May 2011

REMOVING TECHNICAL VARIABILITY IN RNA-SEQ DATA USING CONDITIONAL QUANTILE NORMALIZATION

Kasper D. Hansen
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Rafael A. Irizarry
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Zhijin Wu
Department of Community Health, Section of Biostatistics, Brown University, Zhijin_Wu@brown.edu

Removing technical variability in RNA-seq data using conditional quantile normalization

Kasper D. Hansen[1], Rafael A. Irizarry[1], and Zhijin Wu[2,*]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
[2]Department of Community Health, Section of Biostatistics, Brown University

Abstract

The ability to measure gene expression on a genome-wide scale is one of the most promising accomplishments in molecular biology. Microarrays, the technology that first permitted this, were riddled with problems due to unwanted sources of variability. Many of these problems are now mitigated, after a decade's worth of statistical methodology development. The recently developed RNA sequencing (RNA-seq) technology has generated much excitement in part due to claims of reduced variability in comparison to microarrays. However, we show RNA-seq data demonstrates unwanted and obscuring variability similar to what was first observed in microarrays. In particular, we find GC-content has a strong sample specific effect on gene expression measurements that, if left uncorrected, leads to false positives in downstream results. We also report on commonly observed data distortions that demonstrate the need for data normalization. Here we describe statistical methodology that improves precision by 42% without loss of accuracy. Our resulting conditional quantile normalization (CQN) algorithm combines robust generalized regression to remove systematic bias introduced by deterministic features such as GC-content, and quantile normalization to correct for global distortions.

1 Introduction

High-throughput sequencing technology is currently being used to quantify gene expression levels on a genome-wide scale. This is done by first converting RNA transcripts into cDNA fragments, and then sequencing these fragments to produces millions of sequences of length 35-150 bases

*To whom correspondence should be addressed. Email: zhijin_wu@brown.edu

(bps), referred to as *reads*. Gene expression is quantified by counting the number of these reads that map back to each gene. The conventional wisdom is that this approach is an improvement over microarrays as it is a direct measurement of RNA levels and does not rely on hybridization, a process known for its lack of specificity (Wu *and others*, 2004; Zhang *and others*, 2003; Naef and Magnasco, 2003). Early studies, based on a small number of samples in highly controlled conditions, found that RNA-seq has excellent technical reproducibility (Mortazavi *and others*, 2008; Marioni *and others*, 2008; Bullard *and others*, 2010). Furthermore, in a review article Wang *and others* (2009) claimed that analysis of RNA-seq data does not require "sophisticated normalization". However, as more data became available, problems such as sequence-specific biases were reported (Hansen *and others*, 2010; Li *and others*, 2010; Pickrell *and others*, 2010). Here, we make use of three large, publicly available, RNA-seq data sets to demonstrate that sample specific systematic biases, along with distortions that affect the overall distribution of count data, introduce unwanted variation in RNA-seq data that obscures the underlying biological signal.

RNA-seq technology permits applications not previously possible with microarrays. For example, it is possible to discover new alternative transcription, unannotated transcription or measure transcription for non-coding regions (Sultan *and others*, 2008; Trapnell *and others*, 2010; Wilhelm *and others*, 2010; Perkins *and others*, 2009). Although most RNA-seq publications have focused on discovering new transcripts, determining whether the expression level of a genomic unit (such as a gene, exon, or junction) differs across experimental conditions continues to be an important question in functional genomics. Therefore, to demonstrate the importance of normalization in RNA-seq data we focus on the application of differential expression detection (Bottomly *and others*, 2011; Wu *and others*, 2010; Lefebvre *and others*, 2011; Anders and Huber, 2010; Robinson *and others*, 2010; Eveland *and others*, 2010).

We start by counting the number of reads in predetermined genomic regions for each sample to form gene expression matrices with rows representing genes and columns representing samples as with microarray data. Although our approach is agnostic to the choice of regions, here we used regions selected to correspond to the canonical definitions of genes as defined by the Ensembl database (Flicek *and others*, 2011). Because most tests developed for differential expression testing in microarray data depend on assumptions not necessarily applicable to the count data produced by RNA-seq, alternative statistical methodologies have been proposed (Anders and Huber, 2010; Robinson *and others*, 2010; Robinson and Smyth, 2007, 2008). Similarly, alternative normalization approaches have been proposed. The first normalization approach described in the literature was to simply correct each sample for the number of mapped reads produced for each sample, referred to as *sequencing depth*, and each gene for its length (Mortazavi *and others*, 2008). Because variability in sequencing depth was observed in technical replicates, it was assumed to be a technical artifact and because longer genes are expected to have higher counts, Mortazavi *and others* (2008) defined the widely used "reads per kilobase per million" (RPKM) measure as the number of reads mapped to a gene in a sample divided by the product of the length of the gene in kilobases and the total number of reads mapped in the sample, in millions. Various authors then showed that sequencing depth is not a stable scaling factor and a number of more robust alternatives were

suggested (Bullard *and others*, 2010; Robinson and Oshlack, 2010; Anders and Huber, 2010), with Langmead *and others* (2010) suggesting that there might be a gene-specific linear effect of the sample-specific scaling factor. However, in Section 3 we demonstrate that, even with improved scaling, the use of the RPKM measure is not a general solution to the unwanted variability problem. In Section 2 we describe the datasets used throughout the paper and Section 3 motivates the problem and our solution. In Section 4 we present a useful statistical model and use it to motivate a normalization algorithm. In Section 5 we present results illustrating the improvements made possible by our approach. Finally, in Section 6 we discuss future directions and connections to existing methodology for differential expression detection.

