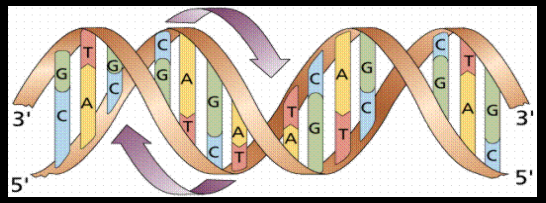
Normalization of RNA-Seq Data: Are the ERCC Spike-In Controls Reliable?

Joint work with
Sandrine Dudoit, Davide Risso and John Ngai,
UC Berkeley.

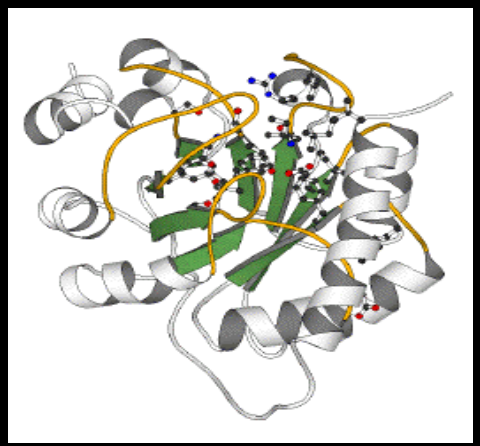
2014 AMSI-SSAI Lecture

The central dogma (not quite these days)

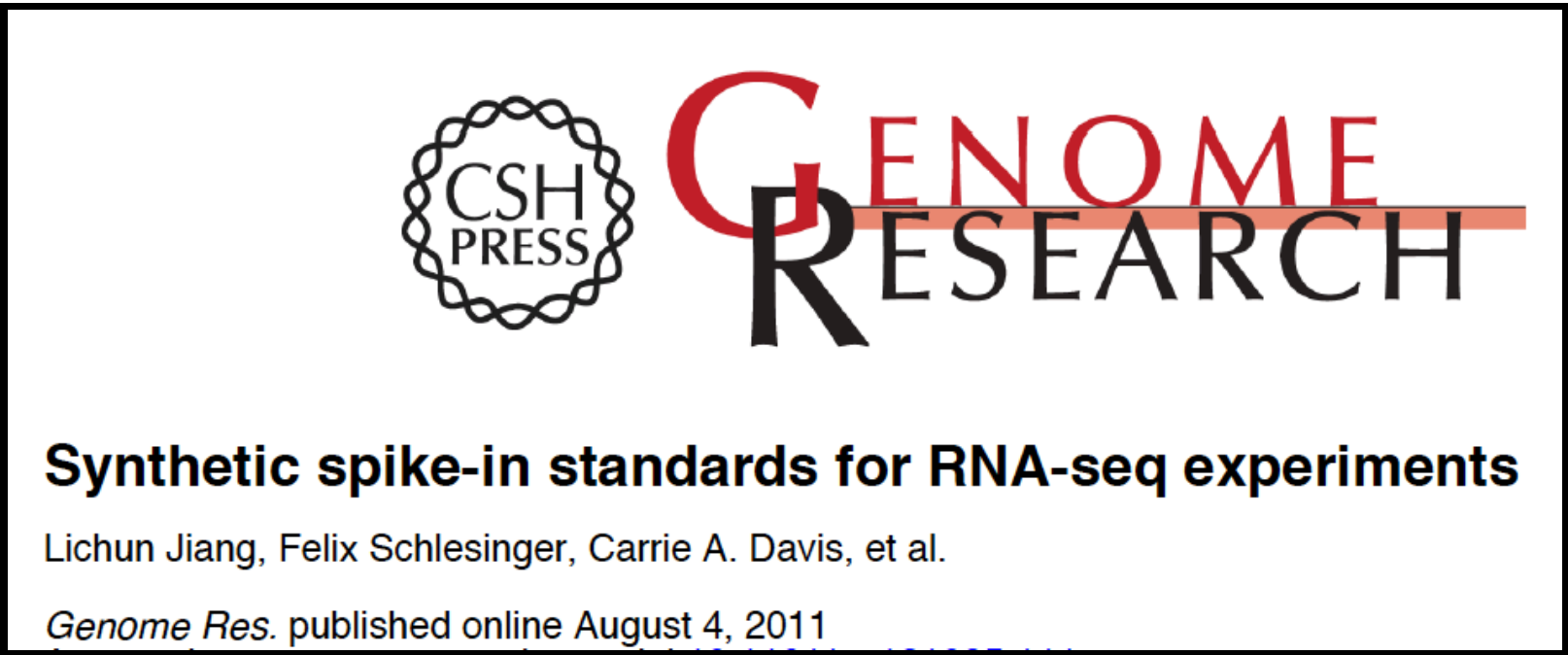


DNA CCTGAGCCAACTATTGATGAA

mRNA CCUGAGCCAACUAUUGAUGAA



Protein PEPTIDE



The External RNA control consortium (ERCC) developed a set of 92 polyadenylated (synthetic or bacterial) transcripts that mimic natural eukaryotic mRNAs:

- 250-2,000 nucleotides in length
- 5%-51% GC content
- Spiked in at various concentrations prior to library prep
- Provide positive and negative controls for RNA-seq

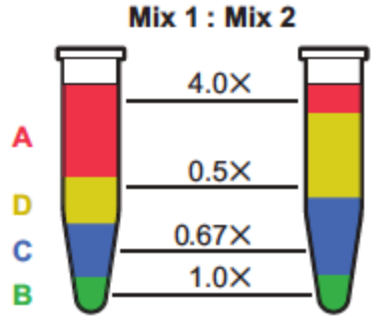
Ambion's two commercial mixes

Start with purified total RNA, poly(A), or rRNA-depleted RNA

↓

Add Spike-In Mix 1 ***or*** Mix 2 to the RNA sample(s)

↓



- They contain the same 92 standards, at different concentrations
- Each group of 23 transcripts span an approx 10^6 concentration range

Today I will

- Evaluate the performance of the ERCC spike-in standards
- Use the spike-ins to evaluate normalization methods that do not use them, and
- See whether we can normalize RNA-seq data using the spike-ins.

We have two very different data sets: zebrafish and SEQC. I'll spend most of my discussion on the first.

The zebrafish project

Broad goal: To investigate mechanisms governing odorant receptor gene expression in zebrafish. More fully, to

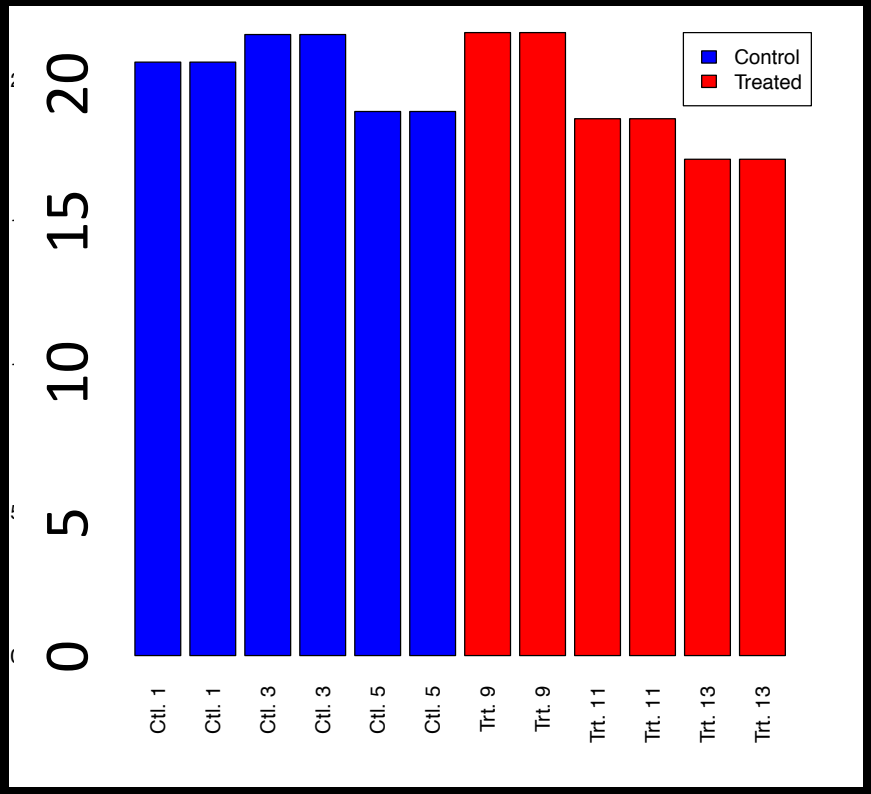
- Study differential expression (DE) between suitable cells from control and gallein-treated zebrafish embryos
- The drug gallein inhibits Gβγ-protein signaling and suppresses olfactory receptor expression.
- The cell were sorted by FACS for GFP fluorescence to identify the subset in which a plasmid was present
- RNA-seq was done using Illumina HiSeq 2000, with sample multiplexing and 100 bp paired-end reads.

The zebrafish dataset

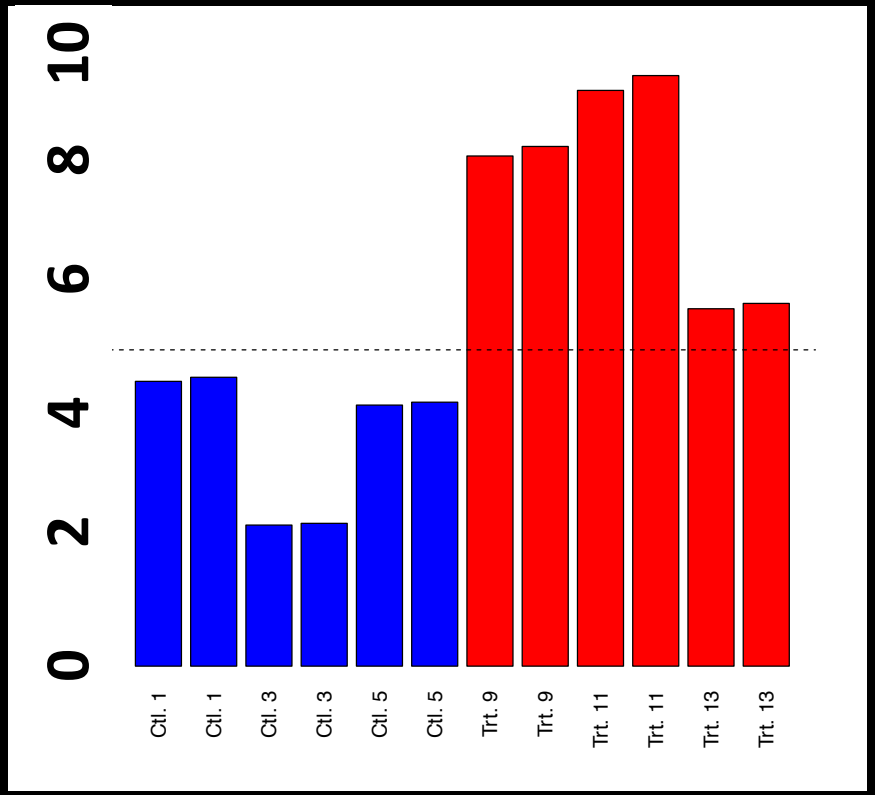
- 3 control and 3 treated pools of zebrafish cells: one library preparation for each pool
- Control and treated libraries are paired by prep date.
- For each of two sequencing runs, a multiplex pool of the 6 libraries sequenced in a single lane (Dec 1 and 20, 2012) :
  2 sample types x 3 libraries x 2 runs = 12 datasets.
- Ambion ERCC Spike-in Mix 1 added to the RNA prior to library prep.
- Technical aspects prior to library prep (e.g. FACS cell sorting) cannot be captured by the spike-in controls.

Fish/ Library	Condition	Library prep. date	Sequencing run date
S1	Control	1/18/2012	12/01/2012 12/20/2012
S3	Control	1/24/2012	12/01/2012 12/20/2012
S5	Control	1/31/2012	12/01/2012 12/20/2012
S9	Treated	1/18/2012	12/01/2012 12/20/2012
S11	Treated	1/24/2012	12/01/2012 12/20/2012
S13	Treated	1/31.2012	12/01/2012 12/20/2012

