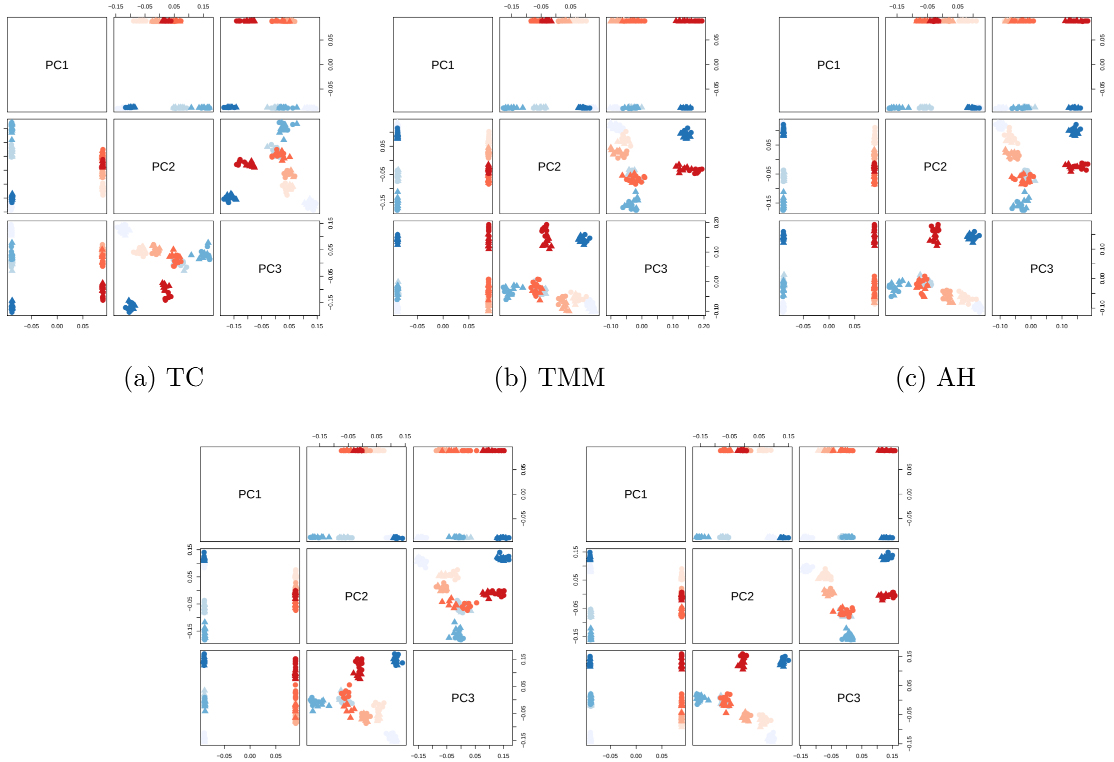
# Using controls for the normalization of RNA-seq data

# Supplementary Figures and Tables

Davide Risso[1], John Ngai[2,3,4], Terence P. Speed[1,5,6] & Sandrine Dudoit[1,7]

[1] Department of Statistics, University of California, Berkeley, California, USA.
[2] Department of Molecular and Cell Biology, University of California, Berkeley, California, USA.
[3] Helen Wills Neuroscience Institute, University of California, Berkeley, California, USA.
[4] Functional Genomics Laboratory, University of California, Berkeley, California, USA.
[5] Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia.
[6] Department of Mathematics and Statistics, The University of Melbourne, Victoria, Austraila.
[7] Division of Biostatistics, University of California, Berkeley, California, USA.



(a) TC

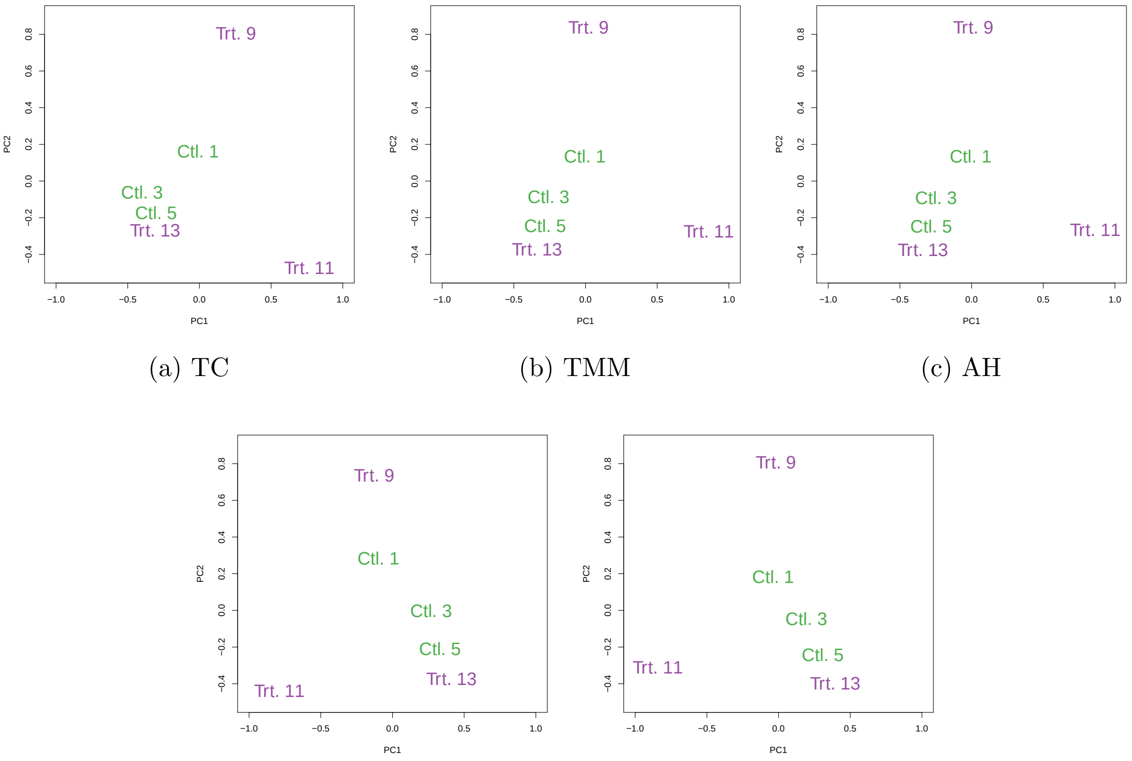
(b) TMM

(c) AH

(d) FQ

(e) CL

Supplementary Figure 1: *Unwanted variation in RNA-Seq data, SEQC dataset.* Scatterplot matrices of first three principal components (PC) for normalized counts (log scale, centered). Each point corresponds to one of the 128 samples. The four Sample A and the four Sample B libraries are represented by shades of blue and red, respectively (16 replicates per library). Circles and triangles represent samples sequenced in the first and second flow-cells, respectively.



