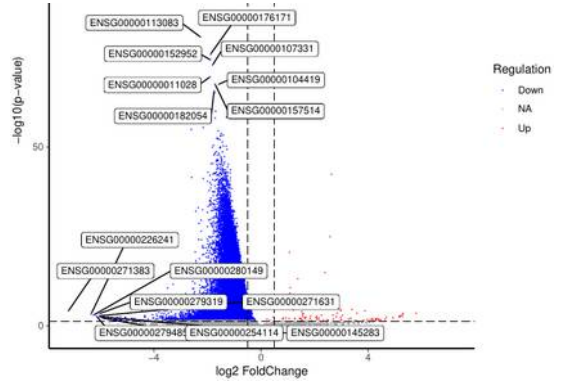
Tycho Kirchner — github.com/tycho-kirchner

Leibniz Institute on Aging –
Fritz Lipmann Institute

Differential gene expression analyses under global transcriptional shifts



Bulk RNAseq in a nutshell

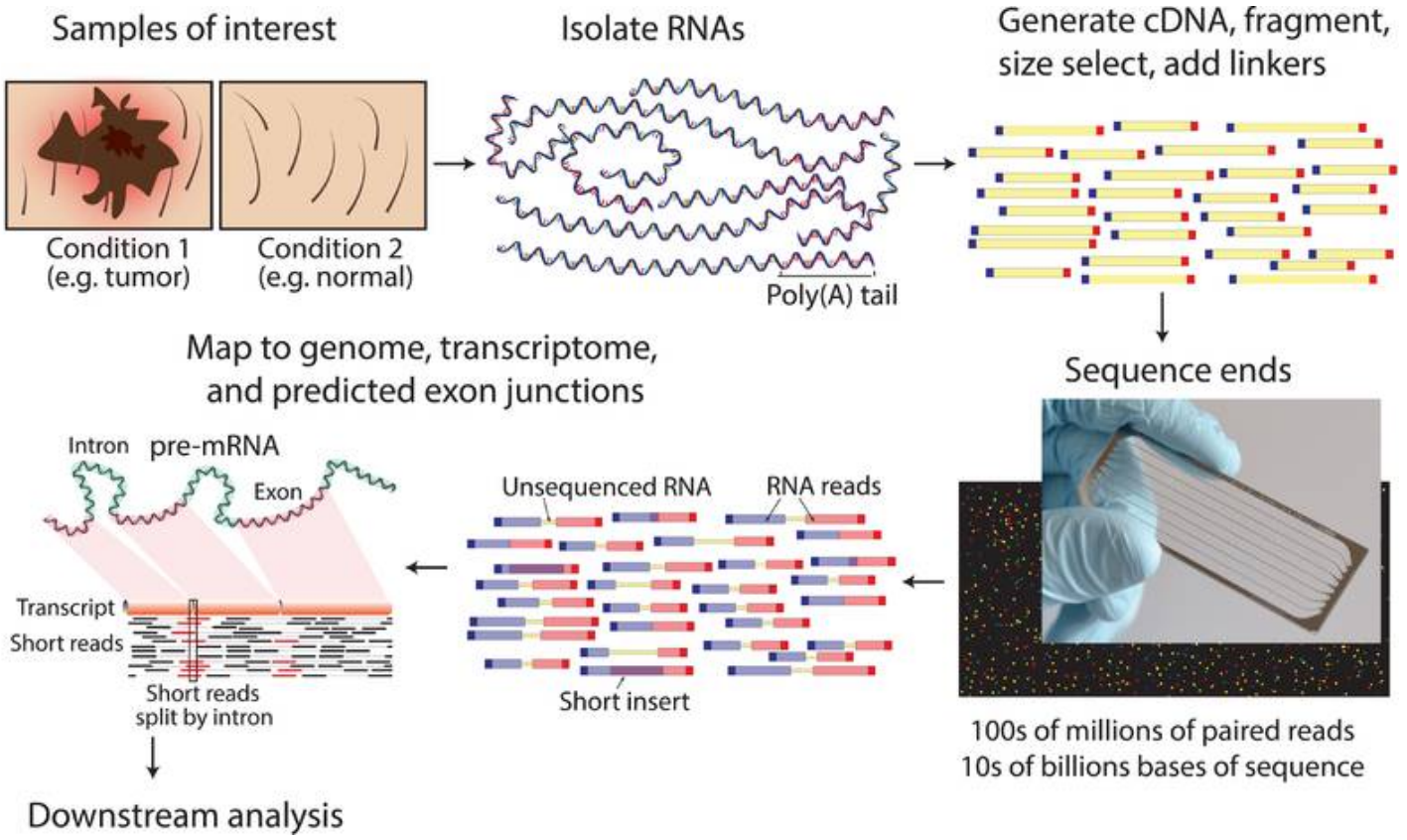


Image from https://en.wikipedia.org/wiki/File:Journal.pcbi.1004393.g002.png

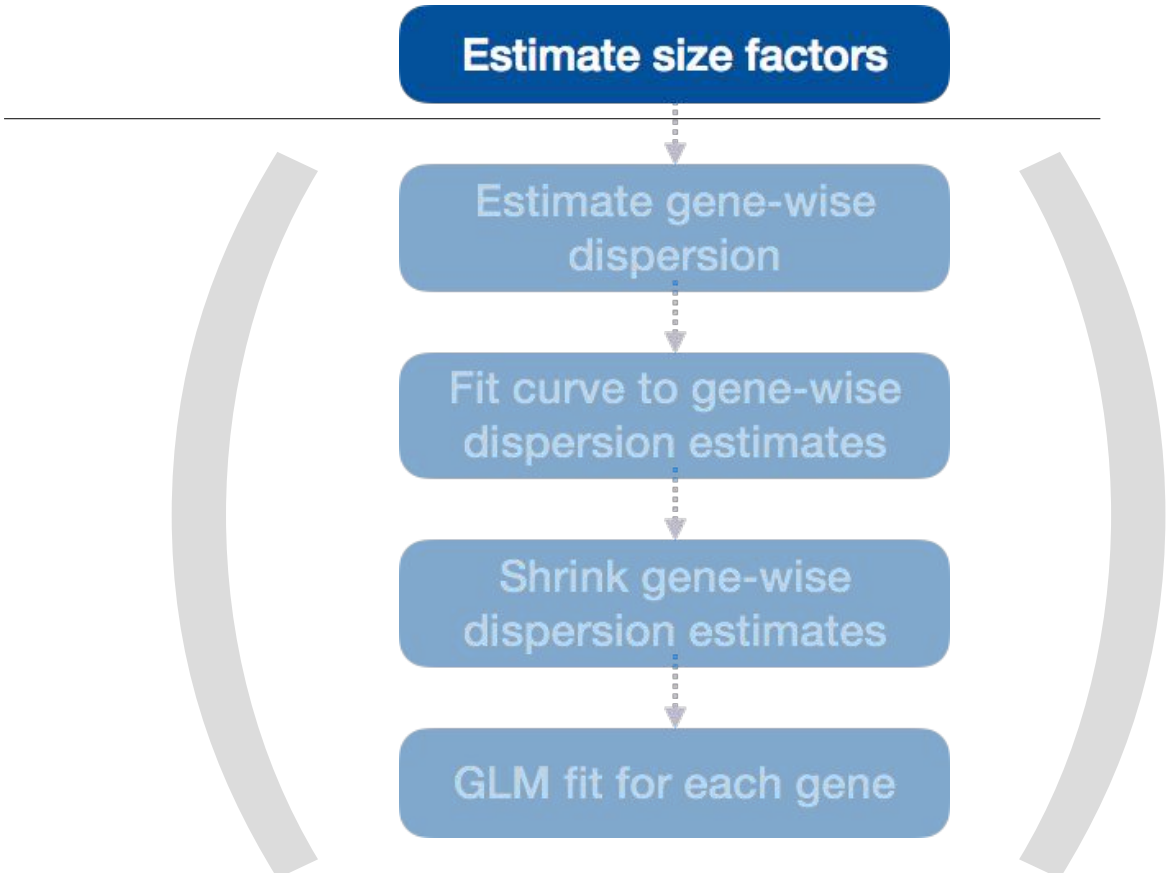
Global transcriptional shutdown or amplification

Shutdown	Amplification
Heat-shock	Cancer
Starvation	Embryonic development
Viral infections	Organ/tissue regeneration

