Title

Normalization of RNA-seq data using factor analysis of control genes or samples.

Permalink
https://escholarship.org/uc/item/4qq3g2hh

Journal
Nature biotechnology, 32(9)

ISSN
1087-0156

Authors
Risso, Davide
Ngai, John
Speed, Terence P
et al.

Publication Date
2014-09-01

DOI
10.1038/nbt.2931

Peer reviewed

Normalization of RNA-seq data using factor analysis of control genes or samples

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

Normalization of RNA-sequencing (RNA-seq) data has proven essential to ensure accurate inference of expression levels. Here, we show that usual normalization approaches mostly account for sequencing depth and fail to correct for library preparation and other more complex unwanted technical effects. We evaluate the performance of the External RNA Control Consortium (ERCC) spike-in controls and investigate the possibility of using them directly for normalization. We show that the spike-ins are not reliable enough to be used in standard global-scaling or regression-based normalization procedures. We propose a normalization strategy, called remove unwanted variation (RUV), that adjusts for nuisance technical effects by performing factor analysis on suitable sets of control genes (e.g., ERCC spike-ins) or samples (e.g., replicate libraries). Our approach leads to more accurate estimates of expression fold-changes and tests of differential expression compared to state-of-the-art normalization methods. In particular, RUV promises to be valuable for large collaborative projects involving multiple laboratories, technicians, and/or sequencing platforms.

Normalization, a crucial step in the analysis of RNA-seq data, has a strong impact on the detection of differentially expressed genes[1–3]. In the last few years, several normalization strategies have been proposed to correct for between-sample distributional differences in read counts, such as differences in total counts (i.e., sequencing depths)[1,4], and within-sample gene-specific effects, such as gene length or GC-content effects[2,5]. Although there have been efforts to systematically compare normalization methods[1,3,6], this important aspect of RNA-seq analysis is still not fully investigated or resolved. In particular, when data arise from complex experiments, involving, for instance, cell sorting, low-input RNA or different batches (e.g., multiple sequencing centers or different read lengths), there may be more to correct for than simply differences in sequencing depths; we refer to such typically unknown nuisance technical effects as unwanted variation.

One largely unexplored direction is the inclusion of spike-in controls in the normalization procedure. Controls have been successfully employed in microarray normalization, for mRNA arrays[7,8] and, more recently, microRNA arrays[9]. One of the advantages of using negative controls in the normalization procedure is the possibility of relaxing the common assumption that the majority of the genes are not differentially expressed between the conditions under study. This assumption can be violated when a global shift in expression occurs between conditions[9–11]; in this case, control-based normalization may be the only option.

Recently, the ERCC developed a set of RNA standards for RNA-seq[12,13]. This set consists of 92 polyadenylated transcripts that mimic natural eukaryotic mRNAs. They are designed to have a wide range of lengths (250–2,000 nucleotides) and GC-contents (5–51%) and can be spiked into RNA samples before library preparation at various concentrations (10^6-fold range). We refer to these standards as ERCC spike-in controls.

Lovén *et al.*[11] have made use of the ERCC spike-in controls in their normalization approach in the context of a global expression shift. Their procedure may be summarized as follows: (i) count the number of cells in each sample; (ii) add the ERCC spike-in sequences to each sample in proportion to the number of cells; (iii) normalize read counts based on cyclic loess robust local regression[14,15] on the spike-in counts. Although their approach does not make any assumptions concerning differences in gene expression between samples, it relies on another equally important assumption: technical effects should affect the spike-ins in the same way as they do the genes. If, for instance, some library preparation step affects spike-in and gene counts differently, then normalization based on the spike-ins may incorrectly adjust the expression measures for the bulk of the genes. Unfortunately, the data set used by the authors to illustrate their approach lacks both technical and biological replication, making it impossible to quantify the extent of variation of the spike-ins and its relation to gene variation[11].

Recently, Qing *et al.*[16] showed that the percentage of RNA-seq reads mapping to the ERCC spike-ins could vary substantially between technical replicate samples and be markedly different from the nominal value. Moreover, the dependence of spike-in read counts on the poly(A) selection protocol (polyA$^+$ versus RiboZero) suggests that poly(A) selection may play a role in spike-in detection. Given the growing interest in the ERCC spike-in standards, it is essential to evaluate their performance, with particular focus on their inclusion in normalization procedures.

[1]Department of Statistics, University of California, Berkeley, Berkeley, California, USA. [2]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, California, USA. [3]Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, California, USA. [4]Functional Genomics Laboratory, University of California, Berkeley, 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, Australia. [7]Division of Biostatistics, University of California, Berkeley, Berkeley, California, USA. Correspondence should be addressed to D.R. (davide.risso@berkeley.edu) or S.D. (sandrine@stat.berkeley.edu).

Received 27 November 2013; accepted 14 May 2014; published online 24 August 2014; doi:10.1038/nbt.2931

In this paper, our aim is twofold. We propose a normalization strategy for RNA-seq, remove unwanted variation (RUV), that uses factor analysis to adjust for nuisance technical effects, based on counts (or residuals counts) for either negative control genes or negative control samples, that is, genes or samples that are not expected to be influenced by the biological covariates of interest. We also evaluate the behavior of the ERCC spike-in standards in two very different data sets, involving different organisms and designs, and explore the possibility of using them as controls for normalization. We show that the spike-ins are not reliable enough to be used in standard global-scaling or regression-based normalization procedures. We further demonstrate that RUV, whether based on controls or not, generally outperforms state-of-the-art normalization approaches in the context of differential expression inference. In particular, it improves upon other control-based methods and is thus promising when relying on controls is the only option (e.g., in case of global expression shift).

RESULTS

Data sets

To evaluate the performance of the ERCC spike-in controls and to validate our RUV normalization strategy, we considered two very different data sets (Online Methods). The first, from the Sequencing Quality Control (SEQC) Consortium[17], compares two commercial RNA samples, Stratagene's Universal Human Reference (UHR) RNA (sample A) and Ambion's Human Brain Reference RNA (sample B). This data set is valuable for assessing normalization methods, as there are several technical replicates for both samples A and B, both at the library preparation (4 libraries for each sample type) and sequencing (16 lanes for each library) levels, and one can rely on external controls from qRT-PCR[18]. However, the absence of biological replication and the extreme difference between sample A and sample B make the data rather artificial and a more realistic and biologically meaningful data set was required to confirm our findings. To this end, we also relied on our recently published RNA-seq data[19] for three treated and three control zebrafish samples, each corresponding to a single FACS (fluorescence-activated cell sorting) run on pools of cells from different fish. Here, cell sorting and library preparation effects are confounded with biological variability between pools of fish cells.

Unwanted variation in RNA-seq data

For both the SEQC and the zebrafish data sets, existing methods did not lead to satisfactory normalization of read counts (**Figs. 1** and **2**). In particular, for the SEQC data set, although the huge biological difference between sample A and sample B was captured by the first principal component, residual library preparation and flow-cell effects were revealed by the second and third principal components (**Fig. 1a**). Upper-quartile normalization successfully adjusted for flow-cell effects (cf. sequencing depth), but not for library preparation effects (**Fig. 1b**).

Figure 2 reveals similar findings for the zebrafish data set and a clear need for normalization. The boxplots of unnormalized relative log expression (RLE) show large distributional differences between replicate libraries (**Fig. 2a**). As for the SEQC data set, upper-quartile normalization was not fully satisfactory and, in particular, failed to capture the excessive variability of library 11 (**Fig. 2b**). Moreover, libraries failed to cluster by treatment in the first two principal components, when considering both unnormalized counts (**Fig. 2c**) and upper-quartile-normalized counts (**Fig. 2d**).

We compared other state-of-the-art normalization methods and found that none led to satisfying results in terms of the removal of library preparation effects for the SEQC data set and clustering of samples by treatment for the zebrafish data set (**Supplementary Figs. 1**–**3**).

Removing unwanted variation through normalization

Building on a previously described method for normalizing microarray data[20,21], we developed RUV as a normalization strategy for RNA-seq data. Briefly, RUV works as follows. Consider a generalized linear model (GLM), where the observed RNA-seq read counts are regressed on both the known covariates of interest (e.g., treatment status) and unknown nuisance variables, that is, factors of unwanted variation (e.g., library preparation). RUV makes use of a subset of the data to estimate the factors of unwanted variation and adjusts for these in the model for differential expression analysis.