(a) TC

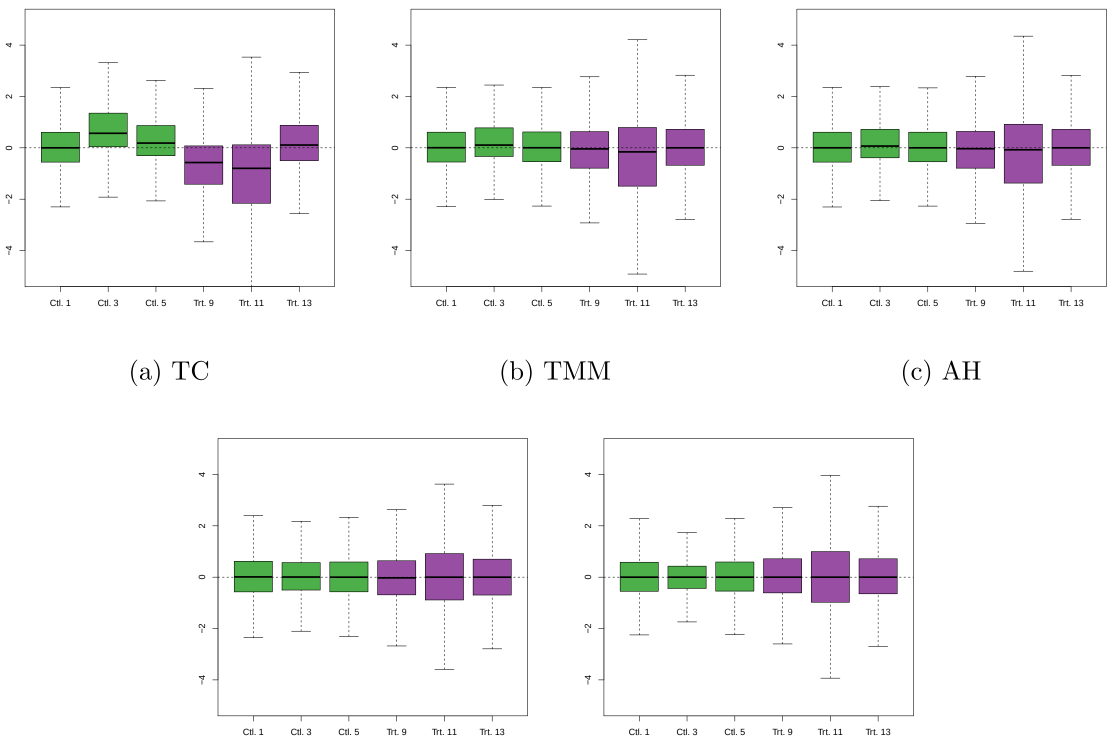
(b) TMM

(c) AH

(d) FQ

(e) CL

Supplementary Figure 2: *Unwanted variation in RNA-Seq data, Zebrafish dataset.* Scatterplots of first two PCs for normalized counts (log scale, centered). Treated libraries are in purple and control libraries in green. We expect libraries to cluster by treatment.