See e.g. also: Percharde M, Bulut-Karslioglu A, Ramalho-Santos M. Hypertranscription in Development, Stem Cells, and Regeneration. Dev Cell. 2017 Jan 9;40(1):9-21. doi: 10.1016/j.devcel.2016.11.010. Epub 2016 Dec 15. PMID: 27989554; PMCID: PMC5225143.

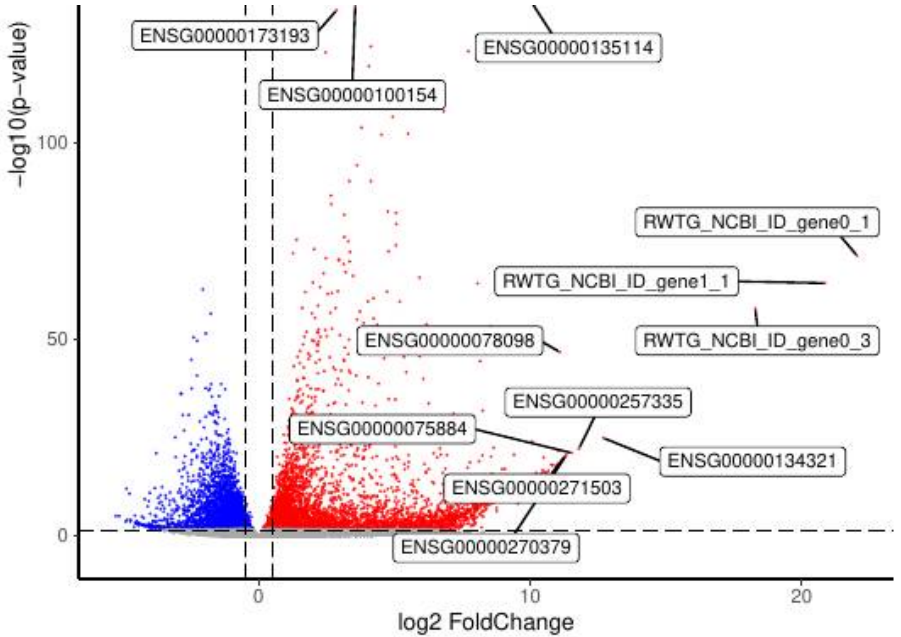
Roadmap

DESeq2's workflow.
Today, the roadtrip ends here



https://hbctraining.github.io/DGE_workshop/img/deseq2_workflow_separate_sf.png

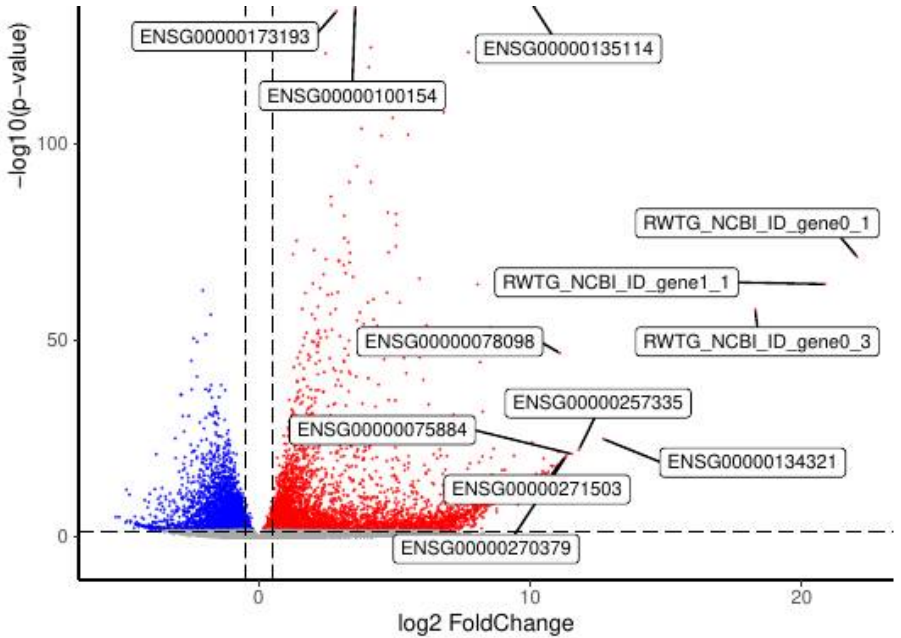
Control vs Rift valley fever virus infected cells - observed vs expected



