	5.71574e+00	-5.18622e+00	2	
4	ENSMUSG0000000031	q1	q2	ОК	1.17518e+01	0.00000e+00	-1.79769e+308	-1.79769e+308	5	
5	ENSMUSG0000000037	q1	q2	NOTEST	0.00000e+00	5.59712e-01	1.79769e+308	1.79769e+308	1	
6	ENSMUSG0000000049	q1	q2	ОК	2.55276e+03	2.46989e+00	-1.00134e+01	5.52937e+00	3	
7	ENSMUSG0000000056	q1	q2	ОК	1.03316e+01	4.08957e+01	1.98488e+00	-1.78927e+00	7	
8	ENSMUSG0000000058	q1	q2	ОК	2.23735e+00	2.25028e+00	8.30915e-03	-8.09860e-03	9	
	1	1	1	1	1					
Display	ed 1000 rows of 37,611 (3	6,611 omitted	i)							
Cons	ole ~/Deskton/wDi di	ff_out/wDi	RNA-seg an:	alvsis/ d	=			_		Files Plots Packages Help
	one pesseoppinon_a		inor ocq an					-		A Clear All
<pre>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.</pre>										
Plots ca (png, jr										n be exported as an image file eg, tiff, bmp, svg or evs) or as
										a pui

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')</pre>
- 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))</pre>
- 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 ENSMUSG0000031138

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)</pre>

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.



* 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





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)