(a) TC

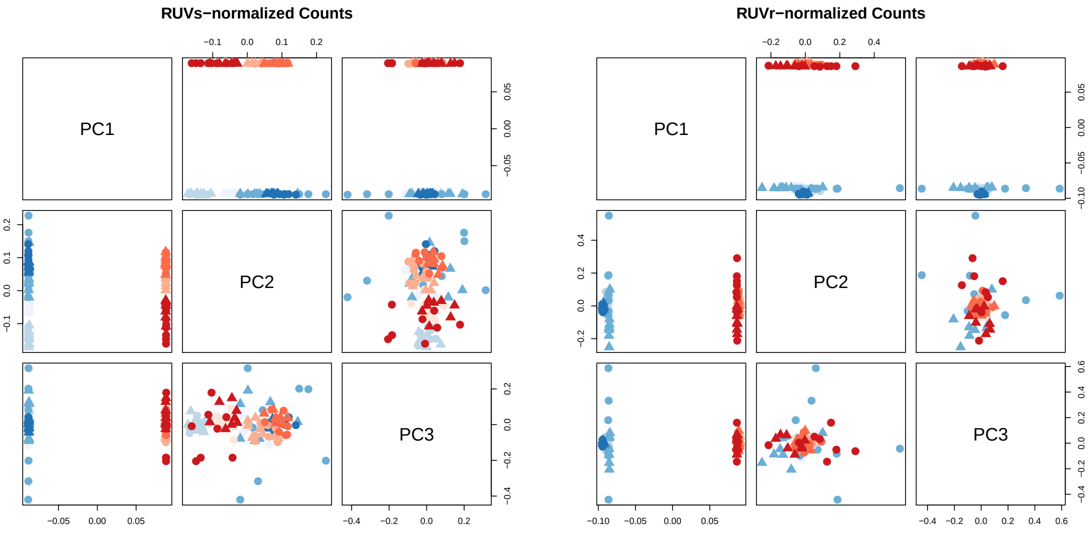
(b) TMM

(c) AH

(d) FQ

(e) CL

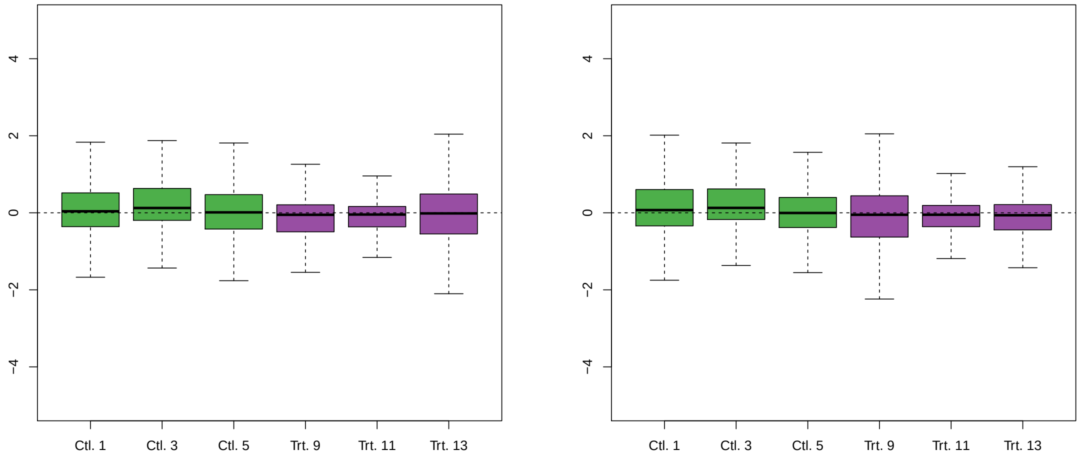
Supplementary Figure 3: *Unwanted variation in RNA-Seq data, Zebrafish dataset.* Boxplots of relative log expression (RLE) for normalized counts. Treated libraries are in purple and control libraries in green. We expect RLE distributions to be centered around zero and as similar as possible to each other.



(a) RUVs

(b) RUVr

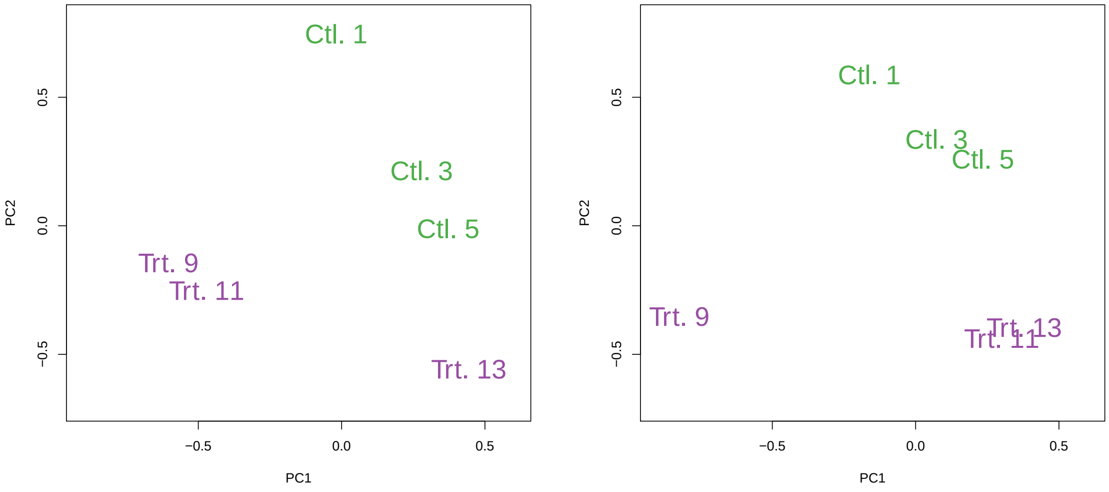
Supplementary Figure 4: *RUVs and RUVr normalization, SEQC dataset.* Scatterplot matrices of first three principal components for normalized counts (log scale, centered).



(a) RUVs

(b) RUVr

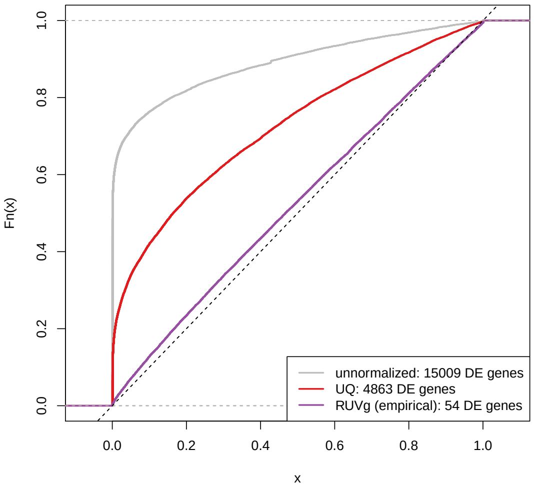
Supplementary Figure 5: *RUVs and RUVr normalization, Zebrafish dataset.* Boxplots of relative log expression for normalized counts.