2 Data Description

We examined the three currently available RNA-seq datasets with the largest number of samples (Pickrell *and others*, 2010; Montgomery *and others*, 2010; Cheung *and others*, 2010). In all three studies, the samples are lymphoblastoid cell lines from unrelated individuals in the HapMap project (International HapMap Consortium, 2003). Montgomery *and others* (2010) sequenced 60 individuals from the CEPH (Utah residents with ancestry from northern and western Europe) population (CEU). Cheung *and others* (2010) sequenced 41 individuals also from the CEU population with 29 in common with Montgomery *and others* (2010). Pickrell *and others* (2010) sequenced 69 individuals from Yoruba in Ibadan, Nigeria. All three studies, hereafter referred to as Montgomery, Pickrell and Cheung, were designed to study the effect of genetics on gene expression and subjects were considered interchangeable. We therefore used these data to assess improvements in precision. The samples that were done in replicate across two studies were particularly useful for this purpose.

To assess accuracy we used samples from Bullard *and others* (2010), in which two samples from the microarray quality control study (MAQC Consortium, 2006) were sequenced. These two samples are Stratagene's universal human reference RNA (UHR) which is a commercial pool of RNA from 15 different cell lines, and Ambion's human brain reference RNA. The same samples have been assayed extensively on microarrays, and we used data from the MAQC Consortium (2006) in which each of the two samples was hybridized to five different Affymetrix U133 Plus 2.0 arrays. The microarrays served as an independent measurement that permitted an assessment of accuracy. This dataset has no biological replicates, and the technical replicates are based on commercially available RNA, making the technical noise smaller than what would be expected from tissue samples.

2.1 Data Processing

For all datasets the original reads were downloaded, mapped, and the gene expression count matrix created as follows. Reads were aligned to the human reference genome sequence (version hg19) using Bowtie (Langmead *and others*, 2009), allowing up to two mismatches. All reads were trimmed from the 3' end to be 35bp long, and for the Montgomery data we used the first read of the paired-end reads. To assign reads to genes we followed essentially the same procedure used by (Bullard *and others*, 2010), except for (a) we determined overlap between a read and a genomic region based on the center base of the trimmed read and not the 5' end and (b) we used union gene representations instead of union-intersection gene representation as discussed in Bullard *and others* (2010). Sequencing depth was determined as the number of reads mapped to the genome.

3 Motivation

The need for a normalization technique more complex than scaling is first motivated by simply noting that the distribution of counts across different samples differs (Figure 1(a)). Since the raw counts are affected by sequencing depth, we also compared the distribution of reads per million (RPM), in each sample to account for differences in sequencing depth (not shown). The locations of the peaks of the RPM densities of these replicates became closer, but both the shape and scale of the distributions still vary between these replicates. Clearly a scaling normalization, that is, a shift in log scale expression, is not sufficient to normalize counts between samples.

Contrary to an early expectation of RNA-seq technology, the number of reads from a given gene is not simply determined by the gene expression level. Rather, certain fragments are preferentially detected in the RNA-seq data acquisition process, leading to nonuniform detection of expression between genes. We refer to this bias in measurement as *counting efficiency*. The best documented example is the effect of the percent of C or G nucleotides in a gene: the so-called *GC-content* effect. GC-content has been shown to influence a number of DNA-related measurements. Examples include gene expression microarrays (Wu *and others*, 2004; Zhang *and others*, 2003; Naef and Magnasco, 2003), copy number arrays (Nannya *and others*, 2005; Carvalho *and others*, 2007), sequencing coverage (Dohm *and others*, 2008) and RNA-seq (Pickrell *and others*, 2010). The difference in counting efficiency between genes means that expression levels cannot be compared between genes directly. A more subtle and detrimental problem is that these systematic biases affect different samples differently, thus even within gene comparison between two samples becomes problematic. In fact Pickrell *and others* (2010) demonstrated that the GC-content effect can change from sample to sample. Here we demonstrate that this appears to be a general problem. In Figure 1(b) we show the distribution of $\log_2$-RPKM for various strata of gene GC-content for two biological replicates from the Montgomery study. For illustration purposes we selected one sample in which a higher GC-content leads to increased counting efficiency, and another in which

there is little impact. This problem has downstream consequences since observed fold changes are obscured by the variability introduced by GC-content effects (Figure 1(c,d)).

Some work has been done to develop methods to address these effects. Pickrell *and others* (2010), the first to notice the sample specific GC-content effect, proposed a sample specific adjustment. They suggested stratifying predefined genomic regions by GC-content and then for each stratum, divide the sample counts by the sum of the counts across all samples. This fraction is considered an enrichment factor for that GC-content stratum, which is then smoothed by GC-content for each sample separately. Counts are then adjusted by the smoothed enrichment factor. Finally, they proposed doing this on the exon level, adding adjusted counts across all exons from a gene in order to obtain gene level adjusted counts. We found two problems with this approach that we decided to improve. First, the enrichment scores are computed for each sample relative to all samples in an experiment, thus this adjustment does not remove the GC-content effect but rather equalizes the effect across samples. As a consequence, adjustments vary depending on what samples are processed together. Second, the GC-content effect is estimated based on the direct summation of counts on different genes in different samples, ignoring the fact that genes with higher expected counts also have greater variance. As a result GC-content effects are not entirely removed (Figure 1(e)). In addition, Roberts *and others* (2011) addresses bias removal within the Cufflinks transcript assembly framework (Trapnell *and others*, 2010) and show improvements in comparisons between sequencing technologies, but does not address variation between biological replicates.