We propose three alternative approaches for estimating the factors of unwanted variation: (i) RUVg uses negative control genes, assumed not to be differentially expressed with respect to the covariates of interest (e.g., ERCC spike-ins); (ii) RUVs uses negative control samples for which the covariates of interest are constant (e.g., centered

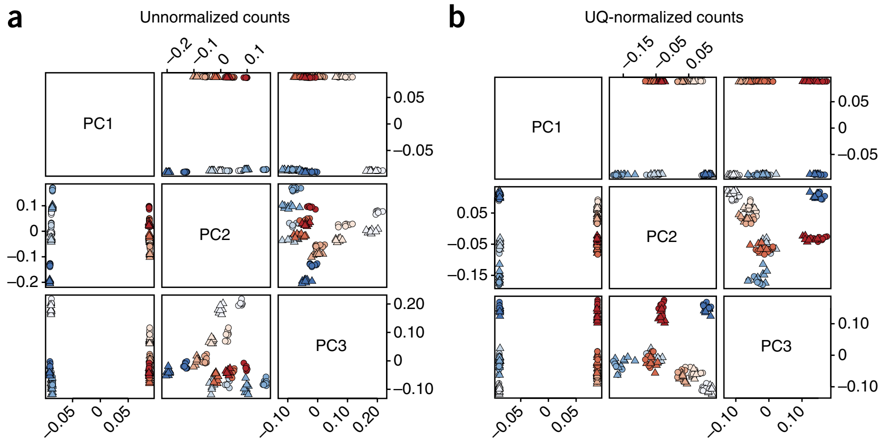

Figure 1 Unwanted variation in the SEQC RNA-seq data set. (**a**) Scatterplot matrix of first three principal components (PC) for unnormalized counts (log scale, centered). The principal components are orthogonal linear combinations of the original 21,559-dimensional gene expression profiles, with successively maximal variance across the 128 samples, that is, the first principal component is the weighted average of the 21,559 gene expression measures that provides the most separation between the 128 samples. Each point corresponds to one of the 128 samples. The four sample A and the four sample B libraries are represented by different shades of blue and red, respectively (16 replicates per library). Circles and triangles represent samples sequenced in the first and second flow-cells, respectively. As expected for the SEQC data set, the first principal component is driven by the extreme biological difference between sample A and sample B. The second and third principal components clearly show library preparation effects (the samples cluster by shade) and, to a lesser extent, flow-cell effects reflecting differences in sequencing depths (within each shade, the samples cluster by shape). (**b**) Same as **a**, for upper-quartile (UQ)-normalized counts. UQ normalization removes flow-cell effects (the circles and triangles now cluster together), but not library preparation effects. All other normalization procedures but RUV behave similarly as UQ (**Supplementary Fig. 1**).

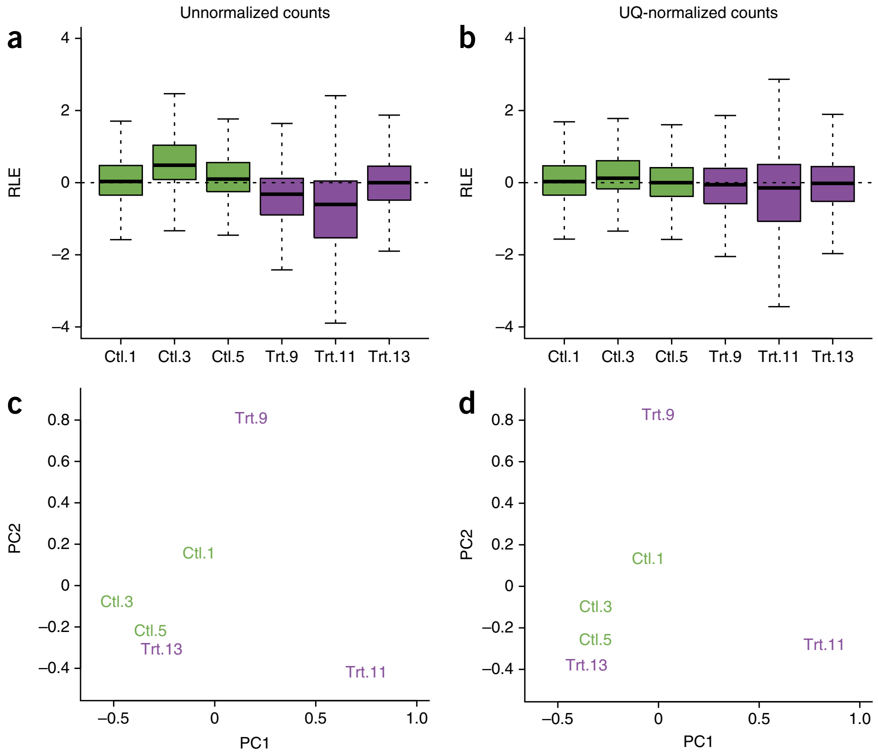

Figure 2 Unwanted variation in the zebrafish RNA-seq data set. (**a**) Boxplots of RLE for unnormalized counts. Purple: treated libraries (Trt); green: control libraries (Ctl). We expect RLE distributions to be centered around zero and as similar as possible to each other. The RLE boxplots clearly show the need for normalization. (The bottom and top of the box indicate, respectively, the first and third quartiles; the inside line indicates the median; the whiskers are located at 1.5 the inter-quartile range (IQR) above and below the box.) (**b**) Same as **a**, for upper-quartile-normalized counts. UQ normalization centers RLE around zero, but fails to remove the excessive variability of library 11. (**c**) Scatterplot of first two principal components for unnormalized counts (log scale, centered). Libraries do not cluster as expected according to treatment. (**d**) Same as **c**, for UQ-normalized counts. UQ normalization does not lead to better clustering of the samples. All other normalization procedures but RUV behave similarly as UQ (**Supplementary Figs. 2** and **3**).

counts for technical replicates of sample A and of sample B in the SEQC data set); (iii) RUVr uses residuals from a first-pass GLM regression of the unnormalized counts on the covariates of interest.

We first applied RUVg to the SEQC and zebrafish data sets using a set of *in silico* empirical control genes (Online Methods and **Fig. 3**); RUVr and RUVs performed similarly (**Supplementary Figs. 4–6**). RUVg effectively reduced library preparation effects for the SEQC data set without weakening the sample A versus B effect (**Fig. 3a**). We also performed differential expression analysis between technical replicates for both sample A (**Fig. 3b**) and sample B (**Supplementary Fig. 7**). In the absence of differential expression, the *P*-value distribution should be as close as possible to the uniform distribution (identity line for the empirical cumulative distribution function in **Fig. 3b**).