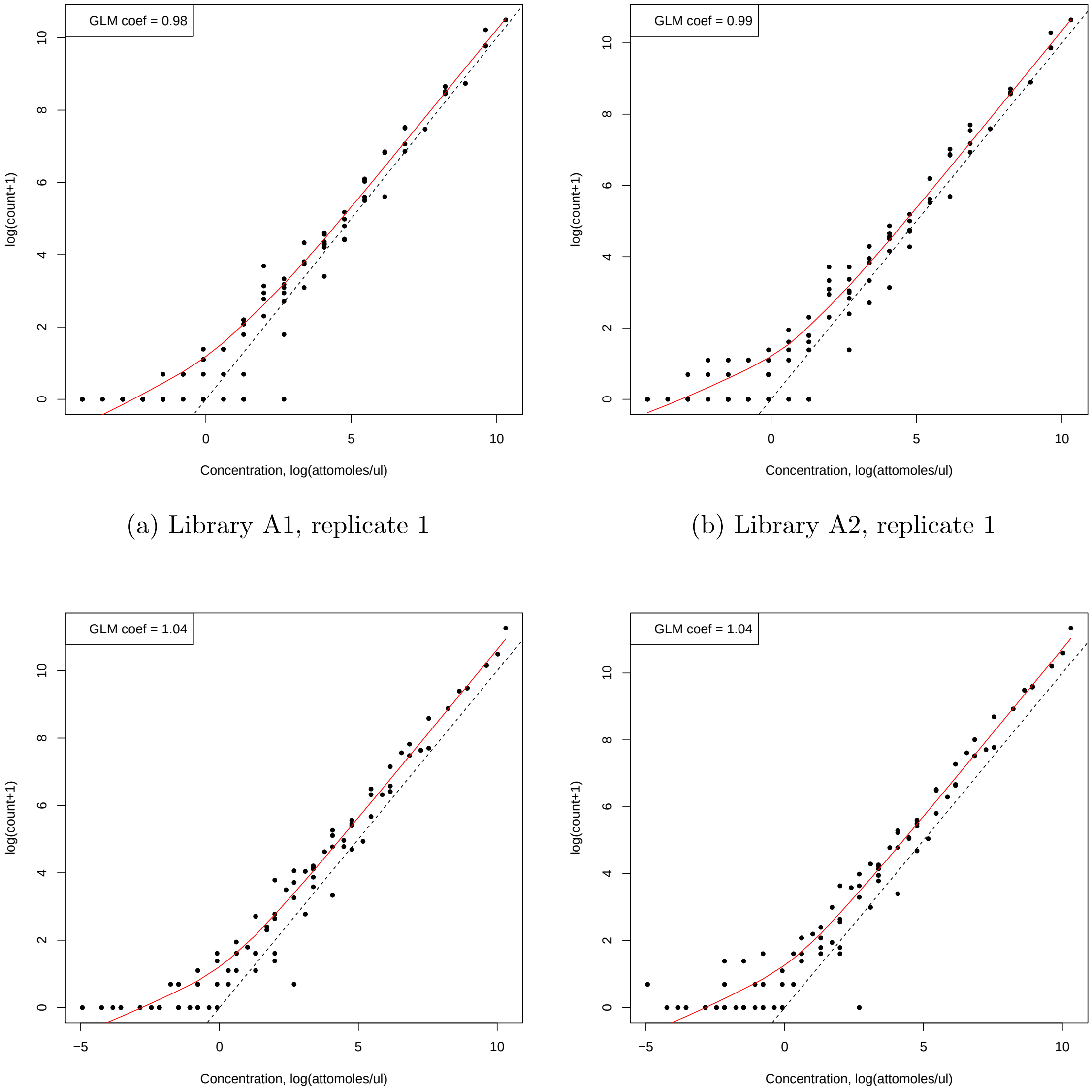
(a) RUVs

(b) RUVr

Supplementary Figure 6: *RUVs and RUVr normalization, Zebrafish dataset.* Scatterplots of first two PCs for normalized counts (log scale, centered).



Supplementary Figure 7: *RUVg normalization using in silico empirical control genes, SEQC dataset.* Empirical cumulative distribution function (ECDF) of $p$-values for tests of DE between Sample B replicates. We expect no DE and $p$-values to follow a uniform distribution, with ECDF as close as possible to the identity line. This is clearly not the case for unnormalized (gray line) and UQ-normalized (red) counts; only with RUVg (purple) do $p$-values behave as expected.



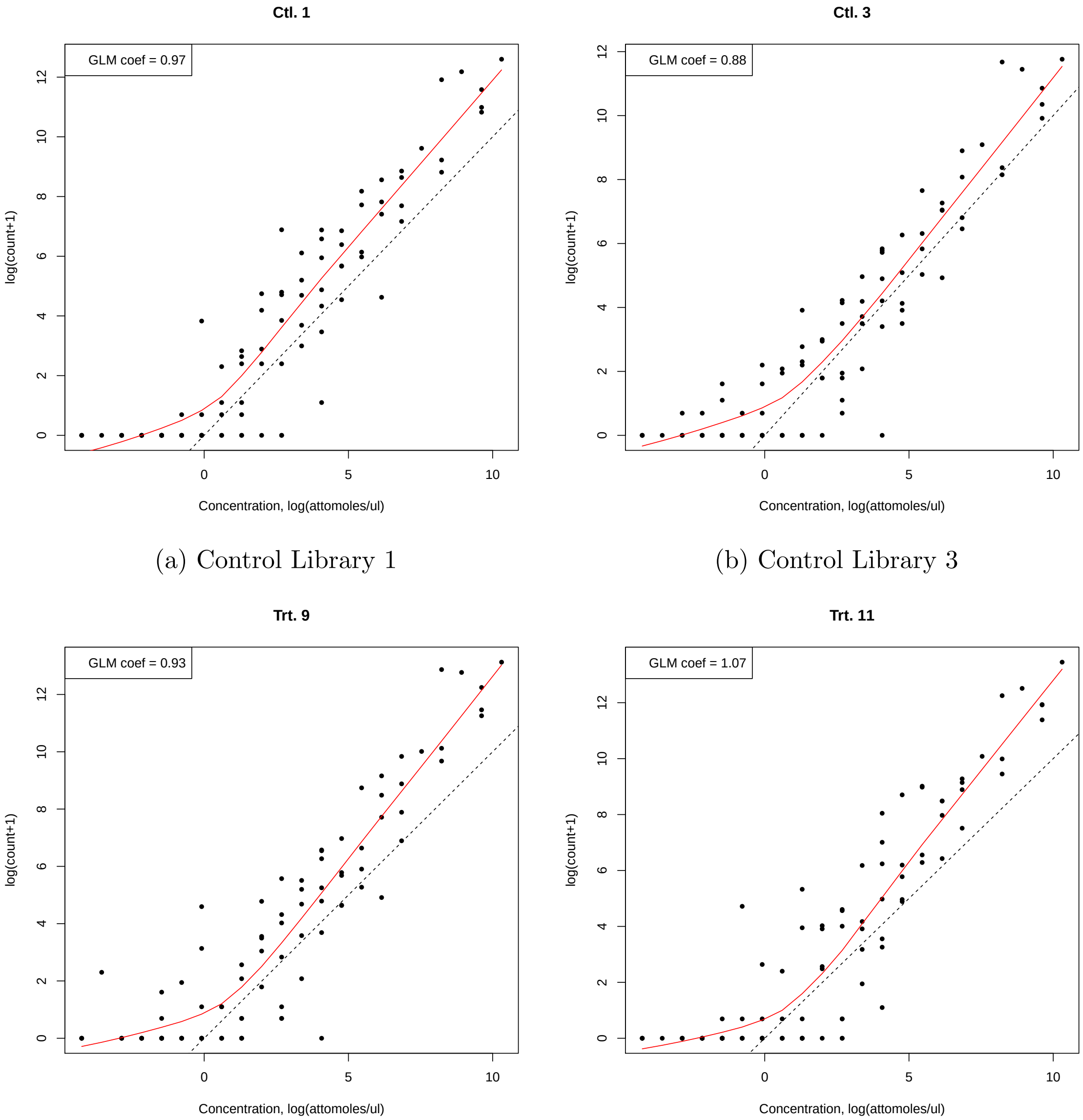
(a) Library A1, replicate 1

(b) Library A2, replicate 1

(c) Library B1, replicate 1

(d) Library B2, replicate 2

Supplementary Figure 8: *Behavior of the ERCC spike-in controls, SEQC dataset.* Scatterplots (log scale) of unnormalized read count vs. concentration (attomole/$\mu$l), for four libraries. All other libraries behave similarly (not shown).



