

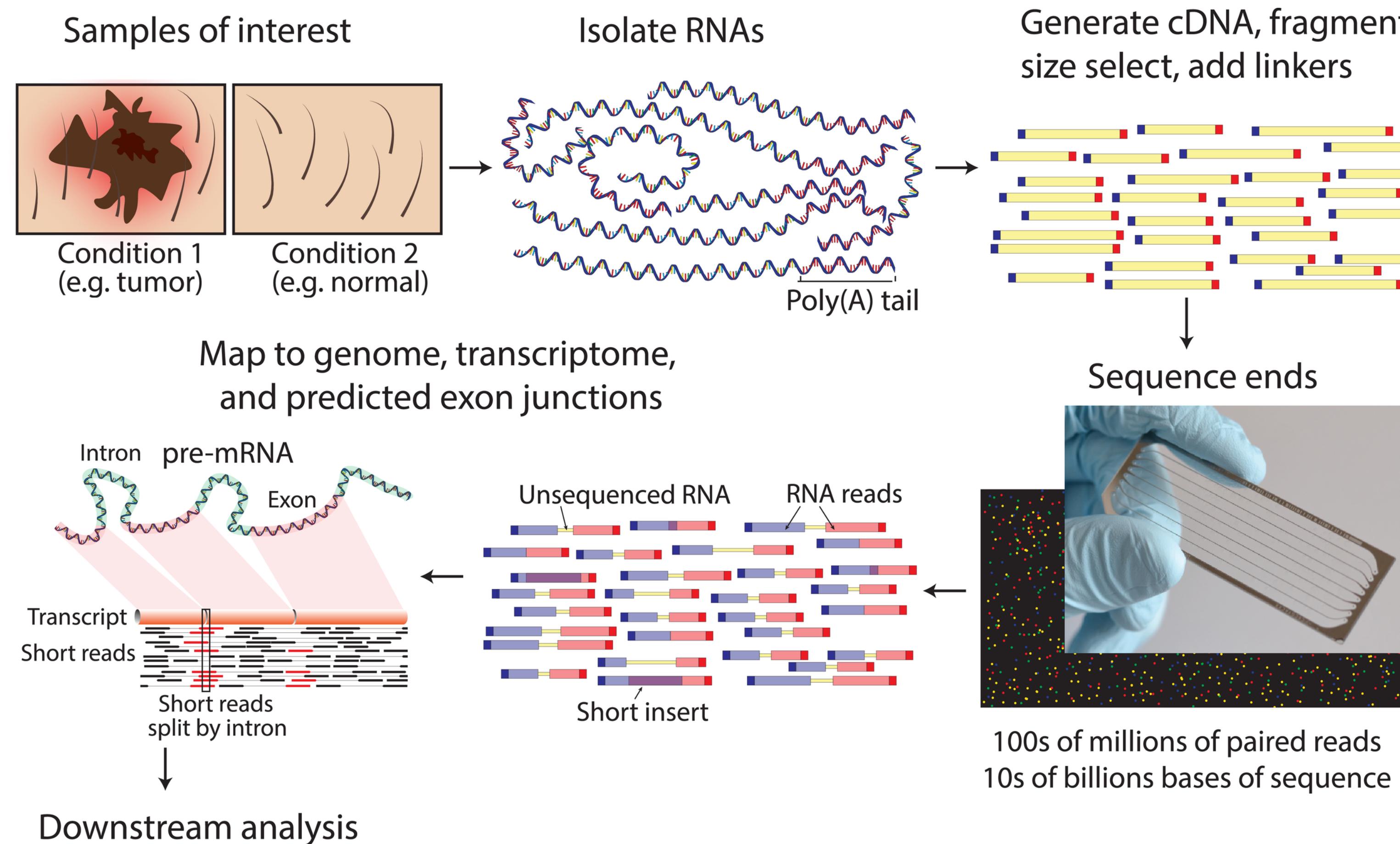
Common analyses with RNA-seq data

Matthew Sutcliffe

2020-06-05

<https://github.com/JanesLab/shared-lab/tree/master/Matt-RNA-seq-analysis-tutorial>

RNA-sequencing



Additional resources

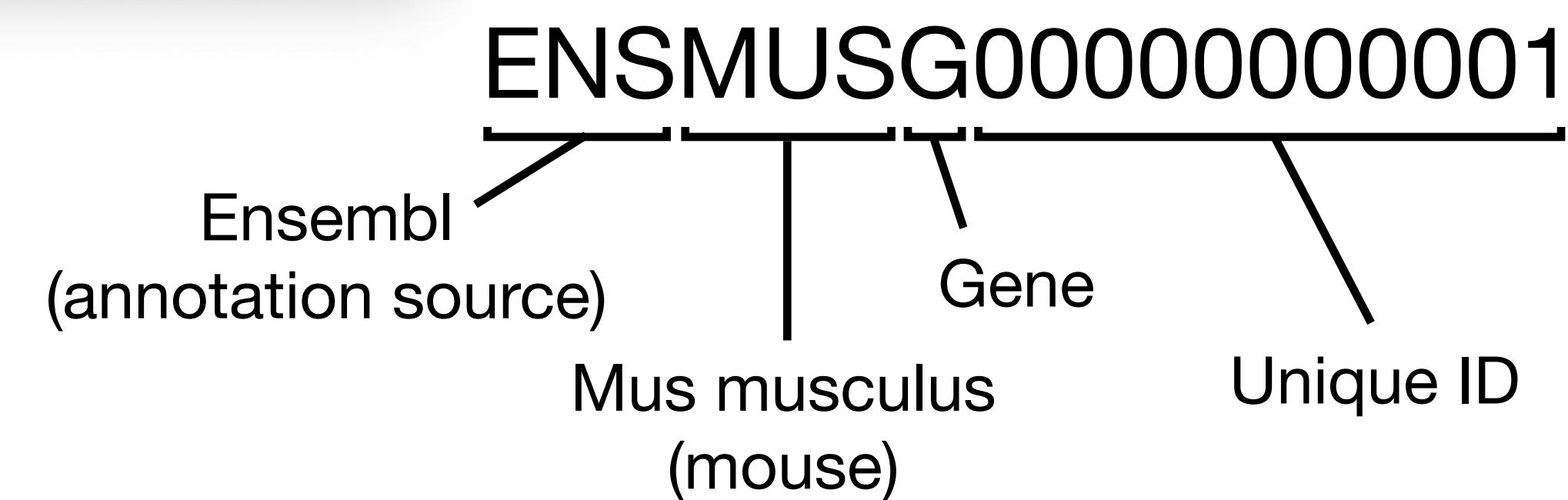
- A brief introduction to RNA-seq at UVA (Shambhavi Singh)
 - [smb://sammas.storage.virginia.edu/JanesLab/Matthew Sutcliffe/Resources/2019_06_05_RNAseqoverview.pptx](smb://sammas.storage.virginia.edu/JanesLab/Matthew%20Sutcliffe/Resources/2019_06_05_RNAseqoverview.pptx)
- RNA-seq alignment tutorial in UNIX
 - <https://github.com/JanesLab/rna-seq-rivanna>
- Video tutorials:
 - Part 1: <https://youtu.be/jsB3fXnSgRQ>
 - Part 2: <https://youtu.be/2QHseQ422I4>
- Intro to Git (Shambhavi Singh) - highly recommended version control system
 - [smb://sammas.storage.virginia.edu/JanesLab/Matthew Sutcliffe/Resources/Sham intro to Git.pdf](smb://sammas.storage.virginia.edu/JanesLab/Matthew%20Sutcliffe/Resources/Sham%20intro%20to%20Git.pdf)
- HarvardX Biomedical Data Science course
 - <http://rafalab.github.io/pages/harvardx.html>