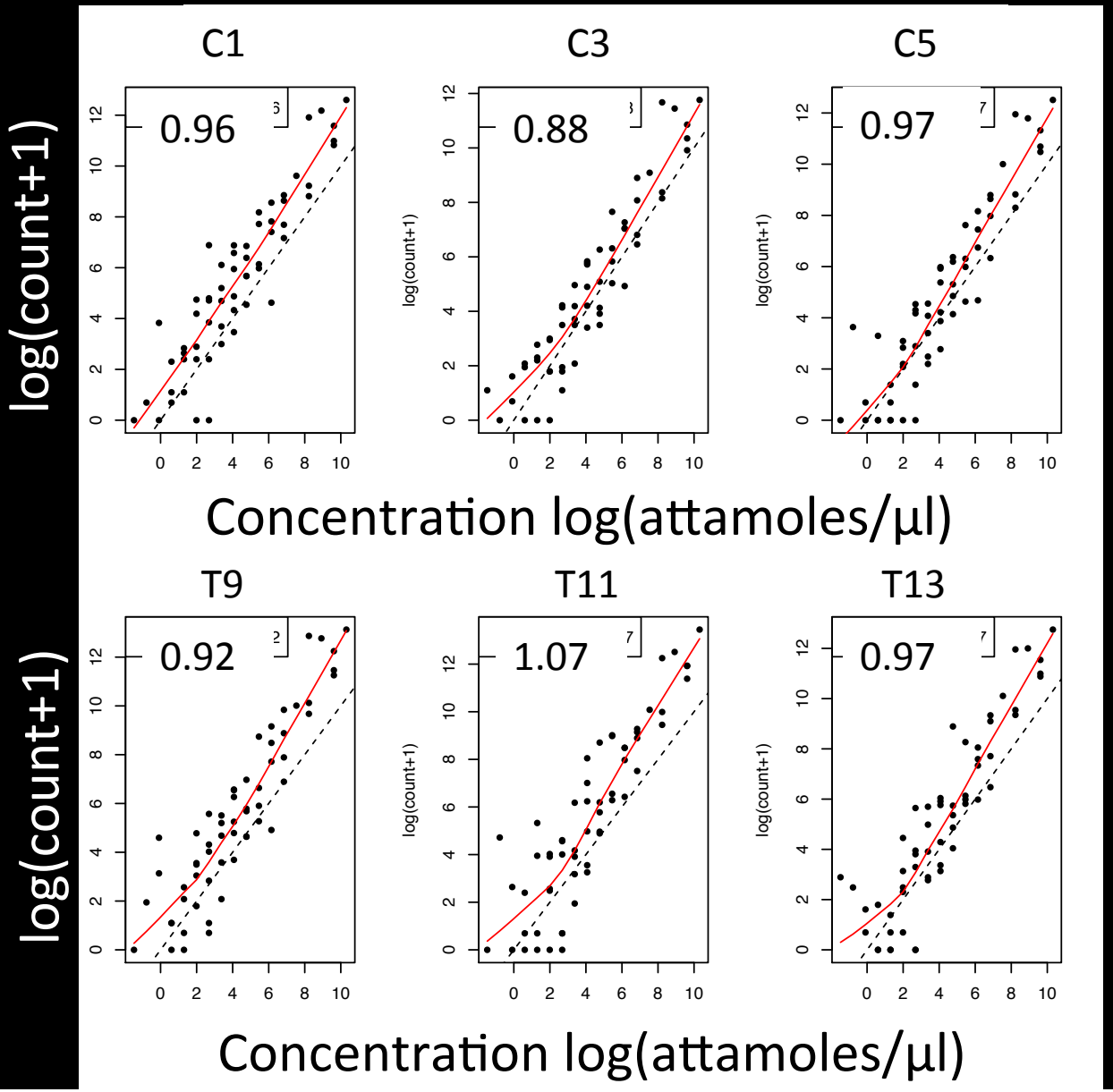
#M of mapped reads



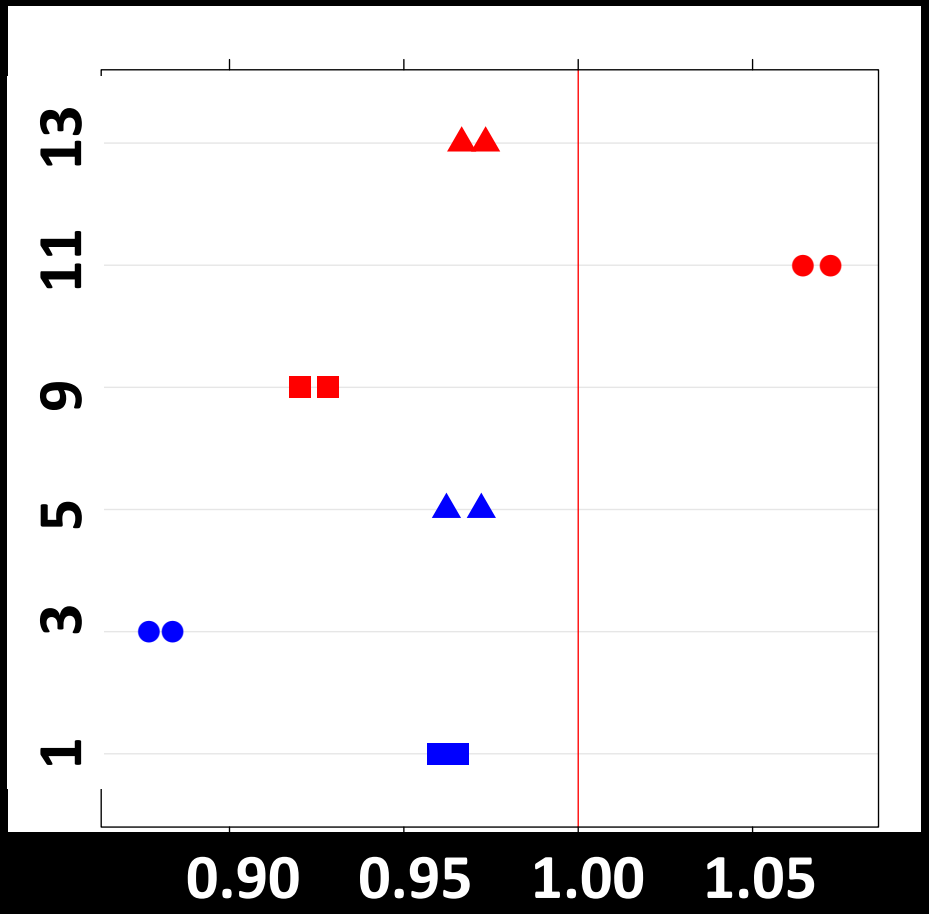
% ERCC spike-ins



ERCC spike-ins: un-normalized read counts vs concentration (log-log)

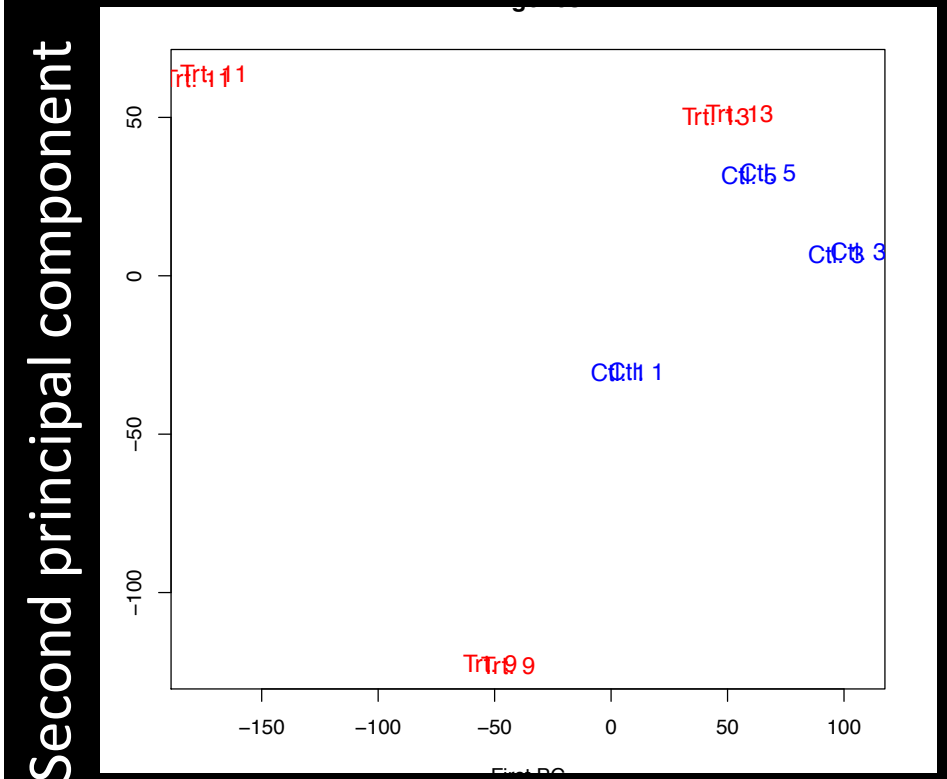


Regression coefficients in loglinear model of un-normalized counts vs concentration

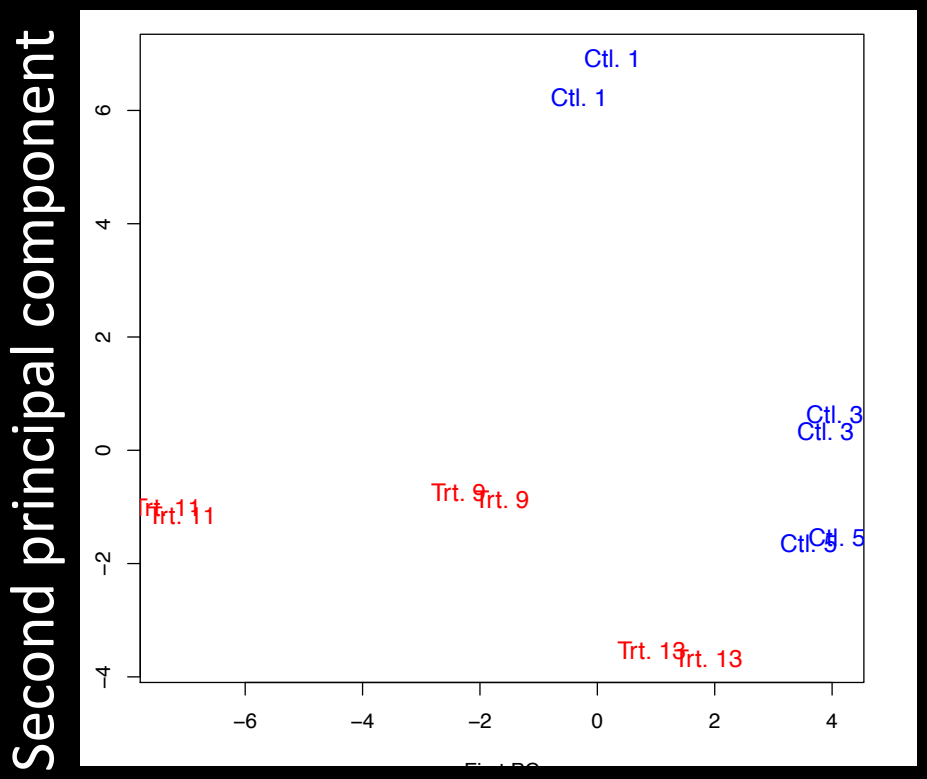


PC2 vs PC1 for the 12 sets

All genes

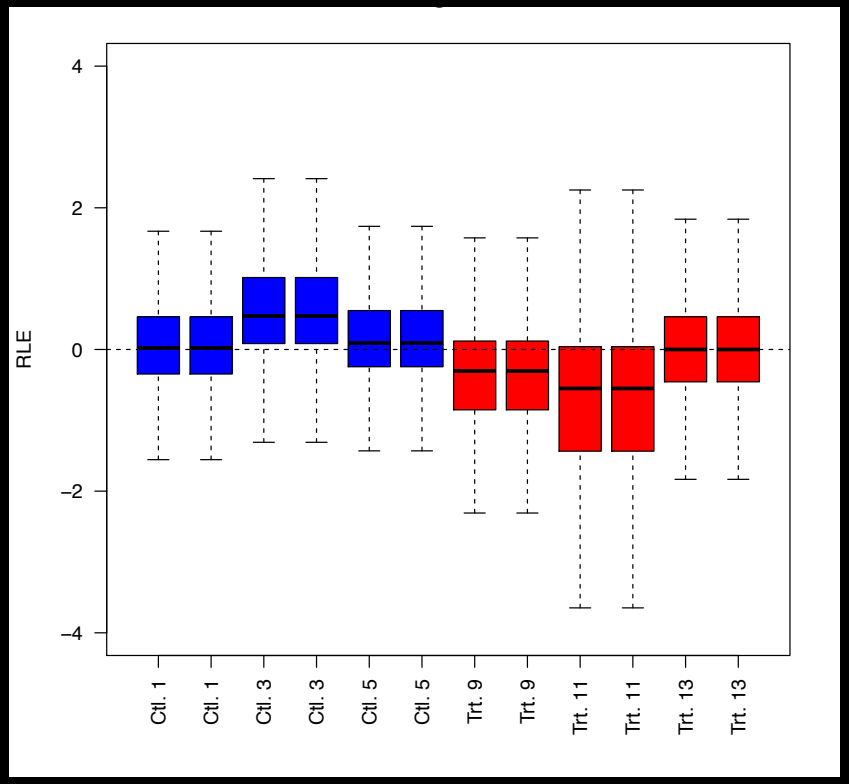


ERCC controls

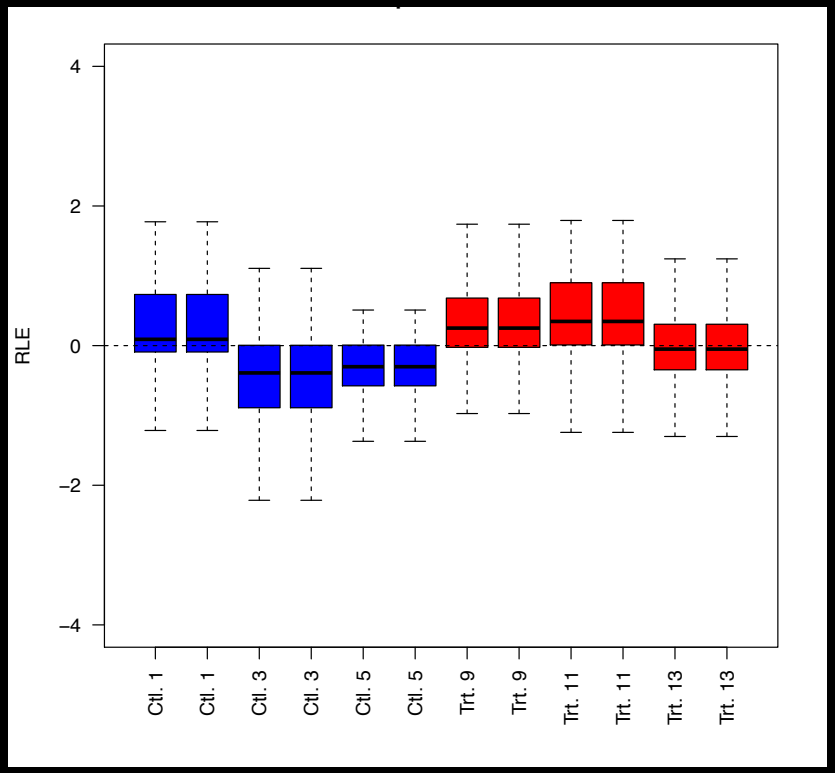


RLE plots of the 12 sets

All genes

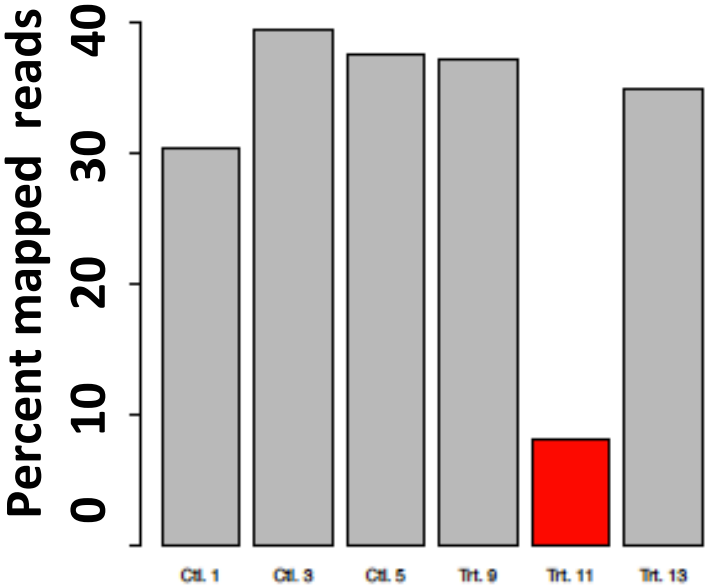


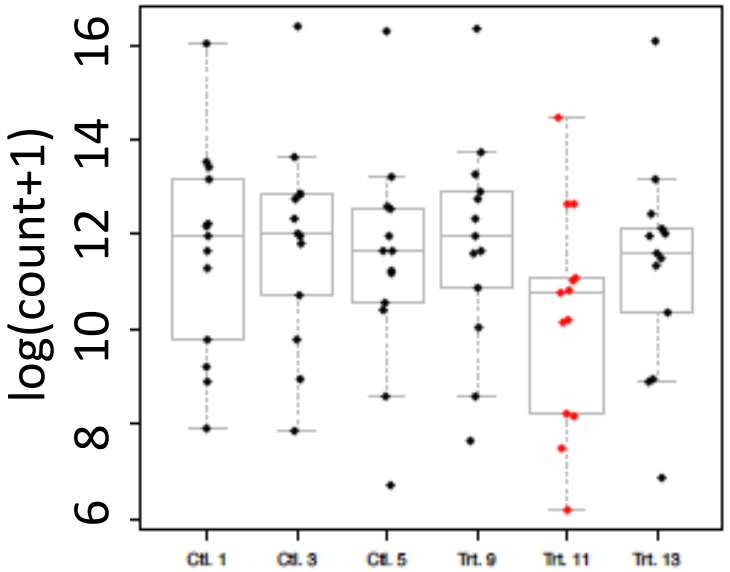
ERCC controls



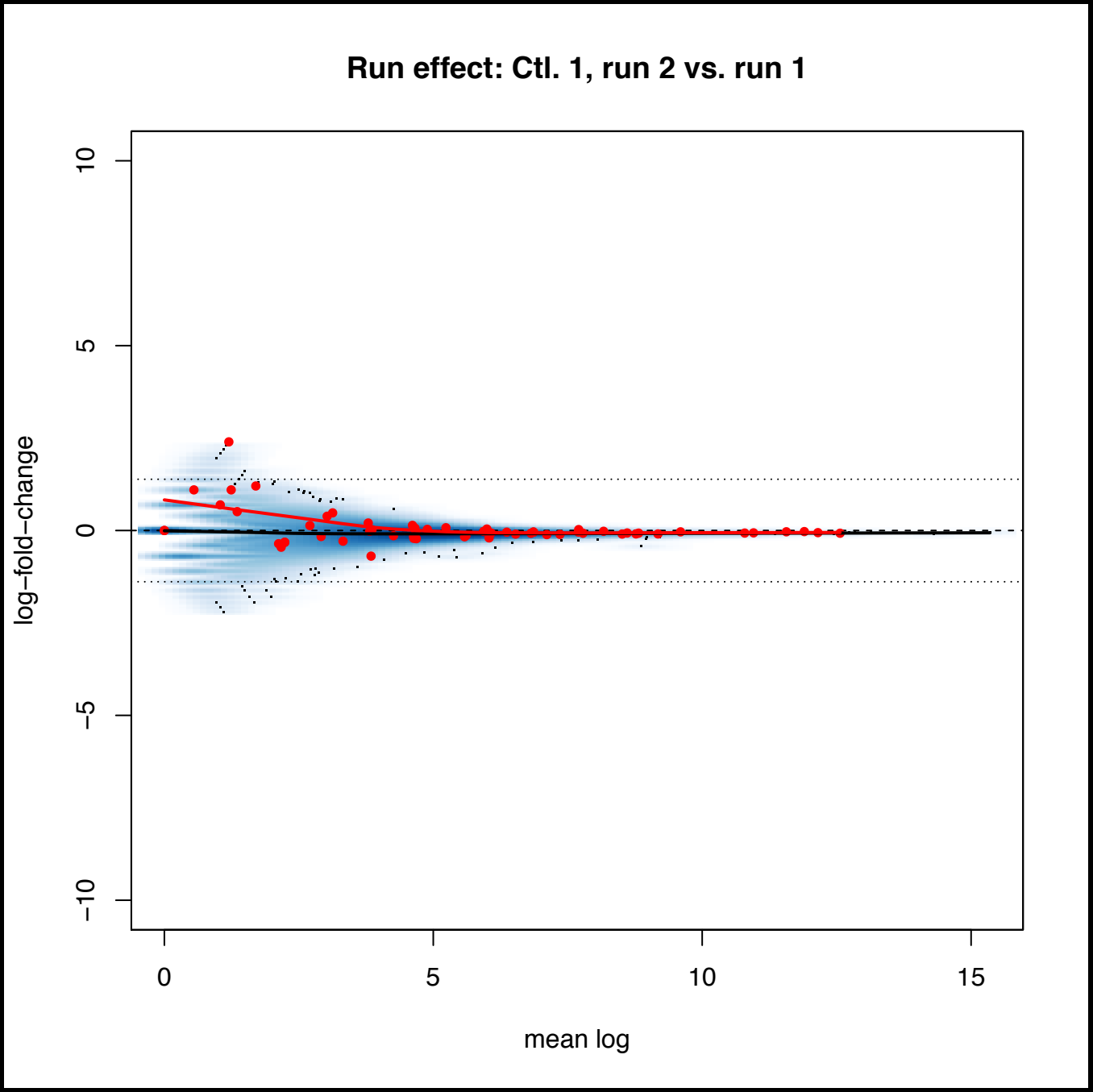
RLE = Relative Log Expression = log(count+1) – median{log(count+1)}