4 Methods

We present a normalization algorithm motivated by a statistical model that accounts for both the need to correct systematic biases and the need to adjust for distributional distortions. We denote the log gene expression level for gene g at sample i with $\theta_{g,i}$, which we consider a random variable. For most g, $\theta_{g,i}$ are independent and identically distributed across i. We assume that the marginal distribution of the $\theta_{g,i}$ is the same for all samples i, and denote it by G. Note that this variability accounts for the difference in gene expression across different genes. The p covariates thought to cause systematic errors are denoted with $\mathbf{X}_g = (X_{g,1}, \ldots, X_{g,p})'$. Examples of covariates considered here are GC-content, gene length, and gene mappability defined as the percentage of uniquely mapping subreads of a gene. To model the observed counts $Y_{g,i}$ for gene g in sample i we write:

$$Y_{g,i} \,|\, \mu_{g,i} \sim \text{Poisson}(\mu_{g,i})$$

with

$$\mu_{g,i} = \exp\left\{ h_i(\theta_{g,i}) + \sum_{j=1}^{p} f_{i,j}(X_{g,j}) \right\}$$

with $f_{i,j}(\bar{X}_{\cdot j}) = 0 \,\forall\, j$ for identifiability. Here, the h_is are non-decreasing functions that account for the fact that count distributions are distorted in non-linear ways across the different samples (Figure 2(a)). The $f_{i,j}$s account for sample dependent systematic biases. Data exploration suggested

that these are smooth functions, so for tractability we model these as (parametric) natural cubic splines with known degrees of freedom and knot locations. If there is no technical variability, h_i is the identity function and $\sum_{j=1}^{p} f_{i,j}(X_{g,j}) = 0$, then the distribution of Y_{gi} for a given i reduces to a G-Poisson mixture.

With the model in place, obtaining normalized counts is equivalent to estimating $\theta_{g,i}$. To do this we needed to estimate the non-parametric h_i functions along with the linear parameters that define the splines. Note that the distribution of the $\theta_{g,i}$ in a sample is determined by the biological system, which varies greatly between species, tissue types and developmental stages. Thus it is unrealistic to restrict it to a particular parametric family of distributions. This makes estimation requiring full likelihood, including maximum likelihood estimation (MLE) and Bayesian approaches unsuitable. In addition, outliers can arise because of either biological activity or technical artifacts. Since both h and f represent the global impact of systematic effects on all genes in general, it is crucial to define estimation procedures that are robust to outliers. We take advantage of the large amount of data for each sample and our parsimonious model to define a stable algorithm which we now motivate and describe.

For any given i, the distribution of $h_i(\theta_{g,i})$ is unspecified and Figure 2(b) shows that values can range from $-\infty$ to 8. First we observe that when $\mu_{g,i}$ is large, $\log(Y_{g,i}) \,|\, \mu_{g,i}$ is approximately normal with mean $\log(\mu_{g,i})$ and variance $1/\mu_{g,i}$. The small variance implies that for large $\mu_{g,i}$

$$\log(Y_{g,i}) \,|\, \mu_{g,i} \approx \log(\mu_{g,i}) = h_i(\theta_{g,i}) + \sum_{j=1}^{p} f_{i,j}(X_{g,j}),$$

showing that for a fixed i and large $\mu_{g,i}$, the distribution of $\log(Y_{g,i})$ is equal to $h_i(\theta_{g,i})$ except for a location shift given by $\sum_{j=1}^{p} f_{i,j}(X_{g,j})$. Even though the shape of $h_i(G)$ is left unspecified, the quantiles of $\log(Y_{g,i})$ shift by $\sum_{j=1}^{p} f_{i,j}(X_{g,j})$. We therefore use quantile regression to estimate the $f_{i,j}$s. To assure the large $\mu_{g,i}$ assumption is satisfied, instead of fixing the quantile choice, we use median regression on a subset of genes with average counts beyond a lower bound.

To estimate the h_is we take advantage of the fact that

$$\mathrm{E}\left\{\log(Y_{g,i}) - \sum_{j=1}^{p} f_{i,j}(X_{g,j})\right\} = h_i(\theta_{g,i})$$

and that the distribution of $\theta_{g,i}$ does not depend on i, to use subset quantile normalization (Wu and Aryee, 2010).

The specifics of our algorithm are as follows:

1. Select a subset of genes with $\bar{Y}_{g,\cdot} > 50$. Then for each i, use median regression on $\log(Y_{g,i})$ to estimate the parameters that define the splines $f_{i,j}$ and determine $\hat{f}_{i,j}$.

2. For each i, apply quantile normalization to $\log(Y_{g,i}) - \sum_{j=1}^{p} \hat{f}_{i,j}(X_{g,j})$ to obtain $\hat{h}_i^{-1}$.

