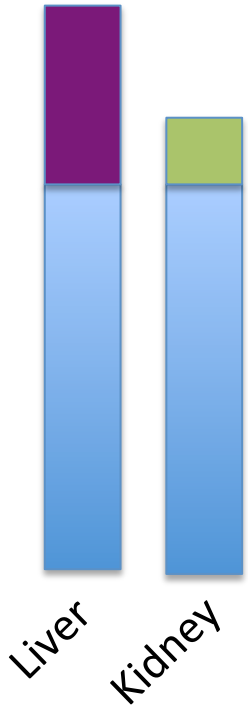
Comparing RNA-Seq Samples

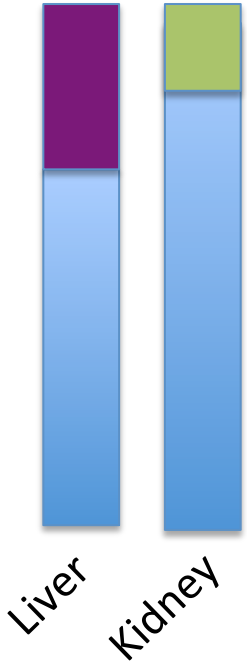
Some Cross-sample Normalization May Be Required

Why cross-sample normalization is important

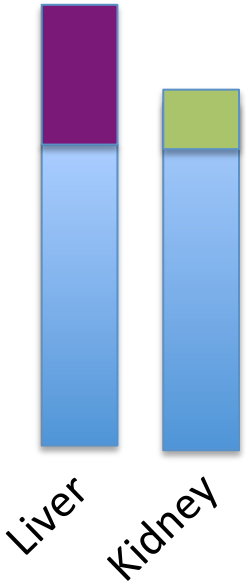
Absolute RNA quantities per cell



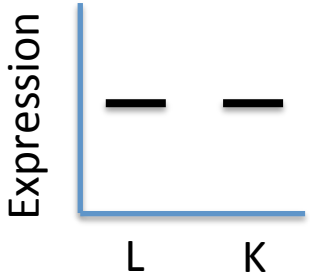
Measured relative abundance via RNA-Seq

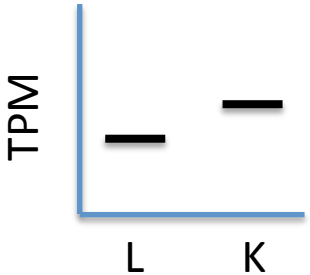


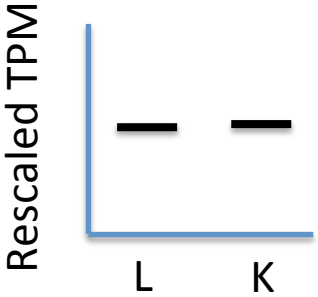
Cross-sample normalized (rescaled) relative abundance



eg. Some housekeeping gene's expression level:







Cross-sample Normalization Required
Otherwise, housekeeping genes look diff expressed due to sample composition differences

Subset of genes highly expressed in liver

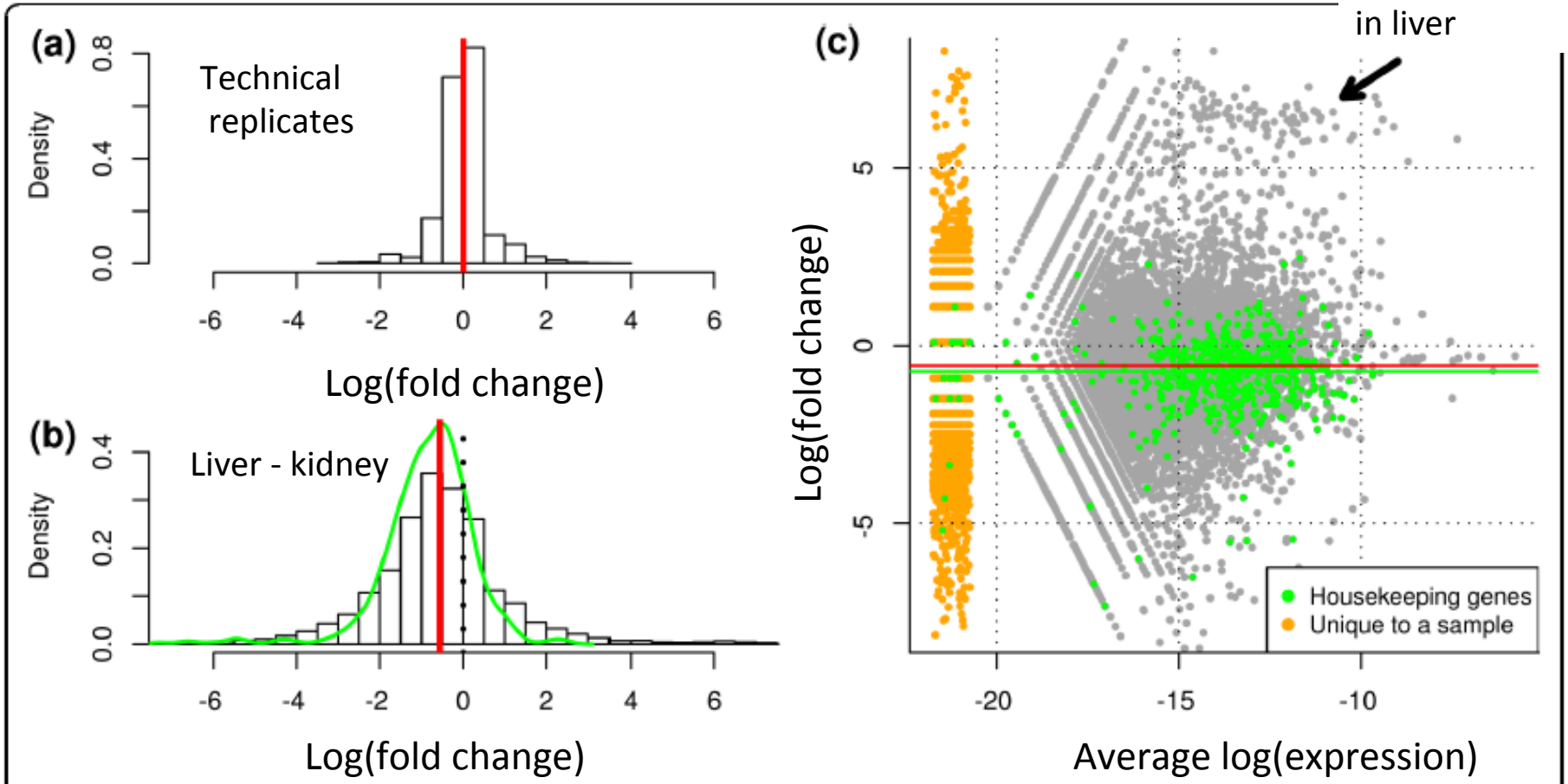
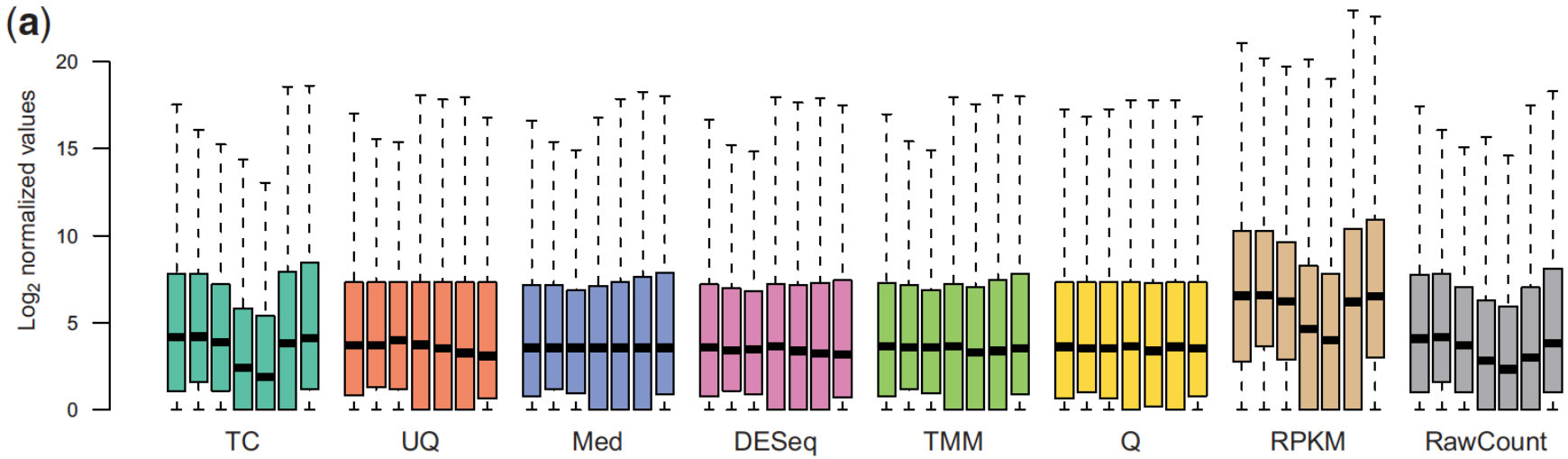


Figure 1 Normalization is required for RNA-seq data. Data from [6] comparing log ratios of **(a)** technical replicates and **(b)** liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of the housekeeping genes. **(c)** An M versus A plot comparing liver and kidney shows a clear offset from zero. Green points indicate 545 housekeeping genes, while the green line signifies the median log-ratio of the housekeeping genes. The red line shows the estimated TMM normalization factor. The smear of orange points highlights the genes that were observed in only one of the liver or kidney

the overall bias in log-fold-changes.

Adapted from: Robinson and Oshlack, Genome Biology, 2010

Normalization methods for Illumina high-throughput RNA sequencing data analysis.