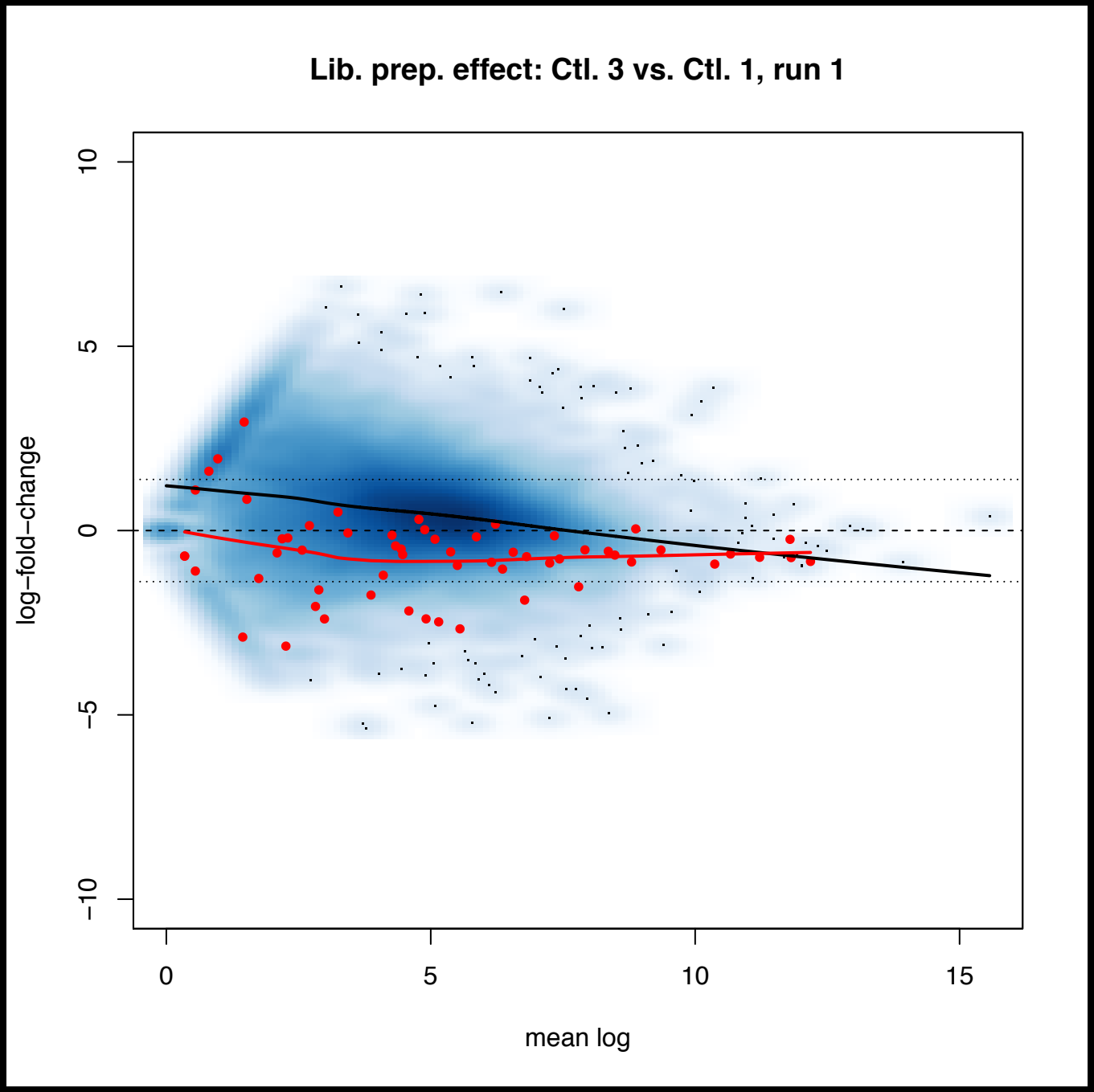
Mitochondrial genes in the 6 samples

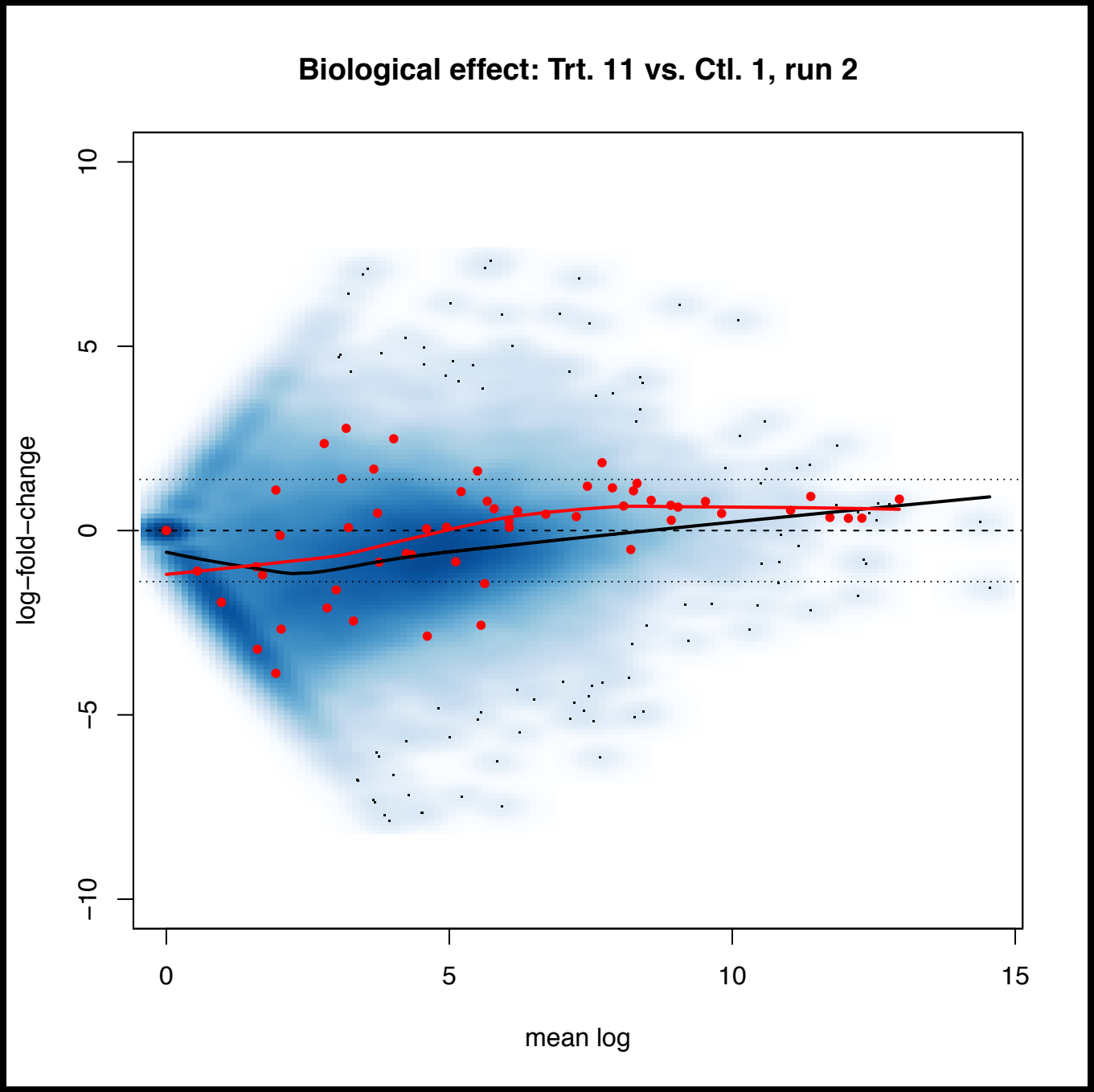




Sample	Ctl.1	Ctl.3	Ctl.5	Trt.9	Trt.11	Trt.13
No. of Cells from FACS	49K	29K	25K	37K	70K	16K
Total RNA (ng)	63	81	52	49	126	31
After polyA+ from 25 ng (pg)	147	115	91	99	145	117


Run effect: Ctl. 1, run 2 vs. run 1
log−fold−change
10
5
0
−5
−10
0
5
10
15
mean log


Lib. prep. effect: Ctl. 3 vs. Ctl. 1, run 1
log−fold−change
10
5
0
−5
−10
0
5
10
15
mean log


Biological effect: Trt. 11 vs. Ctl. 1, run 2
log−fold−change
10
5
0
−5
−10
0
5
10
15
mean log

Summary of ERCC spikes for zebrafish data

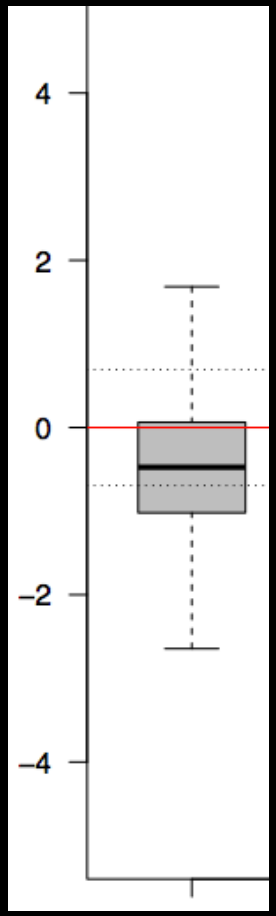
- There is a fair-good linear relationship between (log) read count and concentration, except at the low end
- The % reads mapped to the controls is highly variable between library preparations, and deviates markedly from the nominal proportions (seen before, Qing *et al* 2013)
- Plots of individual counts across samples show high variability for lower concentration spike-ins
- Both the genes and the controls have similar read counts across runs but not library preparations
- The controls do not capture all technical effects (especially library preparation)
- The ERCC controls exhibit a treatment-control difference. Why? Interaction with sample RNA? Different proportions of poly A?

That was all about the ERCC spikes.

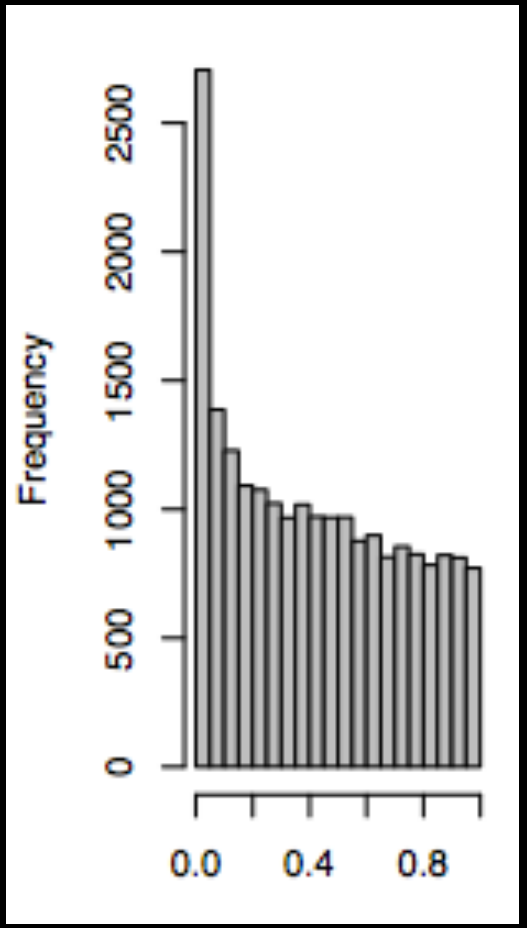
What about the original aim, comparing gallein-treated zebrafish embryos to controls?

There *is* a problem testing *trt v. ctl*, so something is needed: call it *normalization*

Log fold change



THIS IS WITH THE RAW DATA



p-value histogram

Can normalization help us, either without or with the ERCC controls?

We do not discuss within sample (GC- or gene-length) normalization, just between samples.

Between-sample normalization methods

- Total Count (**TC**) = RPKM without the PK
- Upper Quartile (**UQ**), Bullard et al 2010
- Full Quantile (**FQ**), Bullard et al 2010
- Trimmed Mean of M-values (**TMM**), Robinson & Oshlack (2010)
- Relative Log Expression (**AH**), Anders & Huber (2010)
- Cyclic loess (**CL**) on MA-plots of log-counts for pairs, or (not cyclic) on each sample w.r.t. a synthetic reference (when on the spikes, Loven et al 2012)

TC, UQ, TMM and AH all scale linearly

TC: by the sum of the counts;

UQ: by the upper quartile;

TMM: by the weighted mean log-ratio of each sample to the reference (after trimming extremes), where the sample whose UQ is closest to the mean UQ is used as reference;

AH: by the median log-ratio of each sample to the reference, where the geometric mean of all samples is the reference (i.e. using the RLE plot)

Remove Unwanted Variation-2 for RNA-seq

Uses the log-linear model (GLM)

$$log\ E(Y) = W\alpha + X\beta$$

where Y is the matrix of gene-level read counts, X is the design matrix of "wanted variation", and W is the unobserved matrix of "unwanted variation." We estimate W from the negative control genes Y_c based on

$$log\ E(Y_c) = W\alpha_c$$

How we estimate W, and how we get Y_c

As in RUV-2 for microarrays (Gagnon-Bartsch & S, Biostatistics 2012) we use the singular value decomposition

$$\log Y_c = U \Lambda V^T .$$

We estimate $W\alpha_c$ by $U\Lambda_k V^T$, where Λ_k has the first k singular values, and then we estimate W by $U\Lambda_k$.

Negative control genes can be housekeeping, spike-ins or *in silico* (aka empirically determined) controls. Care is needed in this choice (see G-B&S), as it is with k.

Below we take $k=1$, and exclude the 5,000 most DE genes to get empirical controls, or, we use the ERCC spikes.

Control-based normalizations

TC, UQ, TMM, AH, CL and RUV-2 can all be based only on the ERCC spike-in controls: 59 for zebra fish, only 14 for SEQC satisfying our filter.

This gives another set of normalizations.