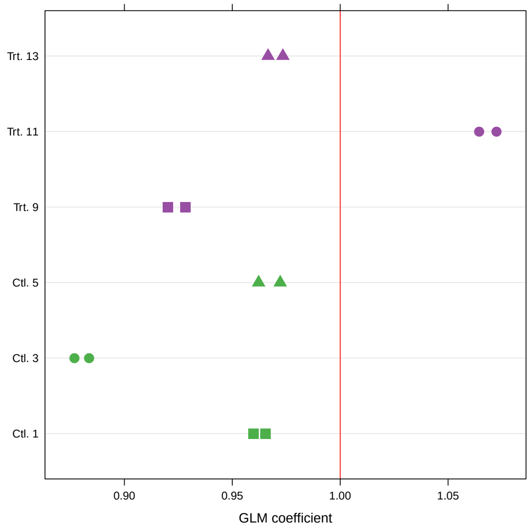
(a) Control Library 1

(b) Control Library 3

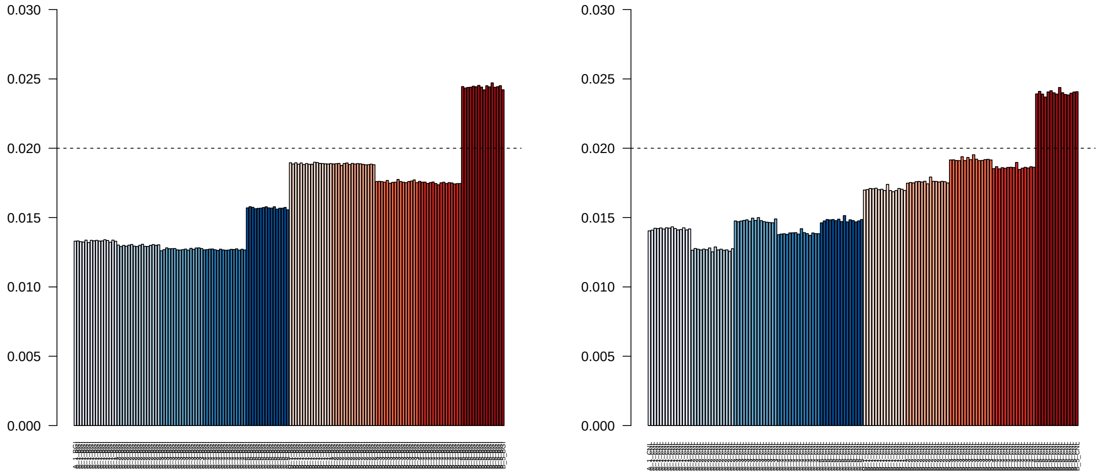
(c) Treatment Library 9

(d) Treatment Library 11

Supplementary Figure 9: *Behavior of the ERCC spike-in controls, Zebrafish dataset.* Scatterplots (log scale) of unnormalized read count vs. concentration (attomole/$\mu$l), for four libraries. All other libraries behave similarly (not shown).



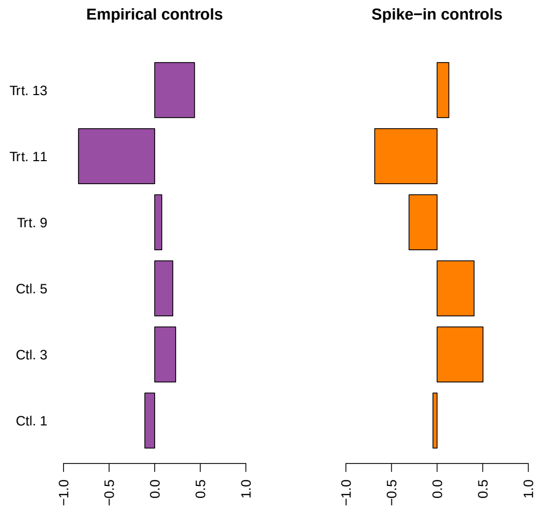
Supplementary Figure 10: *Behavior of the ERCC spike-in controls, Zebrafish dataset.* Log-linear regression coefficients of spike-in read counts on nominal concentrations. Treated libraries are in purple and control libraries in green. Replicate runs from the same library are represented using the same plotting symbol.