From "A comprehensive evaluation of normalization methods for Illumina high throughput RNA sequencing data analysis" Brief Bioinform. 2013 Nov;14(6):671-83
http://www.ncbi.nlm.nih.gov/pubmed/22988256

Differential Expression Analysis

Thx, Charlotte Soneson! ☺

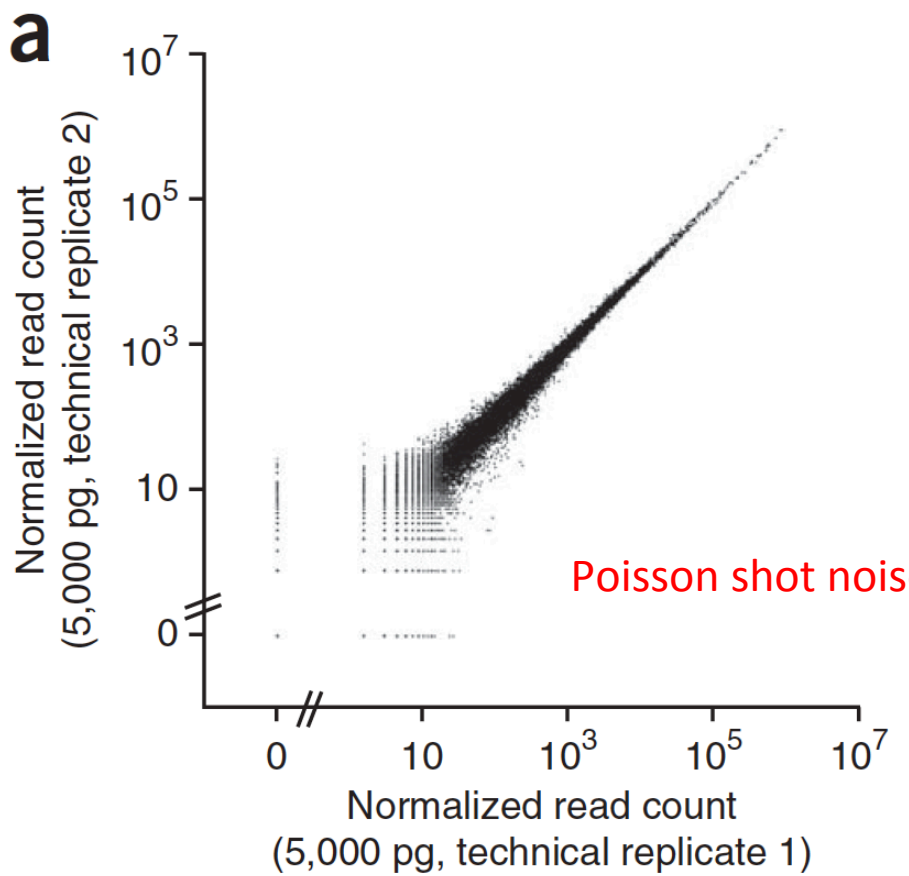
Differential Expression Analysis Involves

- Counting reads mapped to features
- Statistical significance testing

Beware of small counts leading to notable fold changes

	Sample_A	**Sample_B**	**Fold_Change**	**Significant?**
Gene A	1	2	2-fold	No
Gene B	100	200	2-fold	Yes

Variation Observed Between Technical Replicates



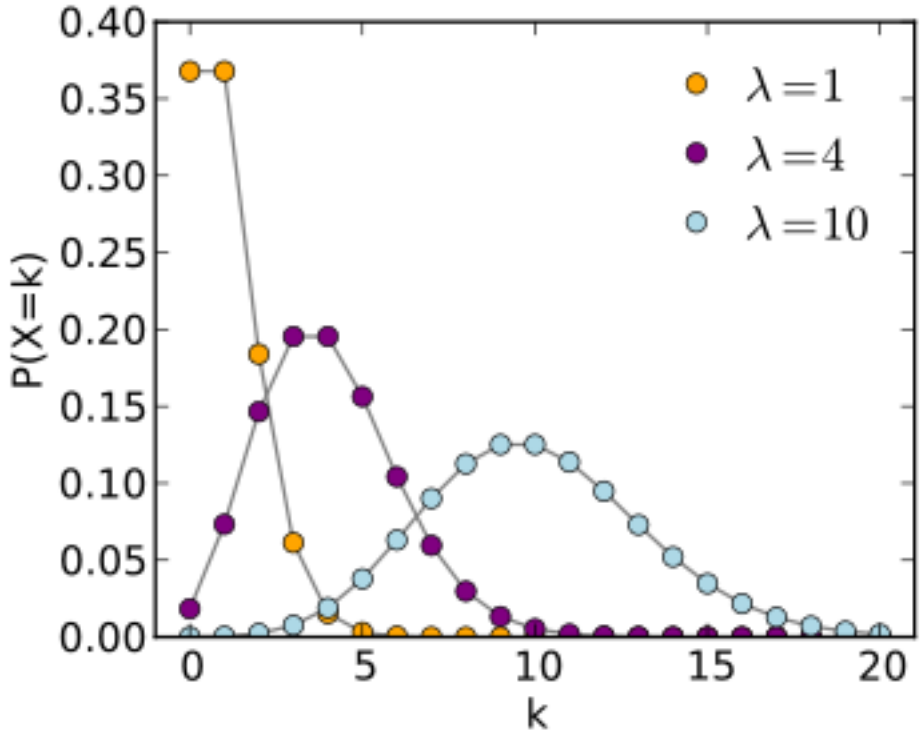
Variation observed is well described by models of random sampling (Poisson Distribution)

Poisson shot noise is high for small counts.

* plot from Brennecke, et al. Nature Methods, 2013

Observed RNA-Seq Counts Result from Random Sampling of the Population of Reads

Technical variation in RNA-Seq counts per feature is well modeled by the Poisson distribution



Mean # fragments

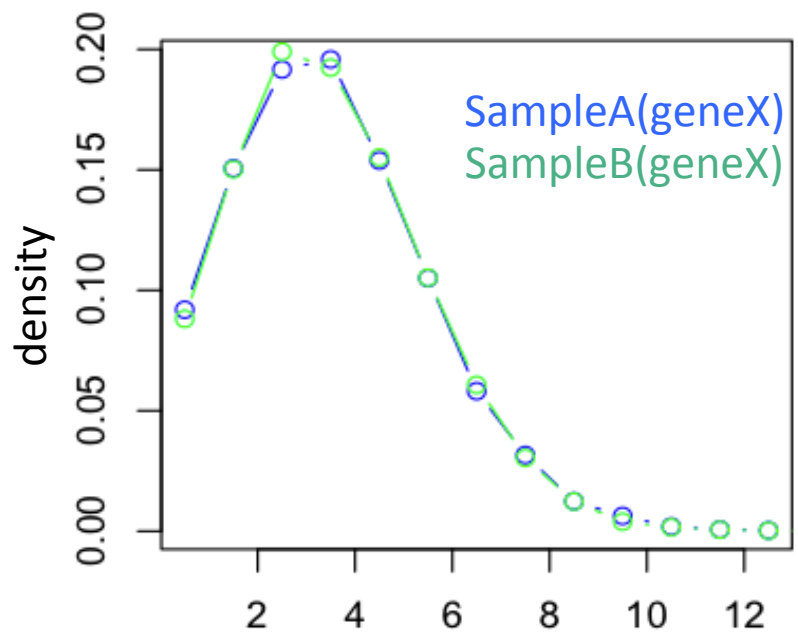
(observed read counts)

See: http://en.wikipedia.org/wiki/Poisson_distribution

Example: One gene*not* differentially expressed

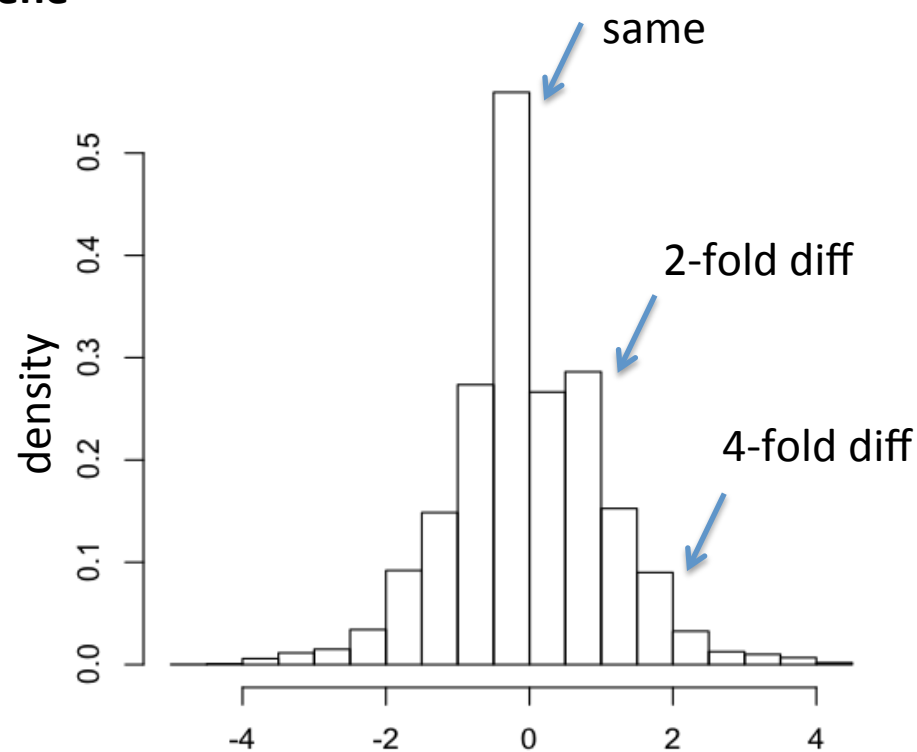
Example: SampleA(gene) = SampleB(gene) = 4 reads

Distribution of observed counts for single gene (under Poisson model)



(k) number of reads observed

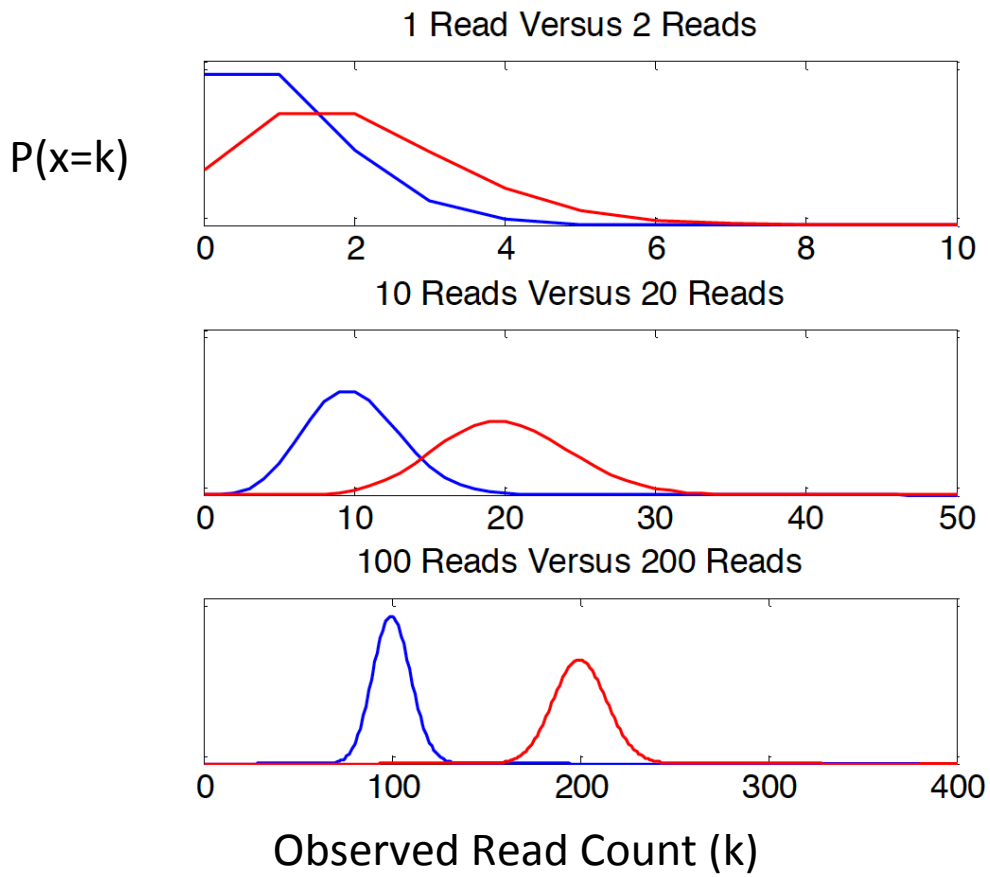
Dist. of $\log_2$(fold change) values



$x = \log_2(\text{SampleA}/\text{SampleB})$

Sequencing Depth Matters

Poisson distributions for counts based on **2-fold** expression differences



No confidence in 2-fold difference. Likely observed by chance.