Only FQ has no analogue here.

Between sample normalization methods

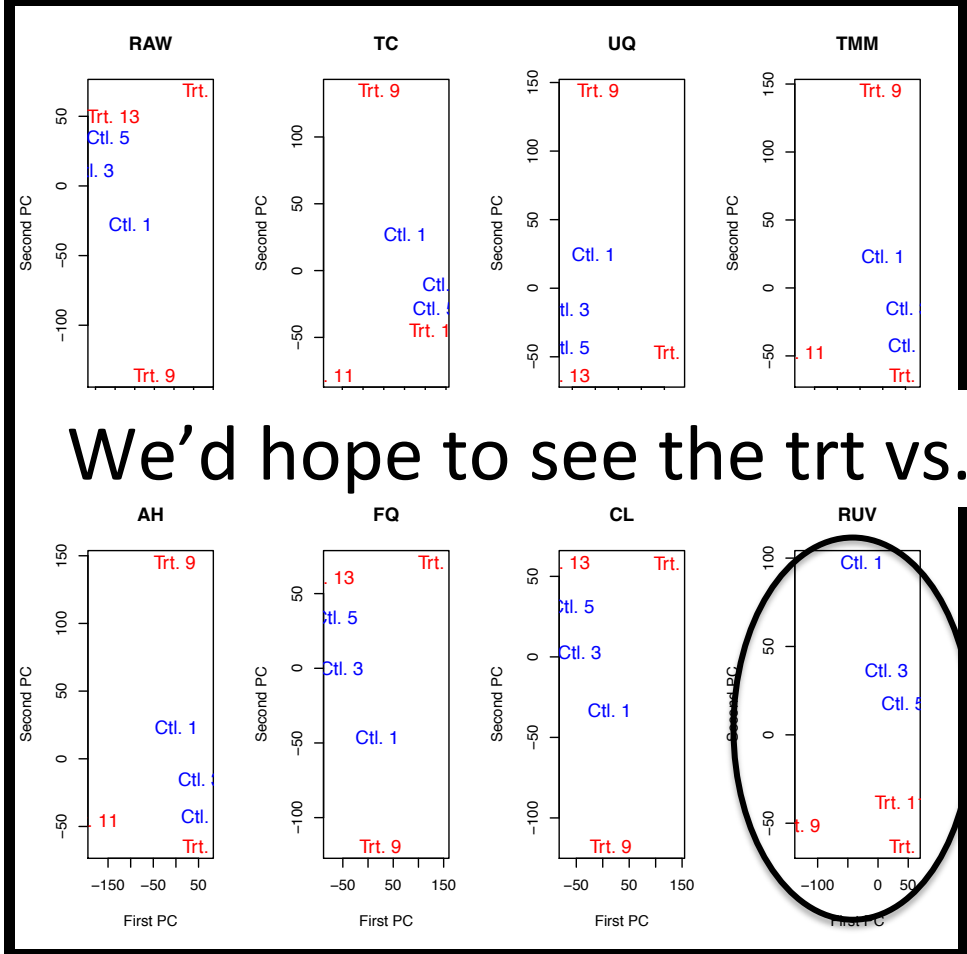
Method	All genes	ERCC negative spike-ins ZF: 59, SEQC: 14
Global-scaling		
Total-count (TC)	✓	✓
Upper-quartile (UQ)	✓	✓
Trimmed Mean of M values (TMM)	✓	✓
Anders and Huber (2010) (AH)	✓	✓
Full-quantile (FQ)	✓	✗
Cyclic-loess (CL)	✓	✓
RUV-2		
ZF	✓ 15,839 in-silico, $k = 1$	✓ $k = 1$
SEQC	✓ 16,500 in-silico, last two of $k = 3$	✓ $k = 2$

Genes were filtered out if there was not ≥ 5 reads in ≥2 samples.

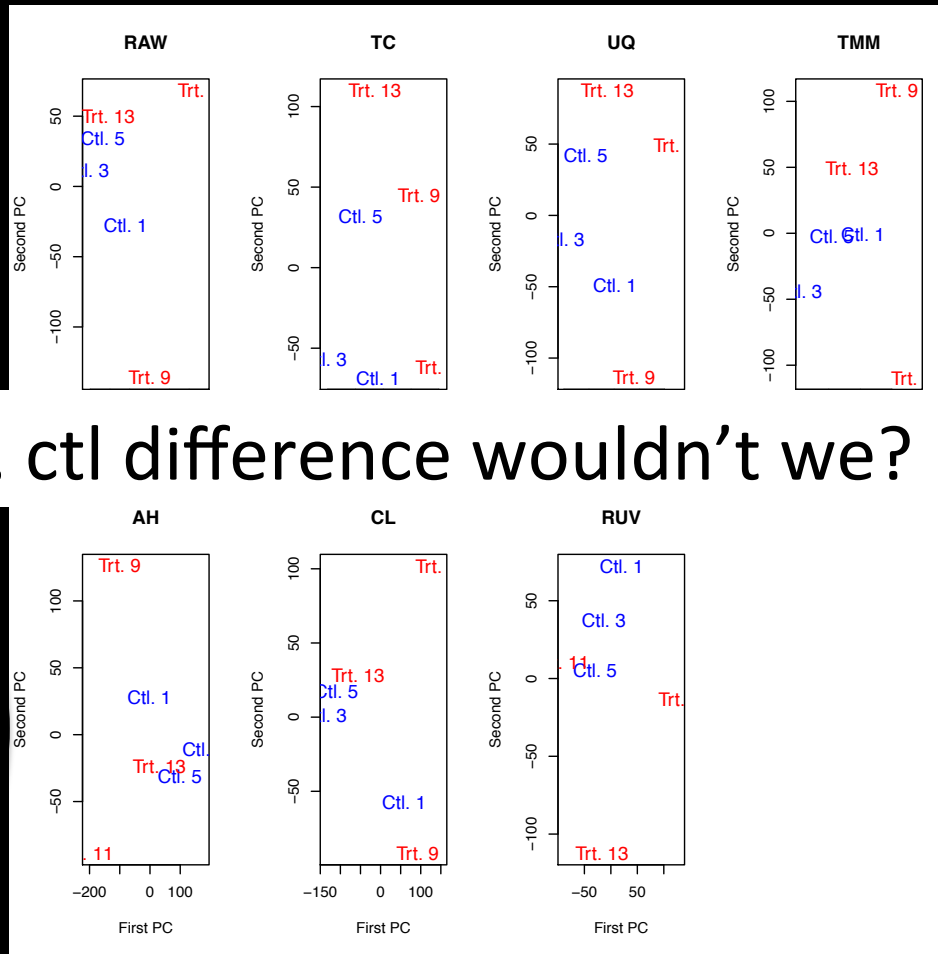
Results of normalizing using all genes, and using just the ERCC controls

PC2 vs PC1 of normalized data

Using all genes

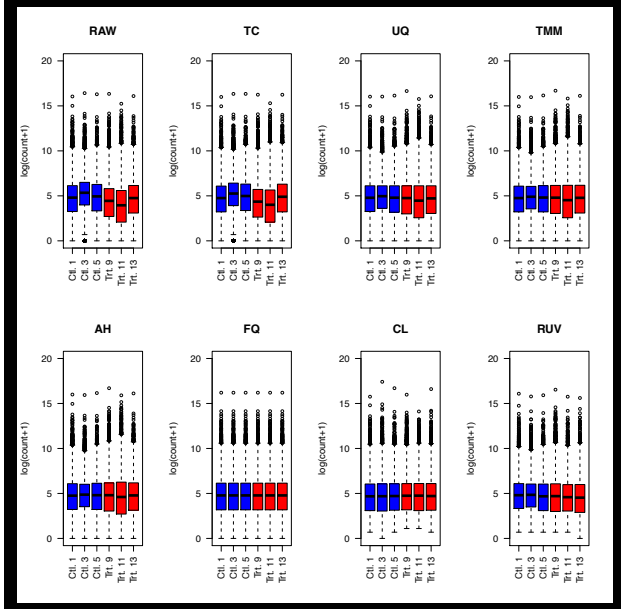


Using ERCC controls



We'd hope to see the trt vs. ctl difference wouldn't we?

Normalized gene log(counts+1)



Using

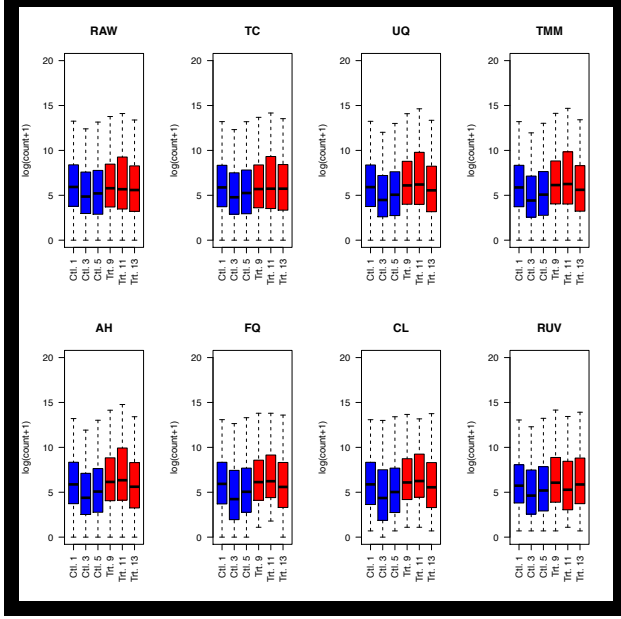
Left: Right:

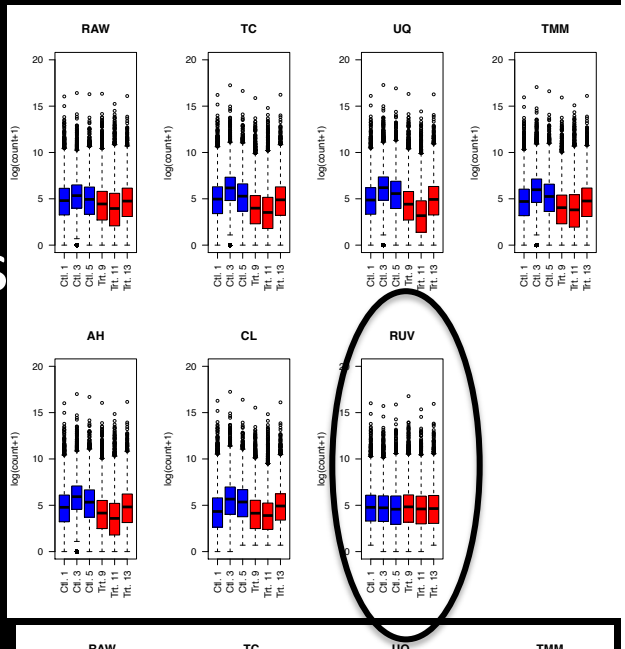
All genes ERCC Spikes

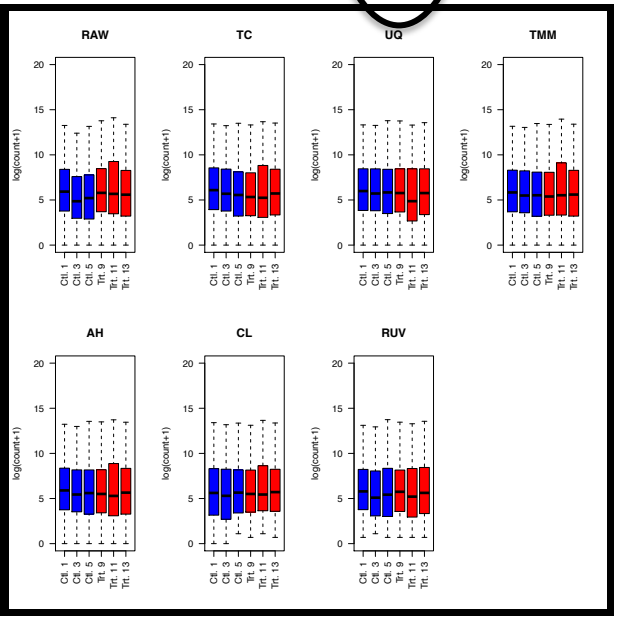
Top:All genes

Applied to

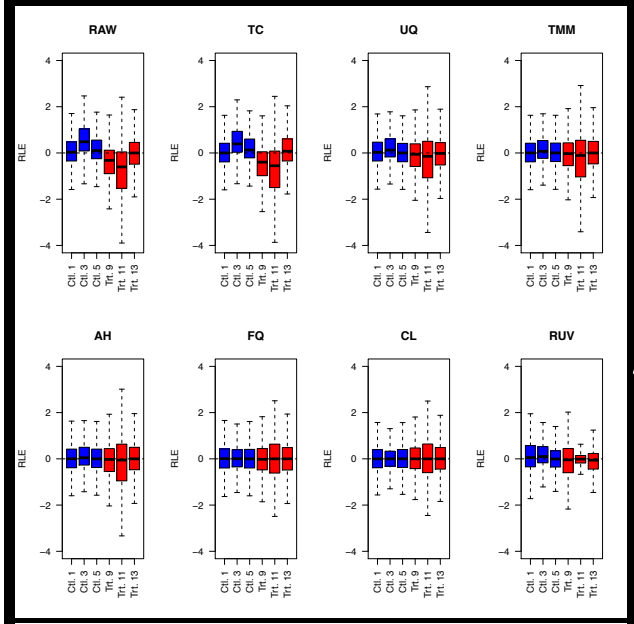
Bottom:ERCC spikes

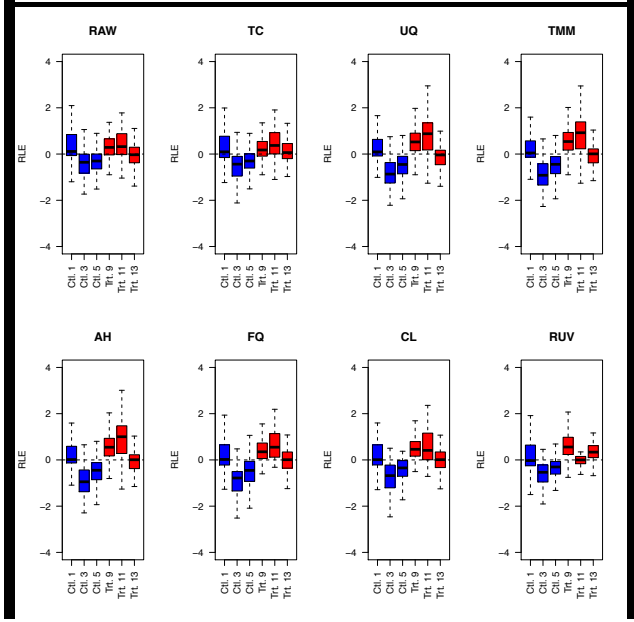






RLE plots of normalized data





Using

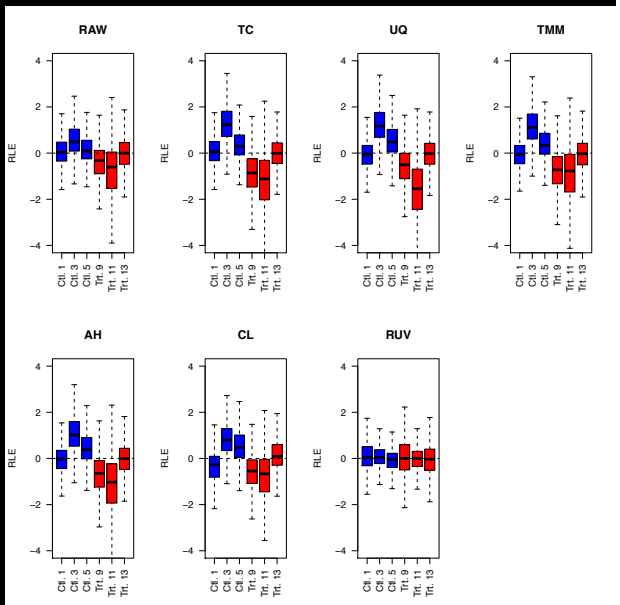
Left: Right:

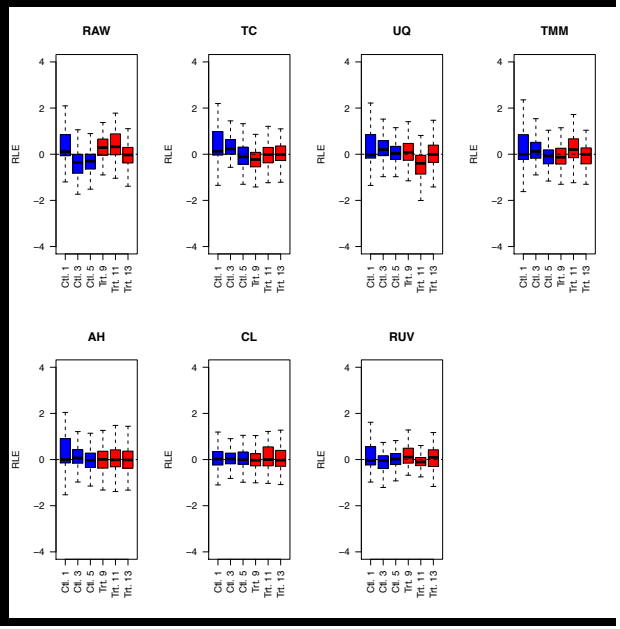
All genes ERCC Spikes

Top: All genes

Applied to

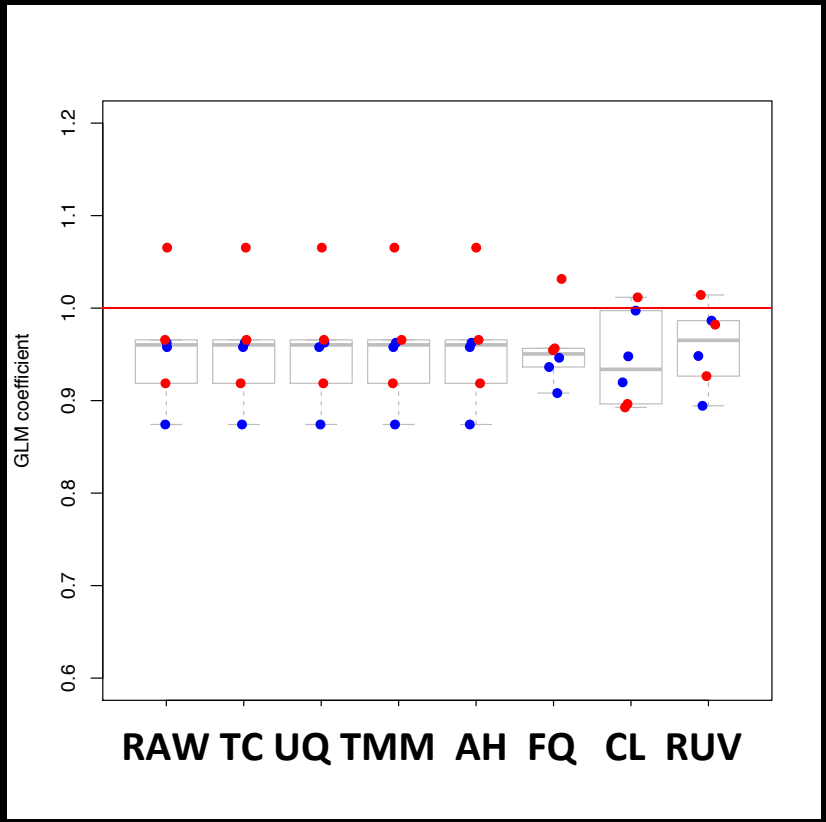
Bottom: ERCC spikes



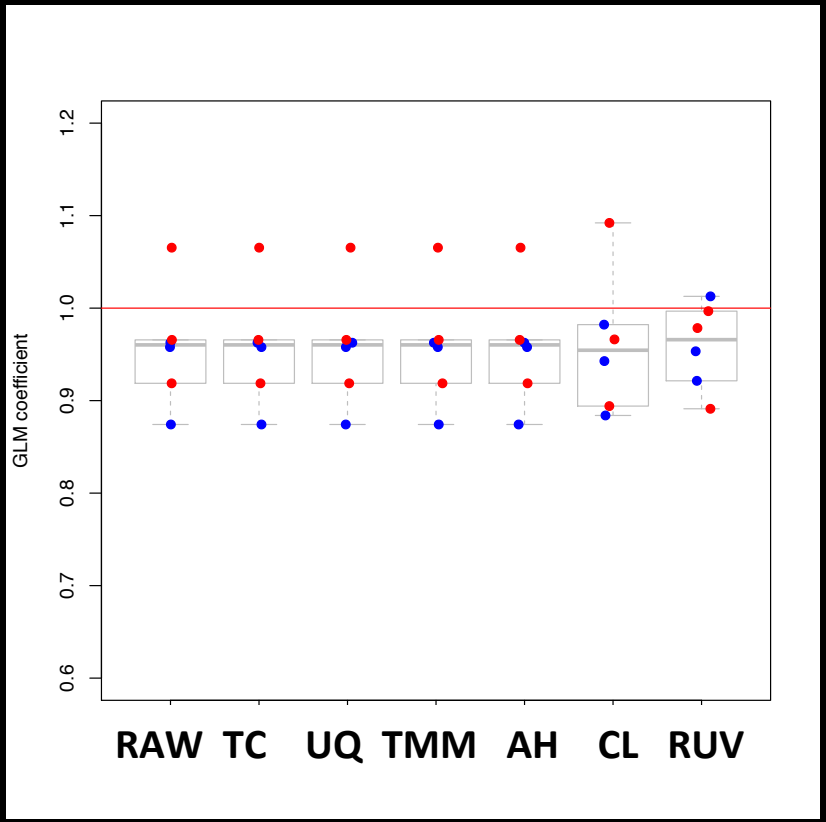


GLM slope of concentration after normalization

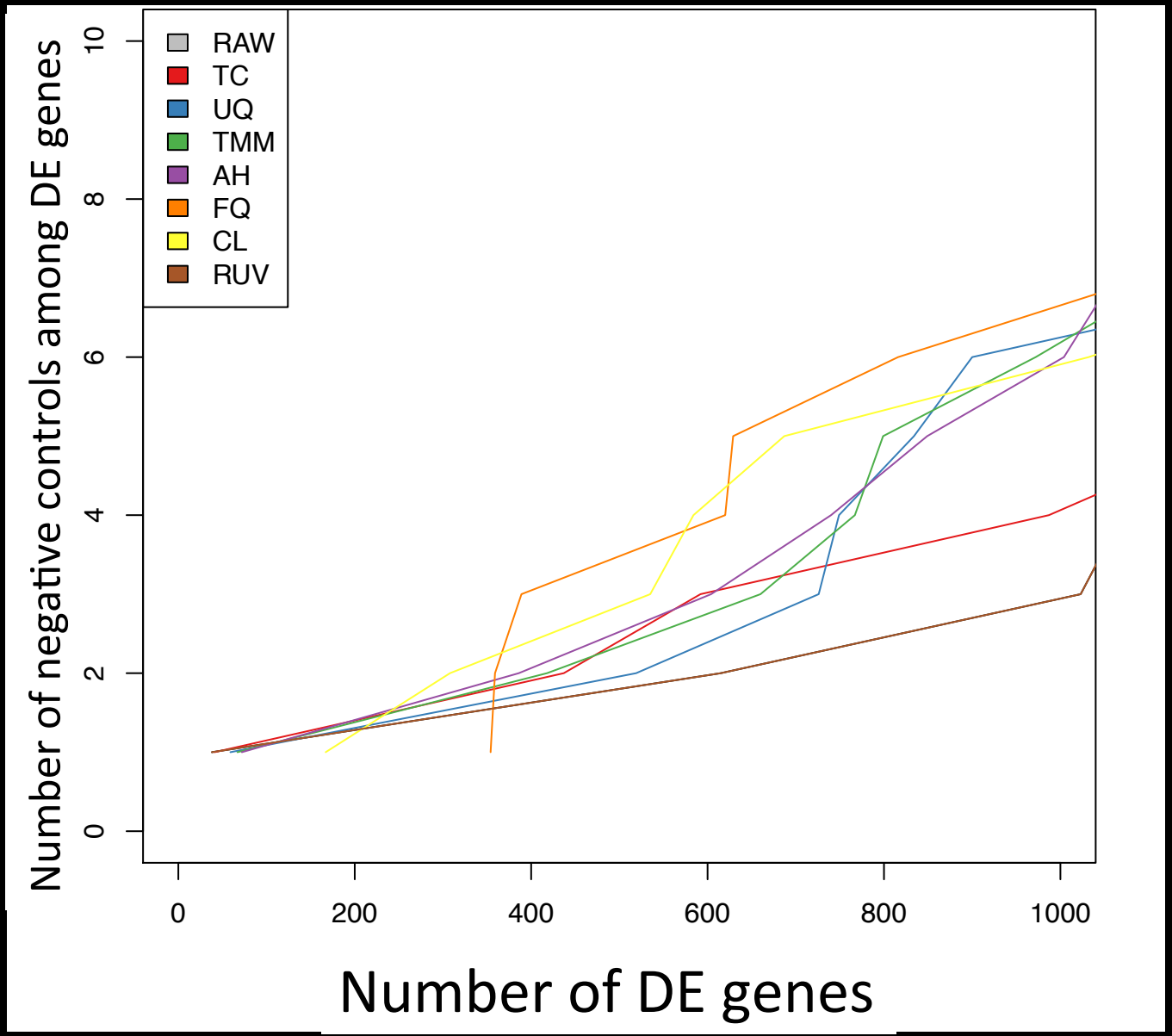
Using all genes



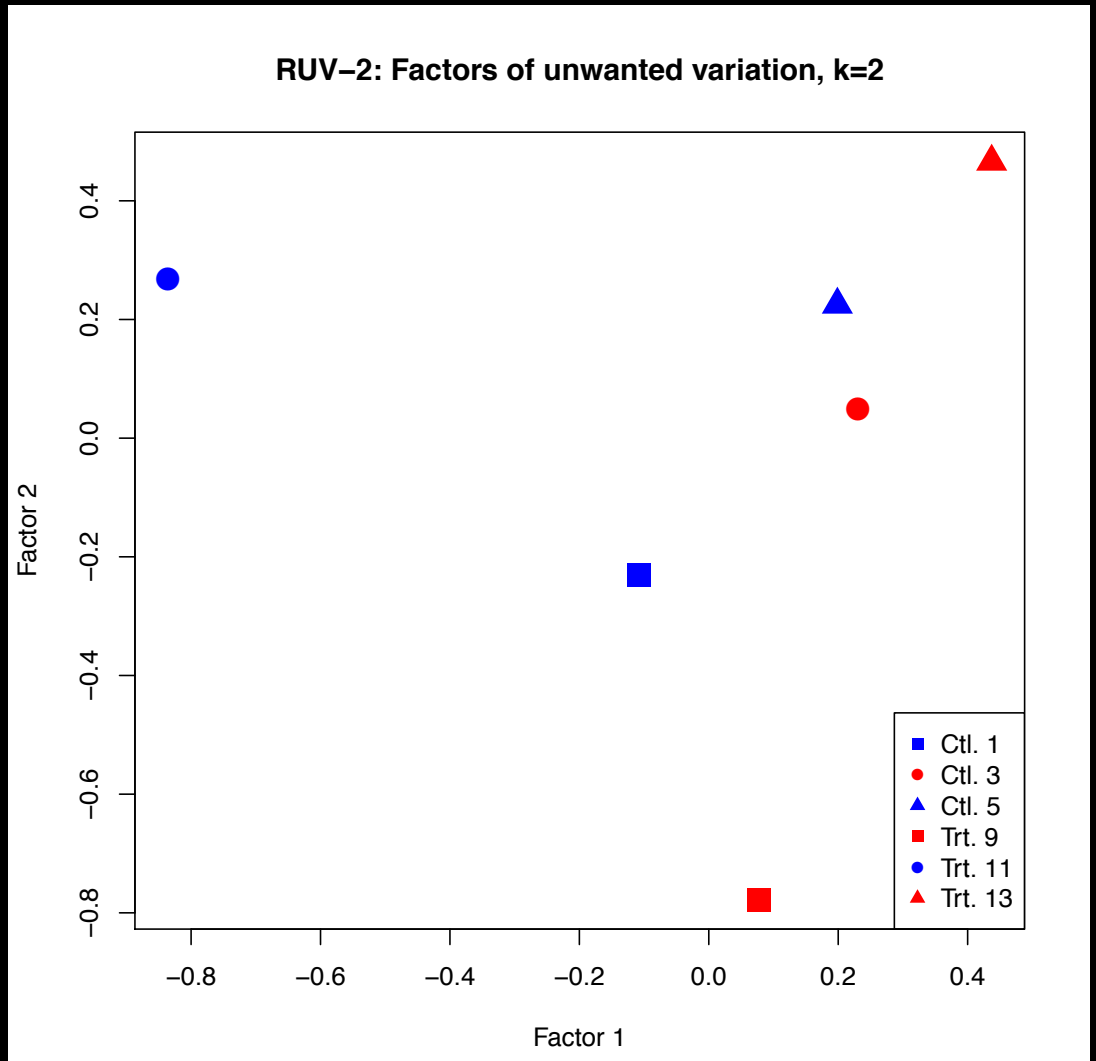
Using ERCC controls



False positive rates across normalizations

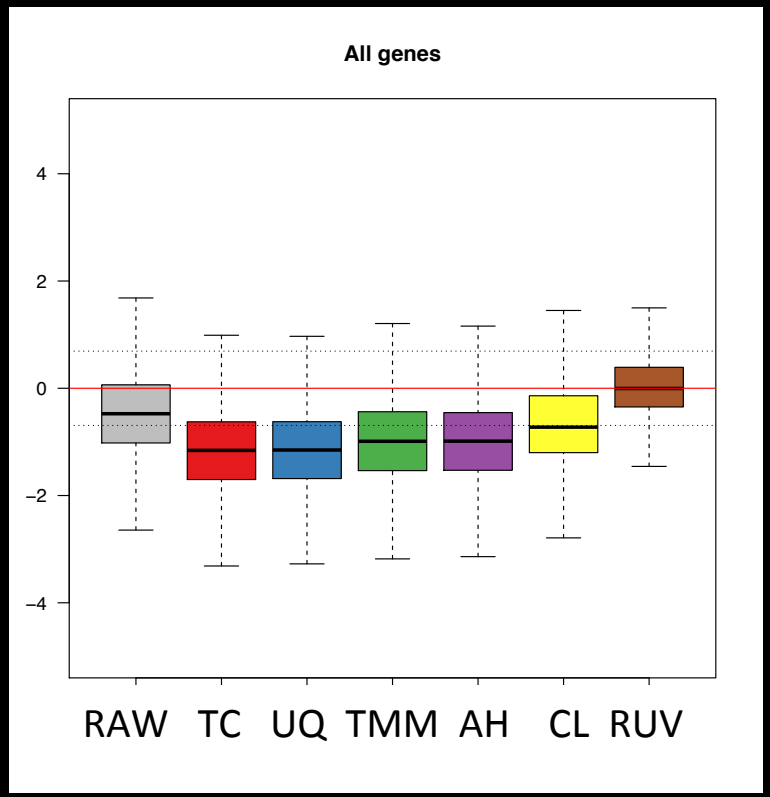


What's RUV-2 doing? Choose k=2.