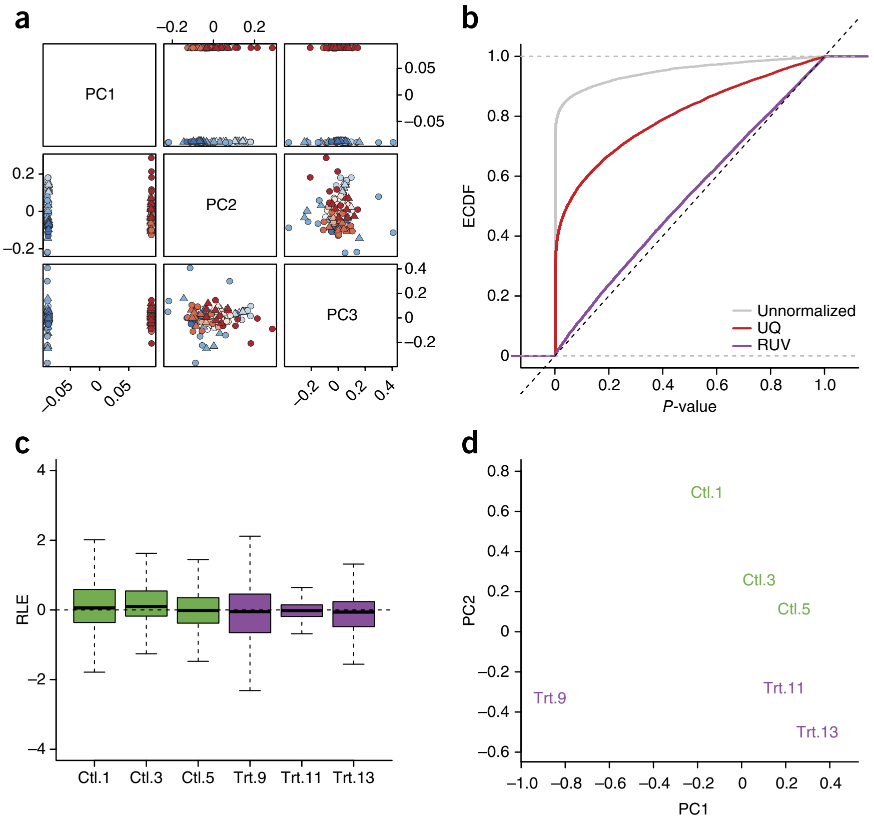

Figure 3 RUVg normalization using *in silico* empirical control genes. (**a**) For the SEQC data set, scatterplot matrix of first three principal components after RUVg normalization (log scale, centered). RUVg adjusts for library preparation effects (cf. **Fig. 1**), while retaining the sample A versus B difference. (**b**) For the SEQC data set, empirical cumulative distribution function (ECDF) of *P*-values for tests of differential expression between sample A replicates (given a value *x*, the ECDF at *x* is simply defined as the proportion of *P*-values ≤ *x*). We expect no differential expression 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 upper-quartile-normalized (red) counts; only with RUVg (purple) do *P*-values behave as expected. (**c**) For the zebrafish data set, boxplots of RLE for RUVg-normalized counts. RUVg shrinks the expression measures for library 11 toward the median across libraries, suggesting robustness against outliers. (The bottom and top of the box indicate, respectively, the first and third quartiles; the inside line indicates the median; the whiskers are located at 1.5 the inter-quartile range above and below the box.) (**d**) For the zebrafish data set, scatterplot of first two principal components for RUVg-normalized counts (log scale, centered). Libraries cluster as expected by treatment.

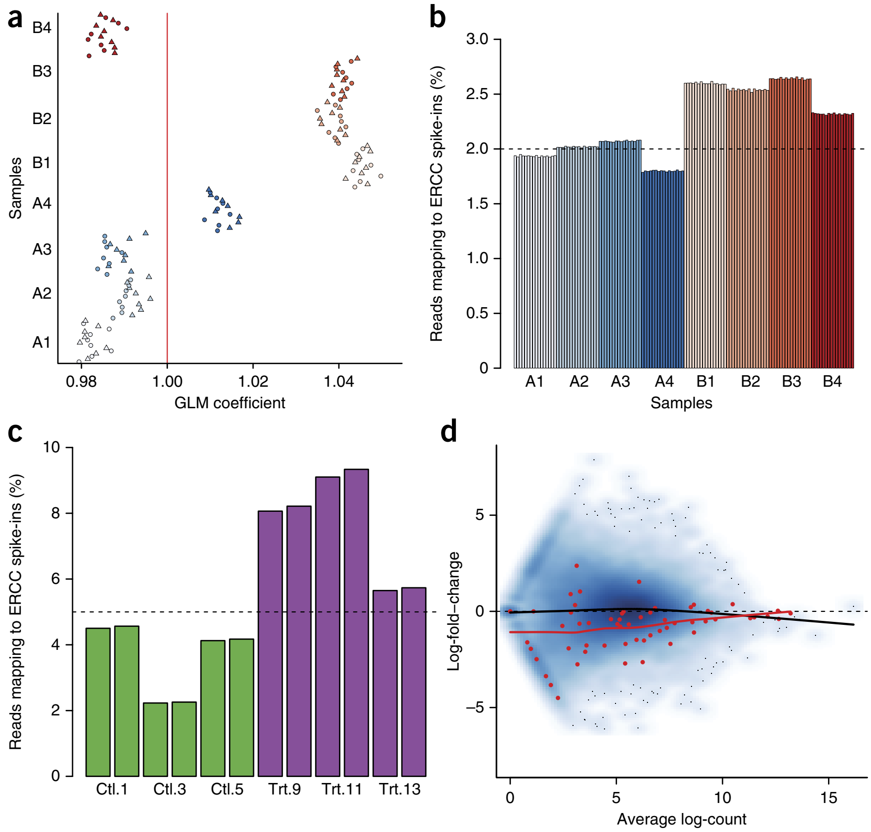

Figure 4 Behavior of the ERCC spike-in controls. (**a**) For the SEQC data set, GLM regression coefficients of spike-in read counts on nominal concentrations. Each point corresponds to one of the 128 samples. The four sample A and the four sample B libraries are represented by different shades of blue and red, respectively (16 replicates per library). Circles and triangles represent samples sequenced in the first and second flow-cells, respectively. There are evident library preparation effects. (**b**) For the SEQC data set, the proportion of reads mapping to the spike-ins deviates markedly from the nominal value (dashed line). There are library preparation effects and troubling sample A versus B effects, which may bias the inference of differential expression. (**c**) For the zebrafish data set, the proportion of reads mapping to the spike-ins deviates markedly from the nominal value (dashed line). Again, there are library preparation and treatment effects (purple: treated libraries (Trt); green: control libraries (Ctl); data for the two runs of each library are displayed in adjacent bars). (**d**) For the zebrafish data set, mean-difference plot of unnormalized counts (log scale) for two control samples (library 5 versus library 1). The shading represents point density and spike-in counts are plotted using red symbols. The lines are the lowess robust local regression[29] fits for genes (black) and spike-ins (red). As expected, log-fold-changes are scattered around the horizontal zero line, indicating that most genes are equally expressed in the two control samples. The negative slope of the black line suggests the need for normalization. The difference between the two lowess fits indicates that, disturbingly, the spike-ins do not behave as the bulk of the genes.

There were substantial library preparation effects for unnormalized counts. These were only attenuated (and not fully removed) by upper-quartile normalization. By contrast, RUVg fully adjusted for library preparation effects. For the zebrafish data set, RUVg down-weighted the effect of outlying library 11 on subsequent analyses (e.g., differential expression), by shifting its read counts towards the median counts across samples, as shown in the RLE boxplots of **Figure 3c**, thus leading to more robust differential expression results (see "Impact on differential expression analysis"). More importantly, RUVg led to better separation between treated and control samples (**Fig. 3d**).

Behavior of the ERCC spike-in controls

The main assumption of RUVg is that one can identify a set of negative control genes, that is, genes whose expression is not influenced by the biological covariates of interest. Although using a set of *in silico* empirical controls worked well in practice (**Fig. 3**), an obvious strategy is to design synthetic negative controls, known a priori not to be influenced by the biological covariates under study. To this end, we explored the possibility of using the recently proposed ERCC spike-in controls in the normalization procedure.

In order for the spike-ins to be trusted for normalization, two conditions must be satisfied: (i) spike-in read counts are not affected by the biological covariates of interest and (ii) the unwanted variation affects spike-in and gene read counts similarly. Note that these assumptions are not limited to our RUV normalization approach and are needed also by other control-based methods[11]; hence, careful exploration of the behavior of the ERCC spike-ins is essential before applying any normalization method that makes use of them.

First, we considered the relationship between the ERCC spike-in counts and their nominal concentrations. Although there was a good linear relationship between log-read count and log-concentration[13] (**Supplementary Figs. 8** and **9**), strong library preparation effects were observed. We used a Poisson GLM to regress the spike-in counts on the nominal concentrations. **Figure 4a** displays the estimates of the regression coefficients for each of the 128 SEQC samples (see **Supplementary Fig. 10** for the zebrafish data set). Ideally, the coefficients should be as close as possible to 1. Replicate samples clustered by library (**Fig. 4a**), suggesting library preparation effects on the spike-in counts.

The proportion of reads mapping to the ERCC spike-ins was highly variable between samples and deviated markedly from the nominal value (**Fig. 4b**,**c**). In addition to the already observed library preparation effects, spike-in counts seemed to be affected by the biological factor of interest, a disturbing observation. In particular, for the SEQC data set, spike-ins consistently received a greater proportion of reads in sample B than in sample A (**Fig. 4b**). This was true for all the sequencing centers (**Supplementary Fig. 11**). Similar patterns were observed for the zebrafish data set (**Fig. 4c**). The proportion of reads mapping to the spike-ins was stable between sequencing runs of the same library, but was very variable between libraries and exhibited a strong treatment effect (being consistently higher in treated than in control samples). These distributional properties of the spike-ins have important implications for inferring differential gene expression. For the zebrafish data set, the mean-difference plot (MD-plot) in **Figure 4d** contrasts read counts for two control fish libraries, for which there was no treatment effect and for which the spike-ins were expected to have log-fold-changes of zero. The distribution of log-fold-changes