High confidence in 2-fold difference. Unlikely observed by chance.

From: http://gkno2.tumblr.com/post/24629975632/thinking-about-rna-seq-experimental-design-for
and from supplementary text of Busby et al., Bioinformatics, 2013

Greater Depth = More Statistical Power

Example: Single gene, reads sampled at different sequencing depths

Reads per sample	Sample A Number of reads	Sample B Number of reads	P-value (Fishers Exact Test)
100,000	1	2	1
1,000,000	10	20	0.099
10,000,000	100	200	**8.0e-09**

Technical vs. Biological Replicates

RNA-Seq Technical replicates aren't essential

(Technical variation is well-modeled by the Poisson distribution)

"We find that the Illumina sequencing data are highly replicable, with relatively little technical variation, and thus, for many purposes, it may suffice **to sequence each mRNA sample only once**" *Marioni et al., Genome Research, 2008*

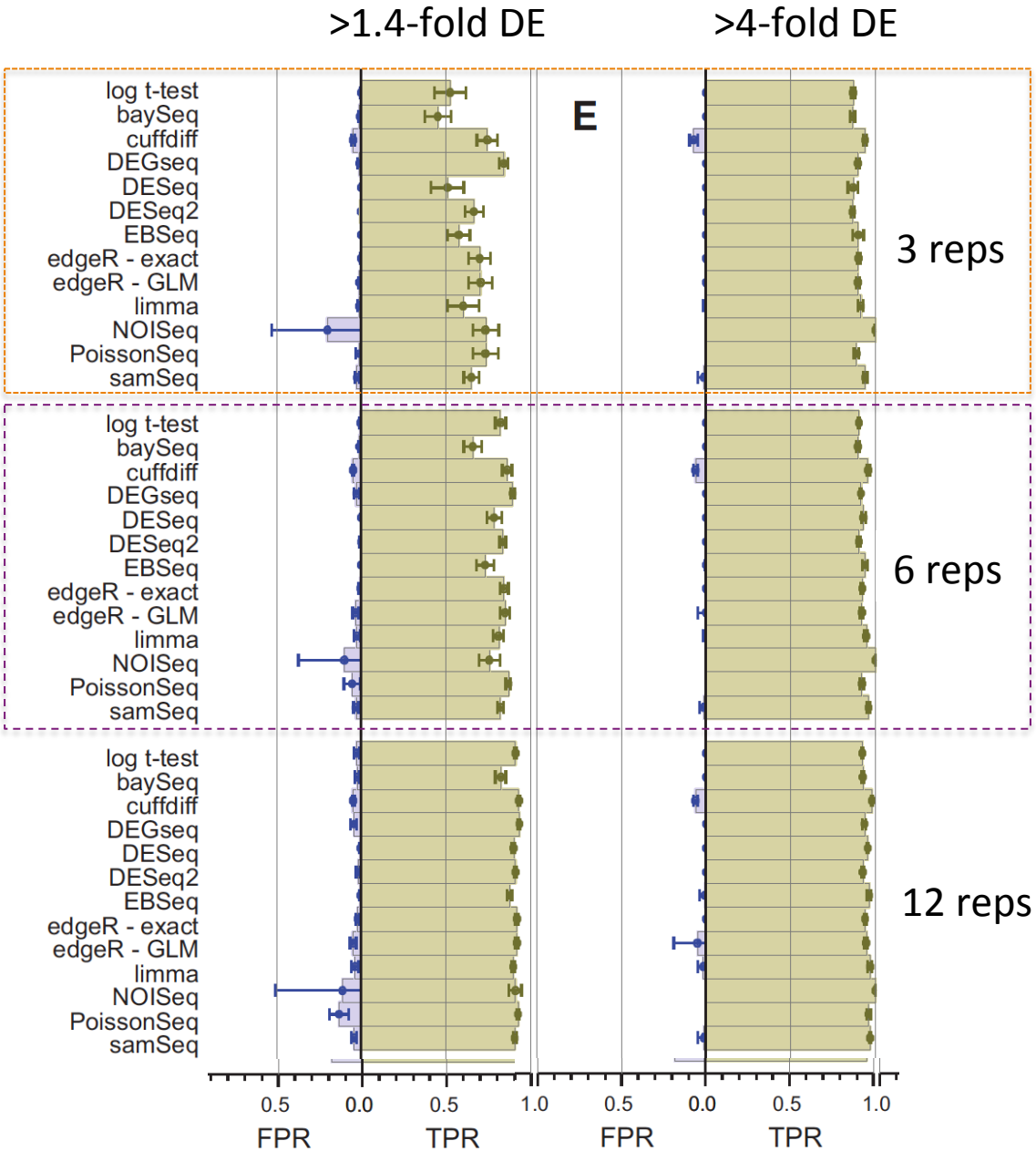
However, biological replicates *ARE* essential

total_variance = technical_variance + biological_variance

(Total variance well-modeled by negative binomial distribution)

"… **at least six biological replicates should be used**, rising to at least 12 when it is important to identify SDE genes for all fold changes." *Schurch et al., RNA, 2016*

DE Accuracy Improves with Higher Biological Replication



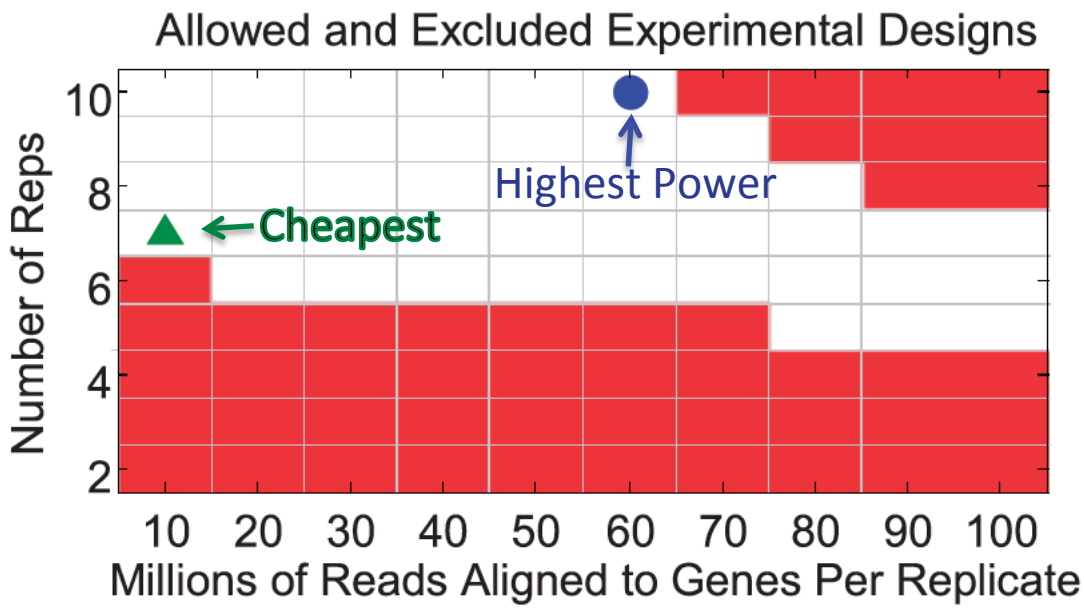
At a minimum, do 3 bio replicates

Respectable goal

*Figure taken and adapted from Schurch et al., RNA, 2016

Planning Experiments:
How many reads and how many replicates?

Input: max total reads, max total replicates, max total $$$



Scotty: http://scotty.genetics.utah.edu/scotty.php

Busby et al., Bioinformatics, 2013

Tools for DE analysis with RNA-Seq

edgeR
ShrinkSeq
DESeq
baySeq
Vsf
Limma/Voom
mmdiff
cuffdiff

ROTS
TSPM
DESeq2
EBSeq
NBPSeq
SAMseq
NoiSeq

(italicized not in R/Bioconductor but stand-alone)

See: http://www.biomedcentral.com/1471-2105/14/91
A comparison of methods for differential expression analysis of RNA-seq data
Soneson & Delorenzi, 2013

Typical output from DE analysis

	logFC	logCPM	PValue	FDR
TRINITY_DN876_c0_g1_i1	-7.15049572793027	10.6197708379285	0	0
TRINITY_DN6470_c0_g1_i1	-7.26777912190146	7.03987604865422	1.687485656951e-287	6.46813252309319e-284
TRINITY_DN5186_c0_g1_i1	-7.85623682454322	9.18570464327063	1.17049180235068e-278	2.99099671894011e-275
TRINITY_DN768_c0_g1_i1	7.72884741150304	9.7514619195169	4.32504881419265e-272	8.28895605240022e-269
TRINITY_DN70_c0_g1_i1	-12.7646078189688	7.86482982471445	3.92853491279431e-253	6.02322972829624e-250
TRINITY_DN1587_c0_g1_i1	-5.89392061881667	9.07366563894607	6.32919557933429e-243	8.08660221852944e-240
TRINITY_DN3236_c0_g1_i1	-7.27029815068473	8.02209568234202	3.64955175271959e-235	3.99678053376405e-232
TRINITY_DN4631_c0_g1_i1	-7.45310693639574	6.91664918183241	4.30540921272851e-229	4.1256583780971e-226
TRINITY_DN5082_c0_g5_i1	-5.33154406167545	10.6977538760467	2.74243356676259e-225	2.33594396920022e-222
TRINITY_DN1789_c0_g3_i1	10.2032564835076	7.32607652700285	1.44273728647186e-213	1.10600240380933e-210
TRINITY_DN4204_c0_g1_i1	4.81030233739325	9.88844409410644	9.27180216086162e-205	6.46160321501501e-202
TRINITY_DN799_c0_g1_i1	-4.22044475626154	6.9937398638711	1.24746518421083e-197	7.96922341846683e-195
TRINITY_DN196_c0_g2_i1	4.60597918494257	9.86878463857276	1.9819997623131e-192	1.16877001368402e-189
TRINITY_DN5041_c0_g1_i1	-4.27126549355785	9.70894399883	1.8930437900069e-185	1.03657669244235e-182
TRINITY_DN1619_c0_g1_i1	-4.47156415953777	9.22535948721718	1.76766063029526e-181	9.03392426122899e-179
TRINITY_DN899_c0_g1_i1	-4.90914328409143	7.93768691394594	1.11054513767547e-180	5.32089939088761e-178
TRINITY_DN324_c0_g2_i1	4.87160837667488	6.84850312231775	2.20092562166991e-179	9.92487989160089e-177
TRINITY_DN3241_c0_g1_i1	-4.77760618069256	7.94111259715689	1.60585457735621e-173	6.83915621667372e-171
TRINITY_DN4379_c0_g1_i1	3.85133572453294	7.23712813663389	3.48140532848425e-164	1.4046554341137e-161
TRINITY_DN1919_c0_g1_i1	4.05998814332136	6.95937301668582	1.8588621194715e-161	7.12501850393425e-159
TRINITY_DN2504_c0_g1_i1	-6.92417817059644	6.20370039359785	2.42022459856956e-160	8.83497227268296e-158