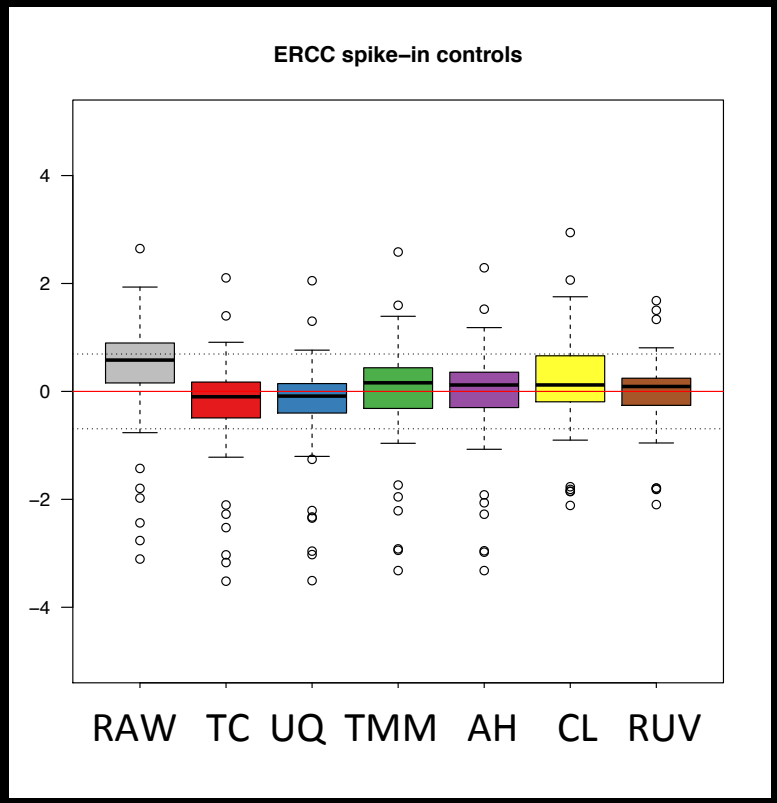


Log fold changes (treated vs control)

Using all genes

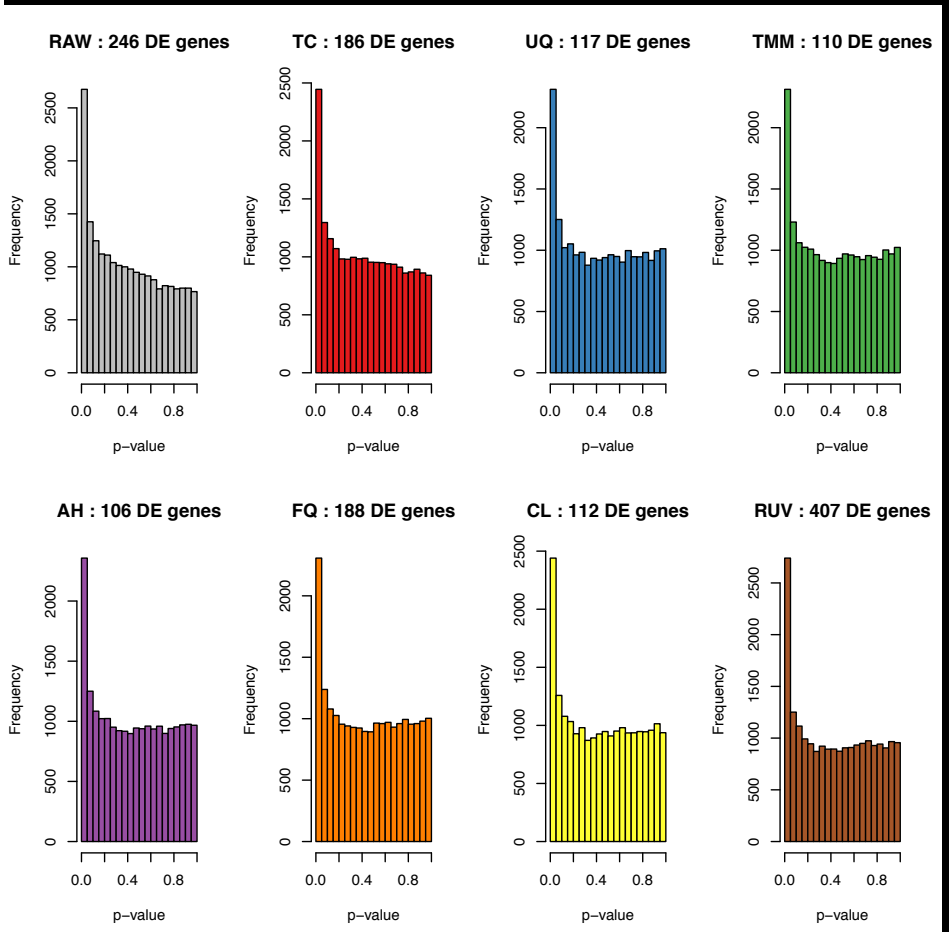


Using ERCC controls

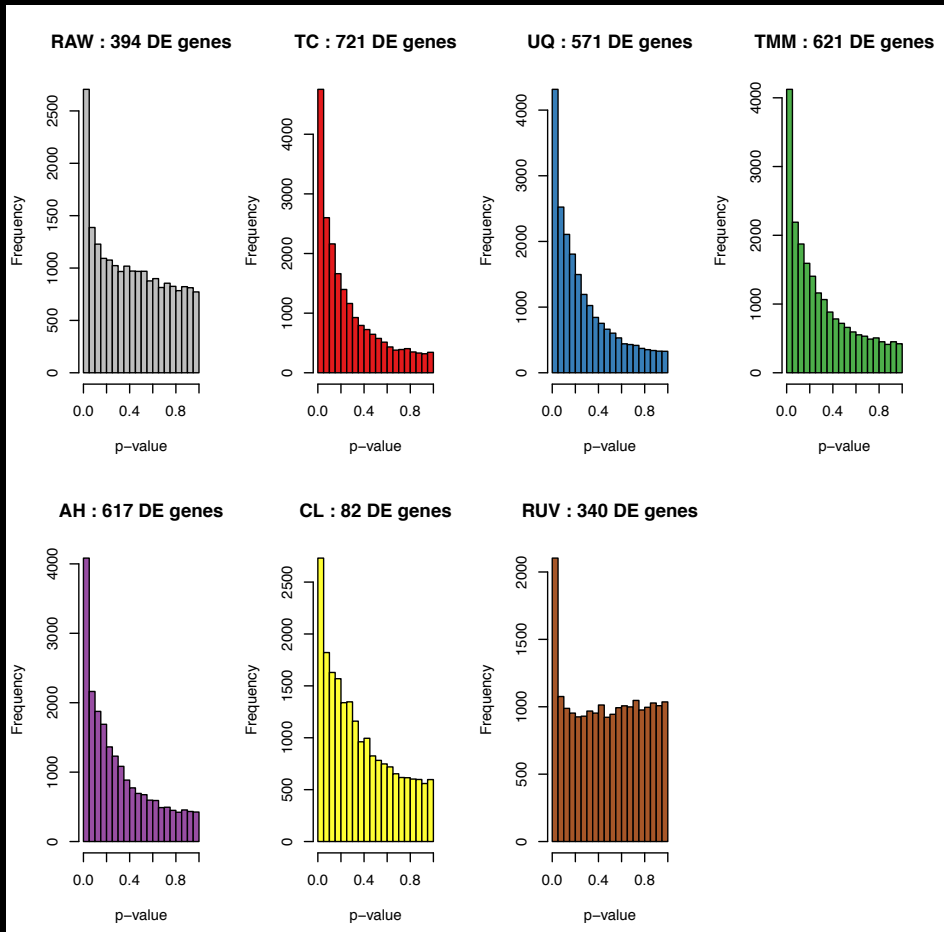


p-value histograms (testing trt v. ctl)

Using all genes



Using ERCC controls



Summary

- The ERCC spike-in controls show a high variability across replicate samples, especially at low concentrations.
- This is possibly due to differences in polyA selection efficiency.
- The ERCC spike-in controls do not fully capture the library preparation effects.
- Thus, they are not effective at benchmarking normalization methods and cannot be used to directly estimate a global normalization factor.
- Similar results were recently reported by Qing et al (2013), where the authors show a different behavior in the ERCC controls between polyA+ and RiboZero protocols.
- RUV-2 leads to surprisingly good results when using the ERCC spike-ins as negative controls and needs to be investigated in more detail.

Now let’s look briefly at the SEQC data

The Sequencing Quality Control (SEQC) project is phase III of the **MicroArray Quality Control Project** (MAQC). It provides datasets to assess the performance of platforms and algorithms. Four different types of biological samples were used, including

Sample A. Stratagene's Universal Human Reference RNA

Sample B. Ambion's human brain reference RNA.

The samples were sequenced at several facilities (17 in total) around the world and with different platforms (Illumina HiSeq 2000, Life Technologies, Roche 454).

Here, we consider Sample A and Sample B sequenced on the Illumina HiSeq 2000 (101-bp paired-end reads) at the Australian Genome Research Facility.

Four libraries were prepared for each of samples A and B.

Multiplex pools of the resulting eight libraries were sequenced in eight lanes on each of two flow-cells, yielding a total of 16 replicates per library and 64 replicates per sample type.

2 samples × 4 libraries × 2 flow-cells × 8 lanes = 128 datasets.

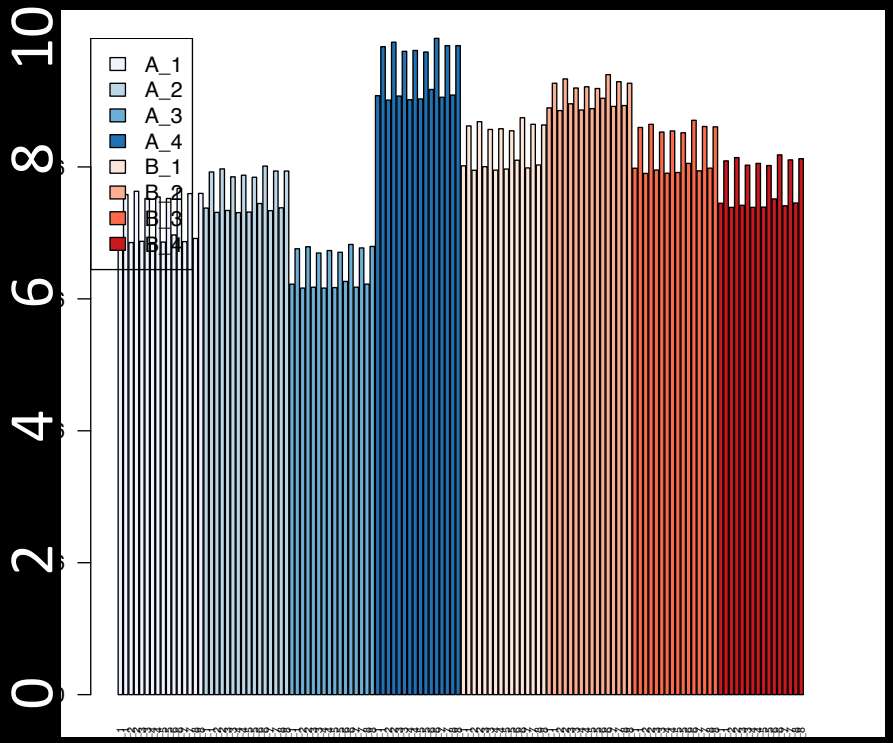
Ambion ERCC Spike-in Mix 1 was added to Sample A, and Mix 2 was added to Sample B prior to library preparation.

Controls in the SEQC data

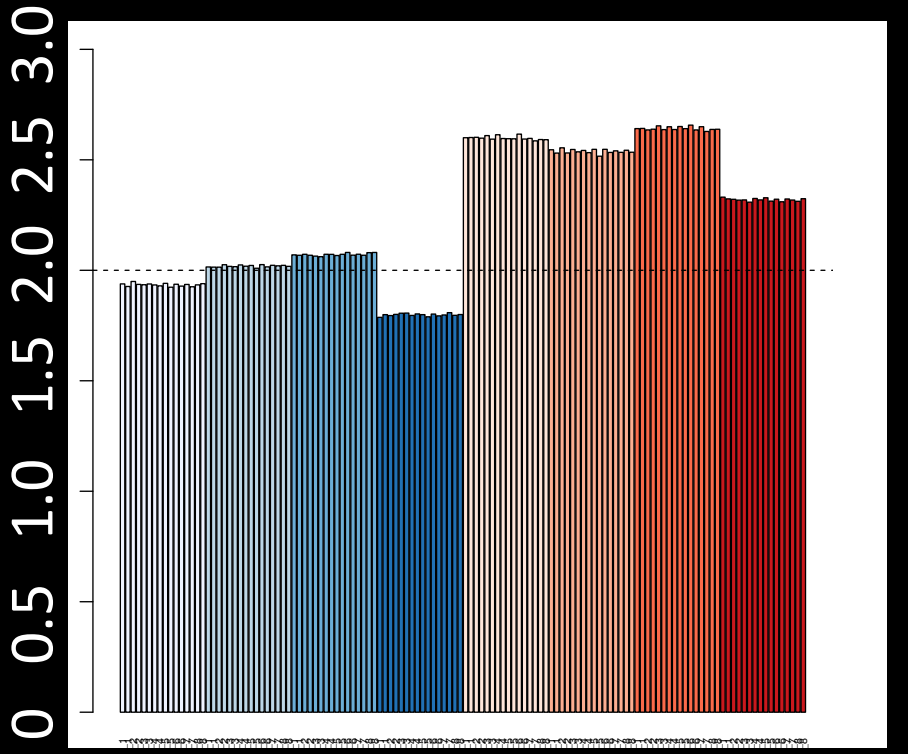
In addition to the internal ERCC spike-in controls, one can use other negative and positive controls, such as the qRT-PCR data (~1,000 genes), and microarray measures from the original MAQC study (Canales *et al,* 2006).

The SEQC datasets allow the assessments of various technical effects (e.g., platform, facility, library preparation, flow-cell). However, as with the original MAQC datasets, the UHR v. Brain comparison is rather limited, as one cannot assess biological effects in the presence of individual variability.