Observed: 3458 down- and 4534 upregulated genes

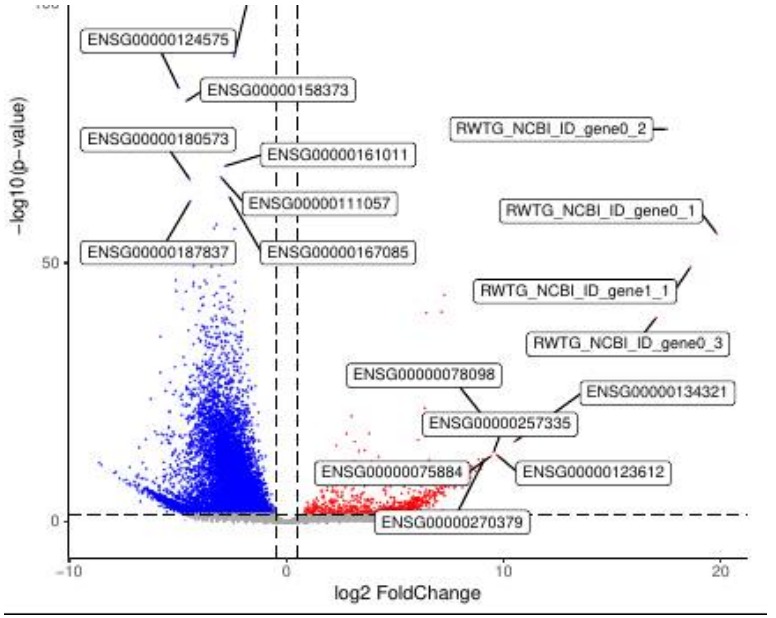
Pinkham et al. (2017): "*The large number of upregulated genes identified in our RNA-seq study was somewhat surprising, given the ability of NSs to suppress host transcription*"

Control vs Rift valley fever virus infected cells - observed vs expected



Observed: 3458 down- and 4534 upregulated genes

DESeq2's default sizefactor estimation

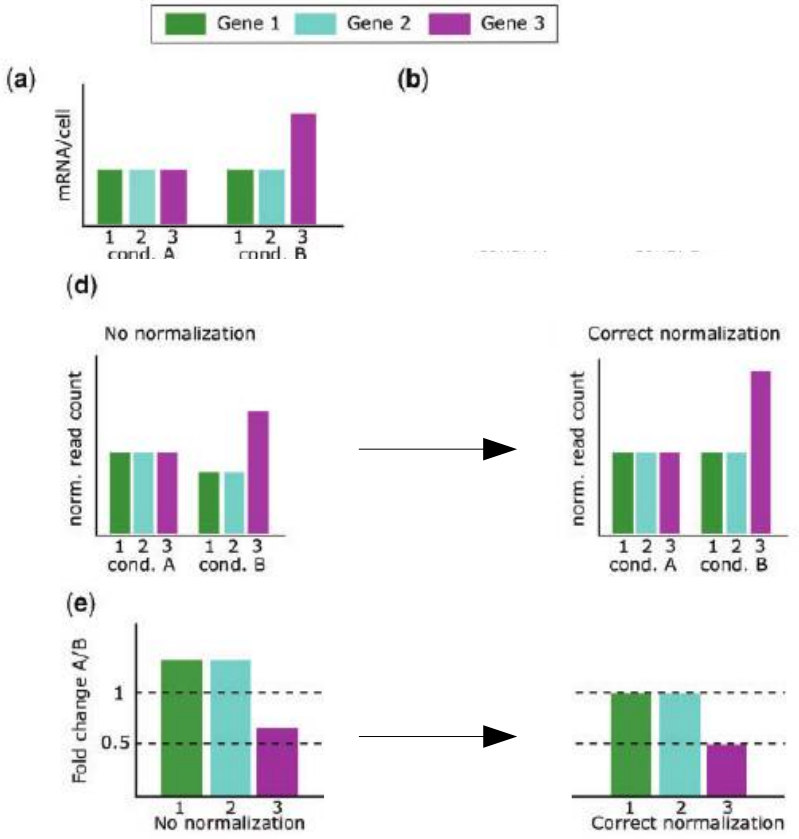


12851 down- and 778 upregulated genes

New sizefactor estimation method

From mRNA/cell to fold changes

Sizefactor estimation



Evans, C., Hardin, J. and Stoebel, D.M. (2018) 'Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions', Briefings in Bioinformatics, 19(5), pp. 776–792. Available at: https://doi.org/10.1093/bib/bbx008.

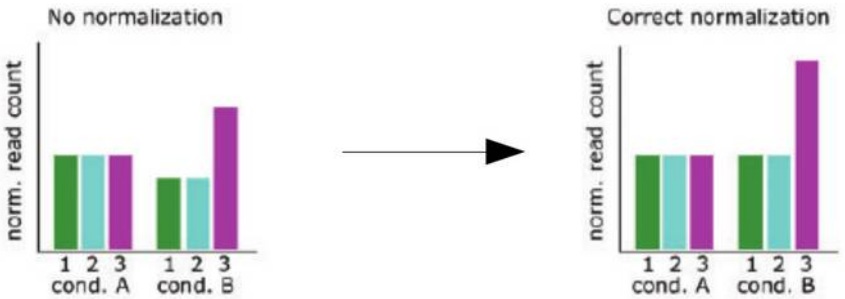
DESeq2's *median of ratios* method

- Each sample is assigned **one** single sizefactor ŝ
- The sizefactors are determined by the *median of ratios* to the geometric mean of all samples.
- (The transcript length shortens out, so DESeq2 doesn't need it)

$$\hat{s}_j = \underset{i}{\text{median}} \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv} \right)^{1/m}}$$