Output of RNA-seq alignments

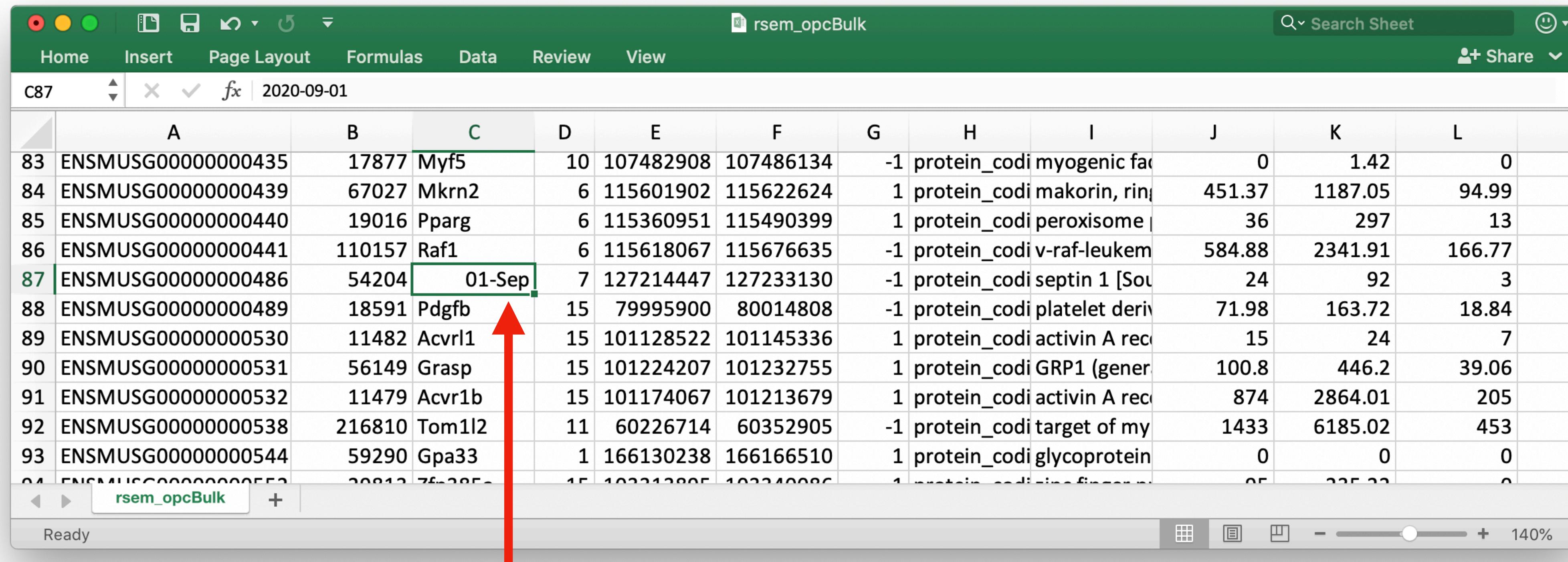
Samples →

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ensgene	entrez	symbol	chr	start	end	strand	biotype	description	opcBulk_12c	opcBulk_12c	opcBulk_12c	opcBulk_12c
2	ENSMUSG000000000001	14679	Gnai3	3	108107280	108146146	-1	protein_codi	guanine nucl	1036	4271	435	760
3	ENSMUSG000000000003	54192	Pbsn	X	77837901	77853623	-1	protein_codi	probasin [So	0	0	0	0
4	ENSMUSG000000000028	12544	Cdc45	16	18780447	18811987	-1	protein_codi	cell division	23	26	6	8
5	ENSMUSG000000000037	107815	Scml2	X	161117193	161258213	1	protein_codi	sex comb on	47	96	6	18
6	ENSMUSG000000000049	11818	Apoh	11	108343354	108414396	1	protein_codi	apolipoprote	2	9	0	1
7	ENSMUSG000000000056	67608	Narf	11	121237253	121255856	1	protein_codi	nuclear prela	309	1237	98	195
8	ENSMUSG000000000058	12390	Cav2	6	17281185	17289115	1	protein_codi	caveolin 2 [S	172	304	29	240
9	ENSMUSG000000000078	23849	Klf6	13	5861482	5870394	1	protein_codi	Kruppel-like	331.01	471	53	307.35
10	ENSMUSG000000000085	29871	Scmh1	4	120405281	120530186	1	protein_codi	sex comb on	648.99	2645.99	216.99	436
11	ENSMUSG000000000088	12858	Cox5a	9	57521274	57532426	1	protein_codi	cytochrome	1481	4013	299.99	997
12	ENSMUSG000000000093	21385	Tbx2	11	85832551	85841948	1	protein_codi	T-box 2 [Sou	5	18	3	5
13	ENSMUSG000000000094	21387	Tbx4	11	85886422	85916097	1	protein_codi	T-box 4 [Sou	4	20.07	1	1
14	ENSMUSG000000000103	22768	Zfy2	Y	2106015	2170409	-1	protein_codi	zinc finger pi	0	0	0	0
15	ENSMUSG000000000120	18053	Ngfr	11	95568818	95587735	-1	protein_codi	nerve growthl	13.07	16.83	13.07	5.65
16	ENSMUSG000000000125	22415	Wnt3	11	103774150	103817957	1	protein_codi	wingless-typ	26	79	5	10
17	ENSMUSG000000000126	216795	Wnt9a	11	59306928	59333552	1	protein_codi	wingless-typ	22	27	6	20
18	ENSMUSG000000000127	14158	Fer	17	63896018	64139494	1	protein_codi	fer (fms/fps	151.75	666.13	63.51	71.36
19	ENSMUSG000000000131	74204	Xpo6	7	126101715	126200501	-1	protein_codi	exportin 6 [S	785.84	2887.91	213.8	564.09
20	ENSMUSG000000000134	209446	Tfe3	X	7762560	7775202	1	protein_codi	transcription	345.99	784.99	62.01	280
21	ENSMUSG000000000142	12006	Axin2	11	108920349	108950783	1	protein_codi	axin 2 [Sourc	277	670	85	165
22	ENSMUSG000000000148	231841	Brat1	5	140705011	140719379	1	protein_codi	BRCA1-assoc	72	182	15	24
23	ENSMUSG000000000149	14673	Gna12	5	140758408	140830431	-1	protein_codi	guanine nucl	547.05	708	81.46	317
24	ENSMUSG000000000154	18400	Slc22a18	7	143473736	143499334	1	protein_codi	solute carrie	3	12	0	1
25	ENSMUSG000000000157	16415	Itgb21	16	96422288	96443619	-1	protein_codi	integrin beta	0	0	0	0
26	ENSMUSG000000000159	72058	Igsvf5	16	96361668	96525580	1	protein_codi	immunoglob	8.32	3.6	0	0
27	ENSMUSG000000000167	72611	Rib1d2	9	50617221	50625000	1	protein_codi	DIH1 domain	5	28	4	11