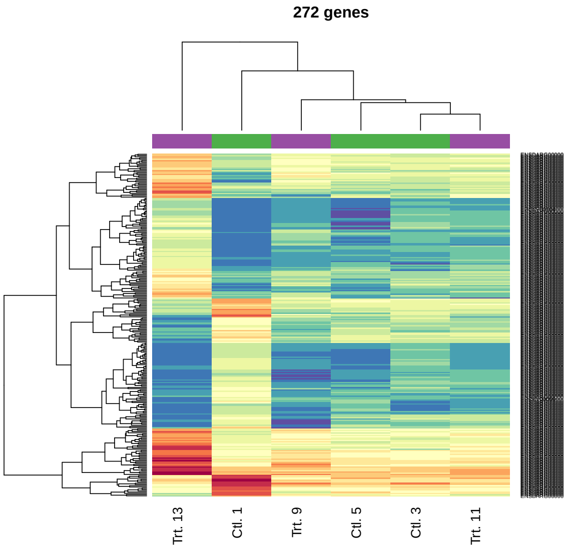
(a) BGI

(b) Cornell

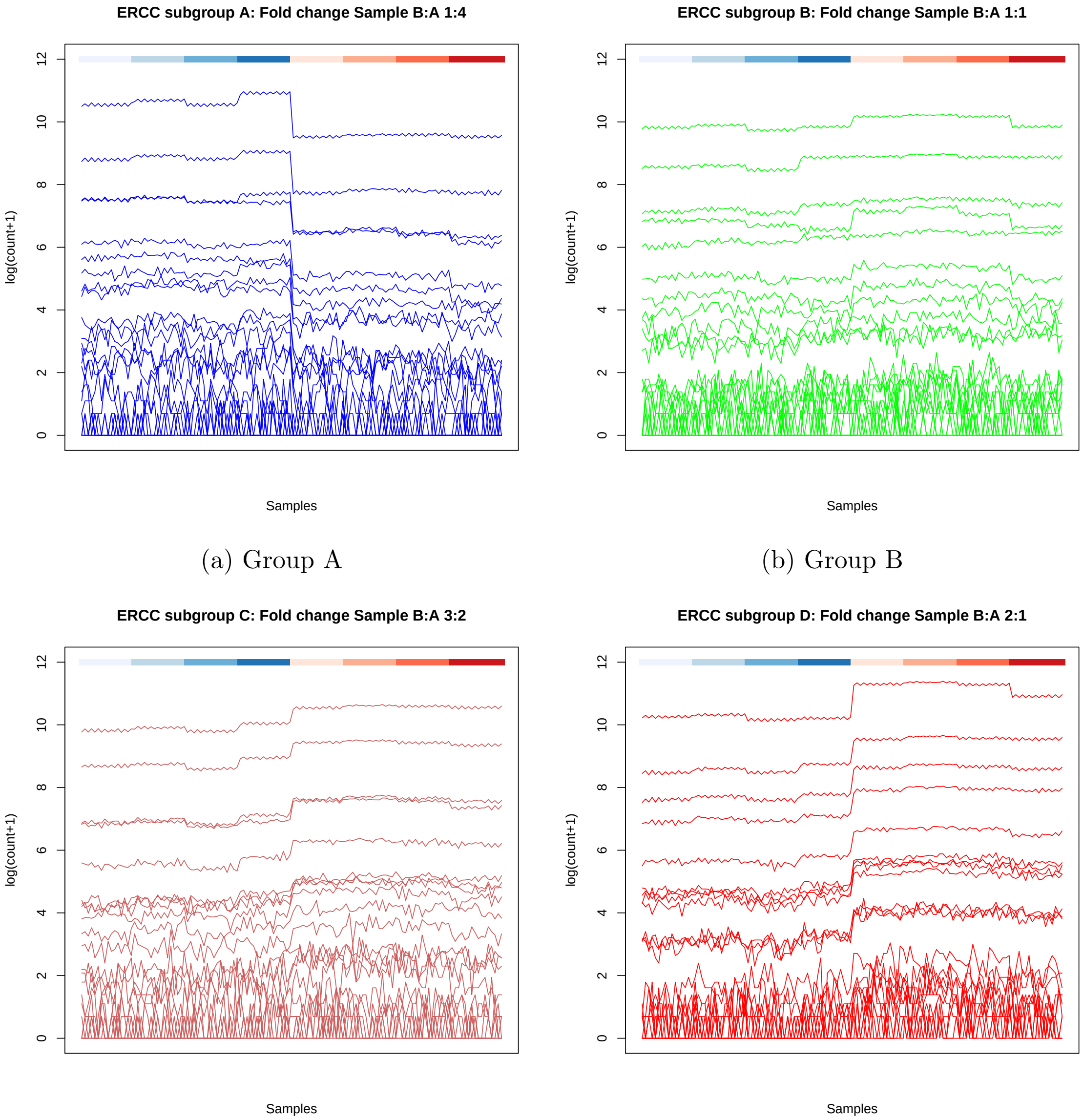
Supplementary Figure 11: *Behavior of the ERCC spike-in controls, SEQC dataset.* Proportion of reads mapping to ERCC spike-ins in different sequencing centers. Note that the last library of each sample (A4 in dark blue and B4 in dark red) was prepared by Illumina (rather than by the sequencing centers).



Supplementary Figure 12: *RUVg normalization, Zebrafish dataset.* Factors of unwanted variation $W$ estimated by RUVg with 15,839 empirical control genes and with the ERCC spike-in controls. The unwanted factors are similar between genes and spike-ins, in the sense that they span similar linear spaces.



Supplementary Figure 13: *Impact of normalization on differential expression analysis, Zebrafish dataset.* Heatmap of expression measures for the 272 genes found DE after RUVg with ERCC spike-in controls but not after UQ normalization. The clustering is not as good as with RUVg based on a set of empirical control genes (Fig. 6e).