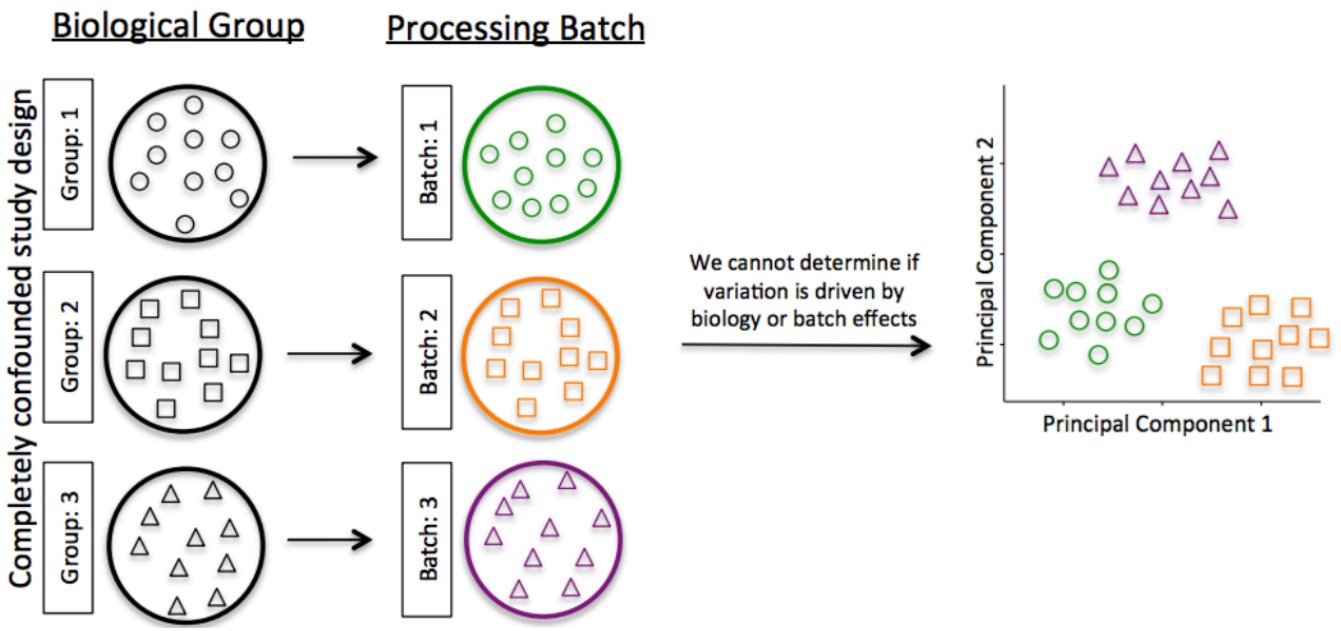
…

Up vs. Down regulated

Avg. expression level

Significance

Avoid Batch Effects



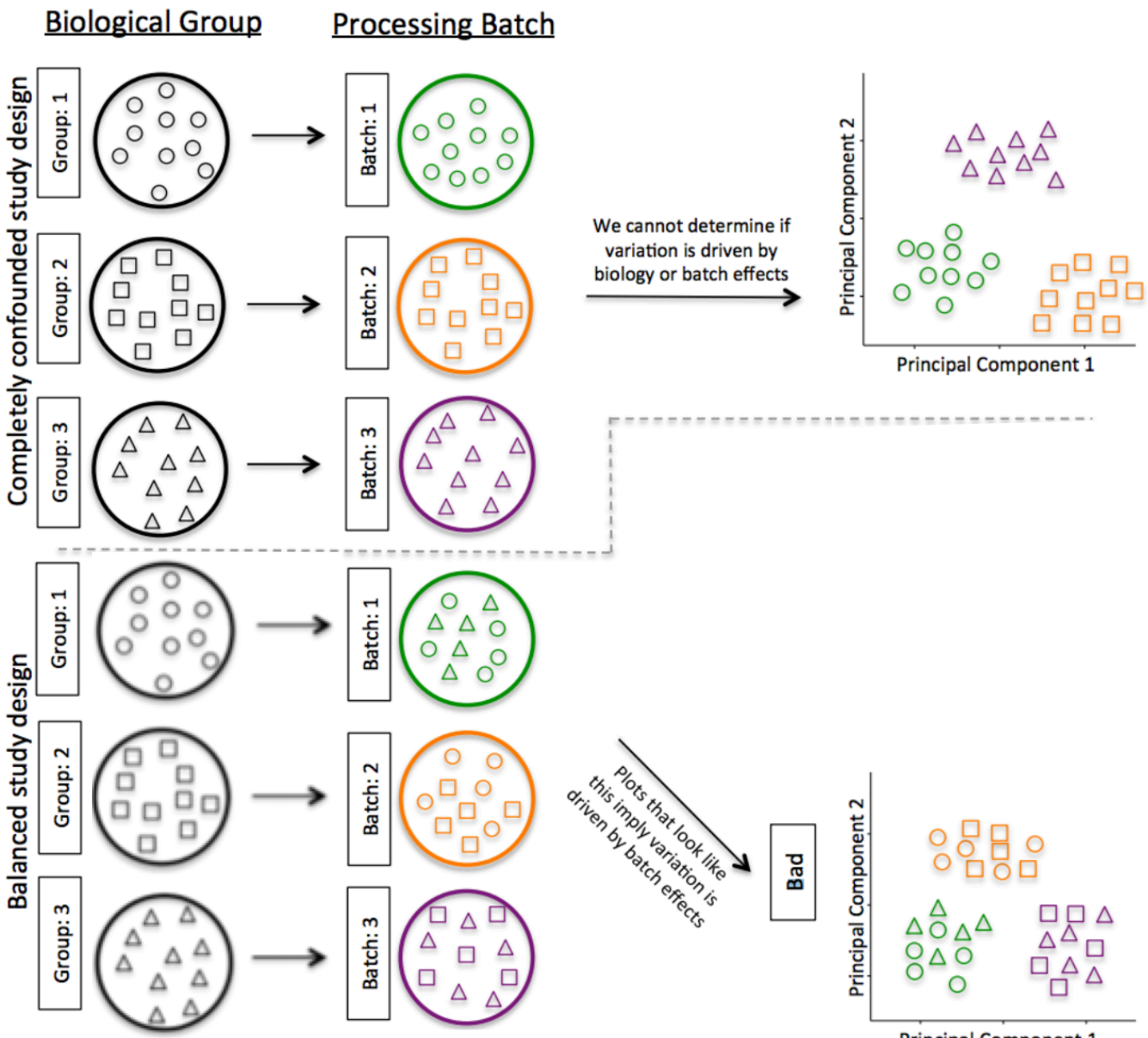
Grouping by Study or Batch?

Batch variable types:

- Times and dates
- Technician processing the samples
- Sequencing machine, or flow cell lane (Illumina)

Adapted from: Stephanie C. Hicks, Mingxiang Teng, Rafael A. Irizarry.
https://www.biorxiv.org/content/early/2015/09/04/025528
On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data.

Avoid Batch Effects



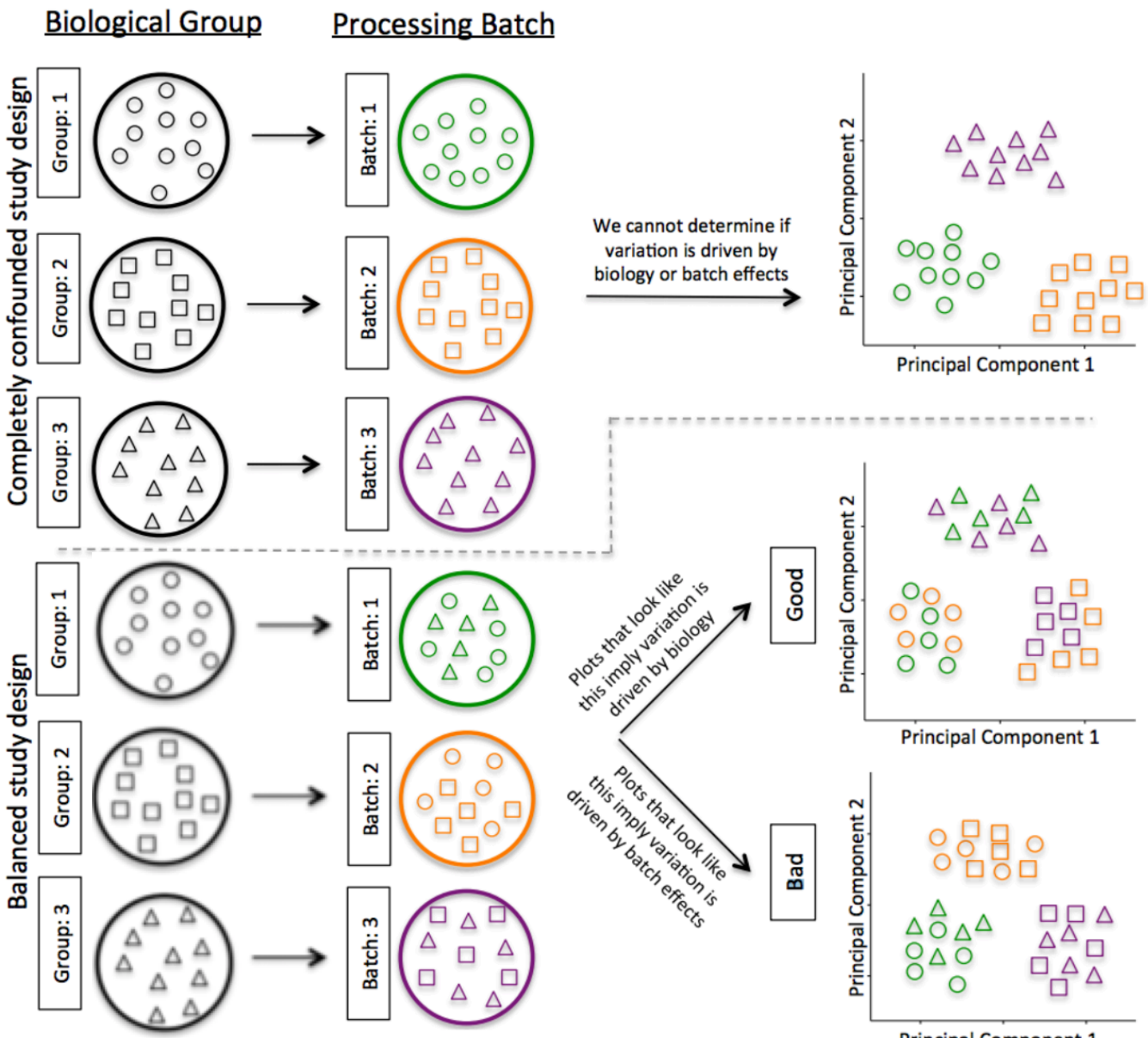
Grouping by Study or Batch?