~34,000 rows (genes)

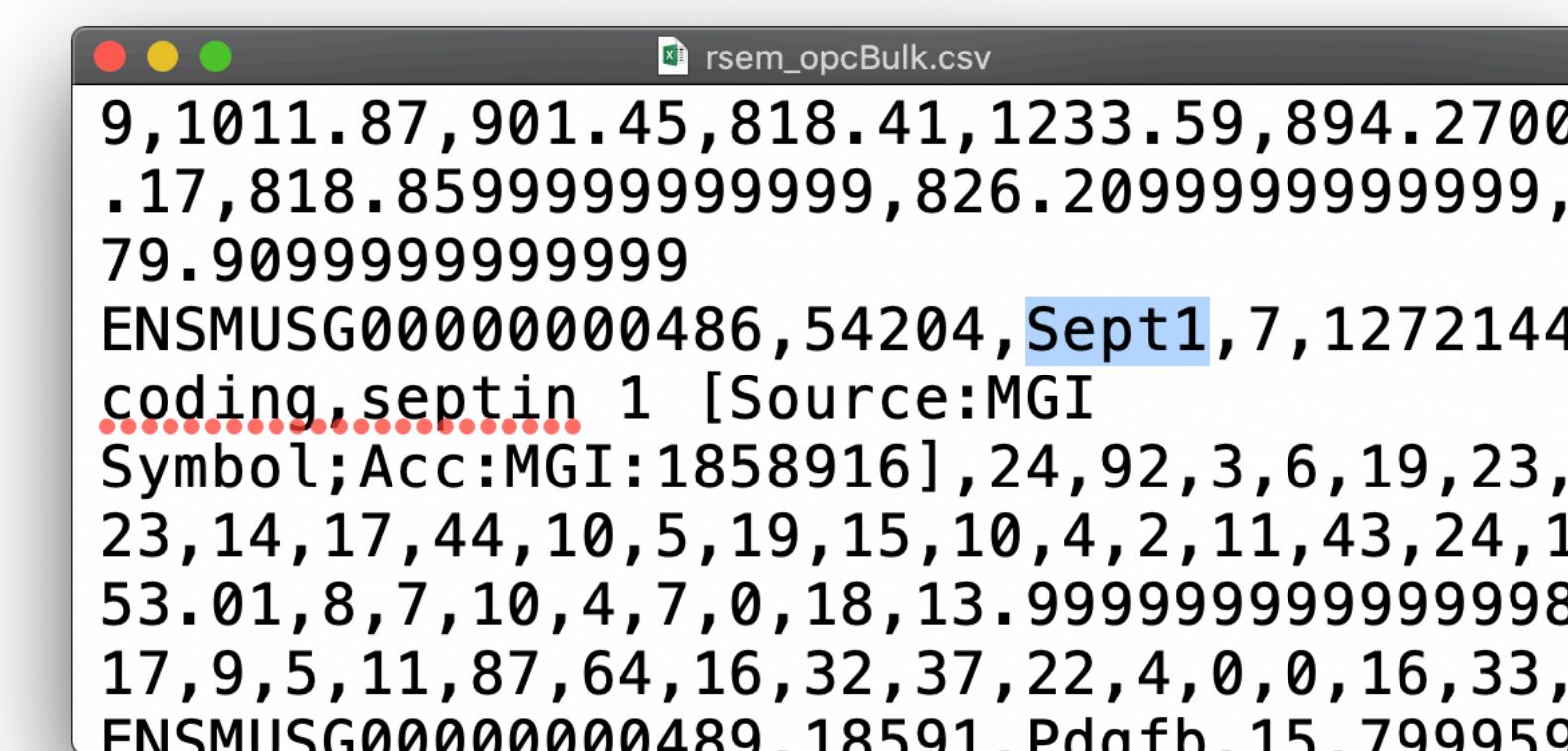


(One of the many reasons) why we don't open these files in Excel



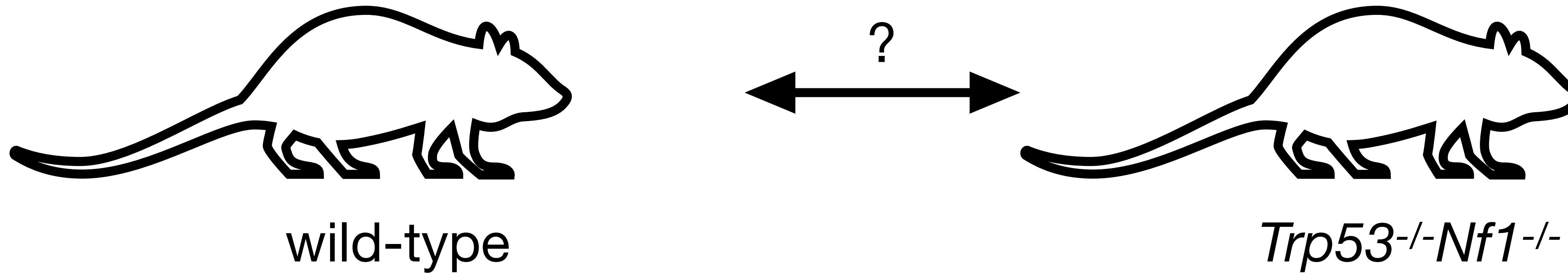
A screenshot of Microsoft Excel showing a data table titled "rsem_opcBulk". The table has columns labeled A through M. Row 87 contains the date "01-Sep" in cell C. A red arrow points from the text "date entry" in the adjacent image to this cell. The Excel ribbon at the top shows tabs for Home, Insert, Page Layout, Formulas, Data, Review, and View. The status bar at the bottom indicates "Ready".

