Detecting and correcting systematic variation in large-scale RNA sequencing data

Sheng Li[1,2,11], Paweł P Łabaj[3,11], Paul Zumbo[1,2,11], Peter Sykacek[3], Wei Shi[4], Leming Shi[5], John Phan[6], Po-Yen Wu[6], May Wang[6], Charles Wang[7], Danielle Thierry-Mieg[8], Jean Thierry-Mieg[8], David P Kreil[3,9] & Christopher E Mason[1,2,10]

High-throughput RNA sequencing (RNA-seq) enables comprehensive scans of entire transcriptomes, but best practices for analyzing RNA-seq data have not been fully defined, particularly for data collected with multiple sequencing platforms or at multiple sites. Here we used standardized RNA samples with built-in controls to examine sources of error in large-scale RNA-seq studies and their impact on the detection of differentially expressed genes (DEGs). Analysis of variations in guanine-cytosine content, gene coverage, sequencing error rate and insert size allowed identification of decreased reproducibility across sites. Moreover, commonly used methods for normalization (cqn, EDASeq, RUV2, sva, PEER) varied in their ability to remove these systematic biases, depending on sample complexity and initial data quality. Normalization methods that combine data from genes across sites are strongly recommended to identify and remove site-specific effects and can substantially improve RNA-seq studies.

The deep sampling capabilities and single-base resolution of RNA-seq have led to its adoption for a variety of studies of the transcriptome, which include many inter-site and large-scale studies such as the ENCODE Project, GEUVADIS, GTEx, the Epigenomics Roadmap, the human Brainspan Project and the Nonhuman Primate Reference Transcriptome Resource. However, it is notable that RNA-seq, just like microarrays, has taken many years to emerge as a trusted and established method, as experiments can suffer from lack of principled experimental design, poor sample quality, inconsistent library preparation or platform-specific measurement biases[1,2]. Indeed, when microarrays started being used to identify biomarkers for drug toxicity and disease, the US Food and Drug Administration (FDA) recognized that an effort was needed to assure data quality and inter-site and inter-platform reproducibility and to this end established the MicroArray Quality Control (MAQC) Consortium[3]. Through the consortium, experimental standards and control RNA samples were developed, along with quality assurance guidelines and standardized microarray procedures[4]. Standards were also developed for data repositories (the minimum information about a microarray experiment, MIAME)[5], along with robust methods for analyzing microarray experiments from multiple sources[6]. These and other efforts have enabled the exploitation of the large publicly available microarray data sets and the subsequent deduction of important biological and clinical insights[7].

The success of MAQC motivated the development of similar guidelines and standards for high-throughput sequencing[8,9], in particular for RNA-seq[10,11], which led to the creation of the FDA Sequencing Quality Control (SEQC)/MAQC-III Consortium and the Association of Biomolecular Resource Facilities (ABRF) studies on Next-Generation Sequencing (NGS). Previous large-scale RNA-seq studies have focused on the variation between lanes and flow cells[12], and considerable progress has been made on reducing batch effects by normalizing GC content bias, fragment bias and the biases of isolation procedures[13–23]. So far, several RNA-seq data quality metrics have been developed[13,22,24,25], and surrogate variable analysis (sva)[26,27] has been applied to RNA-seq and microarray data from individual laboratories to improve expression measures[28]. Recently, a thorough, cross-site examination of Illumina RNA-seq data[29] demonstrated that "laboratory effects" strongly affect GC content and insert size of prepared RNA-seq libraries, and a method proposed to correct for them, probabilistic estimation of expression residuals (PEER)[30], was able to reduce artifacts without having an adverse impact on the detection of expression quantitative trait loci (eQTLs).