DESeq2's *median of ratios* method

Sample A	Sample B	Pseudo-reference (geometric mean)	Ratio of reference to Sample A	Ratio of reference to Sample B	Sample A scaled	Sample B scaled
1	4	2	2/1	2/4	3	$1,\overline{3}$
1	9	3	3/1	3/9	3	3
1	16	4	4/1	4/16	3	$5,\overline{3}$



But: this only works, if most genes are not DE

RNA spike-ins

- The External RNA Control Consortium (ERCC) provides 96 synthetic RNAs with various lengths, and GC content covering a 2^{20} concentration range
- Same amount of spike-ins applied to each sample
- DESeq2 estimates sizefactors based on the spike-ins
- **But**: cells have to be counted ☹
- Spike-ins often unavailable ☹
- *Housekeeping genes* potentially unreliable (inhibition of Pol II) ☹

Invitrogen™

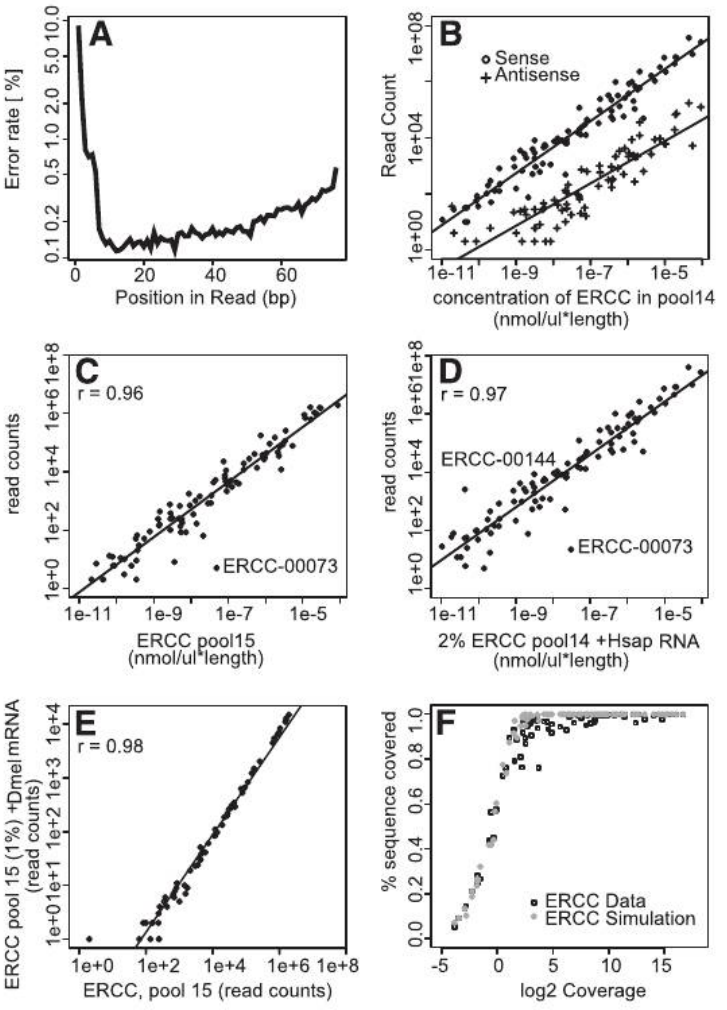
ERCC RNA Spike-In Mix

Catalog number: 4456740

Related applications: Real Time PCR (qPCR) | RNA Sequencing | SOLiD® Next-Generation Sequencing

Technical Support | Customer Service

	Catalog Number	Unit Size	Price (EUR)	Availability	Quantity
✓	4456740	1 kit	**Price:** 1.392,00 **Online Offer:** 1.268,65 *(ends 31-Dec-2023)* **Your Price:** Sign In	...	



Existing tools - qsmooth

Qsmooth quantile-normalizes the counts between the reference quantile of biological replicates or towards the overall reference quantile, depending on variability within and between conditions.

Biostatistics

Issues Advance articles Submit ▾ Purchase Alerts About ▾

Article Navigation

JOURNAL ARTICLE

Smooth quantile normalization FREE

Stephanie C Hicks, Kwame Okrah, Joseph N Paulson, John Quackenbush, Rafael A Irizarry, Héctor Corrada Bravo

Biostatistics, Volume 19, Issue 2, April 2018, Pages 185–198, https://doi.org/10.1093/biostatistics/kxx028

Published: 10 July 2017 **Article history** ▾

PDF Split View Cite Permissions Share ▾

SUMMARY

Between-sample normalization is a critical step in genomic data analysis to remove systematic bias and unwa
variation in high-throughput data. Global normalization methods are based on the assumption that observed
global properties is due to technical reasons and are unrelated to the biology of interest. For example, some me

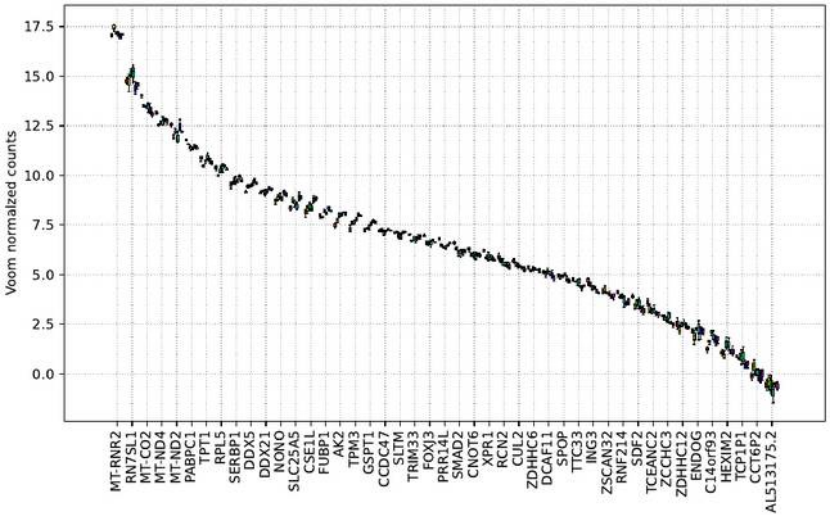
Existing tools - moose²

Research | Open access | Published: 05 September 2017

RNA-sequence data normalization through in silico prediction of reference genes: the bacterial response to DNA damage as case study

Bork A. Berghoff, Torgny Karlsson, Thomas Källman, E. Gerhart H. Wagner & Manfred G. Grabherr