	A	B	C	D	E	F	G	H	I	J	K	L	M
83	ENSMUSG000000000435	17877	Myf5	10	107482908	107486134	-1	protein_codi	myogenic fac	0	1.42	0	
84	ENSMUSG000000000439	67027	Mkrn2	6	115601902	115622624	1	protein_codi	makorin, ring	451.37	1187.05	94.99	28
85	ENSMUSG000000000440	19016	Pparg	6	115360951	115490399	1	protein_codi	peroxisome	36	297	13	
86	ENSMUSG000000000441	110157	Raf1	6	115618067	115676635	-1	protein_codi	v-raf-leukem	584.88	2341.91	166.77	31
87	ENSMUSG000000000486	54204	01-Sep	7	127214447	127233130	-1	protein_codi	septin 1 [Sou	24	92	3	
88	ENSMUSG000000000489	18591	Pdgfb	15	79995900	80014808	-1	protein_codi	platelet deriv	71.98	163.72	18.84	
89	ENSMUSG000000000530	11482	Acvrl1	15	101128522	101145336	1	protein_codi	activin A rece	15	24	7	
90	ENSMUSG000000000531	56149	Grasp	15	101224207	101232755	1	protein_codi	GRP1 (gener	100.8	446.2	39.06	6
91	ENSMUSG000000000532	11479	Acvr1b	15	101174067	101213679	1	protein_codi	activin A rece	874	2864.01	205	
92	ENSMUSG000000000538	216810	Tom1l2	11	60226714	60352905	-1	protein_codi	target of my	1433	6185.02	453	
93	ENSMUSG000000000544	59290	Gpa33	1	166130238	166166510	1	protein_codi	glycoprotein	0	0	0	
94	ENSMUSG000000000552	20812	76-285-	15	102212805	102240086	1	protein_codi	adiponectin	25	225.22	0	



A screenshot of a CSV file titled "rsem_opcBulk.csv". The file contains a single row of data: "9,1011.87,901.45,818.41,1233.59,894.2700 .17,818.859999999999,826.209999999999, 79.909999999999 ENSMUSG0000000486,54204,Sept1,7,1272144 coding,septin 1 [Source:MG Symbol;Acc:MG:1858916],24,92,3,6,19,23, 23,14,17,44,10,5,19,15,10,4,2,11,43,24,1 53.01,8,7,10,4,7,0,18,13.9999999999998 17,9,5,11,87,64,16,32,37,22,4,0,0,16,33, ENSMUSG0000000489,18591,Pdafh,15,799959

Differential expression analysis with DESeq2



Inputs to DESeq2

Counts matrix

RStudio Source Editor

	opcBulk_150dpi_WT_21590_tdTpositive	opcBulk_150dpi_WT_21591_tdTpositive	opcBulk_150dpi_WT_21592_tdTpositive	opcBulk_150dpi_WT_21593_tdTpositive
Gna13	624	571	306	1286
Pbsn	0	0	0	0
Cdc45	10	5	0	15
Scml2	43	60	79	64
Apoh	1	0	0	1
Narf	128	121	51	458
Cav2	160	433	201	295
Klf6	146	166	51	412
Scmh1	242	255	135	1004
Cox5a	842	372	285	1274
Tbx2	4	1	5	27
Tbx4	1	4	0	2
Zfy2	0	0	0	0
Ngfr	7	3	0	30
Wnt3	14	6	0	15
Wntqa2	4	2	1	0

Showing 1 to 16 of 34,623 entries, 13 total columns

Sample description

	genotype
opcBulk_150dpi_WT_21590_tdTpositive	WT
opcBulk_150dpi_WT_21591_tdTpositive	WT
opcBulk_150dpi_WT_21592_tdTpositive	WT
opcBulk_150dpi_WT_21593_tdTpositive	WT
opcBulk_150dpi_WT_22062_tdTpositive	WT
opcBulk_150dpi_WT_22064_tdTpositive	WT
opcBulk_150dpi_Tumor_16713_tdTpositive	Tumor
opcBulk_150dpi_Tumor_16839_tdTpositive	Tumor
opcBulk_150dpi_Tumor_16840_tdTpositive	Tumor
opcBulk_150dpi_Tumor_16844_tdTpositive	Tumor
opcBulk_150dpi_Tumor_21256_tdTpositive	Tumor
opcBulk_150dpi_Tumor_21349_tdTpositive	Tumor
opcBulk_150dpi_Tumor_21354_tdTpositive	Tumor

Resources from the creators of DESeq2

A screenshot of a web browser window displaying a PDF document titled "Beginner's guide to using the DESeq2 package". The browser has a standard OS X interface with red, yellow, and green window control buttons. The address bar shows the URL: <https://bioc.ism.ac.jp/packages/2.14/bioc/vignettes/DESeq2/inst/doc/beginner.pdf>. The PDF content includes the title, authors (Michael Love^{1*}, Simon Anders², Wolfgang Huber²), and affiliations (¹ Department of Biostatistics, Dana Farber Cancer Institute and ² European Molecular Biology Observatory). A sidebar on the left contains a summary of the vignette's purpose and a table of contents. The main body of the PDF is visible, showing the detailed table of contents.

This vignette describes how to compare two or more groups of samples between conditions using the DESeq2 pipeline. It starts by reading count matrices from raw sequencing data, and then moves on to analyzing RNA-Seq or ChIP-Seq data. The vignette provides a step-by-step guide to each of these steps, explaining each step in detail and illustrating it with examples.

Beginner's guide to using the DESeq2 package

Michael Love^{1*}, Simon Anders², Wolfgang Huber²

¹ Department of Biostatistics, Dana Farber Cancer Institute and
² European Molecular Biology Observatory

Beginner's guide to using the DESeq2 package

2

Contents

1 Introduction	2
2 Input data	2
2.1 Preparing count matrices	3
2.2 Aligning reads to a reference	3
2.3 Counting reads in genes	4
2.4 Experiment data	7
2.4.1 The DESeqDataSet, column metadata, and the design formula	7
2.4.2 Starting from SummarizedExperiment	8
2.4.3 Starting from count tables	10
2.5 Collapsing technical replicates	12
3 Running the DESeq2 pipeline	13
3.1 Preparing the data object for the analysis of interest	13
3.2 Running the pipeline	14
3.3 Inspecting the results table	15
3.4 Other comparisons	16
3.5 Multiple testing	17
3.6 Diagnostic plots	19
4 Independent filtering	22

<https://bioc.ism.ac.jp/packages/2.14/bioc/vignettes/DESeq2/inst/doc/beginner.pdf>

In case you get stuck somewhere...

 Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

ASK QUESTION LATEST 200 NEWS 9 JOBS TUTORIALS

Question: Rescaling predictors in DESeq changes outcomes?

 When I'm running a GLM I will sometimes need to rescale predictors (e.g. take z-scores) to make the coefficients of different predictors easier to compare to each other. My understanding is that this doesn't and shouldn't change anything about the model fit except the scale of the coefficient. When I do this in DESeq2, I find that it changes not only the log2FC estimates, but also the p-values (before adjustment). I don't understand how or why that would be happening, or what to do about it.

 Does anyone understand this? Is it expected behaviour? What should I do about it?

Thank you!

deseq2 glm rescale • 49 views

ADD COMMENT • link • Not following ▾   modified 5 days ago by Michael Love ♦ 28k • written 5 days ago by jc.szamosi • 10

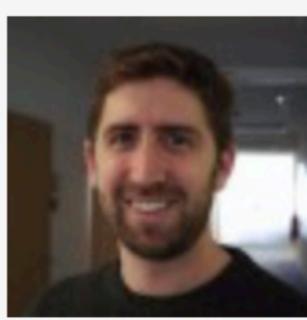
Answer: Rescaling predictors in DESeq changes outcomes?