Yet, to date, there has been no systematic examination of the impact of site-specific bias in detecting DEGs, which is often the primary goal of an RNA-seq experiment. Moreover, there are various proposed means by which to correct for such biases, but the performance of several competing methods has not been systematically characterized. Here we used the controlled experimental design of the standardized SEQC/ABRF samples to test intra- and inter-site reproducibility,

[1]Department of Physiology and Biophysics, Weill Cornell Medical College, New York, New York, USA. [2]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medical College, New York, New York, USA. [3]Chair of Bioinformatics Research Group, Boku University Vienna, Vienna, Austria. [4]Department of Bioinformatics, WEHI, Melbourne, Australia. [5]State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, Schools of Life Sciences and Pharmacy, Fudan University, Shanghai, China. [6]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA. [7]Center for Genomics and Division of Microbiology & Molecular Genetics, School of Medicine, Loma Linda University, Loma Linda, California, USA. [8]National Center for Biotechnology Information (NCBI), Bethesda, Maryland, USA. [9]University of Warwick, Coventry, UK. [10]The Feil Family Brain and Mind Research Institute, New York, New York, USA. [11]These authors contributed equally to this work. Correspondence should be addressed to C.E.M. (chm2042@med.cornell.edu) or D.P.K. (david.kreil@boku.ac.at) or (d.kreil@warwick.ac.uk).

sensitivity and specificity of RNA-seq for pairwise comparisons of samples with varying complexity, representative of different experimental scenarios. We benchmarked two different sequencing platforms (Life Technologies Personal Genome Machine (PGM) and Illumina HiSeq2000) across 20 laboratory sites and assessed a variety of methods for data normalization and bias removal (cqn[14], EDASeq[15], RUV2 (ref. 31), sva[26,27] and PEER[30]). To our knowledge, no other group has reported a cross-platform evaluation of methods for assessing RNA-seq quality and removing variance from data for multi-site, multi-platform reproducibility, which is a prerequisite for reliable conclusions and the integration of measurements and experiments from different laboratories. Finally, this work shows that, although bias-correlation methods can be successful at improving data quality, the degree of impact on the detection of DEGs ranges widely, for which correction methods often make a tradeoff between accuracy and reproducibility.

RESULTS

Experimental data comparing intra- and inter-site variation

The experimental design of the main SEQC and ABRF studies are described in detail elsewhere[32,33]. Briefly, four RNA samples were provided by the SEQC Consortium, A (cancer cell lines), B (brain) and two titrated mixtures of A:B (C and D). Samples C and D represent mixtures of samples A and B at the defined ratios of 3:1 and 1:3, respectively, and thus hold 'built-in truths' of sample mixing ratios. These were sequenced and analyzed by over 20 laboratories and a total of six sequencing platforms. Here we use two RNA-seq platforms from the SEQC and ABRF studies where we had library preparation replicates of each sample at every site: Illumina's HiSeq2000 and Life Technologies PGM. For Illumina, each sample was distributed from a single source to six different primary test sites (ILM1–6) and prepared in quadruplicate at those sites. A fifth library for each sample, prepared at an independent seventh site, was also distributed and sequenced at three test sites (ILM2,3,5). Samples were barcoded and pooled together before sequencing to assess lane and batch effects[7,15], and were then paired-end sequenced (2×100) on two flow cells using Illumina's HiSeq2000 platform. For the PGM platform, samples were prepared in duplicate at three sites and sequenced on two Ion Torrent 318 chips at each site. We first focus on the results from the Illumina platform.