BioData Mining **10**, Article number: 30 (2017) | Cite this article



Benchmark against spiked datasets with transcriptional shifts

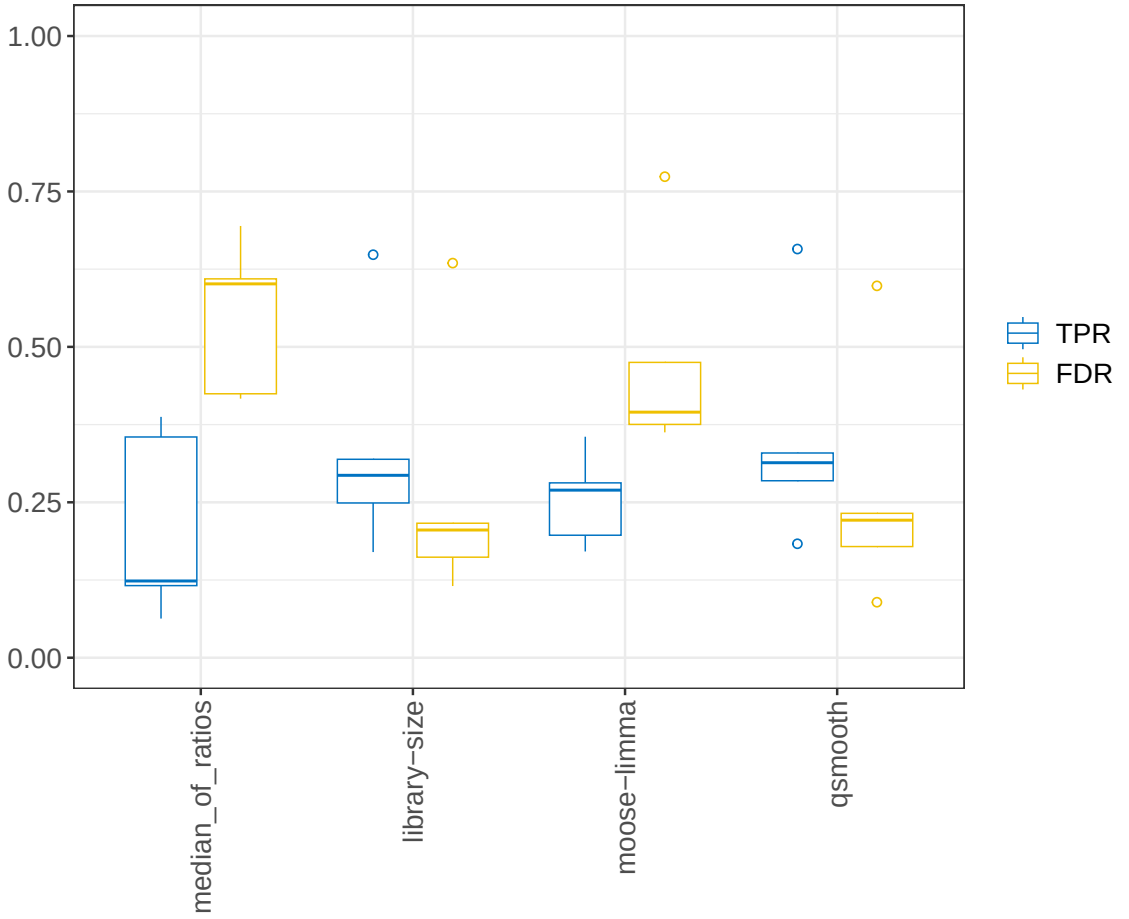
- Buschle et al. (2021) examined Raji-cells expressing the Epstein-Barr virus (EBV)
- Lau et al. (2023) found a large downregulation in topoisomerase I (TOP1) overexpressing Human embryonic kidney 293 cells. TOP1 is known for its role in relieving DNA supercoils for enabling transcription
- Bruno et al. (2020) found that Che-1 depletion induces a global transcription shutoff by reducing histone acetylation. Che-1 is an interactor of RNA polymerase II

*In all publications, cells were actually **counted** (many other authors don't, but use spikes as control for technical bias).*

1. Buschle, A., Mrozek-Gorska, P., Cernilogar, F.M., Ettinger, A., Pich, D., Krebs, S., Mocanu, B., Blum, H., Schotta, G., Straub, T., et al. (2021). Epstein-Barr virus inactivates the transcriptome and disrupts the chromatin architecture of its host cell in the first phase of lytic reactivation. Nucleic Acids Res 49, 3217–3241. 10.1093/nar/gkab099.
2. Lau, M.S., Hu, Z., Zhao, X., Tan, Y.S., Liu, J., Huang, H., Yeo, C.J., Leong, H.F., Grinchuk, O.V., Chan, J.K., et al. (2023). Transcriptional repression by a secondary DNA binding surface of DNA topoisomerase I safeguards against hypertranscription. Nat Commun 14, 6464. 10.1038/s41467-023-42078-9.
3. Bruno, T., De Nicola, F., Corleone, G., Catena, V., Goeman, F., Pallocca, M., Sorino, C., Bossi, G., Amadio, B., Cigliana, G., et al. (2020). Che-1/AATF-induced transcriptionally active chromatin promotes cell proliferation in multiple myeloma. Blood Advances 4, 5616–5630. 10.1182/bloodadvances.2020002566.