3. For each gene g on each sample i, define a *normalization offset* as $\exp[\log(Y_{g,i}) - \hat{h}^{-1}\{\log(Y_{g,i}) - \hat{f}_{i,j}(X_{g,j})\}]$.

The algorithm returns an offset rather than normalized data for two reasons. First, for interpretability we want to preserve the data as counts, i.e. integer numbers. Due to the large sampling error, small counts should be treated with caution thus users of the algorithm benefit from access to these original counts. Second, the most widely used methodology for identifying differentially expressed genes from RNA-seq data model the counts in a way that sampling error from counting process (such as Poisson) and variation in gene expression (θ) are taken into account (Robinson *and others*, 2010; Anders and Huber, 2010). Providing an offset allows direct application of these existing methods which take counts as input and can be easily adapted to adjust for offsets.

While the algorithm allows one to correct for a variety of systematic biases, we have consistently used GC content and gene length. An R package *cqn* implementing the method is being submitted to Bioconductor.

5 Results

Because experimentally controlling for the amount of RNA extracted from a sample is difficult, the total number of counts varies across samples and manifests itself as between sample differences in the locations of the log read-count distributions (Figure 1(a)). This unwanted technical variability is further augmented by the differences in cDNA amplification efficiency (Aird *and others*, 2011) and other technical artifacts and differences in distribution shapes and scales persist after library size is taken into account. Scaling normalization based on more robust estimates of the shift in location (Bullard *and others*, 2010; Robinson and Oshlack, 2010; Anders and Huber, 2010) can provide further improvement, although improvement is limited in the samples we have analyzed (as an example, results for trimmed median of M-values, TMM, from (Robinson and Oshlack, 2010) are shown below). In contrast, our normalization approach (CQN) results in sample distributions with comparable scales and shapes, as discussed below.

To demonstrate the down-stream advantages of our algorithm we first considered comparisons between two samples. For illustrative purposes we selected two samples with very different systematic bias patterns ($f_{i,j}$s). For the assessment, we focused on fold-change as it is considered the basic unit for differential expression analysis. We computed log-fold-change for each gene after both RPKM normalization and CQN and a substantial improvement was observed (Figure 1(e,f)). Specifically, while the RPKM showed a strong dependence between fold-change and GC-content, CQN eliminated it.

The resulting estimates of $\hat{f}_{i,j}$ provided a useful quality assessment since plotting these demonstrated a wide range of GC-content and gene length effects (Figure 3(c) and (d)). In the data sets we analyzed, length effects were more consistent between samples than GC-effects. For many samples the length effect is close to linear with a constant slope for genes shorter than 5000 bp. This result implies that dividing by gene length, as done by the RPKM approach, is suitable in most circumstances. However, we observed that for genes shorter than 1000bp the length effect appears to be stronger, while for genes beyond 5000bp the length effect plateaus. This suggests that dividing by gene length may not always be appropriate. A sample-specific gene length effect may capture sample specific fragmentation bias as well as differences in size selection.

We have illustrated the potential downstream consequences of not normalizing with a comparison of two samples (Figure 1). To demonstrate the advantages of CQN in a study with replicates, we performed a five versus five comparison of biological replicates, between which we expect little difference. Systematic bias was observed in the average log fold changes with a strong dependence on GC content, using standard RPKM (Figure 3(e)). These problems were removed by CQN (Figure 3(f)). The log-fold-variation was noticeably reduced by CQN (Figure 4(a)).

To perform a global assessment of precision, we compared the 29 Hapmap samples processed by both the Cheung and the Montgomery studies. For each gene we computed the mean squared difference between the expression measures from the two technical replicates. Our approach improved precision greatly as shown in Figure 4(b): the median mean squared difference was reduced by 42% after normalization. Note that this shows improvements in across study comparisons.

Finally, to assure that the gains in precision were not achieved by simply reducing overall dynamic range, we assessed accuracy by comparing RNA-seq counts to measurements from an independent technology: microarrays. Specifically, we computed log-fold-change values between UHR and brain and averaged these across all replicates. We did this for both microarrays and sequencing counts and then compared the agreement with microarrays to sequencing counts after RPKM normalization or CQN. We found similar accuracy (Figure 4(c)): between technology correlations were 0.84 using CQN normalization compared to 0.85 using standard RPKM. Indeed, using standard RPKM instead of CQN normalization produced a plot very similar to Figure 4(c) (not shown).

6 Discussion

Unlike previous reports based on small samples, by examining large datasets processed from four different studies, we found RNA-seq data to be greatly affected by bias and systematic errors. Just as with microarrays, we found that lack of normalization can lead to false positives in a differential expression analysis. Particularly, sample specific GC-content effects led to confounding of GC-content and observed log-fold-change values. Apart from correcting for biases such as GC-content, we also found a need to quantile normalize the data to correct for sample-specific distortion. To

remove these unwanted sources of variation we developed a normalization procedure for RNA-seq data that greatly improves precision without affecting accuracy. We demonstrated the improvements with comparisons of two biological samples and a five versus five example. Although in a comparison with many biological replicates the observed sample specific biases may cancel out, large studies are not the norm due to the cost and current optimistic view of the technology's precision (Hansen *and others*, 2011). More importantly, we show a great increase in precision across studies using CQN.