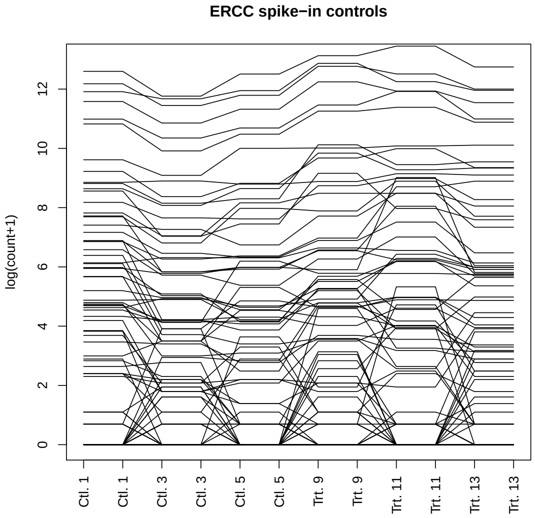
(a) Group A

(b) Group B

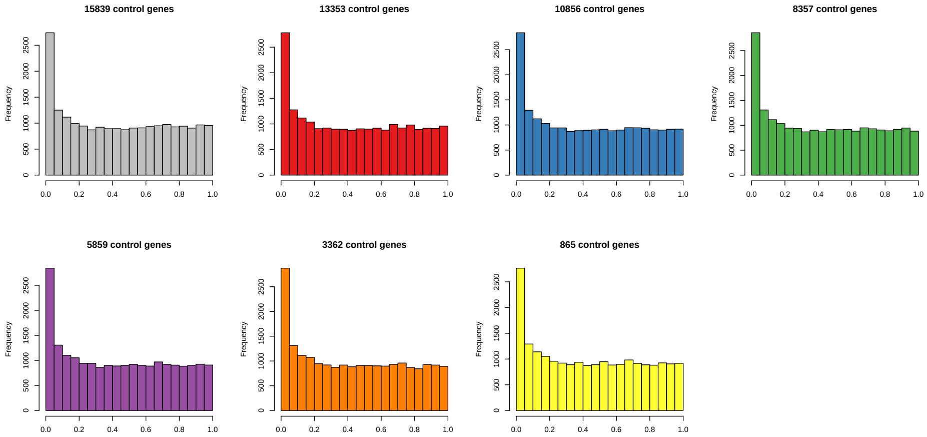
(c) Group C

(d) Group D

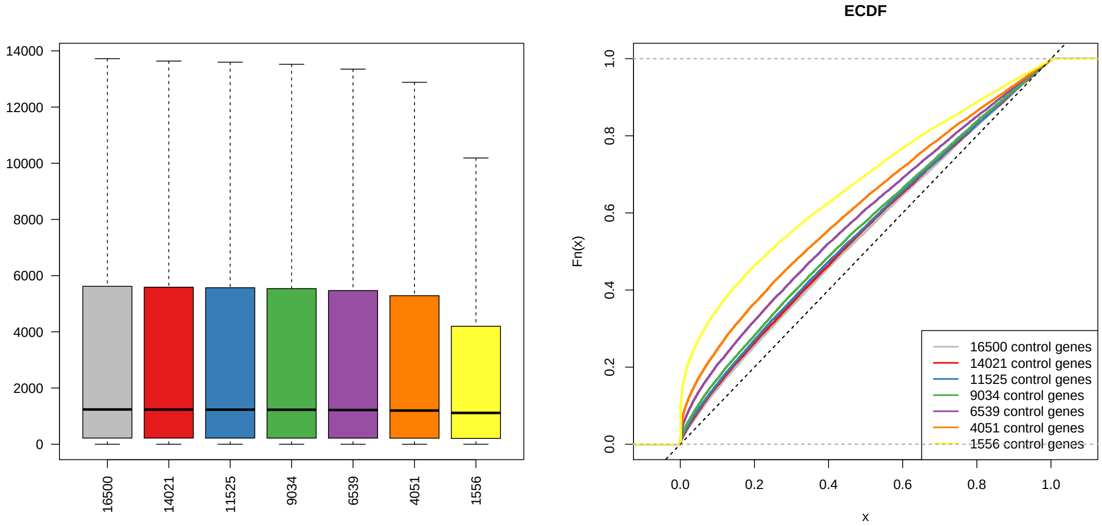
Supplementary Figure 14: *Behavior of the ERCC spike-in controls, SEQC dataset.* Unnormalized log(count+1) for individual ERCC spike-ins across the 128 samples, by ERCC control group. Panel (a): Group A, nominal Sample B:A fold-change of 1:4. Panel (b): Group B, nominal Sample B:A fold-change of 1:1. Panel (c): Group C, nominal Sample B:A fold-change of 3:2. Panel (d): Group D: nominal Sample B:A fold-change of 2:1.



Supplementary Figure 15: *Behavior of the ERCC spike-in controls, Zebrafish dataset.* Unnormalized log(count+1) for individual ERCC spike-ins across the 12 samples. Counts are expected to be constant across samples.



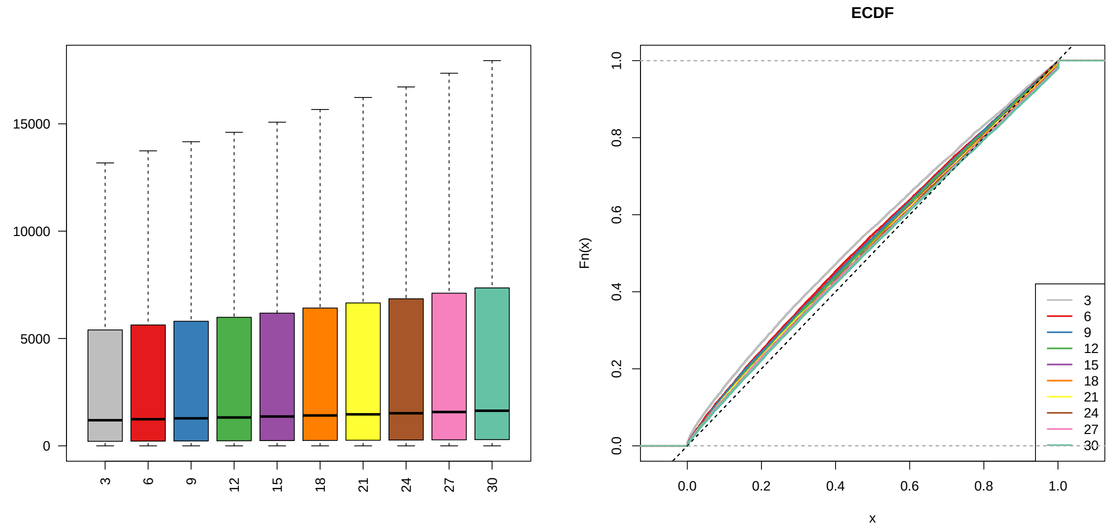
Supplementary Figure 16: *Robustness of RUVg to the choice of empirical control genes, Zebrafish dataset.* Distribution of $p$-values for tests of differential expression between treated and control samples, using RUVg with a decreasing set of in silico empirical control genes, defined as all but the top $L$ DE genes, as ranked by *edgeR* $p$-values for UQ-normalized data. The number of top DE genes $L$ ranges from 5,000 to 20,000, in steps of 2,500. The title of each histogram provides the number $J_c$ of control genes defined as not belonging to the union of the top $L$ DE genes and ERCC spike-ins.