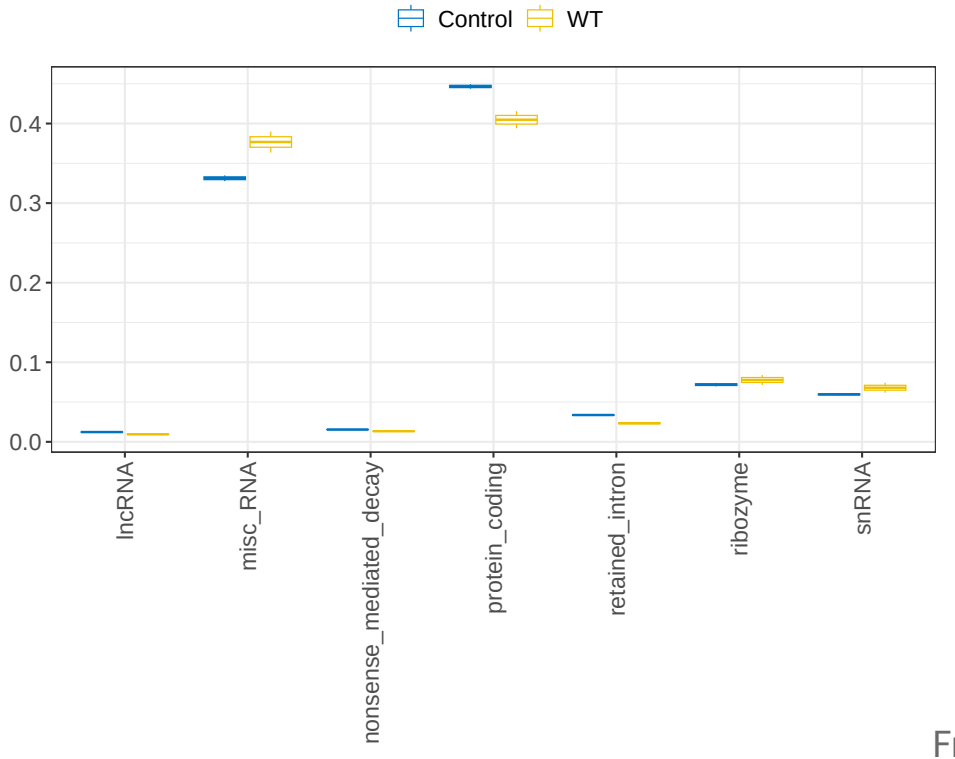
DEGs detection performance by normalization method

Control vs treated compared to spike-ins (padjust < 0.05)



Fragment percentages by ensembl biotype

TOP1



- lncRNAs transcribed by Pol I, II or III [1]
- snRNAs transcribed by Pol II or III

Fragment := mapped read divided by exon length

1.Mattick, J.S., Amaral, P.P., Carninci, P., Carpenter, S., Chang, H.Y., Chen, L.-L., Chen, R., Dean, C., Dinger, M.E., Fitzgerald, K.A., et al. (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol 24, 430–447. 10.1038/s41580-022-00566-8.

Normalization by biotype — basic assumptions

- **The total number of read ncRNA fragments should be *roughly* equal across conditions. If it's higher, in case of a shutdown: likely simply more cells were sequenced**
- At least ***some abundant*** ncRNAs are likely unaffected by the global shift
- ncRNA is presumably less affected by treatment, e.g. it's transcribed by POL I, II and III as opposed to cRNA (Pol II)
- some ncRNAs are quite abundant and stable, such as circular RNAs (mRNA-levels!)[1]
- a greater share of ncRNA is likely non-functional as compared to cRNA[1]

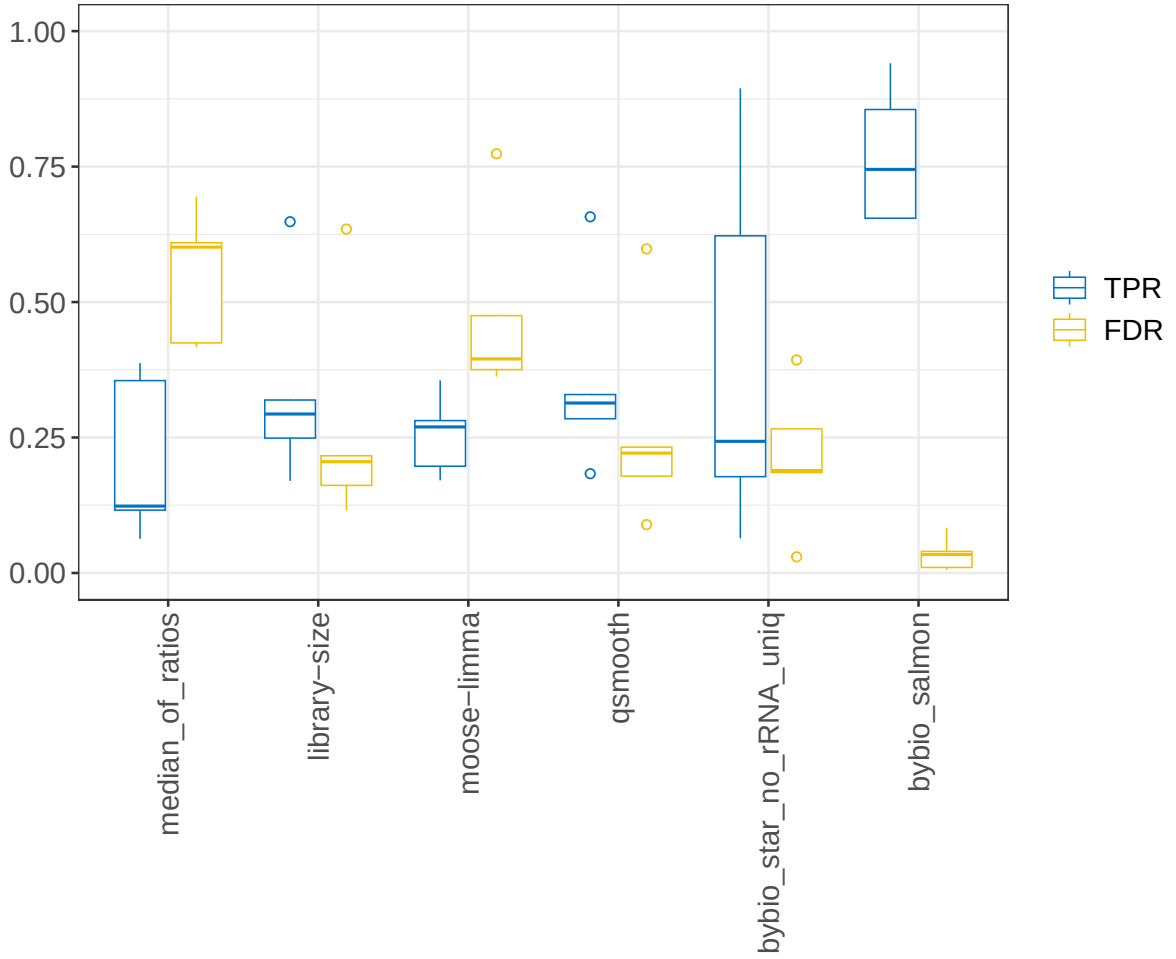
1. Palazzo, A.F., and Lee, E.S. (2015). Non-coding RNA: what is functional and what is junk? Frontiers in Genetics 6.