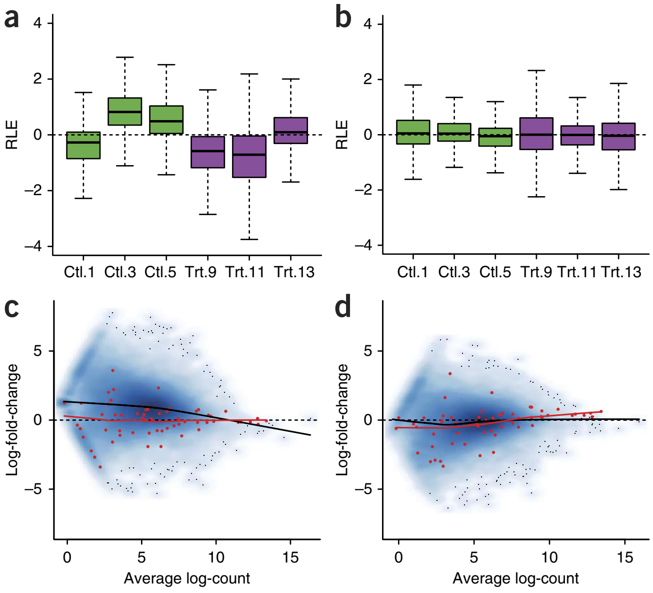

Figure 5 Using the ERCC spike-in controls for normalization, zebrafish data set. (**a**) Boxplots of RLE for cyclic loess-normalized counts (purple: treated libraries (Trt); green: control libraries (Ctl)). The expression measures are clearly not comparable across replicate libraries and cyclic loess based on the spike-ins is not effective at normalizing the counts. (**b**) Boxplots of RLE for RUVg-normalized counts. RUVg based on the spike-ins leads to much more reasonable RLE distributions, similar to those obtained using a set of empirical controls (**Fig. 3c**). (**c**) Mean-difference plot (MD-plot) for cyclic loess-normalized counts (log scale) for the same control samples as in **Figure 4c**. By shifting the spike-in log-fold-changes toward zero, cyclic loess normalization leads to a global shift of the gene log-fold-changes away from zero. For control samples, with no expected differential expression, cyclic loess normalization is likely to bias expression measures. (**d**) MD-plot for RUVg-normalized counts (log scale) for the same control samples as in **Figure 4c**. Log-fold-changes for both the spike-ins and the genes are scattered around the zero line, yielding more realistic expression measures than cyclic loess normalization.

for the spike-ins was markedly different from that of the genes. Using a loess fit on the spike-ins to normalize this pair of samples, in a procedure similar to that of Lovén *et al.*[11], would result in wrongly shifting the gene log-fold-changes upward (**Fig. 5**). Indeed, because we were comparing two control samples, we did not expect this global shift in expression to be real.

Using the ERCC spike-in controls for normalization

Properly behaved spike-ins could be a valuable resource for normalization: by design, their read counts are expected to be constant (or to have known fold-changes) between samples and hence any deviations from nominal fold-changes should reflect nuisance technical effects. One could therefore use functions of the spike-in counts to scale gene-level read counts, using existing procedures such as upper-quartile or trimmed mean of M values (TMM)[4] normalization. Unfortunately, given the troubling behavior of the ERCC spike-ins in our two data sets (**Fig. 4**), global-scaling normalization factors based on these were unrealistic and led to poorly normalized counts (**Supplementary Fig. 3**). Note that similar findings were reported for TMM normalization using a different set of spike-ins[4]. Cyclic loess normalization based on the spike-ins led to similarly poor results (**Fig. 5a**). By contrast, RUVg normalization led to reasonable results when based on the spike-ins (**Fig. 5b**). In particular, cyclic loess normalization unrealistically shifted log-fold-changes upward in the comparison between two control libraries (**Figs. 4d** and **5c**), whereas both spike-in and gene expression log-fold-changes were centered around zero with RUVg (**Fig. 5d**).

The good performance of RUVg compared to global-scaling and regression-based normalization can be explained by the different assumptions underlying each approach. Global-scaling and regression-based normalization methods assume that unwanted technical effects (i.e., between-sample differences excluding biological effects of interest) are roughly the same for genes and spike-ins and are captured by either a single parameter per sample or a regression function between pairs of samples. Such assumptions were clearly violated for our data sets (e.g., **Fig. 4d**). RUVg, on the other hand, only assumes that the factors of unwanted variation estimated from the spike-ins span the same linear space as the factors of unwanted variation W for all of the genes. The effects of the unwanted factors on the counts (i.e., the nuisance parameter α) are gene-specific and reestimated for all of the genes in step 4 of RUVg (see RUVg and equation (1) in Online Methods). These different and more general assumptions seem reasonable for our data sets (**Supplementary Fig. 12**). However, the estimation of W was problematic when based on such a small set of negative controls (only 59 spike-ins). This explains the better performance of RUVg when it was based on a larger set of empirical controls (**Fig. 6**, **Supplementary Figs. 12** and **13**).

Impact on differential expression analysis

One of the most important applications of RNA-seq is in the study of differential gene expression between two or more biological conditions (e.g., treated versus control samples in the zebrafish data set or sample A versus B in the SEQC data set). Normalization has been shown to have a strong impact on the inference of differentially expressed genes[1–3]. To compare RUV to other normalization methods in terms of differential expression results, we exploited the availability of external qRT-PCR controls for the SEQC data set. By viewing qRT-PCR as a gold standard, one can estimate the bias in RNA-seq sample A/sample B expression log-fold-changes by the differences between the RNA-seq and corresponding qRT-PCR estimates.

For the SEQC data set, we observed a slight bias in the unnormalized sample A versus B log-fold-changes (**Fig. 6a**), which suggests the need for normalization, although the balanced design, the large number of technical replicates, and the extreme differences between samples A and B somewhat alleviated the impact of technical effects on measures of differential expression. Upper-quartile normalization based on all genes led to unbiased estimates of log-fold-changes. However, using the ERCC spike-ins for upper-quartile or cyclic loess normalization led to biased estimates. All versions of RUV (with empirical or spike-in controls) yielded unbiased estimates. The receiver operating characteristic (ROC) curves led to similar conclusions (**Fig. 6b**), although the extreme power of the differential expression tests (resulting from the large sample sizes and extreme differences between samples A and B) made it difficult to distinguish between methods. Indeed, even unnormalized counts led to a reasonable ROC curve, despite their biased fold-change estimates (**Fig. 6a**).

In the absence of a gold standard for the zebrafish data set, one can nonetheless examine the distribution of P-values for tests of differential expression between treated and control samples. Ideally, one expects a uniform distribution for the bulk of non-differentially expressed genes, with a spike at zero corresponding to a few differentially expressed genes. This was indeed the case for upper-quartile normalization based on all genes and all RUV versions (**Fig. 6c**). However, upper-quartile and cyclic loess normalization based on