#M of mapped reads

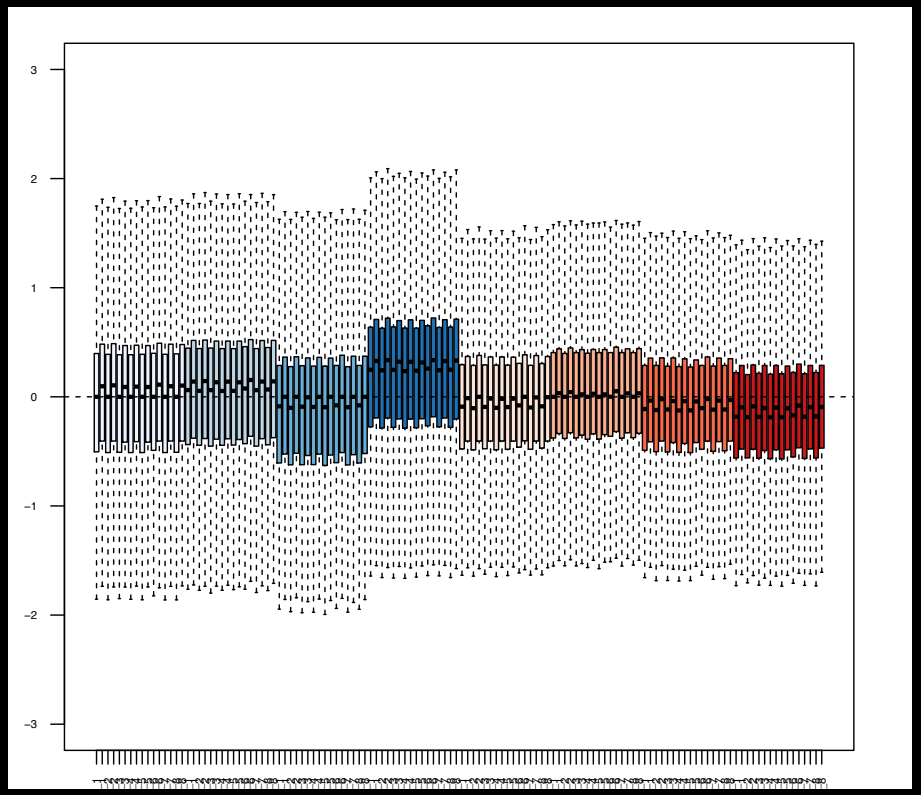


% ERCC spike-ins

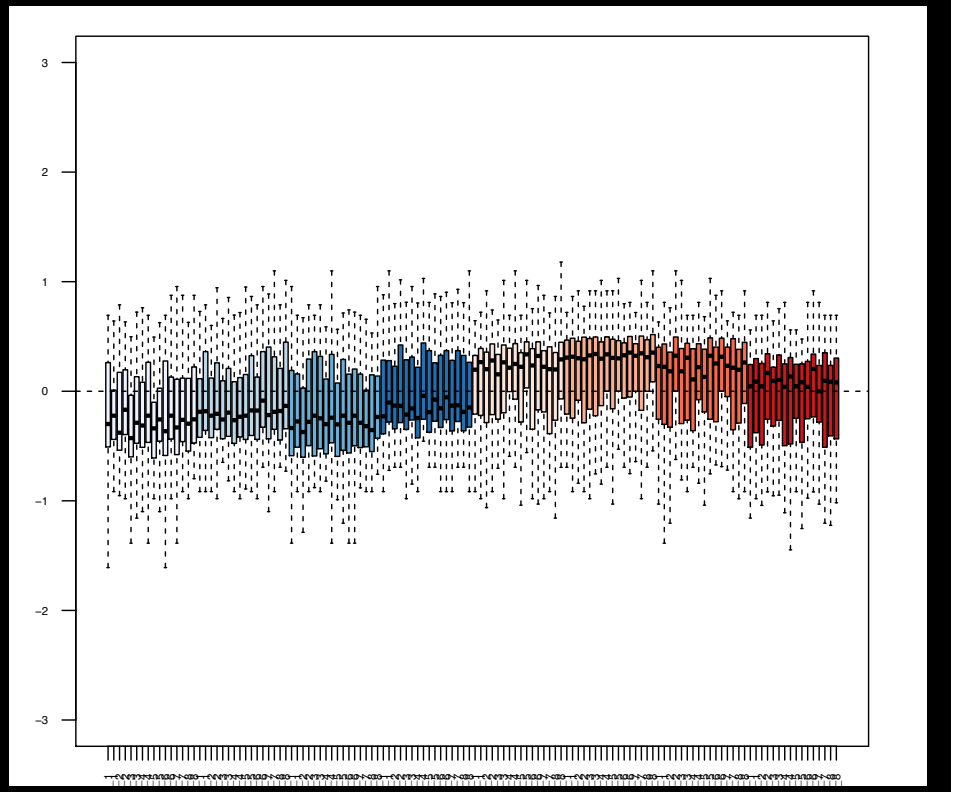


RLE plots of SEQC data

All genes

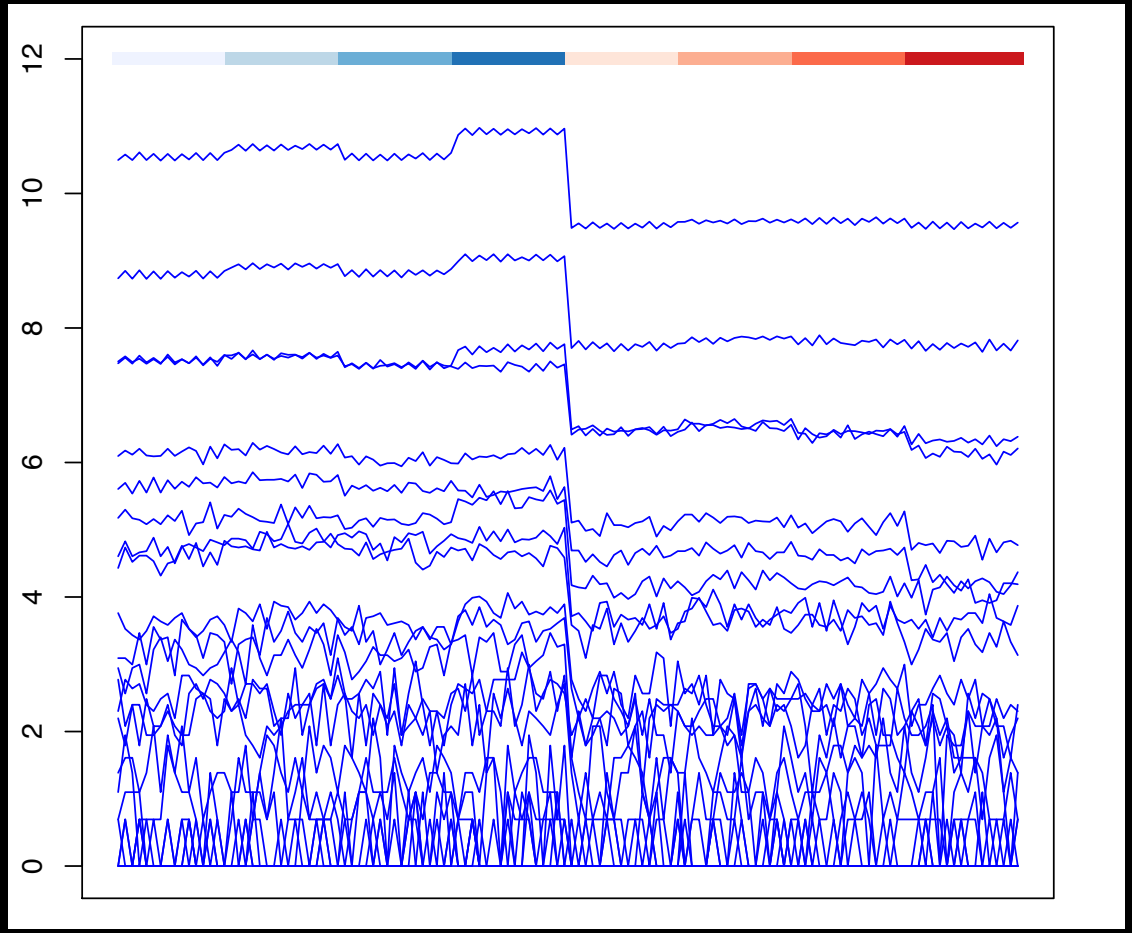


ERCC spike-ins

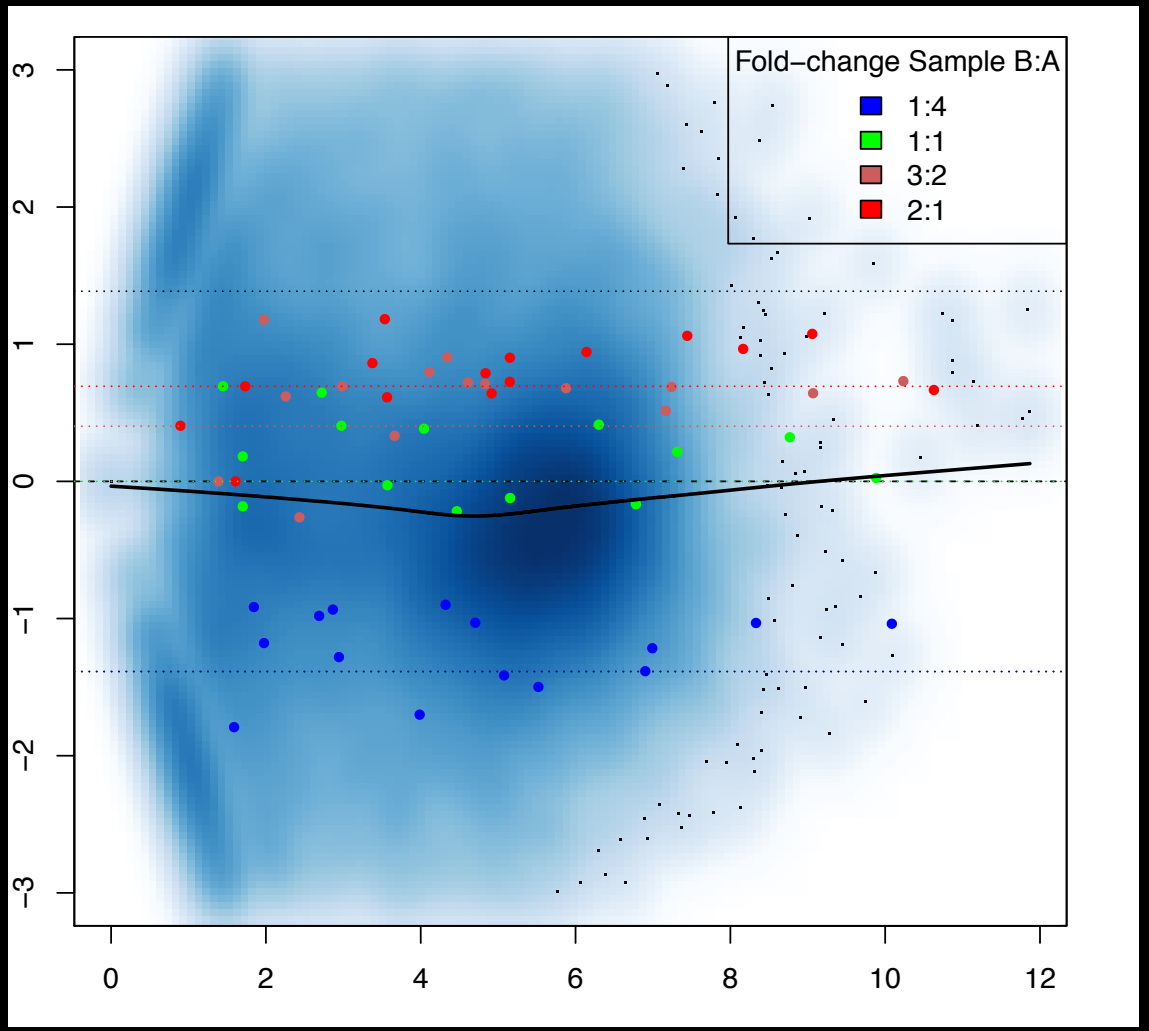


These data need normalization

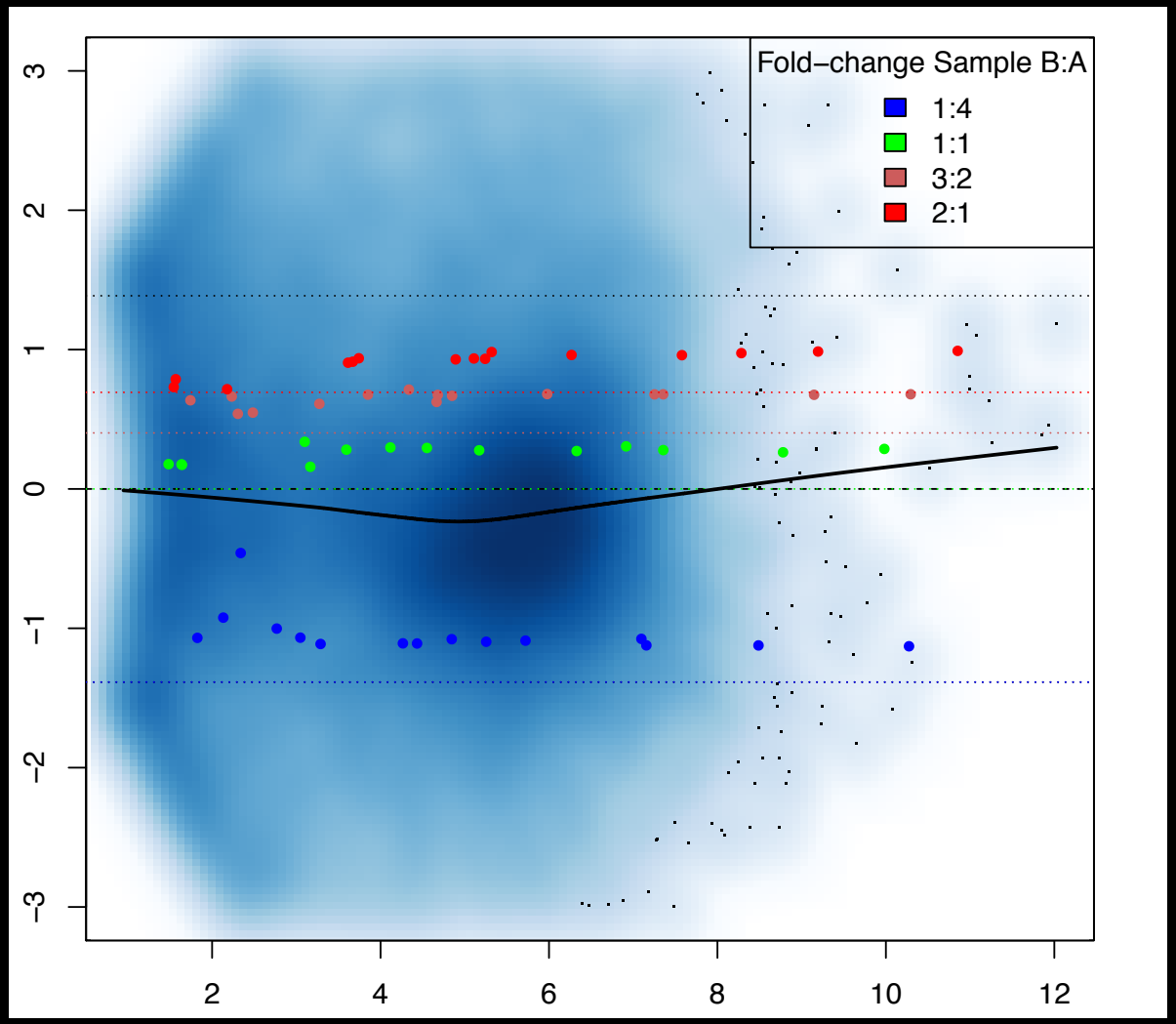
Plot of ERCC gene subgroup A where the fold-change between samples A and B 4:1



Biological differences: B4 vs A (F2, lane 8)



Biological differences: B vs A, all lanes



Summary of ERCC spikes for SEQC data
(exactly the same as for the zf data!)

- There is a fair-good linear relationship between (log) read count and concentration, except at the low end
- The % reads mapped to the controls is highly variable between library preparations, and deviates markedly from the nominal proportions (seen before, Qing *et al* 2013)
- Plots of individual counts across samples show high variability for lower concentration spike-ins
- Both the genes and the controls have similar read counts across runs but not library preparations
- The controls do not capture all technical effects (especially library preparation)
- The ERCC controls exhibit a treatment-control difference. Why?

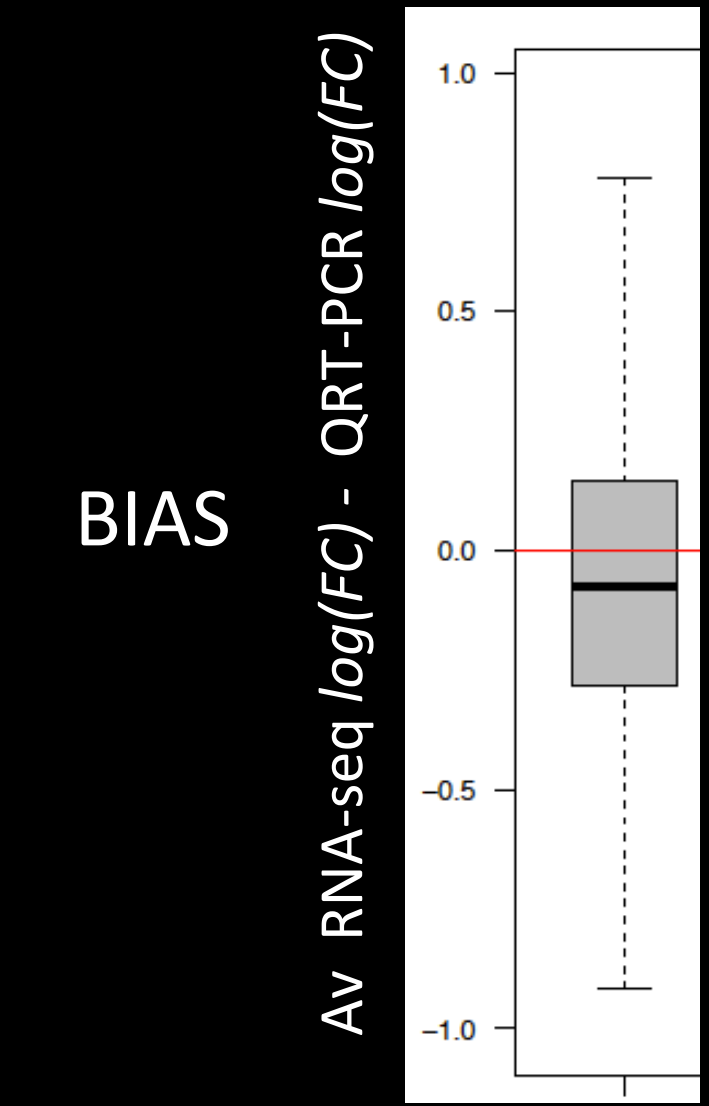
Now let’s turn to the A vs B comparisons

Some issues with the SEQC data

- Samples A and B are ***so*** different, it is not easy to identify negative controls for RUV-2 to estimate W
- We have other ways to estimate W, one involving residuals, another differences between replicates
- Even the ERCC spike mixes exhibit A-B differences
- Samples A and B are so different, it is hard to see differences in discrimination between the different normalizations in ROC curves
- Nevertheless, they are there, as we see next.