DEGs detection performance by normalization method

Control vs treated
compared to spike-ins

bybio_salmon:

- mapped to whole transcriptome (including rRNA)
- salmon keeps multi-mapping reads
- transcript-level length used instead of, e.g., canonical length
- sizefactors based on the salmon-counts were applied to STAR uniquely mapped and rRNA-filtered counts



Perspectives

- Test on more datasets (awaiting ethics approval for a large spiked “cancer dataset”)
- Test on polyA RNAseq (currently only rRNA depleted total-RNA)
- Check performance on spiked data without transcriptional shifts → preliminary benchmarks seem to perform well
- Evaluate implications of *slight* overall changes in global transcription levels.

Acknowledgments

Thanks to the whole Hoffman group, especially

Konstantin
Robert
Steve

And thanks for your attention!

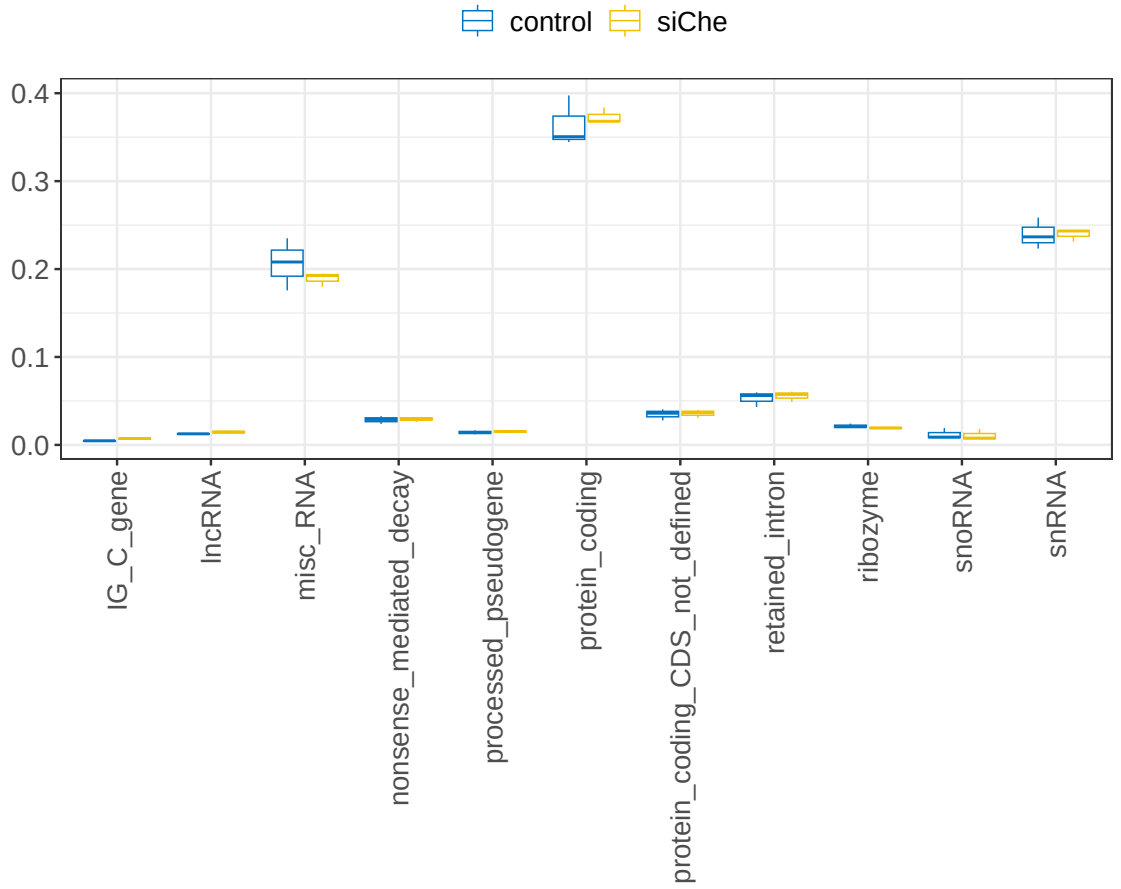
Appendix – Normalization algorithm for **different** ncRNA fragmentcounts across conditions

- Normalize by transcript length and scale all samples to the same total number of fragments
- Ignore cRNAs (as well as pseudogenes and some other protein-associated biotypes)
- Remove the most variant ncRNA's **within** replicates. As the variance is dependent on the mean, these will be predominantly highly expressed genes which are already unstable in one condition
- Calculate a per-sample pseudo-reference by geometric mean of the remaining sum of fragments
- Calculate scaling factors for each sample to that pseudo-reference
- Apply the scaling factors to the whole sample (including cRNA's) and run DESeq2 with disabled sizefactor estimation

Appendix – Normalization algorithm for **equal** ncRNA fragmentcounts across conditions