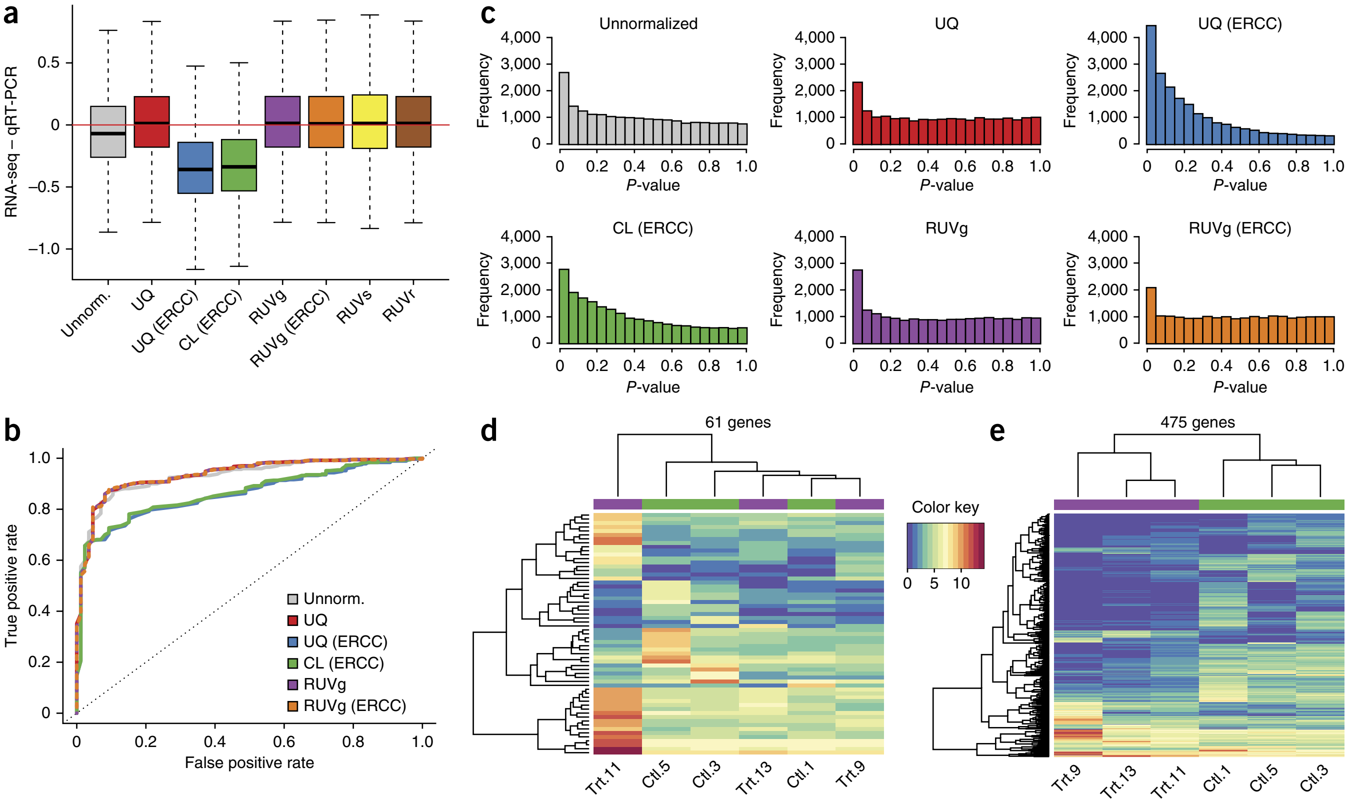

Figure 6 Impact of normalization on differential expression analysis. (**a**) For the SEQC data set, difference between RNA-seq and qRT-PCR estimates of sample A/sample B log-fold-changes, that is, bias in RNA-seq when viewing qRT-PCR as the gold standard. All RUV versions lead to unbiased log-fold-change estimates; cyclic loess (CL) normalization based on the ERCC spike-ins leads to severe bias. (**b**) For the SEQC data set, ROC curves using a set of 370 positive and 86 negative qRT-PCR controls as gold standard. RUVg (based on either empirical or spike-in controls) and UQ normalization perform slightly better than no normalization. UQ normalization based on spike-ins performs similarly to no normalization and CL normalization based on spike-ins performs the worst. (**c**) For the zebrafish data set, distribution of *edgeR* *P*-values for tests of differential expression between treated and control samples. UQ and CL normalization based on spike-ins lead to distributions far from the expected uniform. (**d**) For the zebrafish data set, heatmap of expression measures for the 61 genes found differentially expressed between control (Ctl) and treated (Trt) samples after UQ but not after RUVg normalization. Clustering of samples is driven by outlying library 11. (**e**) Heatmap of expression measures for the 475 genes found differentially expressed after RUVg but not after UQ normalization. Samples cluster as expected by treatment.

the ERCC spike-ins led to a distribution of *P*-values very far from uniform. Finally, the heatmaps of **Figure 6d**,**e** confirm the robust nature of RUVg (cf. **Fig. 3c**). The 61 genes identified as differentially expressed by upper-quartile normalization but not by RUVg were driven solely by the extreme expression of library 11, as indicated by the hierarchical clustering (**Fig. 6d**). On the other hand, the 475 genes identified as differentially expressed by RUVg but not by upper-quartile normalization yielded a more balanced clustering, reflecting the treatment effect (**Fig. 6e**). These heatmaps and the scatterplot of the first two principal components in **Figure 3d** suggest that RUVg led to a more realistic and robust list of differentially expressed genes than other methods.

DISCUSSION

Normalization is an essential, yet often overlooked, aspect of RNA-seq data analysis. As RNA-seq has become the assay of choice for measuring gene expression levels, the availability of data from large collaborative projects (such as The Cancer Genome Atlas[22] and ENCODE[23]) has grown dramatically in the last few years. With such projects employing multiple library preparation protocols (e.g., poly(A)$^+$, total RNA) and sequencing platforms, and with the sequencing technology evolving quickly (cf. read length, paired- versus single-end reads), many sources of unwanted variation can affect read counts. Normalization procedures must therefore be able to adjust for often unknown and more complex effects than simple differences in sequencing depths.

We have used the two very different SEQC and zebrafish data sets to illustrate the misbehavior of the ERCC spike-in controls. Disturbingly, individual spike-in read counts were highly variable compared to their nominal concentrations (**Supplementary Figs. 14** and **15**), the overall proportion of reads mapping to the spike-ins was also highly variable and deviated markedly from the nominal proportion[16] (**Fig. 4b**,**c**), and technical effects (e.g., library preparation effects) were different for the spike-ins than for the bulk of the genes (**Fig. 4d**). We have also demonstrated the need for careful normalization and proposed a normalization strategy, RUV, which adjusts for nuisance technical effects by performing factor analysis on counts (or residual counts) for suitable sets of control genes or samples. The different RUV versions generally outperformed state-of-the-art normalization approaches and led to more accurate estimates of expression fold-changes and tests of differential expression (**Fig. 6**). For the SEQC data set, upper-quartile normalization led to good differential expression results (**Fig. 6a**,**b**), even though it failed to adjust for library preparation effects (**Fig. 1b**). Such behavior is due to the extreme difference between sample A and sample B and is not generalizable to more biologically relevant data sets, where the effects of interest are more subtle and comparable in magnitude to the unwanted technical effects. This was confirmed by the zebrafish data set, where RUV led to better results than upper-quartile normalization in terms of clustering and differentially expressed gene lists (**Figs. 3d** and **6e**). Although RUV performed more robustly when applied to a set of empirical control genes or, when feasible, a set of replicate samples, it was the only method that gave reasonable results when using the ERCC spike-ins (**Fig. 5**).

In this study, our three proposed RUV approaches performed equally well. However, they rely on different assumptions and the validity of these assumptions for the data at hand should guide the choice of the method (Online Methods and **Supplementary Table 1**). RUVg assumes that one can identify a set of negative control genes (e.g., housekeeping genes or spike-ins) that are not affected by the biological covariates of interest and are affected by the factors of unwanted variation in the same way as the rest of the genes. This is essentially the discrete version of RUV-2 (refs. 20,21). RUVr, similarly to previously proposed microarray methods[24], does not make this assumption; in fact, one can use all of the genes to normalize the data with this version. RUVs stands in the middle. Formally, one still needs a set of negative control genes for the estimation of the unwanted factors, but RUVs is much less sensitive to poorly chosen control genes than is RUVg. Indeed, we found that RUVs worked well in practice when using all genes as negative controls. However, both RUVr and RUVs assume that the unwanted factors are uncorrelated with the covariates of interest. This assumption is usually reasonable, but it is not met when, for example, all treated samples are in one batch and all control samples in another. In this case, RUVr and RUVs will not remove the unwanted variation, whereas RUVg should still work, provided it is based on a reliable set of control genes[20,21]. Although RUVs on all genes should perform well if the unwanted factors are not too correlated with the covariates of interest, it can only account for variation that occurs within replicate groups; for example, it can capture library preparation effects only if the replicate groups include multiple libraries. This has implications for experimental design: technical replication at the library preparation level can facilitate normalization and is a good investment in large sequencing projects, especially when multiple centers or platforms are involved[25]. Although we have focused on normalization in the context of differential expression, the RUV approach can be adapted to other settings such as cluster analysis[26].

Internal and external controls are essential for the analysis of high-throughput data and spike-in sequences have the potential to help researchers better adjust for unwanted technical factors. With the advent of single-cell sequencing[27], the role of spike-in standards should become even more important, both to account for technical variability[28] and to allow the move from relative to absolute RNA expression quantification. It is therefore essential to ensure that spike-in standards behave as expected and to develop a set of controls that are stable enough across replicate libraries and robust to both differences in library composition and library preparation protocols.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. GEO: GSE53334 and GSE47792.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the SEQC Consortium for granting us early access to the SEQC pilot data, L. Jacob for his help with the RUV methodology and its software implementation, and J. Gagnon-Bartsch, J. Choi, and W. Shi for helpful discussions. J.N. was supported by a grant from the National Institute on Deafness and Other Communication Disorders. T.P.S. was supported by a National Health and Medical Research Council (NHMRC) Australia Fellowship.