Grouping by Batch

Adapted from: Stephanie C. Hicks, Mingxiang Teng, Rafael A. Irizarry.
https://www.biorxiv.org/content/early/2015/09/04/025528
On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data.

(Explore Batch Removal Techniques)

Avoid Batch Effects



Grouping by Study or Batch?

Grouping by Study

Grouping by Batch

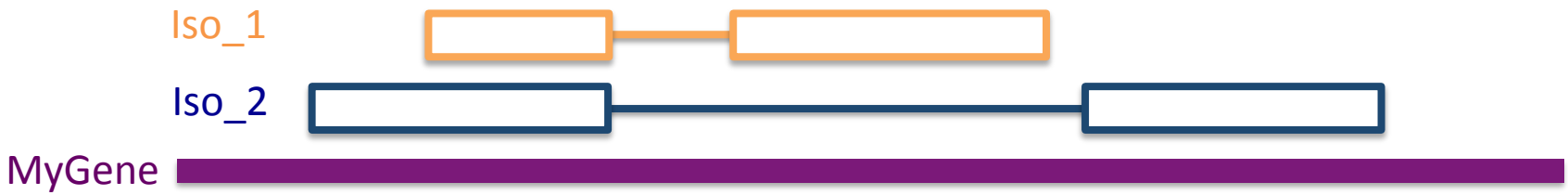
Adapted from: Stephanie C. Hicks, Mingxiang Teng, Rafael A. Irizarry.
https://www.biorxiv.org/content/early/2015/09/04/025528
On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data.

(Explore Batch Removal Techniques)

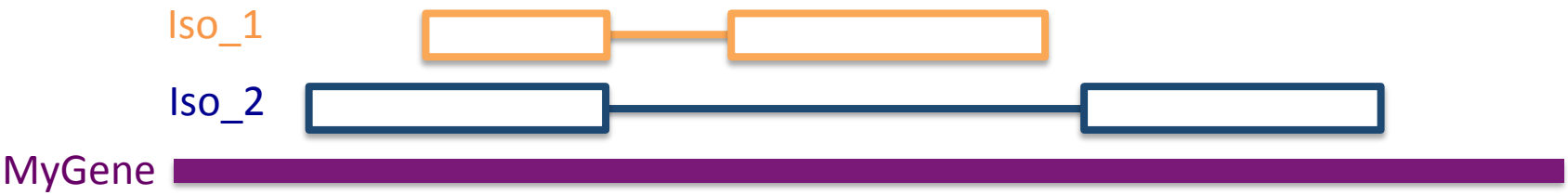
Flavors of Differential Expression Analyses

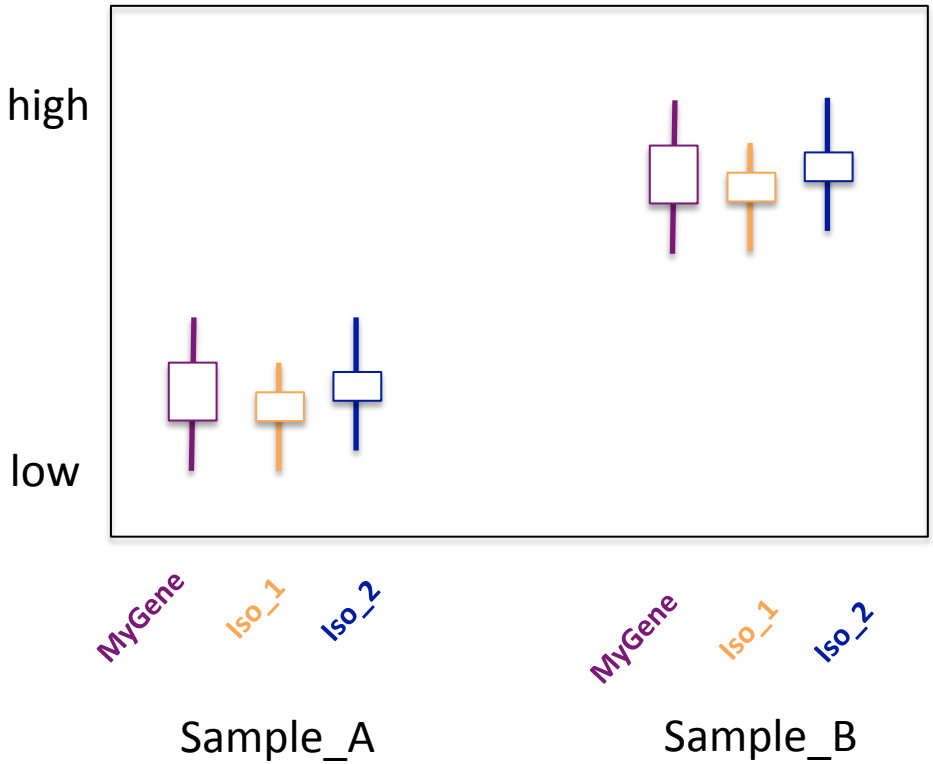
- Differential Gene Expression (DGE)
- Differential Transcript Expression (DTE)
- Differential Transcript Usage (DTU)
- Differential Exon Usage (DEU)

Differential Gene Expression (DGE) and Differential Transcript Expression (DTE) (Example 1)



Differential Gene Expression (DGE) and Differential Transcript Expression (DTE) (Example 1)

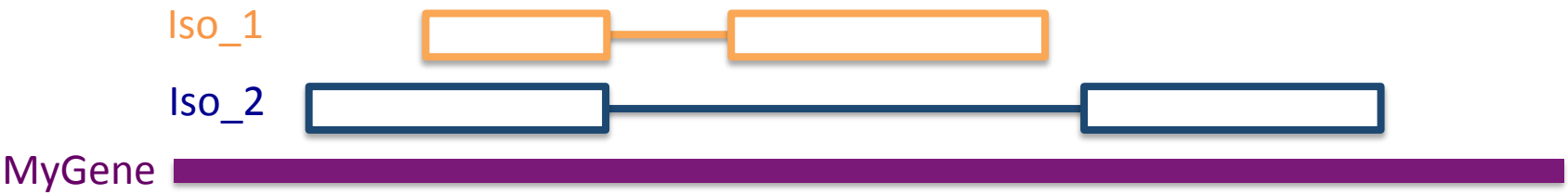


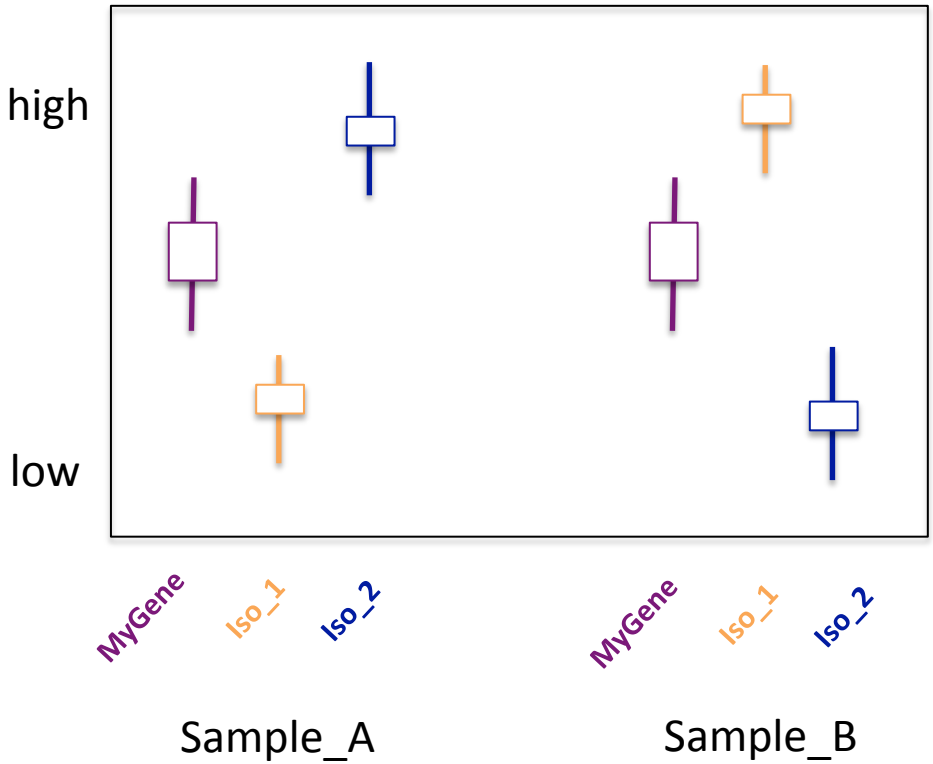


Feature	Diff Expressed?
MyGene	Yes
Iso_1	Yes
Iso_2	Yes

Diff. Transcript Usage ? (eg. Isoform switching) — No

Differential Gene Expression (DGE) and Differential Transcript Expression (DTE) (Example 2)

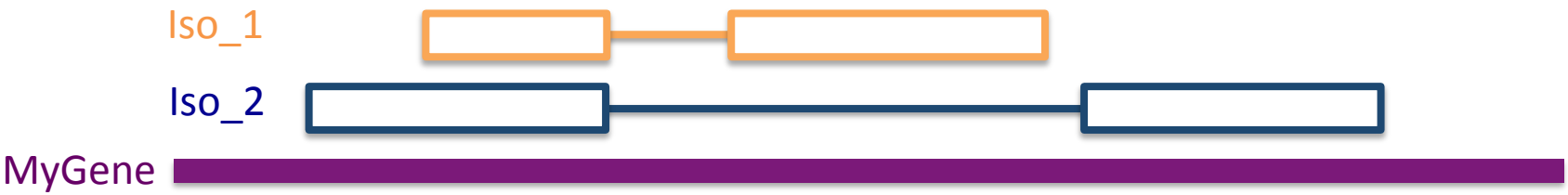


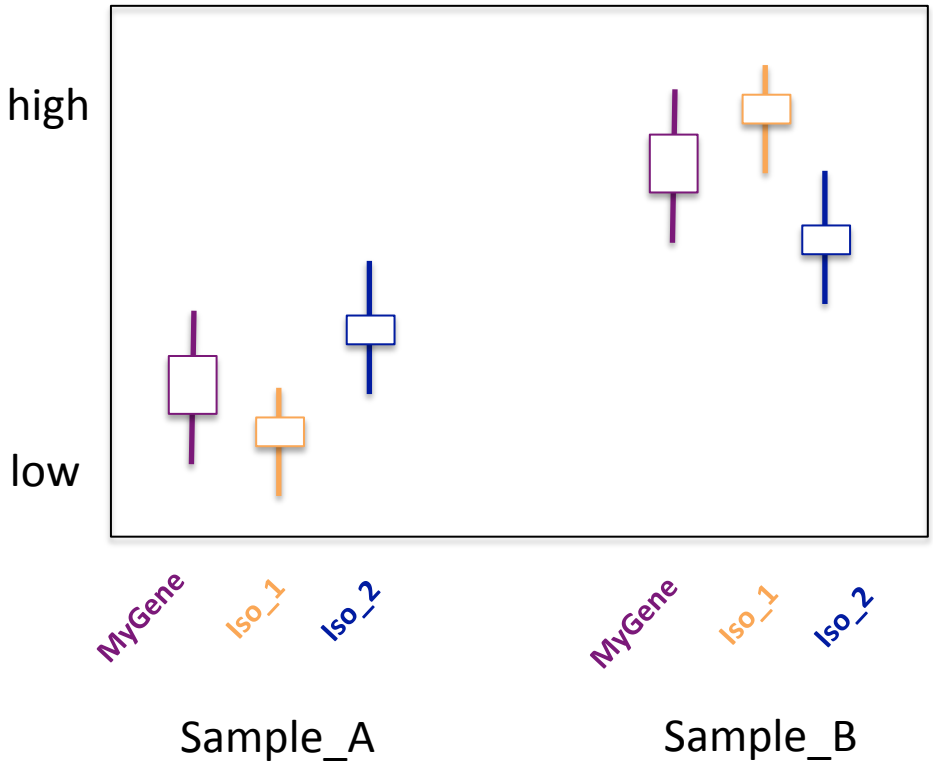


Feature	Diff Expressed?
MyGene	No
Iso_1	Yes
Iso_2	Yes

Diff. Transcript Usage ? (eg. Isoform switching) Yes

Differential Gene Expression (DGE) and Differential Transcript Expression (DTE) (Example 3)