 Having predictors with very different scales can make the fitting numerically unstable, which is known in linear modeling, and it's typically recommended to not use a design matrix with very different scale of predictors.

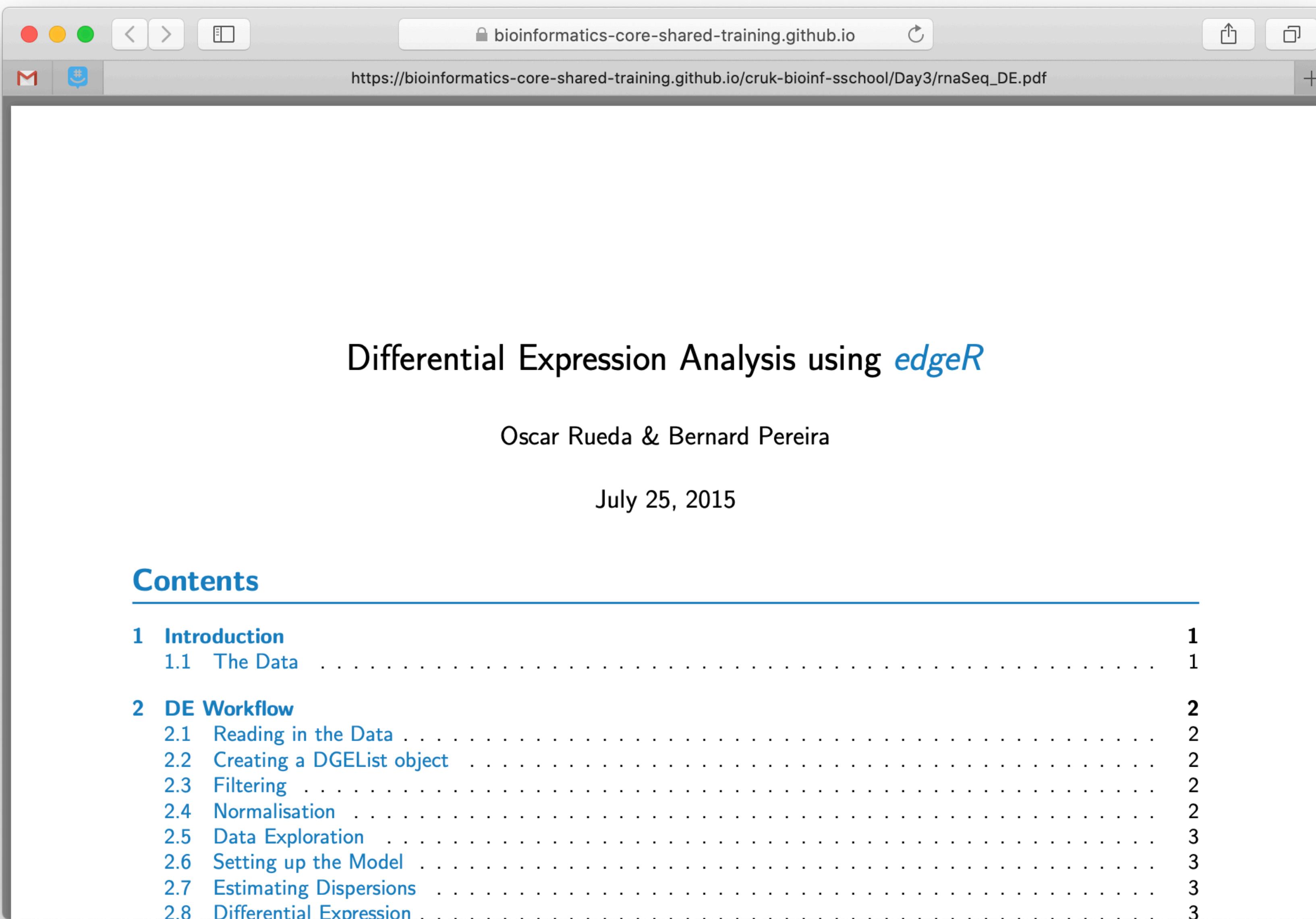
 0 Because I noticed some users putting in very small and very large values in the design, DESeq2 gives a message to scale the predictors first since v1.26 (so current and previous release give this message).