AUTHORS CONTRIBUTION

D.R., S.D. and T.P.S. developed the statistical methods; D.R. and S.D. analyzed the data; J.N. designed the zebrafish experiment; D.R. and S.D. wrote the manuscript; all authors read and approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Bullard, J., Purdom, E., Hansen, K. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
2. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12**, 480 (2011).
3. Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683 (2013).
4. Robinson, M.D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
5. Hansen, K.D., Irizarry, R.A. & Zhijin, W. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).
6. Sun, Z. & Zhu, Y. Systematic comparison of RNA-Seq normalization methods using measurement error models. *Bioinformatics* **28**, 2584–2591 (2012).
7. Yang, Y.H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15 (2002).
8. Oshlack, A., Emslie, D., Corcoran, L.M. & Smyth, G.K. Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. *Genome Biol.* **8**, R2 (2007).
9. Wu, D. *et al.* The use of miRNA microarrays for the analysis of cancer samples with global miRNA decrease. *RNA* **19**, 876–888 (2013).
10. Risso, D., Massa, M.S., Chiogna, M. & Romualdi, C. A modified LOESS normalization applied to microRNA arrays: a comparative evaluation. *Bioinformatics* **25**, 2685–2691 (2009).
11. Lovén, J. *et al.* Revisiting global gene expression analysis. *Cell* **151**, 476–482 (2012).
12. Baker, S.C. *et al.* The external RNA controls consortium: a progress report. *Nat. Methods* **2**, 731–734 (2005).
13. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
14. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
15. Cleveland, W.S. & Devlin, S.J. Locally weighted regression: an approach to regression analysis by local fitting. *JASA* **83**, 596–610 (1988).
16. Qing, T., Yu, Y., Du, T. & Shi, L. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. *Sci. China Life Sci.* **56**, 134–142 (2013).
17. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* doi:10.1038/nbt.2957 (24 August 2014).
18. Canales, R.D. *et al.* Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.* **24**, 1115–1122 (2006).
19. Ferreira, T. *et al.* Silencing of odorant receptor genes by G Protein $\beta\gamma$ signaling ensures the expression of one odorant receptor per olfactory sensory neuron. *Neuron* **81**, 847–859 (2014).
20. Gagnon-Bartsch, J. & Speed, T. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539–552 (2012).
21. Gagnon-Bartsch, J., Jacob, L. & Speed, T.P. Removing unwanted variation from high dimensional data with negative controls. Tech. Rep. 820, Department of Statistics, University of California, Berkeley (2013).
22. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
23. ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306**, 636–640 (2004).
24. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).
25. 't Hoen, P. *et al.* Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31**, 1015–1022 (2013).
26. Jacob, L., Gagnon-Bartsch, J. & Speed, T.P. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. Tech. Rep. 818, Department of Statistics, University of California, Berkeley (2013).
27. Tang, F., Lao, K. & Surani, M.A. Development and applications of single-cell transcriptome analysis. *Nat. Methods* **8**, S6–S11 (2011).
28. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
29. Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots. *JASA* **74**, 829–836 (1979).

ONLINE METHODS

Data sets. *Zebrafish data set.* All procedures were conducted in compliance with US federal guidelines in an AAALAC-accredited facility and were approved by the UC Berkeley Office of Animal Care and Use. Cell pools were created from zebrafish (*Danio rerio*), TgOMP-Gal4;UASGCaMP1.6, mixed sex, 5 days post-fertilization. Olfactory sensory neurons were isolated from three pairs of gallein-treated and control embryonic zebrafish pools and purified by fluorescence-activated cell sorting (FACS)[19]. Each RNA sample was enriched in poly(A)$^+$ RNA from 10–30 ng total RNA and 1 μL (1:1,000 dilution) of Ambion ERCC ExFold RNA Spike-in Control Mix 1 was added to 30 ng of total RNA before mRNA isolation. cDNA libraries were prepared according to manufacturer's protocol. The six libraries were sequenced in two multiplex runs on an Illumina HiSeq2000 sequencer, yielding approximately 50 million 100-base-pair (bp) paired-end reads per library.

We considered for mapping a custom reference sequence, defined as the union of the zebrafish reference genome (Zv9, downloaded from Ensembl[30], v. 67) and the ERCC spike-in sequences (http://tools.invitrogen.com/downloads/ERCC92.fa). Reads were mapped with TopHat[31] (v. 2.0.4, default parameters and supplying the Ensembl GTF annotation through the -G option). Gene-level read counts were obtained using the htseq-count Python script (http://www-huber.embl.de/users/anders/HTSeq/) in the "union" mode and Ensembl (v. 67) gene annotation. After verifying that there were no run-specific biases (data not shown), we used the sums of the counts of the two runs as the expression measures for each library. Genes/spike-ins with more than five reads in at least two libraries were retained, resulting in a total of 20,806 (out of 32,561) expressed genes and 59 (out of 92) "present" spike-ins.

FASTQ files containing the unmapped reads are publicly available in GEO with the accession number GSE53334.

SEQC data set. The third phase of the MicroArray Quality Control (MAQC) project, also known as the Sequencing Quality Control[17] (SEQC) project, aims to assess the technical performance of high-throughput sequencing platforms by generating benchmarking data sets. The design includes four different sample types, namely samples A, B, C and D. Sample A is Stratagene's universal human reference (UHR) RNA; sample B is Ambion's human brain reference RNA; samples C and D are mixes of samples A and B, in a 3:1 and 1:3 ratio, respectively. The four reference samples were sent to several sequencing centers around the world and sequenced using different platforms. Here, we focus on sample A and sample B sequenced at the Australian Genome Research Facility (AGRF) using the Illumina HiSeq2000. Four libraries were prepared for each of sample A and B and multiplex pools of the resulting 8 barcoded libraries were sequenced in 8 lanes of 2 flow-cells, yielding a total of 16 (technical) replicates per library and 64 replicates per sample type. Prior to library preparation, Ambion ERCC ExFold RNA Spike-in Control Mix 1 and Mix 2 were added to sample A and sample B RNA, respectively, in a proportion of 50 µl per 2,500 µl of total RNA. The data consist of an average of 10 million 100-bp paired-end reads per sample.

We considered for mapping a custom reference sequence, defined as the union of the human reference genome (GRCh37, downloaded from Ensembl, v. 69) and the ERCC spike-in sequences. Reads were mapped with TopHat (v. 2.0.6, default parameters and supplying the Ensembl GTF annotation through the -G option). Gene-level read counts were obtained using the htseq-count Python script in the "union" mode and Ensembl (v. 69) gene annotation. Genes/spike-ins with more than five reads in at least ten samples were retained, resulting in a total of 21,559 (out of 55,933) expressed genes and 59 (out of 92) present spike-ins.

In addition to the internal ERCC spike-in positive and negative controls, we used external qRT-PCR positive and negative controls from the original MAQC study[18]. As in our previous work[1,2], among the genes assayed by qRT-PCR, we considered only those that match a unique Ensembl gene, are called present in at least three out of each of the four sample A and sample B qRT-PCR runs, and have standard errors across the eight runs not exceeding 0.25. We found 698 qRT-PCR genes in common with the RNA-seq filtered genes and used this subset to compare expression measures between the two assays. The sample A/sample B expression log-fold-change of a gene is estimated by the log-ratio between the average of the four qRT-PCR measures of sample A and the average of the four measures of sample B.

FASTQ files containing the unmapped reads are publicly available in GEO with the accession number GSE47792. Additional details on the SEQC data set (e.g., blinding, randomization, and statistical power) are available in ref. 17.

ERCC spike-in controls. The External RNA Control Consortium (ERCC)[12] developed a set of 92 polyadenylated transcripts that mimic natural eukaryotic mRNAs. Ambion commercializes two ERCC spike-in mixes, ERCC ExFold RNA Spike-in Control Mix 1 and Mix 2. The two mixes contain the same set of 92 spike-in standards, but at different concentrations. This allows the design of experiments in which the spike-ins can be used both as positive and negative controls. In particular, the spike-ins are divided into four groups of 23 transcripts each, spanning a 10^6-fold concentration range, with approximately the same transcript size and GC-content distributions. The first group has an expected fold-change of 4:1 between the two mixes (Mix 1:Mix2); the second group has an expected fold-change of 1:1 (negative controls); the third and fourth groups have expected fold-changes of 2:3 and 1:2, respectively. (See the white paper at http://tools.invitrogen.com/content/sfs/manuals/cms_086340.pdf for additional details.)

In the zebrafish data set, Mix 1 was used for all samples, so that all spike-ins can be used as negative controls. In the SEQC data set, Mix 1 was added to sample A and Mix 2 to sample B, so that 23 spike-ins can be used as negative controls and 69 as positive controls (23 over-represented and 46 under-represented in sample A).