The datasets used for evaluating our method consists of the largest number of biological replicates in the literature. Unlike the situation for microarrays, there is no reference dataset with biological replicates available, in which the expression changes are known for a subset of genes. Thus we were limited to using different datasets for evaluating precision and accuracy.

Normalization methods developed earlier, including the RPKM measure and the various forms of modified scaling normalization (Bullard *and others*, 2010; Robinson and Oshlack, 2010; Anders and Huber, 2010), considered the sample effect as common for all genes, and thus only one scaling factor is estimated in a sample. Although RPKM takes gene length into account, the effect of this length covariate is considered static and constant for all samples. By studying four different RNA-seq datasets we found that these assumptions do not always hold. In practice, the GC-content effect may vary substantially between samples and the same is true to a lesser degree for gene length. In general, our approach to quantifying systematic biases was useful for quality assessment as it identified specific particularly problematic samples (Figures 3(c,d)).

Instead of returning a normalized version of the data, our procedure returns a gene specific normalization offset. This allows direct adaptation of existing methodology with a generalized linear model structure. For example, Robinson *and others* (2010) use an offset that is sample specific but common to all genes in a sample, a value similar to library size but estimated from the data. This offset essentially makes the mean M value in the MA plot equal to zero. We have demonstrated that this approach does not remove GC-content effects. However, by incorporating our offset into their method, this problem is easily solved. Anders and Huber (2010) model counts with a negative binomial distribution in which the mean of a gene is similarly proportional to a sample specific offset that represents sequencing depth. This model can also be easily adapted by replacing this size factor by the offset estimated by CQN.

Quantile normalization has been widely applied to microarray data and shown to have excellent performance compared with competing nonlinear normalization methods (Bolstad *and others*, 2003). We found that when there are known and measurable confounders, as in the RNA-seq example, estimating and removing their effects before quantile normalization provides further advantage. For the results presented here, we considered only two covariates, GC-content and gene length, but our model permits the inclusion of others: for example, mappability or more elaborate sequence effects. Although the biochemical and technical mechanisms for the inconsistent systematic biases between samples are not fully explained, these biases can be estimated and adjusted

for because of the high throughput nature of RNA-seq technology. Even in situations in which direct quantile normalization has been considered useful, we suggest using exploratory plots (Figures 3(c,d)) before applying quantile normalization.

Acknowledgements

This work was supported by National Institutes of Health [grant number R01HG004059] and National Science Foundation [grant number DBI-1054905]. *Conflict of Interest:* None declared.

References

AIRD, DANIEL, ROSS, MICHAEL G, CHEN, WEI-SHENG, DANIELSSON, MAXWELL, FENNELL, TIMOTHY, RUSS, CARSTEN, JAFFE, DAVID B, NUSBAUM, CHAD AND GNIRKE, ANDREAS. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**(2), R18.

ANDERS, SIMON AND HUBER, WOLFGANG. (2010). Differential expression analysis for sequence count data. *Genome Biology* **11**(10), R106.

BOLSTAD, B.M., IRIZARRY, R.A., ÅSTRAND, M. AND SPEED, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**(2), 185–193.

BOTTOMLY, D., WALTER, N.A.R., HUNTER, J.E., DARAKJIAN, P., KAWANE, S., BUCK, K.J., SEARLES, R.P., MOONEY, M., MCWEENEY, S.K. AND HITZEMANN, R. (2011). Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays. *PloS One* **6**(3), e17820.

BULLARD, JAMES H, PURDOM, ELIZABETH, HANSEN, KASPER DANIEL AND DUDOIT, SANDRINE. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94.

CARVALHO, B., BENGTSSON, H., SPEED, T.P. AND IRIZARRY, R.A. (2007). Exploration, normalization, and genotype calls of high-density oligonucleotide snp array data. *Biostatistics* **8**(2), 485.

CHEUNG, VIVIAN G, NAYAK, RENUKA R, WANG, ISABEL XIAORONG, ELWYN, SUSANNAH, COUSINS, SARAH M, MORLEY, MICHAEL AND SPIELMAN, RICHARD S. (2010). Polymorphic Cis- and Trans-Regulation of Human Gene Expression. *PLoS Biology* **8**(9), e1000480.

DOHM, J.C., LOTTAZ, C., BORODINA, T. AND HIMMELBAUER, H. (2008). Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic Acids Research* **36**(16), e105.

EVELAND, A.L., SATOH-NAGASAWA, N., GOLDSHMIDT, A., MEYER, S., BEATTY, M., SAKAI, H., WARE, D. AND JACKSON, D. (2010). Digital gene expression signatures for maize development. *Plant physiology* **154**(3), 1024.

FLICEK, P., AMODE, M.R., BARRELL, D., BEAL, K., BRENT, S., CHEN, Y., CLAPHAM, P., COATES, G., FAIRLEY, S., FITZGERALD, S. *and others*. (2011). Ensembl 2011. *Nucleic Acids Research* **39**(suppl 1), D800.

HANSEN, KASPER DANIEL, BRENNER, STEVEN E AND DUDOIT, SANDRINE. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research* **38**(12), e131.

HANSEN, KASPER D., WU, ZHIJIN, IRIZARRY, RAFAEL A. AND LEEK, JEFFREY T. (2011). Sequencing technology does not eliminate biological variability. *Nature Biotechnology*. In press.

INTERNATIONAL HAPMAP CONSORTIUM. (2003). The International HapMap Project. *Nature* **426**(6968), 789–796.