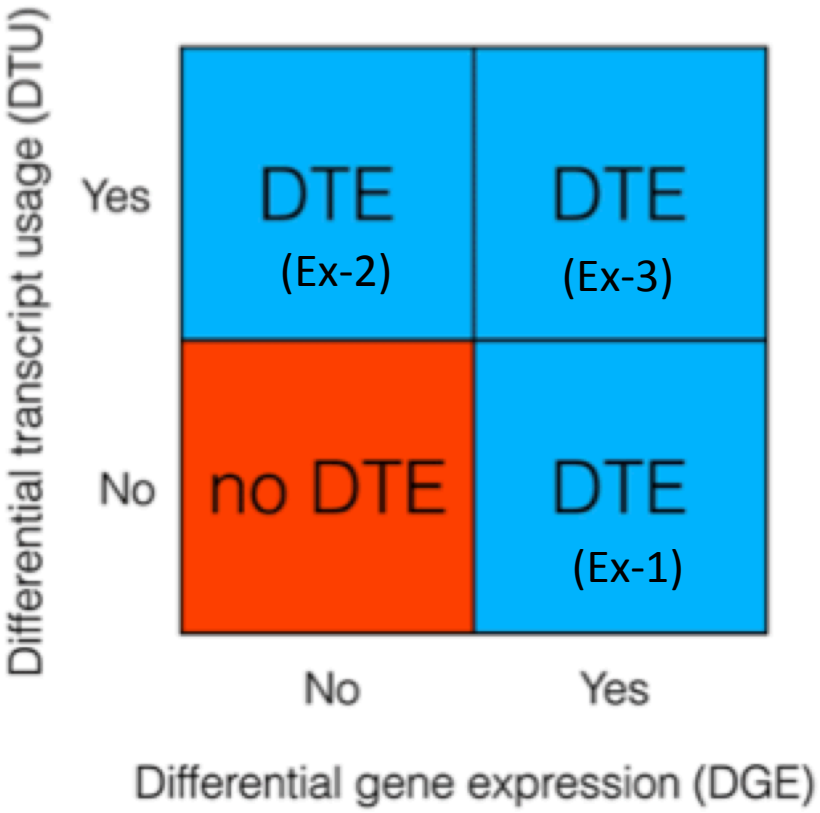




Feature	Diff Expressed?
MyGene	Yes
Iso_1	Yes
Iso_2	Yes

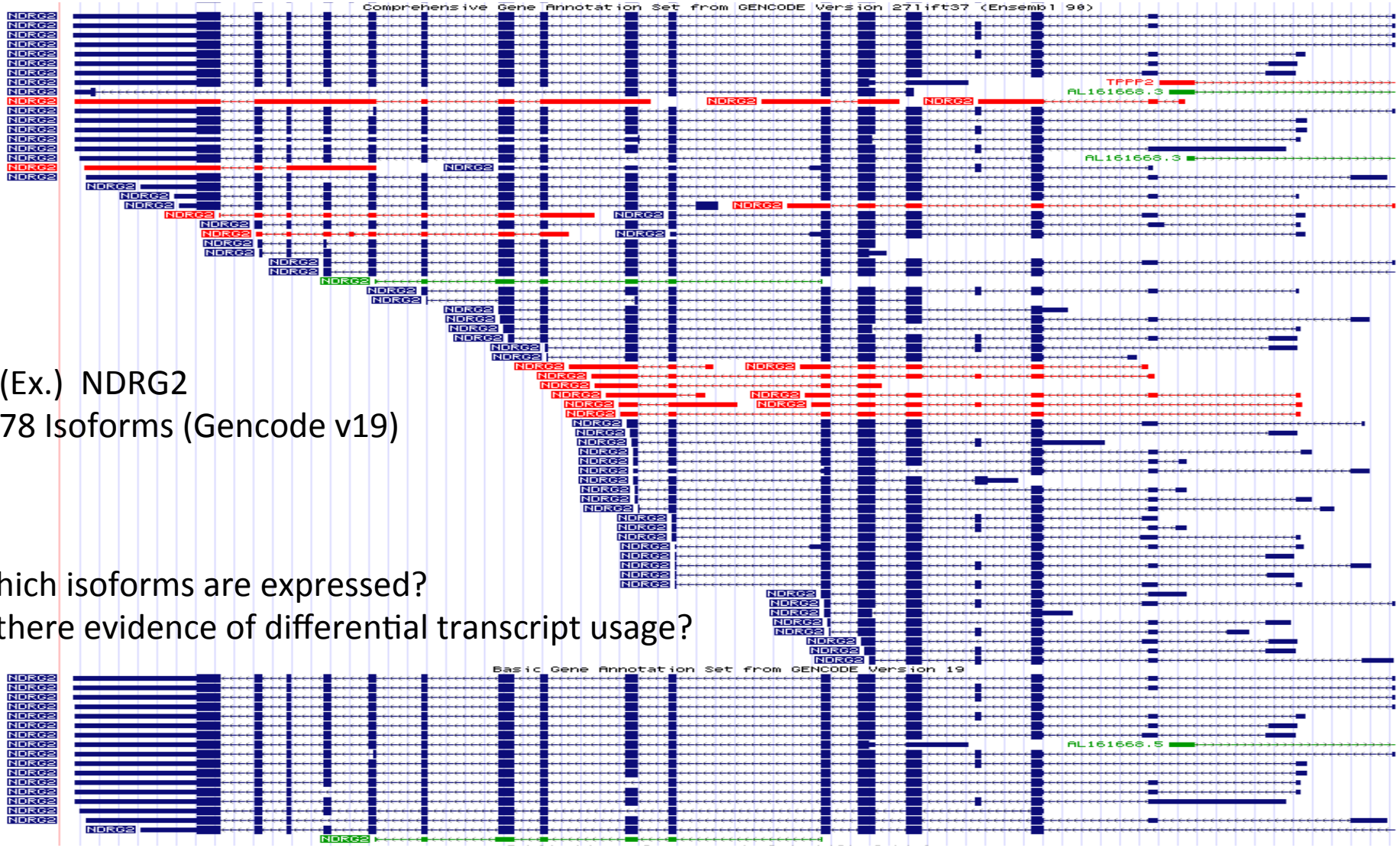
Diff. Transcript Usage ? (eg. Isoform switching) Yes

Differential Transcript Usage(DTU)
vs
Differential Gene Expression (DGE)
vs.
Differential Transcript Expression (DTE)



Soneson C, Love MI and Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2]. F1000Research 2016, 4:1521 (doi: 10.12688/f1000research.7563.2) F1000Research

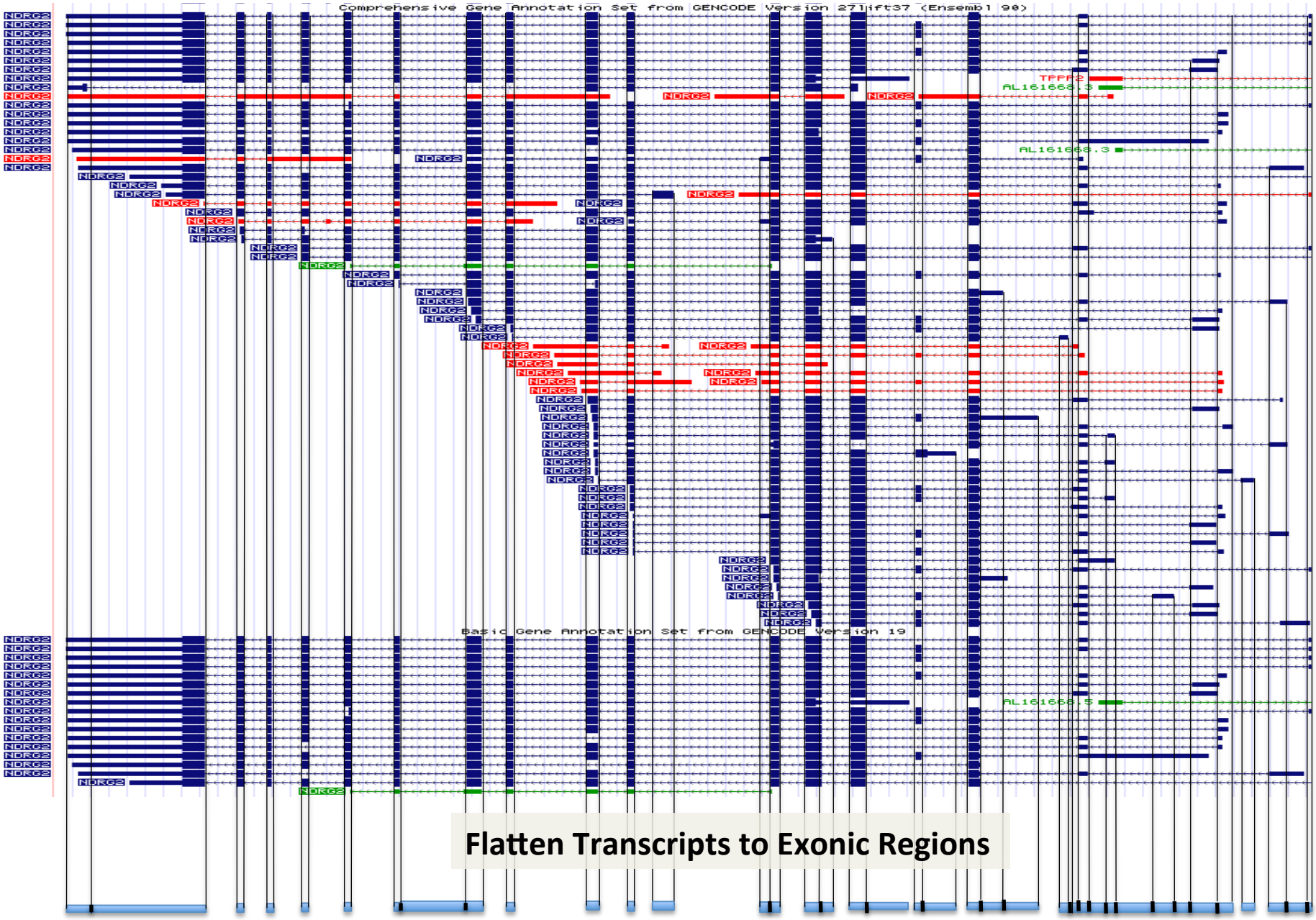
High Confidence Differential Transcript Expression is Difficult to Attain With Many Candidate Isoforms