Some normalization of the SEQC data is needed for the A vs B comparison

BIAS



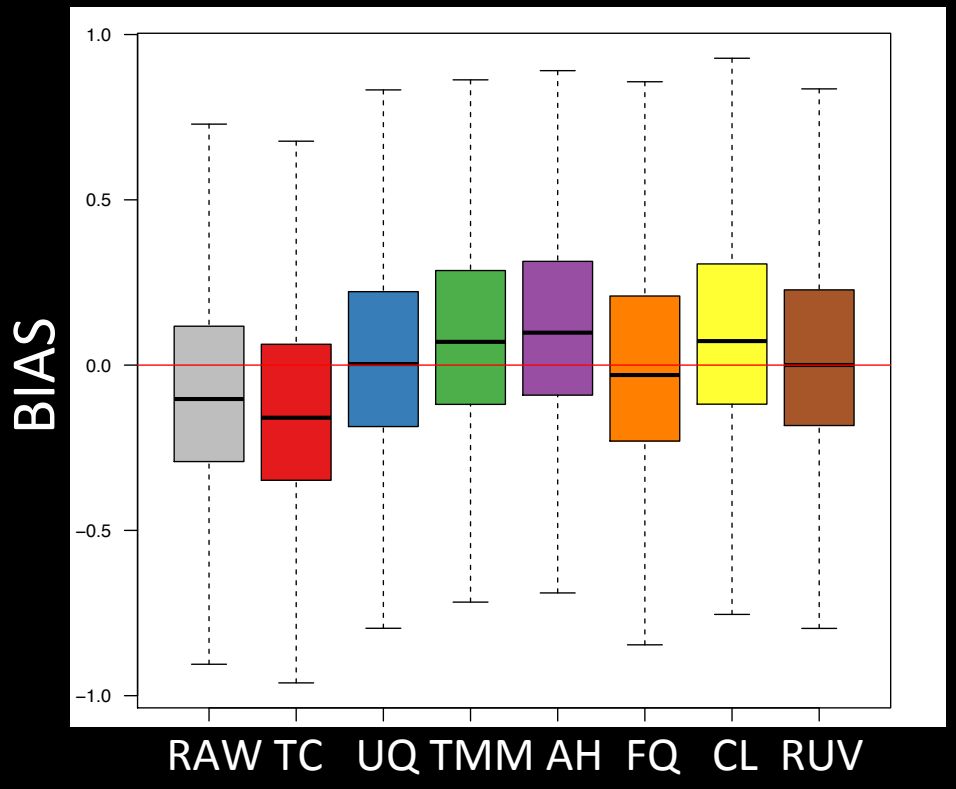
THIS IS WITH THE RAW DATA

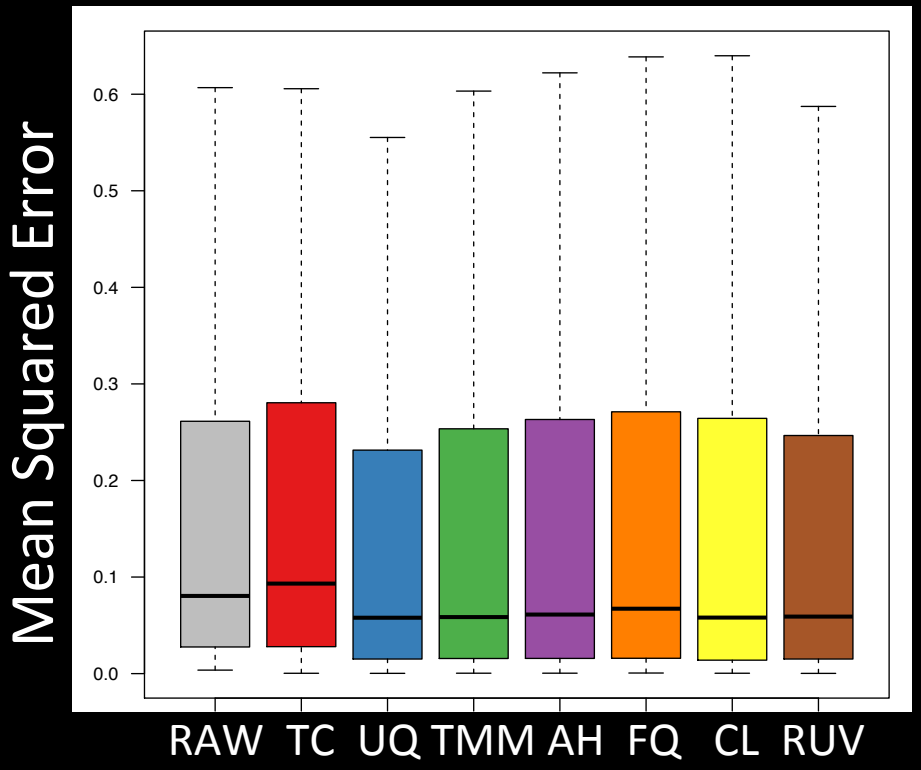
Comparison of normalization methods

- It's hard to tell using ROC curves, as they all look pretty good.
- To compare normalization methods we use the A v. B log fold changes from qRT-PCR data available for ~1,000 genes as truth.
- We consider 10 random subsets of 4 A and 4 B replicates from the original dataset, in order to compute Bias and Mean Squared Error (MSE).

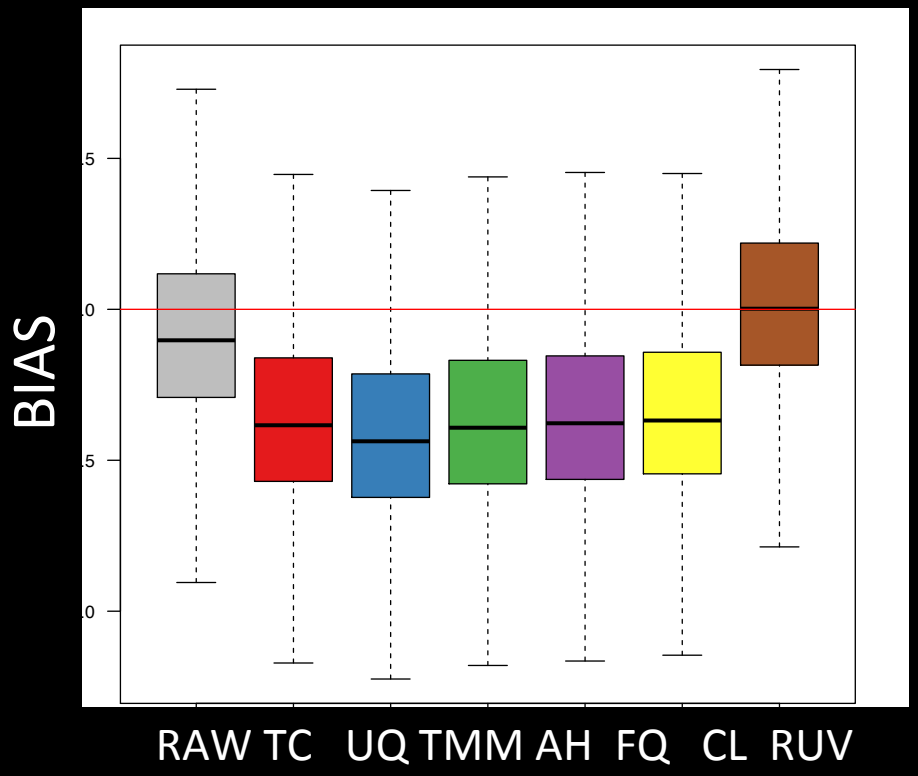
Normalizing using all genes
Bias and MSE of logFC estimates

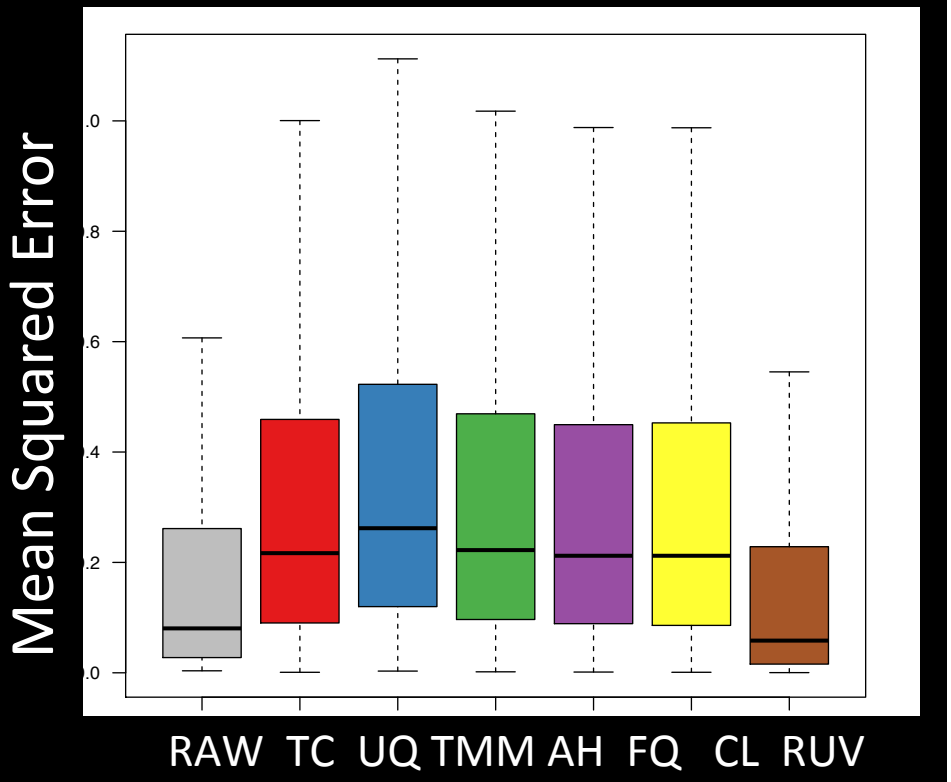
BIAS = Av{RNA-seq *log(FC)* - QRT-pcr *log(FC)*}





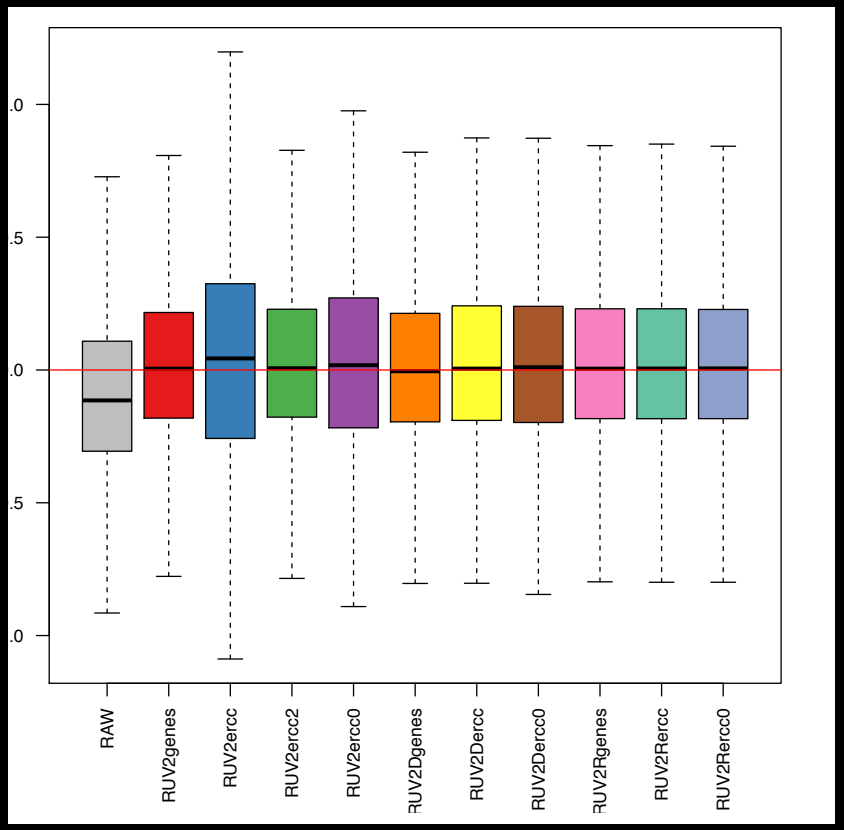
Normalizing using ERCC controls
Bias and MSE of logFC estimates



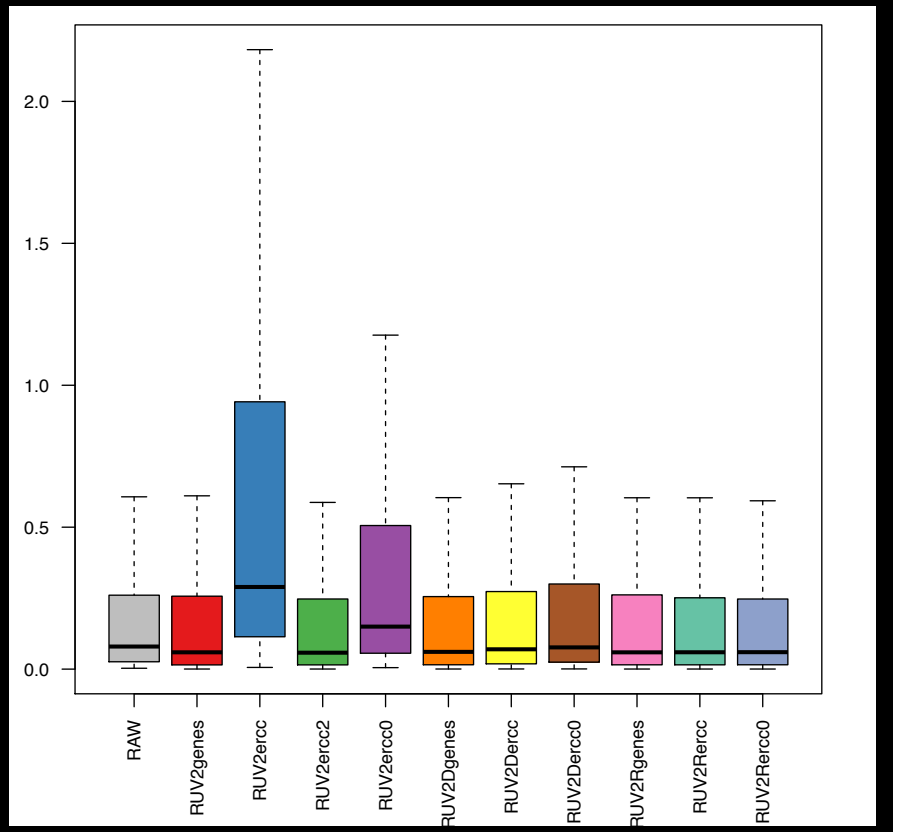


Normalizing using different versions of RUV
Bias and MSE

BIAS



Mean Squared Error



Conclusion from these two studies

- Don't normalize using the ERCC controls,

or, if you, must,

- Use one of the RUV approaches,

but even then,

- You are probably better off normalizing using all suitable genes.

- We need to look at more datasets including the ERCC, to see how broadly our conclusions apply.

Acknowledgements

Sandrine Dudoit, Davide Risso and John Ngai

UC Berkeley

They did all the work!

Johann Gagnon-Bartsch (UC Berkeley) & Laurent Jacob (CNRS, Lyon), the main RUV team.