(a)

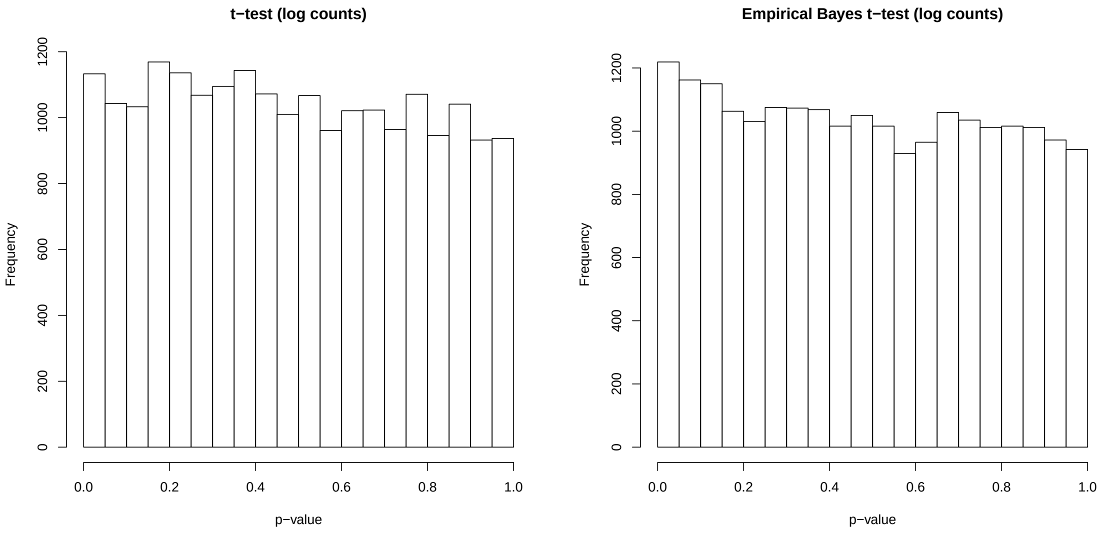
(b)

Supplementary Figure 17: *Robustness of RUVg to the choice of empirical control genes, SEQC dataset.* Panel (a): Boxplots of between-sample to within-sample sums of squares for Sample A vs. B comparison. Panel (b): Empirical cumulative distribution function of *edgeR* $p$-values for tests of differential expression between Sample A replicates. Counts are normalized using RUVg with a decreasing set of in silico empirical control genes, defined as all but the top $L$ DE genes, as ranked by *edgeR* $p$-values for UQ-normalized data. The number of top DE genes $L$ ranges from 5,000 to 20,000, in steps of 2,500. The number below each boxplot (Panel (a)) and annotating each line (Panel (b)) is the number $J_c$ of control genes defined as not belonging to the union of the top $L$ DE genes and ERCC spike-ins.



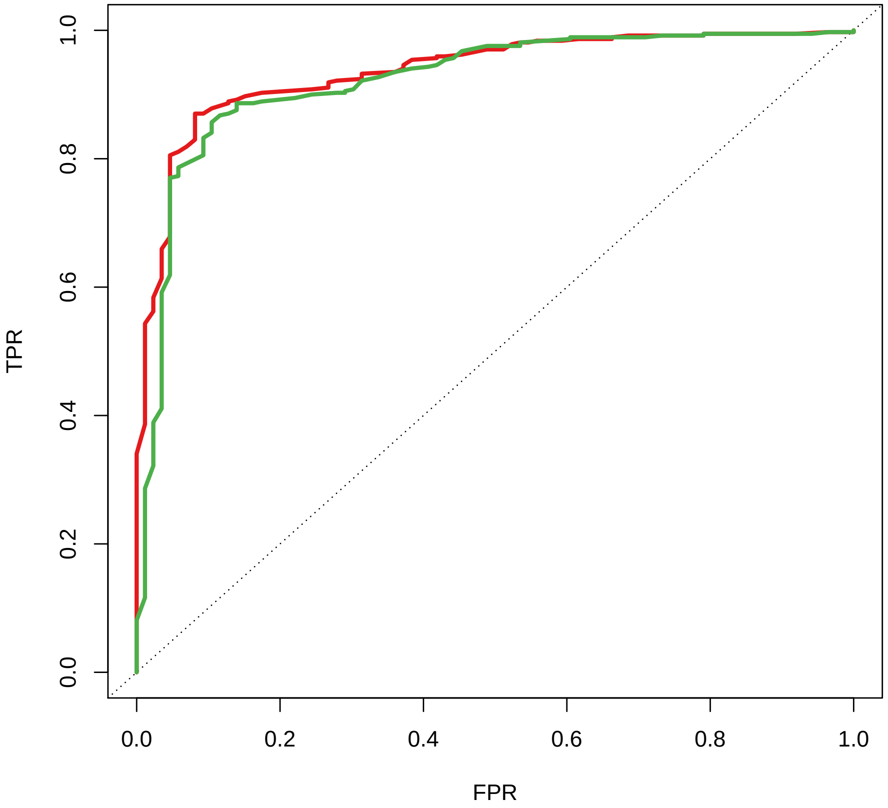
(a) (b)

Supplementary Figure 18: *Robustness of RUVg to the number $k$ of unwanted factors, SEQC dataset.* RUVg based on empirical controls, with varying number $k$ of factors of unwanted variation. Panel (a): Between-sample to within-sample sums of squares for Sample A vs. B comparison. Panel (b): Empirical cumulative distribution function of $p$-values for tests of DE between Sample A replicates.



(a) (b)

Supplementary Figure 19: *LM-based RUVg on log counts, Zebrafish dataset.* Distribution of $p$-values for tests of differential expression between treated and control samples, with log counts normalized by a linear model version of RUVg (i.e., standard RUV-2) and using 15,839 empirical control genes. Panel (a): $t$-test. Panel (b): Empirical Bayes moderated $t$-test from *limma*.



Supplementary Figure 20: *LM-based RUVg on log counts, SEQC dataset.* Receiver operating characteristic (ROC) curves for test of differential expression between Sample A and Sample B, using a set of 370 positive and 86 negative qRT-PCR controls as gold standard. Red and green curves correspond, respectively, to GLM-based RUVg and LM-based RUVg, with the same set of 16,500 empirical control genes.

Table 1: *Assumptions and applicability of RUV normalization procedures.*

| **Assumptions** | **RUVg** | | **RUVr** | **RUVs** |
|---|---|---|---|---|
| | **Spike-ins** | **Empirical** | | |
| **Negative control genes with common unwanted factors** $W$ | Yes | Yes | All genes | Yes<br>(robust) |
| **Replicate/negative control samples** | No | No | No | Yes |
| **Known design matrix** $X$ | No | Yes | Yes | No |
| **Unwanted factors** $W$ **uncorrelated with factors of interest** $X$ | Better | Better | Yes<br>(robust) | Yes<br>(robust) |
| **Applicability** | More general | DE only | DE only | More general<br>Only accounts for variation within replicate groups |