(Ex.) NDRG2
78 Isoforms (Gencode v19)

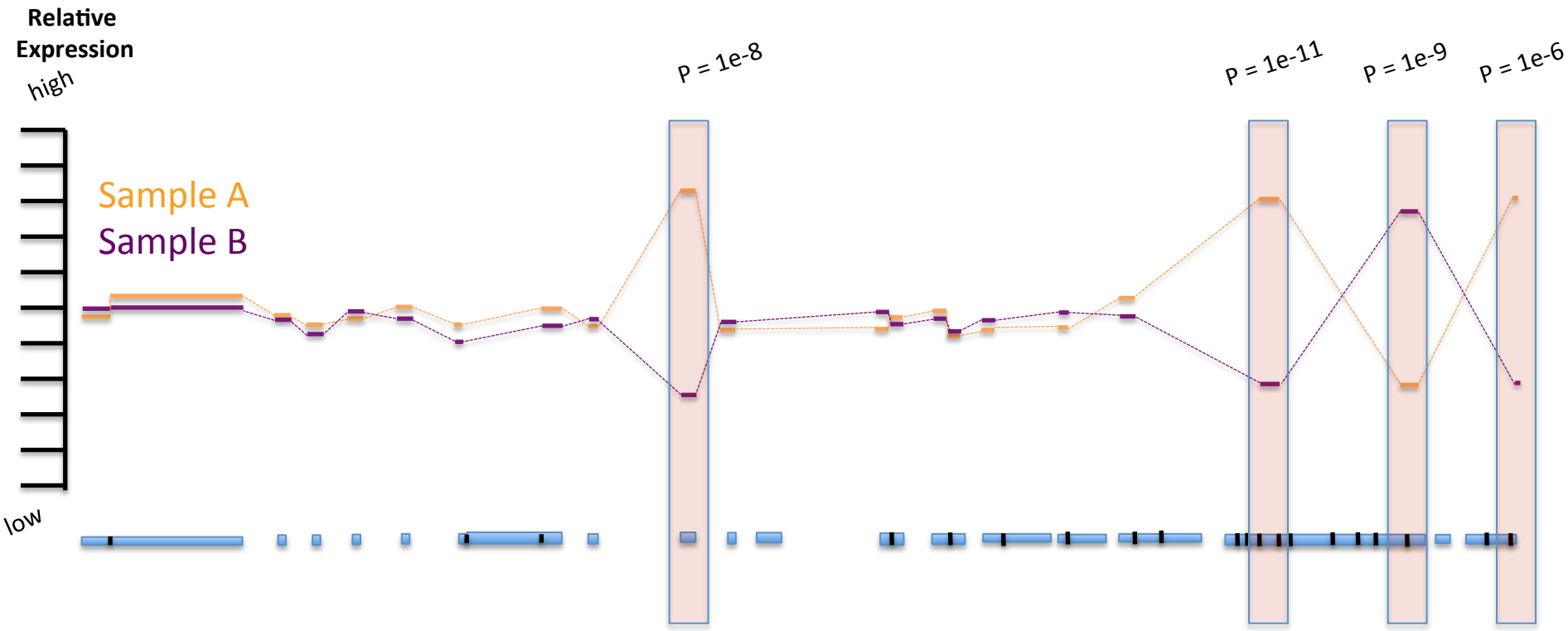
Which isoforms are expressed?
Is there evidence of differential transcript usage?

Measure Differential Transcript Usage (DTU) via Differential Exon Usage (DEU)



Flatten Transcripts to Exonic Regions

Measure Differential Transcript Usage (DTU) via Differential Exon Usage (DEU)



GENOME RESEARCH
CSHL Press | Journal Home | Subscriptions | eTOC Alerts | BioSupplyNet

Genome Res. 2012 Oct; 22(10): 2008–2017. PMCID: PMC3460195
doi: 10.1101/gr.133744.111

Detecting differential usage of exons from RNA-seq data

Simon Anders,[1,2] Alejandro Reyes,[1] and Wolfgang Huber

Averaged Replicates

Each Replicate

Flattened gene structure:

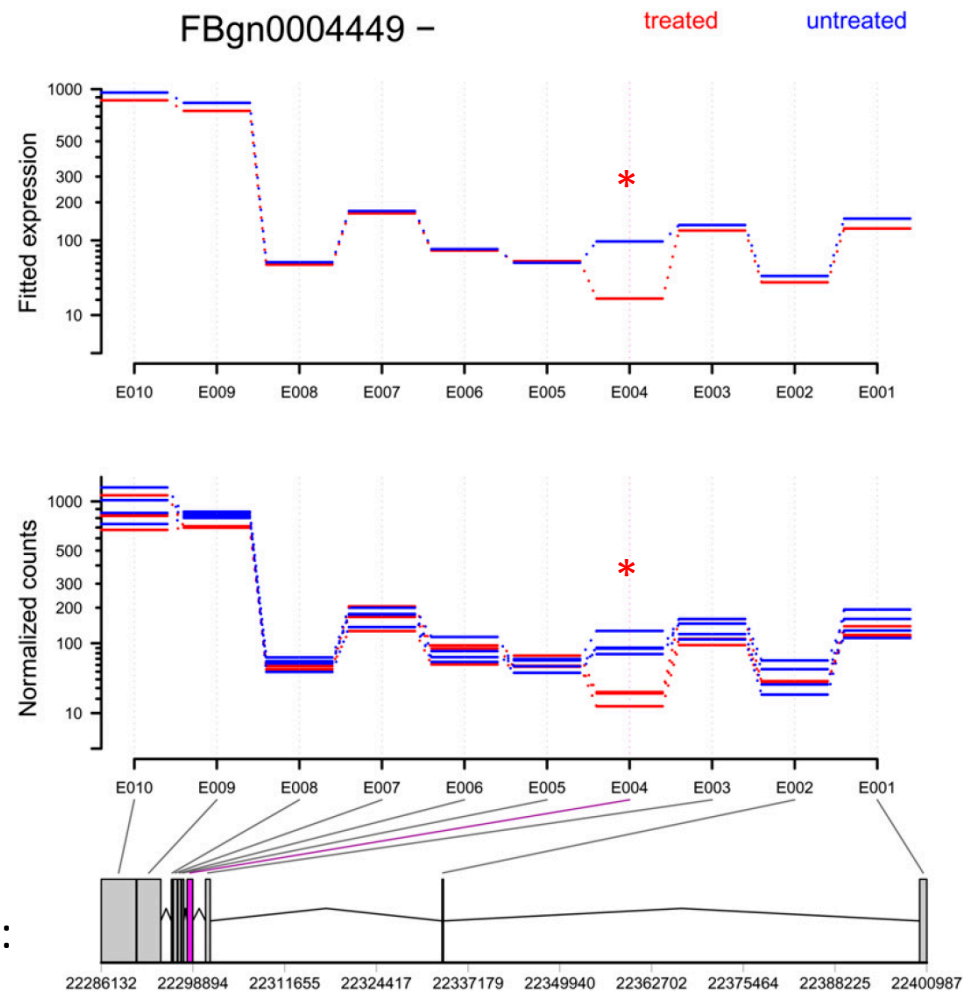


Figure 3. The treatment of knocking down the splicing factor *pasilla* affects the fourth exon (counting bin E004) of the gene *Ten-m* (CG5723). (*Top* panel) Fitted values according to the linear model; (*middle* panel) normalized counts for each sample; (*bottom* panel) flattened gene model. (Red) Data for knockdown samples; (blue) control.

Enabling Differential Transcript Usage Analysis for De novo Transcriptome Assemblies

Davidson *et al. Genome Biology* (2017) 18:148
DOI 10.1186/s13059-017-1284-1

Genome Biology

METHOD

Open Access

SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes

Nadia M. Davidson[1,2*], Anthony D. K. Hawkins[1] and Alicia Oshlack[1,2*]

Enabling Differential Transcript Usage Analysis for De novo Transcriptome Assemblies

Davidson *et al. Genome Biology* (2017) 18:148
DOI 10.1186/s13059-017-1284-1

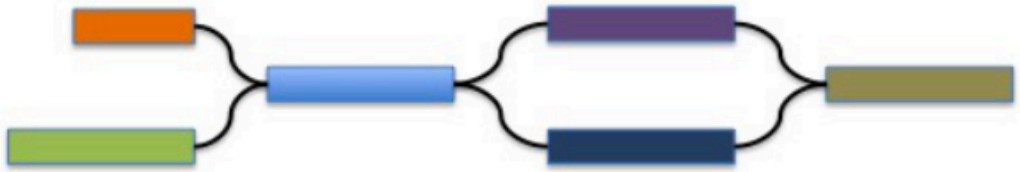
Genome Biology

METHOD Open Access

SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes

Nadia M. Davidson[1,2*], Anthony D. K. Hawkins[1] and Alicia Oshlack[1,2*]

Transcript splice graph:



Similar method and protocols now integrated into Trinity:
https://github.com/trinityrnaseq/trinityrnaseq/wiki/SuperTranscripts

Enabling Differential Transcript Usage Analysis for De novo Transcriptome Assemblies

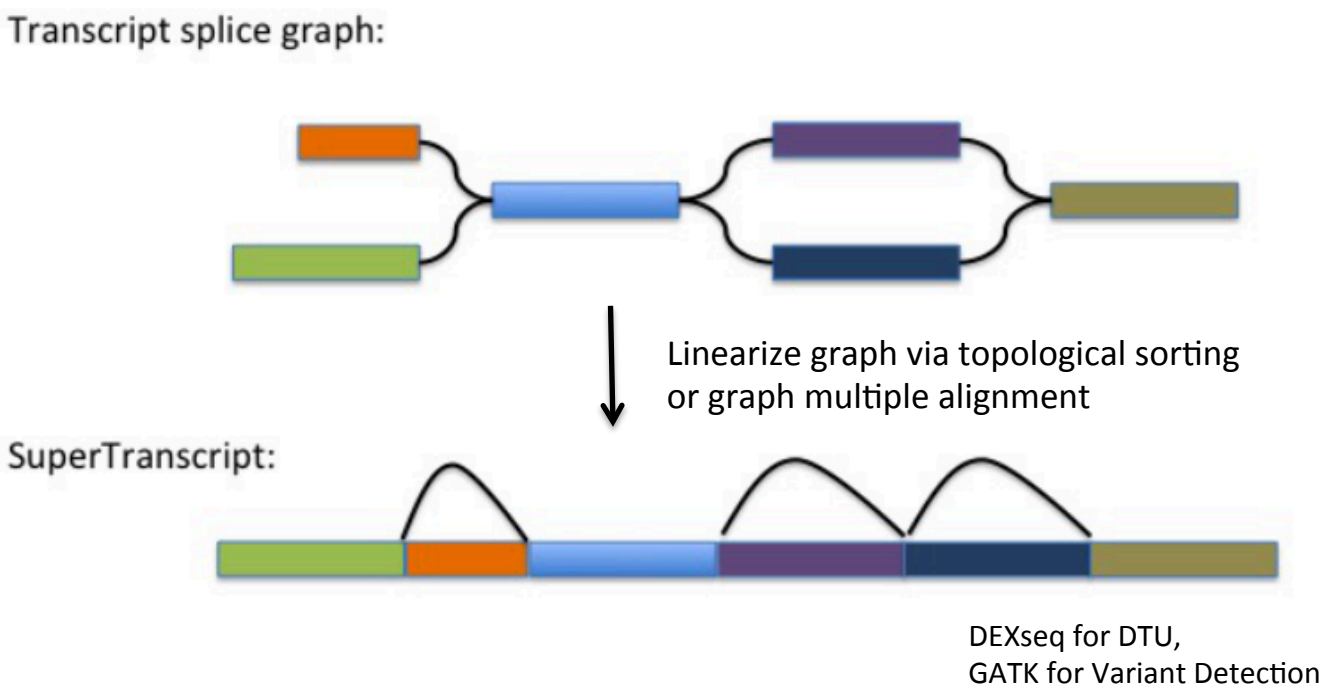
Davidson *et al. Genome Biology* (2017) 18:148
DOI 10.1186/s13059-017-1284-1