ADD COMMENT • link

written 5 days ago by Michael Love ♦ 28k

 5 days ago by Michael Love ♦ 28k
United States

Alternative to DESeq2: edgeR

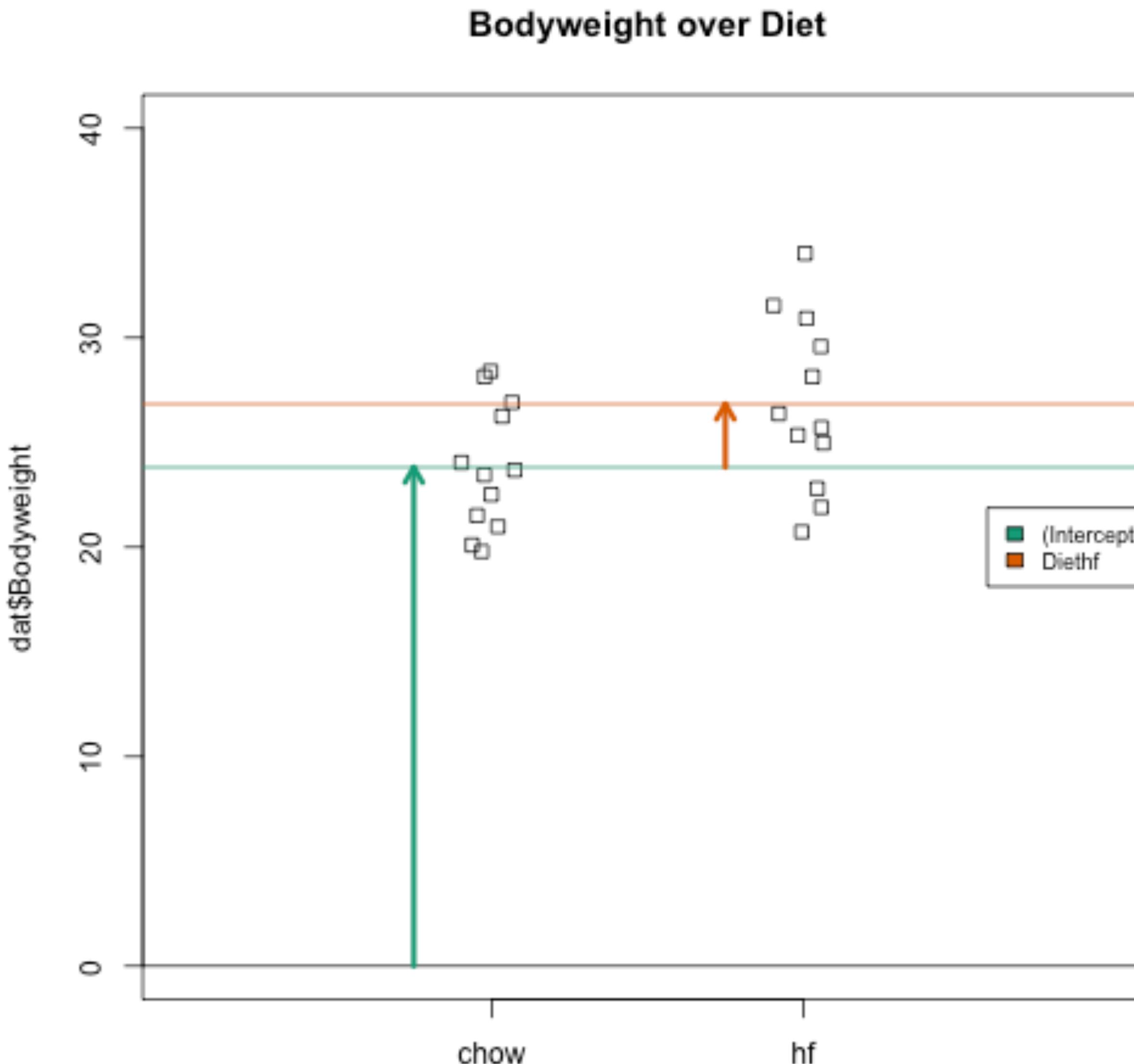


A screenshot of a web browser window displaying a PDF document titled "Differential Expression Analysis using *edgeR*". The browser has a standard OS X interface with a menu bar and a toolbar. The address bar shows the URL: https://bioinformatics-core-shared-training.github.io/cruk-bioinf-sschool/Day3/rnaSeq_DE.pdf. The PDF content includes the title, authors (Oscar Rueda & Bernard Pereira), and date (July 25, 2015). Below this is a "Contents" section with a table of contents.

	1
1 Introduction	1
1.1 The Data	1
2 DE Workflow	2
2.1 Reading in the Data	2
2.2 Creating a DGEList object	2
2.3 Filtering	2
2.4 Normalisation	2
2.5 Data Exploration	3
2.6 Setting up the Model	3
2.7 Estimating Dispersions	3
2.8 Differential Expression	3

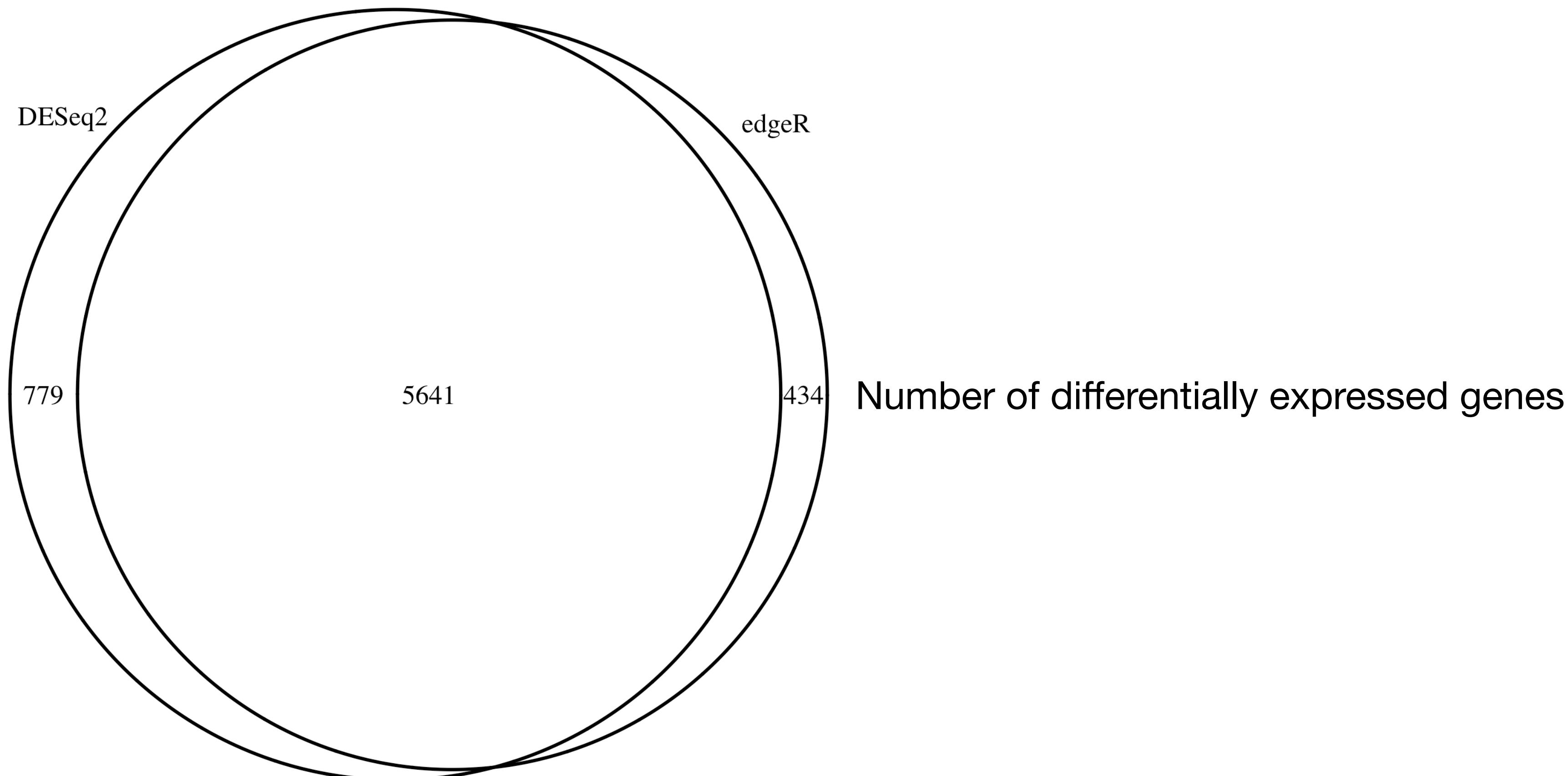
<https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf>

Creating a study design matrix for edgeR



	(Intercept)	dataset_info\$genotypeWT
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	0
8	1	0
9	1	0
10	1	0
11	1	0
12	1	0
13	1	0