LANGMEAD, BENJAMIN, HANSEN, KASPER D AND LEEK, JEFFREY T. (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biology* **11**(8), R83.

LANGMEAD, BEN, TRAPNELL, COLE, POP, MIHAI AND SALZBERG, STEVEN L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**(3), R25.

LEFEBVRE, G., DESFARGES, S., UYTTEBROECK, F., MUNOZ, M., BEERENWINKEL, N., ROUGEMONT, J., TELENTI, A. AND CIUFFI, A. (2011). Analysis of HIV-1 expression level and sense of transcription by high-throughput sequencing of the infected cell. *Journal of Virology*, JVI–00252.

LI, JUN, JIANG, HUI AND WONG, WING H. (2010). Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology* **11**(5), R50.

MAQC CONSORTIUM. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* **24**(9), 1151–1161.

MARIONI, J C, MASON, C E, MANE, S M, STEPHENS, M AND GILAD, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18**(9), 1509–1517.

MONTGOMERY, STEPHEN B, SAMMETH, MICHA, GUTIERREZ-ARCELUS, MARIA, LACH, RADOSLAW P, INGLE, CATHERINE, NISBETT, JAMES, GUIGO, RODERIC AND DERMITZAKIS, EMMANOUIL T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**(7289), 773–777.

MORTAZAVI, ALI, WILLIAMS, BRIAN A, MCCUE, KENNETH, SCHAEFFER, LORIAN AND WOLD, BARBARA. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**(7), 621–628.

NAEF, F. AND MAGNASCO, M.O. (2003). Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Physical Review E* **68**(1), 011906.

NANNYA, Y., SANADA, M., NAKAZAKI, K., HOSOYA, N., WANG, L., HANGAISHI, A., KUROKAWA, M., CHIBA, S., BAILEY, D.K., KENNEDY, G.C. *and others*. (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer research* **65**(14), 6071.

PERKINS, T.T., KINGSLEY, R.A., FOOKES, M.C., GARDNER, P.P., JAMES, K.D., YU, L., ASSEFA, S.A., HE, M., CROUCHER, N.J., PICKARD, D.J. *and others*. (2009). A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. *PLoS Genetics* **5**(7), e1000569.

PICKRELL, JOSEPH K, MARIONI, JOHN C, PAI, ATHMA A, DEGNER, JACOB F, ENGELHARDT, BARBARA E, NKADORI, EVERLYNE, VEYRIERAS, JEAN-BAPTISTE, STEPHENS, MATTHEW, GILAD, YOAV AND PRITCHARD, JONATHAN K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**(7289), 768–772.

ROBERTS, ADAM, TRAPNELL, COLE, DONAGHEY, JULIE, RINN, JOHN L AND PACHTER, LIOR. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology* **12**(3), R22.

ROBINSON, MARK D, MCCARTHY, DAVIS J AND SMYTH, GORDON K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1), 139–140.

ROBINSON, MARK D AND OSHLACK, ALICIA. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**(3), R25.

ROBINSON, MARK D AND SMYTH, GORDON K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**(21), 2881–2887.

ROBINSON, MARK D AND SMYTH, GORDON K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9**(2), 321–332.

SULTAN, M., SCHULZ, M.H., RICHARD, H., MAGEN, A., KLINGENHOFF, A., SCHERF, M., SEIFERT, M., BORODINA, T., SOLDATOV, A., PARKHOMCHUK, D. *and others*. (2008). A

global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**(5891), 956.

TRAPNELL, COLE, WILLIAMS, BRIAN A, PERTEA, GEO, MORTAZAVI, ALI, KWAN, GORDON, VAN BAREN, MARIJKE J, SALZBERG, STEVEN L, WOLD, BARBARA J AND PACHTER, LIOR. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**(5), 511–515.

WANG, ZHONG, GERSTEIN, MARK AND SNYDER, MICHAEL. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**(1), 57–63.

WILHELM, B.T., MARGUERAT, S., GOODHEAD, I. AND BÄHLER, J. (2010). Defining transcribed regions using RNA-seq. *Nature protocols* **5**(2), 255–266.

WU, ZHIJIN AND ARYEE, MARTIN J. (2010). Subset quantile normalization using negative control features. *Journal of Computational Biology* **17**(10), 1385–1395.

WU, Z., IRIZARRY, R.A., GENTLEMAN, R., MARTINEZ-MURILLO, F. AND SPENCER, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association* **99**(468), 909–917.

WU, Z.J., MEYER, C.A., CHOUDHURY, S., SHIPITSIN, M., MARUYAMA, R., BESSARABOVA, M., NIKOLSKAYA, T., SUKUMAR, S., SCHWARTZMAN, A., LIU, J.S. *and others*. (2010). Gene expression profiling of human breast tissue samples using SAGE-Seq. *Genome Research* **20**(12), 1730.

ZHANG, L., MILES, M.F. AND ALDAPE, K.D. (2003). A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotechnology* **21**(7), 818–821.