Genome Biology

METHOD Open Access

SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes

Nadia M. Davidson[1,2*], Anthony D. K. Hawkins[1] and Alicia Oshlack[1,2*]



DEXseq for DTU,
GATK for Variant Detection

Similar method and protocols now integrated into Trinity:
https://github.com/trinityrnaseq/trinityrnaseq/wiki/SuperTranscripts

Further Pushing the Envelope with RNA-Seq Analysis

Audoux *et al. Genome Biology* (2017) 18:243
DOI 10.1186/s13059-017-1372-2

Genome Biology

METHOD Open Access

DE-kupl: exhaustive capture of biological variation in RNA-seq data through *k*-mer decomposition

Jérôme Audoux[1], Nicolas Philippe[2,3], Rayan Chikhi[4], Mikaël Salson[4], Mélina Gallopin[5], Marc Gabriel[5,6], Jérémy Le Coz[5], Emilie Drouineau[5], Thérèse Commes[1,2] and Daniel Gautheret[5,6*]

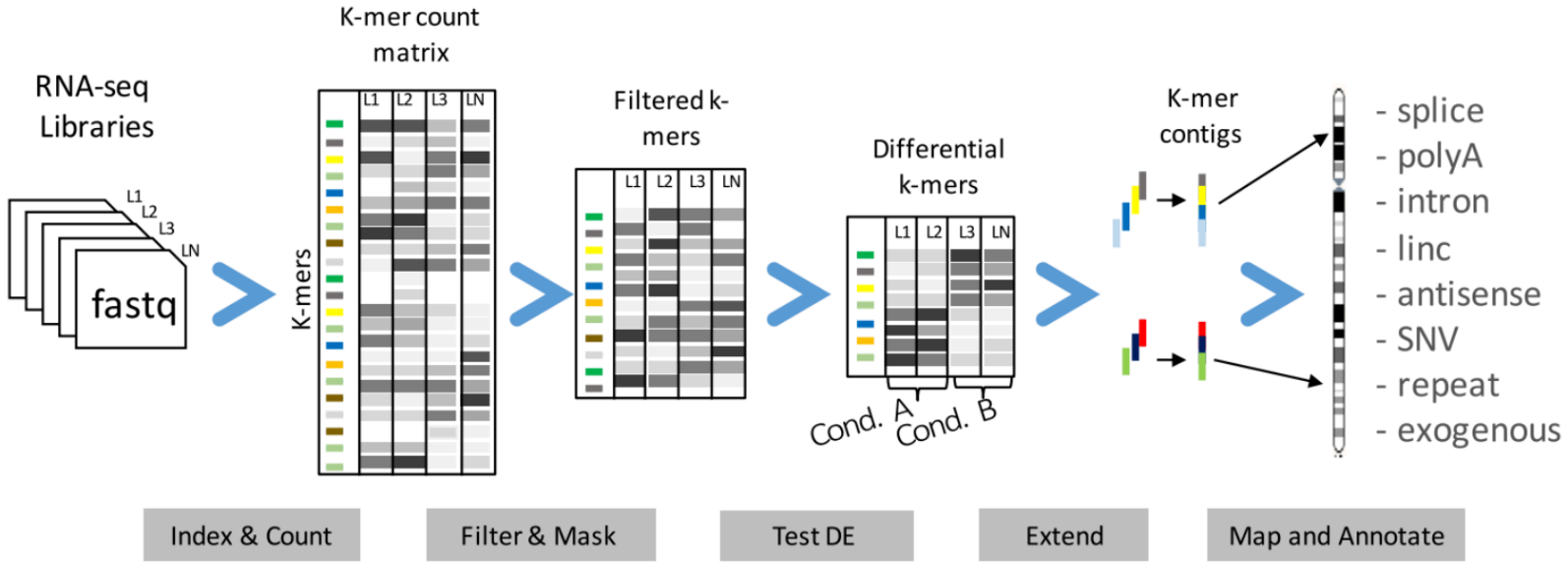
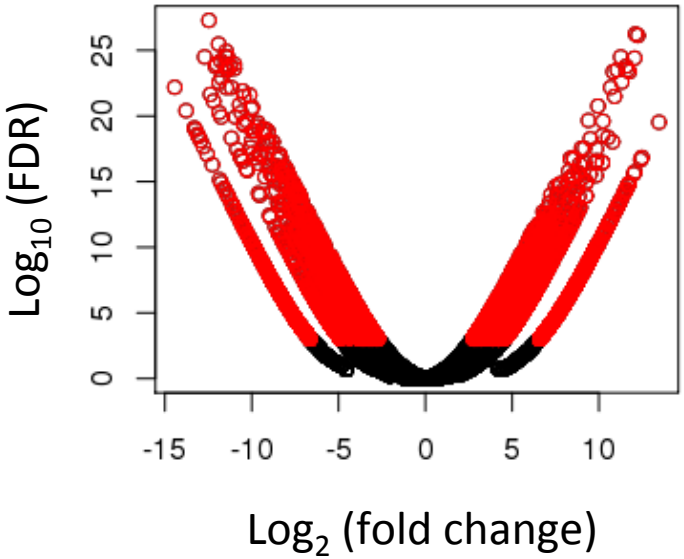


Fig. 4 The DE-kupl pipeline for the discovery and analysis of differentially expressed *k*-mers. First, Jellyfish is applied to count *k*-mers in all libraries. *k*-mers counts are then joined into a count matrix and filtered for low recurrence and matching to the reference transcriptome. Normalization factors are computed from raw *k*-mer counts and the differential expression procedure is applied. Finally, overlapping differentially expressed *k*-mers are extended into contigs and annotated based on their alignment to the reference and overlap with annotated genes

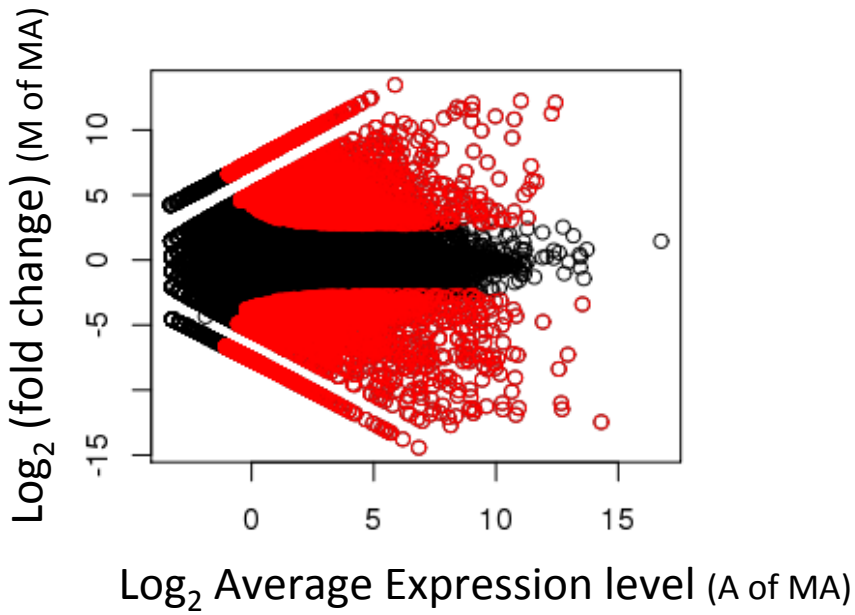
Visualization of DE results and Expression Profiling

Plotting Pairwise Differential Expression Data

Volcano plot
(fold change vs. significance)

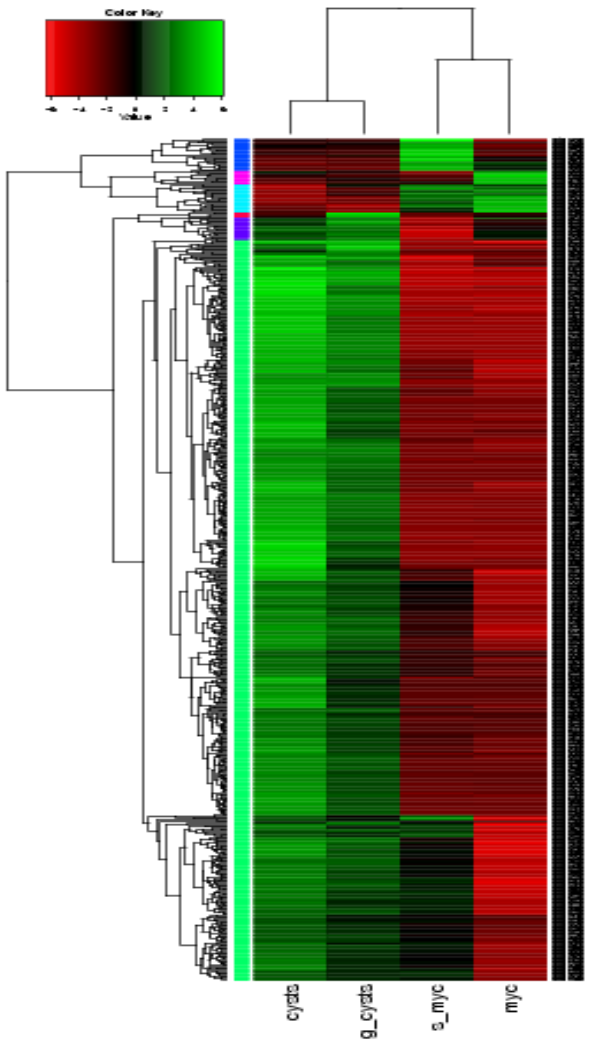


MA plot
(abundance vs. fold change)



Significantly differently expressed transcripts have FDR <= 0.001
(shown in red)

Comparing Multiple Samples



Heatmaps provide an effective tool for navigating differential expression across multiple samples.

Clustering can be performed across both axes:

-cluster transcripts with similar expression patters.

-cluster samples according to similar expression values among transcripts.

Examining Patterns of Expression Across Samples

Can extract clusters of transcripts and examine them separately.