DESeq2 and edgeR produce similar results

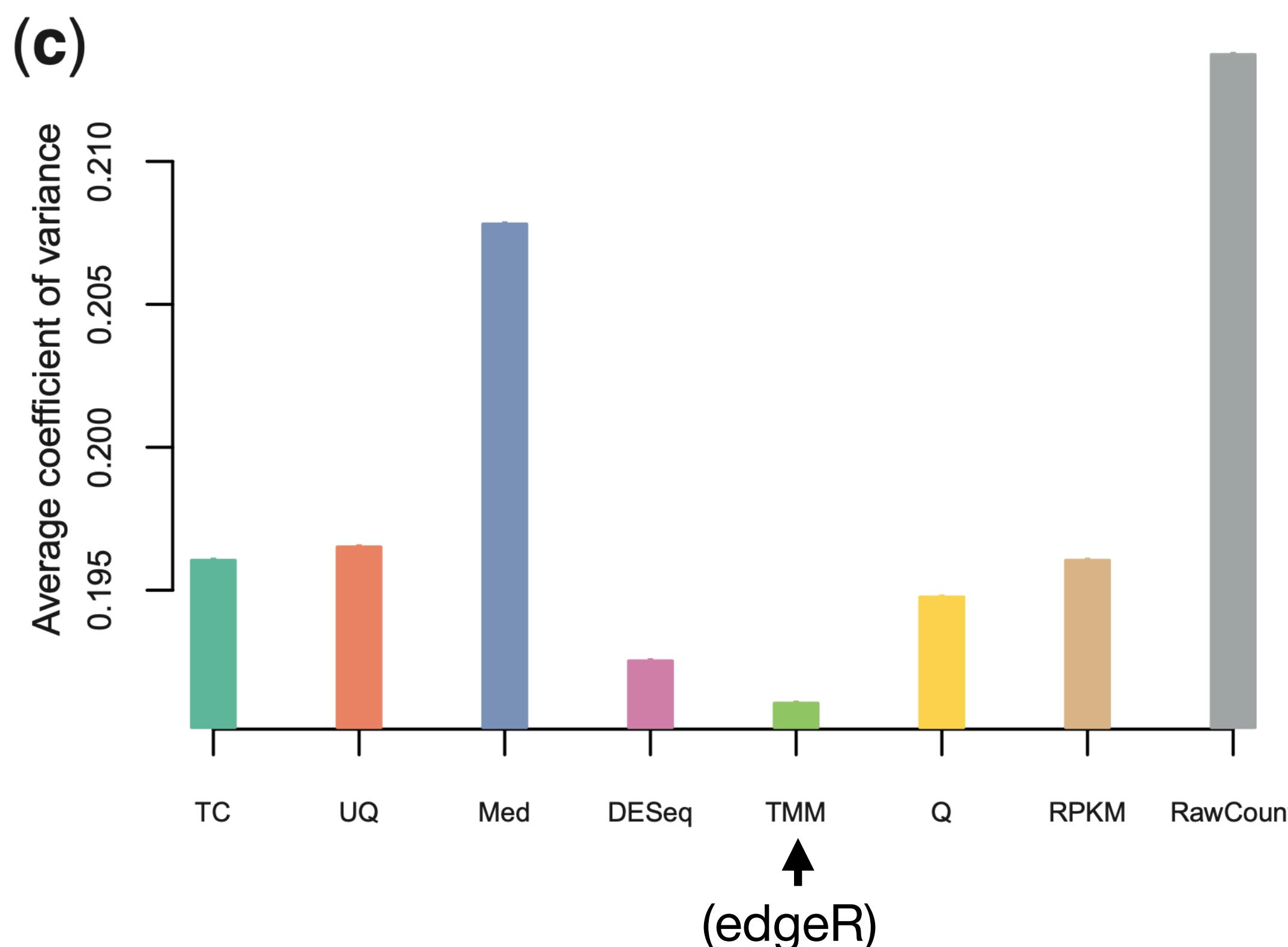


DESeq2 vs. edgeR vs. other: which should I use?

- Either (or both)
 - Using the intersection of multiple methods could produce a higher confidence DE gene list
- There is some evidence that edgeR's normalization method is superior to that of DESeq*



<https://xkcd.com/927/>



- DESeq: median ratio of geometric mean of all genes
- edgeR: “ratio of RNA production uses a weighted trimmed mean of the log expression ratios (trimmed mean of M values (TMM))”

Gene set enrichment analysis (GSEA)

The screenshot shows a web browser window for the GSEA | MSigDB | Investigate Gene Sets page at gsea-msigdb.org. The page has a dark blue header with the GSEA logo and navigation links for Home, Downloads, Molecular Signatures Database (which is active), Documentation, Contact, and Team. A sidebar on the left contains links for MSigDB Home, About Collections, Browse Gene Sets, Search Gene Sets, Investigate Gene Sets (which is active), View Gene Families, and Help. The main content area is titled "Investigate Gene Sets" and features logos for UC San Diego and the Broad Institute. It provides instructions for gaining further insight into gene sets through tools like computing overlaps, displaying expression profiles, categorizing members by gene families, and investigating in NDEx. A citation note and a "show gene families" button are also present. The right side lists compendia expression profiles and the NDEx Biological Network Repository.

GSEA | MSigDB | Investigate Gene Sets

logged in as mds5cg@virginia.edu
logout

GSEA Home Downloads Molecular Signatures Database Documentation Contact Team

► MSigDB Home
► About Collections
► Browse Gene Sets
► Search Gene Sets
► Investigate Gene Sets (active)
► View Gene Families
► Help

Investigate Gene Sets

UC San Diego BROAD INSTITUTE

Gain further insight into the biology behind a gene set by using the following tools:

- ▶ **compute overlaps** with other gene sets in MSigDB ([more...](#))
- ▶ **display the gene set expression profile** based on a selected compendium of expression data ([more...](#))
- ▶ **categorize** members of the gene set by gene families ([more...](#))
- ▶ further investigate the gene set in the online **biological network repository NDEx** ([more...](#))

To cite your use of this page, please reference this website and also the MSigDB (see [Citing MSigDB](#) on the main MSigDB page)

Input Gene Identifiers
(case sensitive)

Compute Overlaps
[about the MSigDB collections]

- H: hallmark gene sets
- C1: positional gene sets
- C2: curated gene sets
 - CGP: chemical and genetic perturbations
 - CP: Canonical pathways
 - CP:BIOCARTA: BioCarta gene sets
 - CP:KEGG: KEGG gene sets
 - CP:PID: PID gene sets
 - CP:REACTOME: Reactome gene sets
- C3: regulatory target gene sets
 - MIR: microRNA targets
 - MIR:MIR_Legacy: Legacy microRNA targets
 - MIR:MIRDB: MIRDB microRNA targets

Compendia Expression Profiles

- GTEx compendium
- Human tissue compendium (Novartis)
- Global Cancer Map (Broad Institute)
- NCI-60 cell lines (National Cancer Institute)