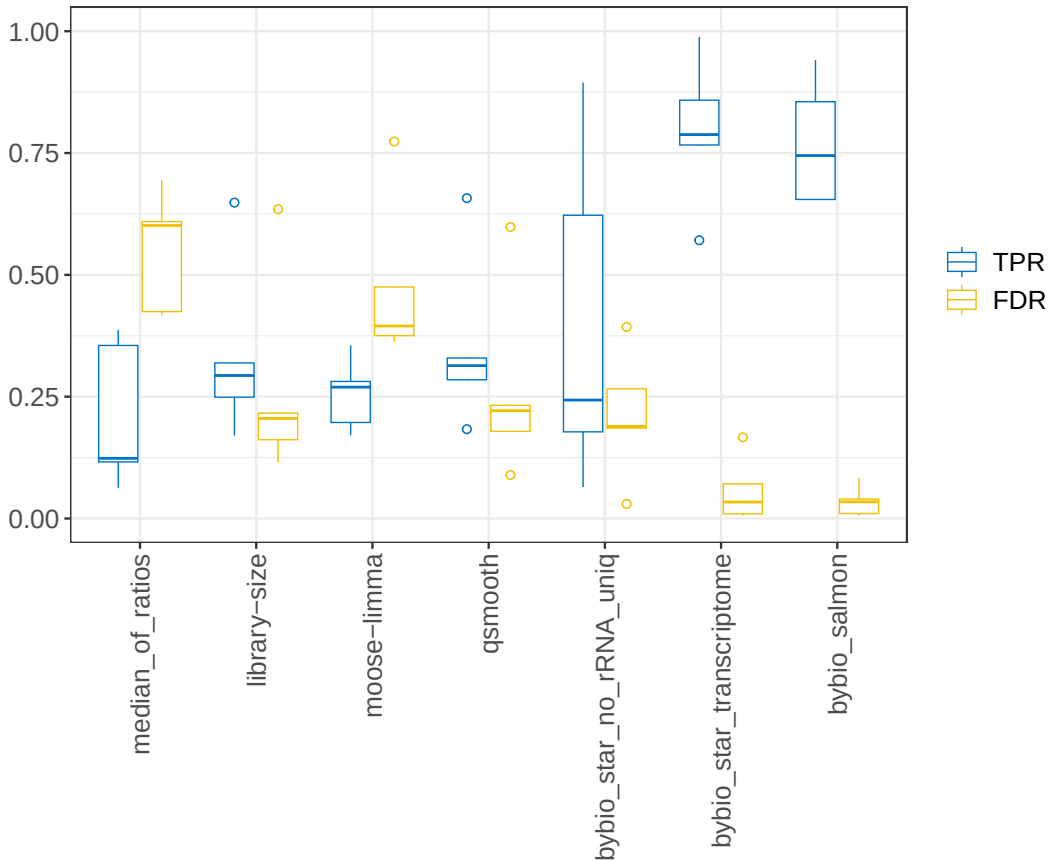
- If total ncRNA fragmentcounts are already similar across conditions, scaling by these makes no sense
- Instead, perform the same procedure as for the different ncRNA fragmentcounts on the slide before, but instead of a pseudo-reference by geometric mean, fall back to DESeq2's *median of ratios* method — but only for the ncRNAs.

Appendix – No global fragment differences in Che1



→ Thus, we fallback to *median of ratios* normalization of ncRNA

Appendix – star on transcriptome with canonical transcript length



Appendix – RNA molecules per cell

Type	Percent of total RNA by mass	Molecules per cell	Average size (kb)	Total weight picograms/cell	Notes	Reference
rRNAs	80 to 90	3–10 × 10^6 (ribosomes)	6.9	10 to 30		Blobel and Potter (1967), Wolf and Schlessinger (1977), Duncan and Hershey (1983)
tRNA	10 to 15	3–10 × 10^7	<0.1	1.5 to 5	About 10 tRNA molecules /ribosome	Waldron and Lacroute (1975)
mRNA	3 to 7	3–10 × 10^5	1.7	0.25 to 0.9		Hastie and Bishop (1976), Carter et al. (2005)
hnRNA (pre-mRNA)	0.06 to 0.2	1–10 × 10^3	10*	0.004 to 0.03	Estimated at 2–4% of mRNA by weight	Mortazavi et al. (2008), Menet et al. (2012)
Circular RNA	0.002 to 0.03	3–20 × 10^3	~0.5	0.0007 to 0.005	Estimated at 0.1–0.2% of mRNA**	Salzman et al. (2012), Guo et al. (2014)
snRNA	0.02 to 0.3	1–5 × 10^5	0.1–0.2	0.008 to 0.04		Kiss and Filipowicz (1992), Castle et al. (2010)
snoRNA	0.04 to 0.2	2–3 × 10^5	0.2	0.02 to 0.03		Kiss and Filipowicz (1992), Cooper (2000), Castle et al. (2010)
miRNA	0.003 to 0.02	1–3 × 10^5	0.02	0.001 to 0.003	About 10^5 molecules per 10 pg total RNA	Bissels et al. (2009)
7SL	0.01 to 0.2	3–20 × 10^4	0.3	0.005 to 0.03	About 1–2 SRP molecules/100 ribosomes	Raue et al. (2007), Castle et al. (2010)
Xist	0.0003 to 0.02	0.1–2 × 10^3	2.8	0.0001 to 0.003		Buzin et al. (1994), Castle et al. (2010)
Other lncRNA	0.03 to 0.2	3–50 × 10^3	1	0.002 to 0.03	Estimated at 1–4% of mRNA by weight	Mortazavi et al. (2008), Ramsköld et al. (2009), Menet et al. (2012)

**The size for the average unspliced pre-mRNA is 17 kb; however, most pre-mRNAs are partially spliced at any given time, and the average size of hnRNA is estimated at 10 kb (Salditt-Georgieff et al., 1976).*

***Based on the finding that 1–2% of all mRNA species generate circular RNA, which is present at 10% of the level of the parental mRNA.*

https://www.frontiersin.org/files/Articles/127231/fgene-06-00002-HTML/image_m/fgene-06-00002-t001.jpg

Appendix – qPCR as confirmation?

OAS/RNASE L: SENSING VIRAL PAMP TRIGGERS GLOBAL RNA DEGRADATION AND TRANSLATIONAL ARREST

Degrading viral genomes presents one potent method of antiviral activity; digesting viral genetic material ensures that no further steps in replication can occur. However, the challenge lies within being able to control RNA degradation to ensure cellular survival or limit destruction within the host. The OAS/RNase L pathway is activated upon sensing the PAMP of dsRNA, serving two functions: sensing viral intruders and inhibiting viral replication by degrading RNA almost indiscriminately, inducing global translational arrest.

Note: Pinkham et al. “confirmed” some of the RNAseq results using the ΔΔCt method and 18 S ribosomal RNA!

Activation of RNase L quickly arrests global translation. This rapid translational arrest is traditionally attributed to degradation of transcripts involved in host translation machinery, as evidenced by degradation of 28S and 18S rRNA upon RNase L activation. However, closer examination revealed

1.Yang, E., and Li, M. (2020). All About the RNA: Interferon-Stimulated Genes That Interfere With Viral RNA Processes. Frontiers in Immunology 11. 10.3389/fimmu.2020.605024.