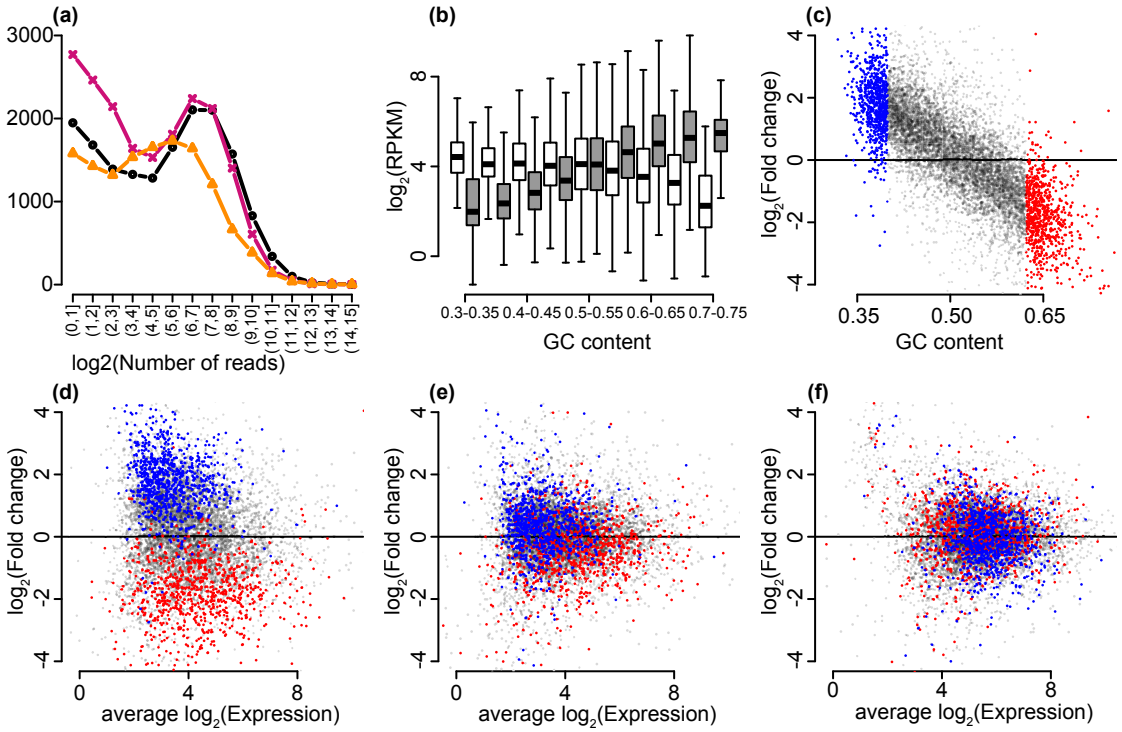

Figure 1. Exploratory plots. (a) The points show the frequency of counts in the bins shown on the x-axis. The three colors represent three samples (NA12812, NA12874, NA11993) from the Montgomery data. (b) $\log_2$ RPKM values are stratified by GC-content for two biological replicates from the Montgomery data (NA11918, NA12761) and are summarized by boxplots. The two samples are distinguished by the two colors. Genes with average (across all 60 samples) $\log_2$ RPKM values below 2 are not shown. (c) Log-fold changes between RPKM values from the two samples and the same genes shown in (b) were computed and are plotted against GC-content. Red is used to show the genes with the 10% highest GC-content and blue is used to show the genes with the 10% lowest GC-content. (d) RPKM log-fold-changes are plotted against average $\log_2$ counts for the samples and genes shown in (b), with the same color coding as in (c). (e) As (d) but from values corrected using the method proposed by Pickrell *and others* (2010). (f) As (d) but for values normalized using our approach (see Methods).

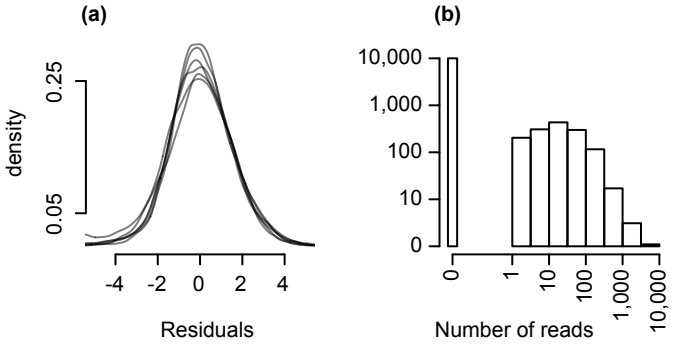

Figure 2. Empirical distributions. (a) Empirical density estimates of $\log(Y_{g,i}) - \hat{f}_{i,j}(X_{g,j})$ are shown for six samples from the Montgomery data. (b) A histogram of counts in a single sample for genes with a GC content of $45\% \pm 1\%$ and with a length between 500bp and 2,000bp is shown.