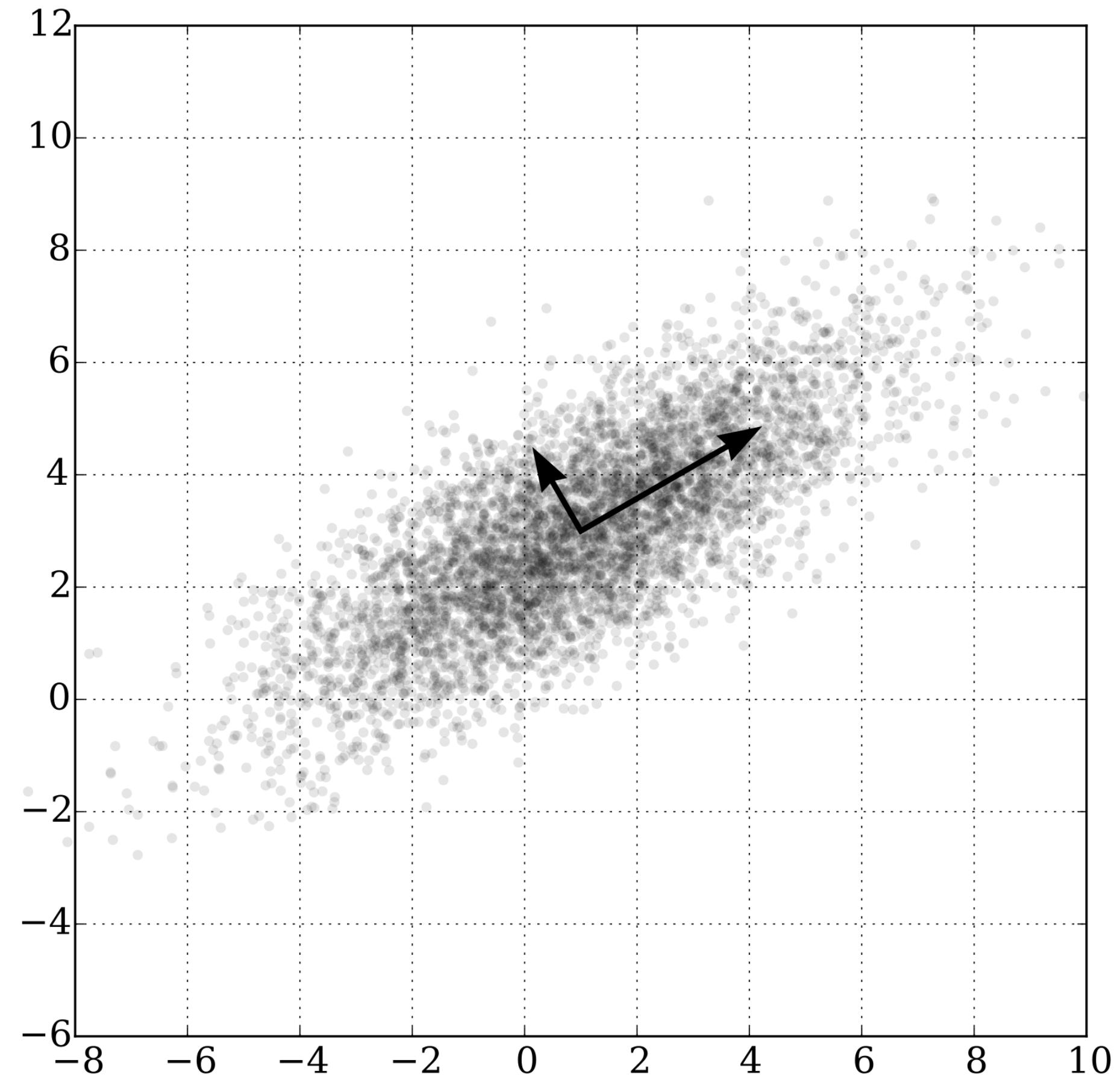
[display expression profile](#)

Gene Families
[show gene families](#)

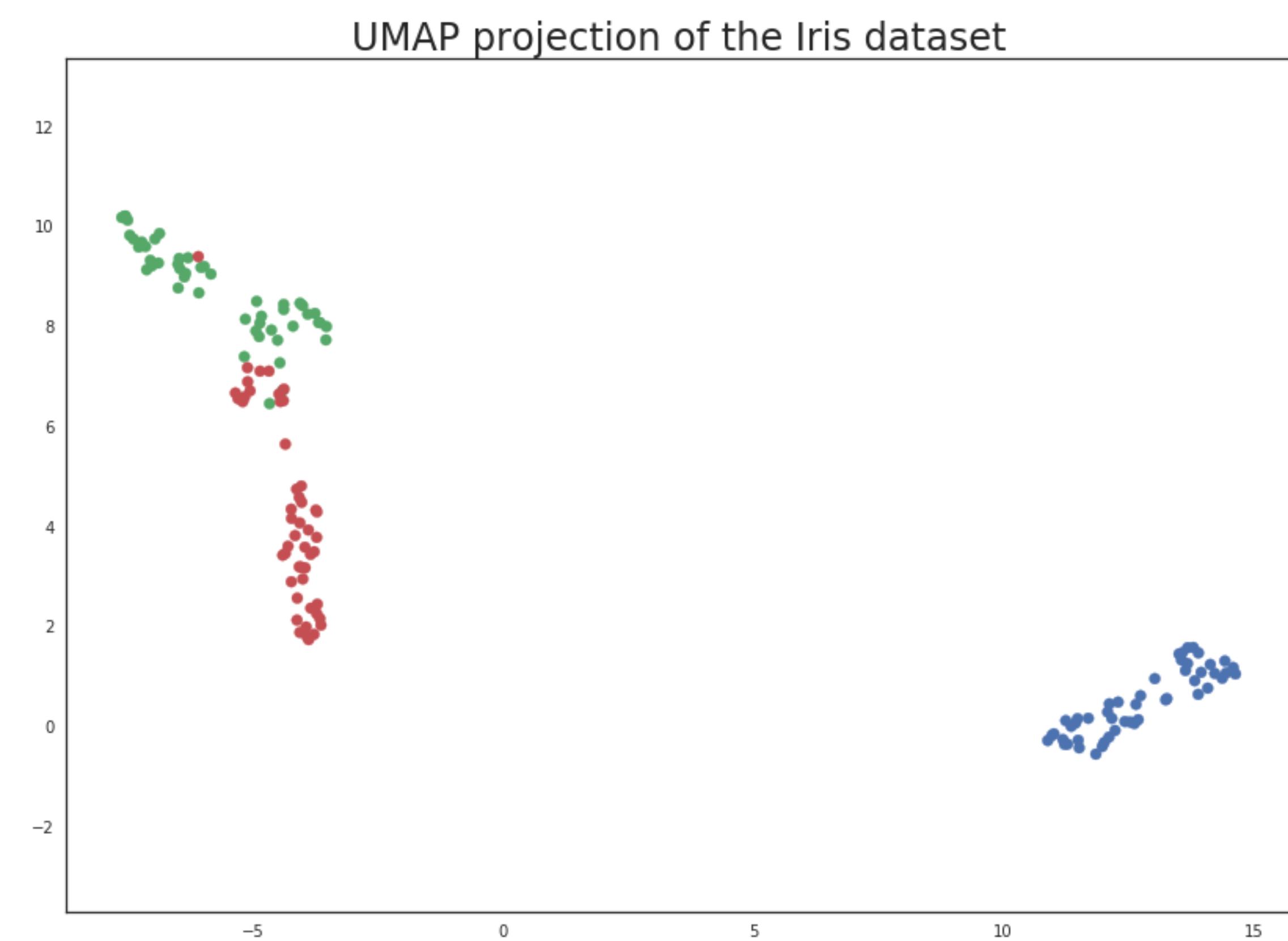
NDEx Biological Network Repository

PCA and UMAP

PCA: linear dimensionality reduction



UMAP: nonlinear dimensionality reduction



When googling problems in R:

- Websites where you will commonly find answers
 - Stack Overflow
 - Bioconductor forums
 - R-bloggers