RUV normalization. Gagnon-Bartsch *et al.*[20,21] developed a method for normalizing (continuous) microarray data coined RUV-2, for remove unwanted variation in two steps. Here, we propose the following extensions of the RUV approach to normalize discrete RNA-seq data. For n samples and J genes, consider the log-linear regression model

$$\log E\,[Y|\; W, X, O] = W\alpha + X\beta + O, \tag{1}$$

where Y is an $n \times J$ matrix containing the observed gene-level read counts, X is an $n \times p$ matrix corresponding to the p covariates of interest/factors of "wanted variation" (e.g., treatment status) and β its associated $p \times J$ matrix of parameters of interest, W is an $n \times k$ matrix corresponding to hidden factors of "unwanted variation" and α its associated $k \times J$ matrix of nuisance parameters, and O is an $n \times J$ matrix of offsets that can either be set to zero or estimated with some other normalization procedure (such as upper-quartile normalization). The matrix X is assumed to be known a priori. For instance, in the usual two-class comparison setting (e.g., treated versus control samples), X is an $n \times 2$ design matrix with a column of ones corresponding to an intercept and a column of indicator variables for the class of each sample (e.g., 0 for control and 1 for treated)[32]. The matrix W is an unobserved random variable and α, β, and k are unknown parameters. The simultaneous estimation of W, α, β and k is infeasible. For a given k, we consider instead the three approaches below to estimate the factors of unwanted variation W.

Unlike previously proposed normalization procedures, RUV can be used to simultaneously normalize read counts ($W\alpha$ term in equation (1)) and infer differential expression ($X\beta$ term), using standard techniques for GLM regression. Normalized counts can also be obtained separately as the residuals from regression of the original counts on the unwanted factors. Note, however, that removing $W\alpha$ from the original counts bears the risk of removing part of the effect of X (ref. 21).

RUVg—RUV with negative control genes.

1. Assume one can identify a set of J_c negative control genes, i.e., non-differentially expressed genes, for which $\beta_c = 0$ and $\log E[Y_c|W,X,O] = W\alpha_c + O_c$, where the subscript c denotes the restriction of matrices to the set of J_c control genes.
2. Define $Z = \log Y - O$ and Z^* as the column-centered version of Z (i.e., the columns of Z^* have zero mean).
3. Perform the singular value decomposition (SVD) of Z_c^*, that is, $Z_c^* = U\Lambda V^T$, where U is an $n \times n$ orthogonal matrix with columns the left singular vectors of Z_c^*, V a $J_c \times J_c$ orthogonal matrix with columns the right singular vectors, and Λ an $n \times J_c$ rectangular diagonal matrix of singular values (at most min (n, J_c) distinct non-zero singular values). For a given k, estimate

doi:10.1038/nbt.2931

$W\alpha_c$ by $\widehat{W\alpha_c} = U\Lambda_k V^T$ and W by $\hat{W} = U\Lambda_k$, where Λ_k is the $n \times J_c$ rectangular diagonal matrix obtained from Λ by retaining only the k largest singular values and setting other diagonal entries to zero (drop null columns to obtain $\hat{W}$).

4. Substitute $\hat{W}$ into equation (1) for the full set of J genes and estimate both α and β by GLM regression.
5. (Optionally) Define normalized read counts as the residuals from ordinary least squares (OLS) regression of Z on $\hat{W}$.

This is essentially the discrete version of RUV-2 (refs. 20,21). The key assumption is that one can identify a set of negative control genes, as detailed below. However, RUV-2 has been found to be quite sensitive to the choice of control genes[20,21,26]. We therefore consider the following two adaptations, which either do not require negative control genes (RUVr) or are more robust to the choice of controls (RUVs).

RUVr—RUV with residuals.

1. Compute an $n \times J$ matrix of residuals E from a first-pass GLM regression of the counts Y on the covariates of interest X (model in equation (1) without $W\alpha$ term), e.g., deviance residuals. The counts may be either unnormalized or normalized with a method such as upper-quartile normalization.
2. Perform the singular value decomposition of the residual matrix, $E = U\Lambda V^T$, and estimate the unwanted factors W by the $n \times k$ matrix $\hat{W} = U\Lambda_k$. Proceed as in steps 4 and 5 of the control gene version of RUV.

RUVs—RUV with replicate/negative control samples.

1. Assume one has replicate samples for which the biological covariates of interest are constant. Then, their count differences behave like those of negative control samples, as they contain no effects of interest. Let $r(i) \in \{1,\ldots,R\}$ denote the replicate group to which sample i belongs; if i does not belong to any replicate group, set $r(i) = 0$. For example, for the SEQC data set, the 64 (= 4 libraries × 2 flow-cells × 8 lanes) replicates of sample A and of sample B each form a replicate group.
2. Column-center the counts within each set of replicate samples, that is, replace the original counts $Y_{i,j}$ by $Y_{i,j} - \bar{Y}_{r(i),j}$, where
$$\bar{Y}_{r,j} = \sum_i I(r(i) = r)Y_{i,j} / \sum_i I(r(i) = r).$$
Let Y_d denote the resulting $n_d \times J$ matrix of column-centered counts for the
$$n_d = \sum_i I(r(i) \neq 0)$$
replicate samples. Then $\log E[Y_d|W,X,O] = W_d\alpha + O_d$, where W_d is $n_d \times k$, α is $k \times J$, and O_d is $n_d \times J$.
3. Perform the singular value decomposition $Z_d^\star = U\Lambda V^T$ (where $Z_d^\star$ is defined as in step 2 of RUVg) and estimate the nuisance parameter α by the $k \times J$ matrix $\hat{\alpha} = \Lambda_k V^T$ obtained by retaining only the k largest singular values. Here, $k \leq \min(n_d, J)$, the upper-bound for the number of distinct non-zero singular values.
4. Estimate the unwanted factors W by OLS regression of Z_c, for all n original samples and a set of J_c negative control genes, on $\hat{\alpha}_c$, $\hat{W} = Z_c\hat{\alpha}_c^T(\hat{\alpha}_c\hat{\alpha}_c^T)^{-1}$. Proceed as in steps 4 and 5 of the control gene version of RUV.

RUV assumptions and scope. Here, we detail the main assumptions and scope of the three proposed RUV approaches. This information is summarized in **Supplementary Table 1**.

1. *Negative control genes with common unwanted factors: RUVg and RUVs.* There exists a set of negative control genes (e.g., empirical or spike-in controls, chosen as indicated below) whose read counts are not influenced by the covariates of interest ($\beta_c = 0$) and for which the estimated factors of unwanted variation span the same linear space as the factors of unwanted variation for all of the genes ($\log E[Y_c|W,X,O] = W\alpha_c + O_c$).
   Interpretation. By modeling the unwanted variation as in equation (1) with the term $W\alpha$ and reestimating α in step 4 using all the genes, RUVg allows gene-specific nuisance effects α. The RUVg assumption is therefore different and more general than the assumptions of global-scaling and regression-based normalization methods, which require unwanted technical effects to be roughly the same for the controls and for the rest of the genes and to be captured by either a single parameter per sample or a regression function between pairs of samples. This is particularly relevant when using the ERCC spike-in controls for normalization purposes.
   Robustness. In practice, this assumption can be relaxed for RUVs, as the method performs well even when its step 4 is based on all genes, provided that the unwanted factors W are not too correlated with the covariates of interest X (ref. 26).
2. *Replicate/negative control samples: RUVs.* There exists a set of negative control samples, that is, samples whose read counts are not influenced by the biological covariates of interest. Such a set can be created easily by computing differences between (technical) replicate samples for which the biological covariates of interest are constant.
   Interpretation. RUVs can only account for variation that occurs within replicate groups, e.g., it can capture library preparation effects only if the replicate groups include multiple libraries.
3. *Known matrix X: RUVg(empirical controls) and RUVr.*
   Interpretation. This assumption is essential for RUVr in order to compute residuals from a first-pass GLM regression of the counts on the covariates of interest. It is needed for RUVg only when there are no a priori known negative control genes and one relies on empirical controls from a first-pass differential expression analysis. The main consequence of this assumption is that RUVg (empirical controls) and RUVr are applicable only to classical differential expression settings (e.g., treatment versus control comparison) and not to clustering (where X is unknown) or time-course (where X is only partially known and model selection may be involved) problems.
4. *Unwanted factors uncorrelated with covariates of interest: RUVr and RUVs.* The unwanted factors W are uncorrelated with the covariates of interest X.
   Interpretation. This assumption is natural for any regression-based method.
   Robustness. In practice, both RUVr and RUVs perform well with modest correlation between W and X.