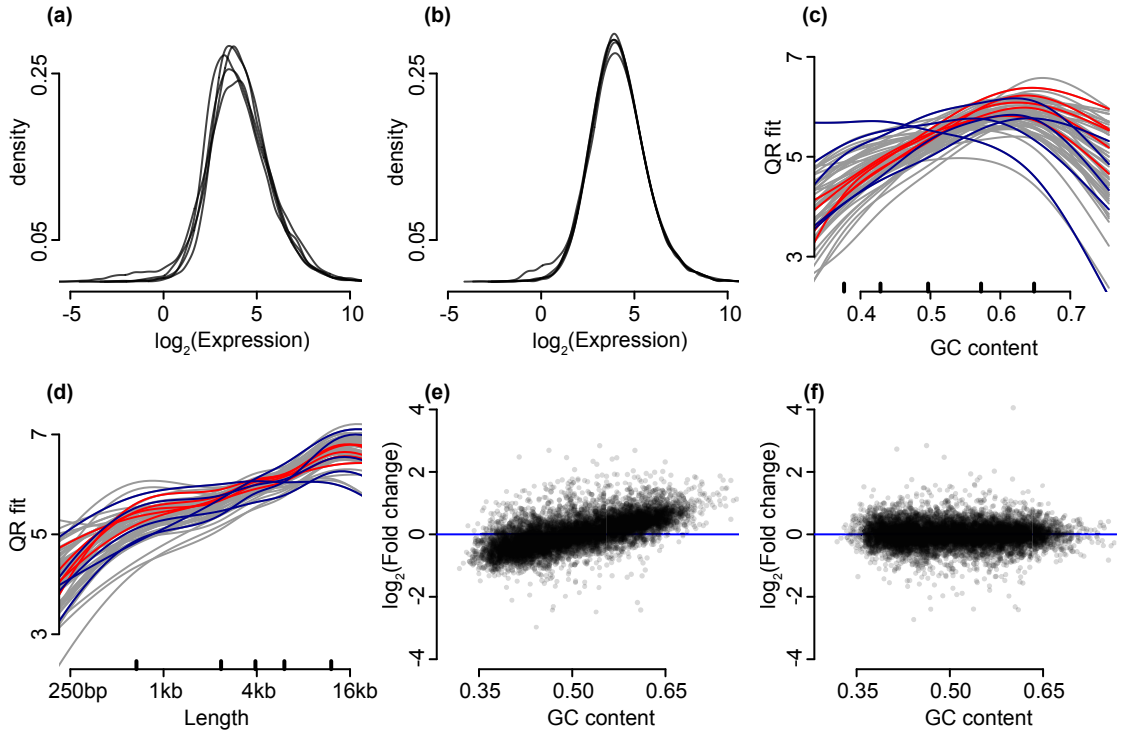

Figure 3. Results from normalizing 60 samples. In these plots we only show genes with a length greater than 100bp and an average (across all 60 samples) standard $\log_2$-RPKM of 2 or greater. (a) Empirical density estimates of $\log_2$-RPKM for five different biological replicates from the Montgomery data are shown. (b) As (a) but CQN normalized expression values on the $\log_2$-scale are shown. (c) The estimated GC-content effect are shown as curves for all 60 biological replicates in the Montgomery study. We created a five versus five comparison using the samples highlighted in blue (group 1) and red (group 2). (d) as (c) but curves are shown for the gene length effect instead of GC-content. (e) Average log-fold-change is plotted against GC-content. Here we used RPKM values and compared group 2 to group 1. (f) Average log-fold-change is plotted against GC-content using CQN normalized expression measures.

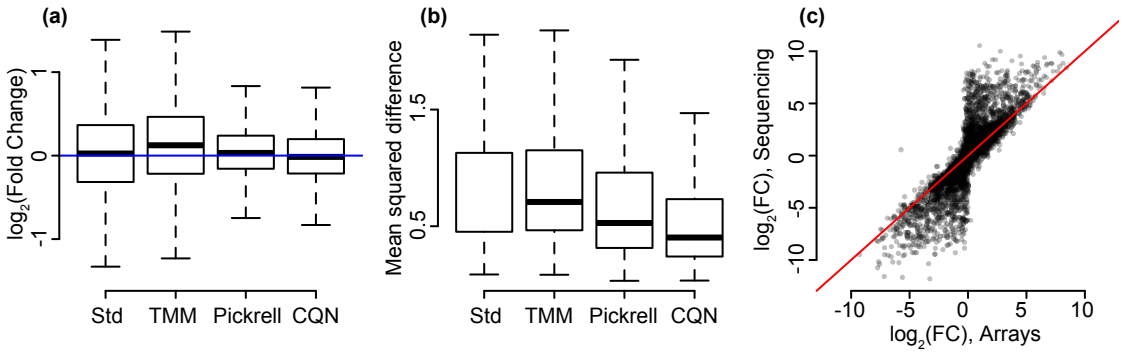

Figure 4. Improved precision provided by CQN on comparisons across studies. (a) We show boxplots of the estimated log fold change between the two groups of five samples (the same two groups as in Figure 3) from the Montgomery data using standard RPKM, expression values normalized by TMM (trimmed median of M-values, the method proposed in Robinson and Oshlack (2010)), the method proposed in Pickrell *and others* (2010), and CQN. We show genes with length greater than 100bp and average (across all samples) $\log_2$ RPKM greater or equal to two. (b) We normalized the 29 samples assayed in both Montgomery and Cheung. For each gene we computed the mean squared difference between the expression measure based on the Montgomery and the Cheung data. The boxplots show the distribution of these precision measures for the highly expressed genes, for each of the four choices of normalization: standard RPKM , TMM, the method proposed in Pickrell *and others* (2010) and CQN. We show genes with length greater than 100bp and average (across all samples) $\log_2$ RPKM greater or equal to two. (c) For the MAQC data we obtained fold change estimates between UHR and brain based on RNA-Seq and microarrays. For RNA-seq we used two samples. For the microarrays we used a five versus five comparison. The microarray data was normalized using RMA and the RNA-seq data was normalized by CQN.