The residual version RUVr does not need the negative control gene assumption and is suited to situations where the effects of interest are much larger than the unwanted variation (e.g., SEQC data set, see **Fig. 1**). It is similar to previously presented microarray methods[24,33]. The replicate sample version RUVs is adapted to the SEQC data set, with large library preparation effects and replicate libraries for each biological condition, and, to a lesser extent, to the zebrafish data set, where one has three libraries per biological condition.

Choice of negative control genes. The main assumption of RUVg is that one can identify a set of negative control genes. Several types of negative controls could be used, including housekeeping genes, spike-in sequences (e.g., ERCC), or "*in silico*" empirical controls such as the J_c least significantly differentially expressed genes based on a first-pass differential expression analysis performed prior to RUVg normalization.

Interestingly, one can relax the negative control gene assumption by requiring instead the identification of a set of J_c positive or negative controls, for which the value of β_c is known a priori but need not be zero. Then, $X\beta_c$ is known and one can perform the singular value decomposition of $\log Y_c - X\beta_c - O_c$ to estimate W as in step 3 of RUVg above. Steps 4 and 5 remain the same. This allows us to make full use of all 92 ERCC spike-in controls for the SEQC data set.

In this study, we consider two different sets of controls for both the zebrafish and the SEQC data sets: (i) a set of empirical controls, defined as all but the top 5,000 differentially expressed genes, as ranked by *edgeR P*-values for UQ-normalized counts (15,839 genes for the zebrafish data set and 16,500 genes for the SEQC data set); (ii) the 59 ERCC spike-in controls called present. **Supplementary Figures 16** and **17** show that RUVg is robust to the set of empirical control genes.

Choice of number of factors of unwanted variation. The main tuning parameter of RUV is the number of factors of unwanted variation, k. The choice of k should be guided by considerations that include sample size, extent of technical effects captured by the first k factors and extent of differential expression[20,21].

doi:10.1038/nbt.2931

For instance, the small sample size ($n = 6$) for the zebrafish data set only allows the estimation of one or two factors of unwanted variation. Here, we set $k = 1$. The SEQC data set has a much greater sample size ($n = 128$) and more factors can be considered. Here, we set $k = 6$; for the RUVg version, we drop the first unwanted factor, as it captures the biological factor of interest, and retain the next $k = 6$ factors. **Supplementary Figure 18** shows that RUV is robust to the choice of k.

Linear model version of RUV. Although GLM are a natural choice for count data and have been successfully applied to address a broad range of questions in RNA-seq[34,35], a simpler alternative is to consider a linear model (LM) for some suitable transformation of the read counts (e.g., logarithmic transformation). Such an LM-based version of RUVg reduces to RUV-2 (refs. 20,21). Additionally, using a linear model allows approaches such as RUV-4 and RUV-inv (ref. 21).

Supplementary Figures 19 and **20** show that LM-based RUVg on log-counts does not perform as well as our proposed GLM-based RUVg. In particular, although LM-based RUVg seems effective at removing the unwanted variation (cf. uniform distribution of P-values in **Supplementary Fig. 19**), it does not yield enough power to detect any differentially expressed genes, neither when using a standard t-test nor when using an empirical Bayes moderated t-test (*limma*[36]).

Other normalization methods. We compare our RUV approach to the following normalization procedures.

Global-scaling normalization scales gene-level counts by a single factor per sample, such as the per-sample total read count (TC), a.k.a., Reads Per Kilobase of exon model per Million mapped reads, or RPKM[37], a housekeeping gene count (e.g., POLR2A), a quantile of the per-sample count distribution[1] (e.g., upper-quartile, UQ) or other robust summaries obtained by relating each sample to a reference sample (e.g., the trimmed mean of M values (TMM)[4] and the approach of Anders and Huber (AH)[35]).

In full-quantile (FQ) normalization[1,14], all quantiles of the gene count distributions are matched between samples. Specifically, for each sample, the distribution of sorted read counts is matched to a reference distribution defined in terms of a function of the sorted counts (e.g., median) across samples.

In loess normalization[7,11,15], loess robust local regression fits are performed for mean-difference plots of log-counts for pairs of samples, e.g., all possible pairs as in cyclic loess or each sample paired with a synthetic reference obtained by averaging counts across samples.

When a reasonable number of negative controls are available and behave as desired across samples, these could be used directly as part of the normalization procedure, e.g., scaling counts by the upper-quartile of the ERCC spike-in counts or fitting a loess regression only on the spike-ins.

In the main comparison, we focused on four RUV procedures (RUVg using empirical control genes or the ERCC spike-ins, RUVr using all genes, and RUVs using all genes), upper-quartile normalization (using all genes or only the spike-ins) and cyclic loess normalization (using only the spike-ins). All other methods led to very similar results as upper-quartile normalization, as shown in **Supplementary Figures 1–3**.

Evaluation criteria. *Relative log expression.* A particularly useful transformation of read counts is their relative log expression (RLE), defined, for each gene, as the log-ratio of a read count to the median count across samples. Comparable samples should have similar RLE distributions that are centered around zero. Unusual RLE distributions could reveal suspicious samples (e.g., problematic library preparation) or batch effects.

Differential expression analysis. To compare normalization procedures in terms of their impact on differential expression results, we consider the negative binomial GLM of *edgeR*[34], with tag-wise dispersion. Upper-quartile normalization is performed through an offset using the *calcNormFactors* function. RUV normalization is performed by including the estimated W matrix in the GLM. Cyclic loess and upper-quartile normalization using the ERCC spike-ins are performed by directly providing the *offset* argument to the *glmFit* function. Differentially expressed genes are identified by likelihood ratio tests for the effects of interest: for the zebrafish data set, treatment effect; for the SEQC data set, sample A versus B effect and, in the null experiment of **Figure 3b**, library preparation effect. A gene is declared differentially expressed if the associated null hypothesis is rejected at a false discovery rate (FDR)[38] of 0.05.

Bias. In order to evaluate bias in log-fold-change estimation, one needs to know the true value of the expression fold-change. For the SEQC data set, one can use the estimate of the sample A/sample B fold-change from qRT-PCR as the true value, since qRT-PCR is often considered as a gold standard for producing accurate estimates of expression levels. The RNA-seq estimated fold-change is the ratio of the average of the normalized counts for the 64 sample A replicates to the average of the normalized counts for the 64 sample B replicates. For a given gene, bias is then estimated as the difference between the estimated log-fold-changes from the two technologies.

Receiver operating characteristic curves. For the SEQC data set, the qRT-PCR measures are used as gold standard to determine "true" differential expression and derive receiver operating characteristic (ROC) curves for the various normalization methods. As in previous work[1], we divide the genes assayed by qRT-PCR into three sets, "non-differentially expressed", "differentially expressed" and "no-call", based on whether their absolute expression log-fold-change is less than 0.2, greater than 1 or falls within the interval [0.2, 1], respectively. We ignore the no-call genes when determining true/false positives/negatives. False positives (FP) are defined as genes declared differentially expressed by RNA-seq (*edgeR* FDR adjusted P-value less than 0.05) but not by qRT-PCR. True negatives (TN) are defined as genes declared non-differentially expressed by both RNA-seq and qRT-PCR. True positives (TP) are declared differentially expressed by both RNA-seq and qRT-PCR. The true positive rate (TPR) is then defined as the number of TP divided by the number of differentially expressed genes according to qRT-PCR. The false positive rate (FPR) is defined analogously as the ratio of the number of FP to the number of non-differentially expressed genes according to qRT-PCR.

Software implementation. RUV is implemented in the open-source R package *RUVSeq*, with source code freely available through the Bioconductor Project[39] (http://www.bioconductor.org/packages/devel/bioc/html/RUVSeq.html) and as **Supplementary Software**. Gene-level counts for the zebrafish data set are provided in the Bioconductor R package (http://www.bioconductor.org/packages/devel/data/experiment/html/zebrafishRNASeq.html).

30. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res.* **40**, D84–D90 (2012).
31. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
32. McCullagh, P. & Nelder, J. *Generalized Linear Models* (Chapman and Hall, New York, 1989).
33. Listgarten, J., Kadie, C., Schadt, E.E. & Heckerman, D. Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl. Acad. Sci. USA* **107**, 16465–16470 (2010).
34. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
35. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
36. Smyth, G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 3 (2004).
37. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
38. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc., B* **57**, 289–300 (1995).
39. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).

doi:10.1038/nbt.2931