Identical inter-site replicates show high rates of false positives

Ideally, expression values generated from identical samples at different sites should show little (and random) variation across sites. Thus, we can compare each sample to itself across the six test sites by pairwise calling of all differentially expressed genes (**Fig. 1a**) to generate an empirical measure of the false-positive rate for all four samples (**Fig. 1b**)—that is, all DEG calls represent false positives. However, we observed many differentially expressed genes at varied fold-change (FC, 1.5–2.0) and false-discovery rate thresholds (FDR, 0.05–0.001) using the limma-voom package. At the most lenient FC (1.5) and FDR (0.05), the number of false-positive DEGs detected was as high as 9,602 (mean = 2,823, s.d. = 3,527, including both changes up and down), or ~20% of all genes (**Fig. 1b**). As the stringency of the FC and FDR thresholds increased, the number of false-positive DEGs decreased; although even at fairly stringent thresholds (FC > 2.0 and FDR < 0.001), the number of DEGs detected was still as high as 3,135 (mean = 739, s.d. = 1,089), representing up to 8% of all genes. When we examined the inter-site DEG false-positive rates for several other analysis pipelines (WHAM[34], Mapsplice[35], Novoalign[36], Cufflinks[23,37,38] and HTSeq[39]) we found similarly high false-positive
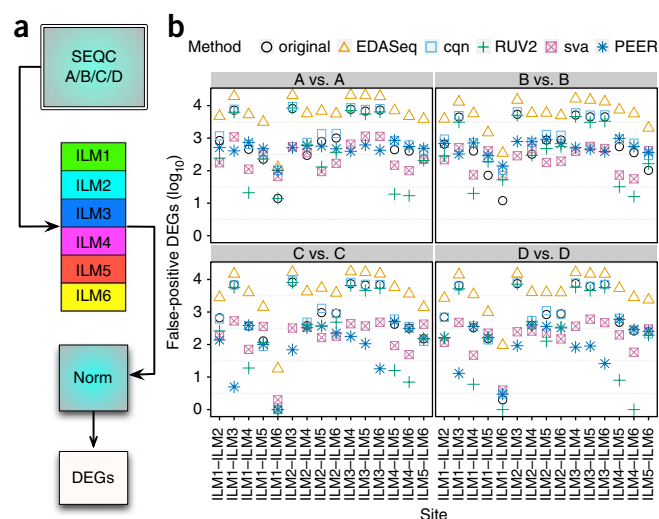

Figure 1 Inter-site normalization and false-positive DEGs. (**a**) Schematic plot of RNA-seq data from all four samples (A,B,C,D) and six sites (ILM1–6), followed by normalization and calling of all pairwise differentially expressed genes (DEGs). (**b**) Intersite false-positive DEGs, by comparing the four replicate libraries made for a particular sample at one Illumina site to the replicates of the same sample from the other five sites, shown for all samples (A versus A, B versus B, C versus C, D versus D). We compare six normalization methods: original (standard limma-voom processing only), and with additional processing by EDASeq, cqn, RUV2, sva, PEER (bar color). Thresholds used for DEG calls: FDR: 0.05, FC: 2.0. One site (ILM3) showed the most false positives before correction, although other sites also showed thousands of false-positive DEGs.

rates, regardless of the analysis pipeline or read alignment methods used (**Supplementary Fig. 1**).

To remove these false positives, we tested several established methods for normalization of RNA-seq data (cqn[14], EDASeq[15], RUV2 (ref. 31), sva[26,27] and PEER[30]), and we observed highly variable results. Some methods (specifically sva and PEER) that leveraged all data across all sites were quite successful at ameliorating the high rate of false positives (**Fig. 1**, **Supplementary Fig. 2**), removing 85.1% to 87.7% of the original total false-positive DEGs. The application of RUV2 with ERCC spike-ins (RUV2-ERCC), which tries to remove confounding factors based on a control set of synthetic RNAs assessed across sites, removes just 20% of false positives on average, but is more effective for sites that already have relatively low false-positive rates (ILM4, ILM5). Notably, neither applying GC bias correction tools (cqn[14] and EDASeq[15]) to individual sites nor changing read counts to only use 3′-UTRs was effective at decreasing the number of inter-site false-positive DEGs (**Fig. 1b** and **Supplementary Fig. 3**); in most cases, these methods actually increased the number of false positives.

Inter-site DEG reproducibility varies by site and sample

However, any method for improving the false-positive rate for DEG detection (A versus A) needs to also be examined in the context of the true positives (validated DEGs), and we sought to determine the pre-normalization relationship between false-positive DEGs, true positives and sites with high false positives (e.g., ILM3). We examined the repeatability and reproducibility of gene expression measures between the different samples with varying levels of complexity (A versus B, and their 3:1 and 1:3 titrations of C and D). We used several analyses to establish the inter-site accuracy of DEG detection: the correlation of measured gene expression profiles, DEG detection within and
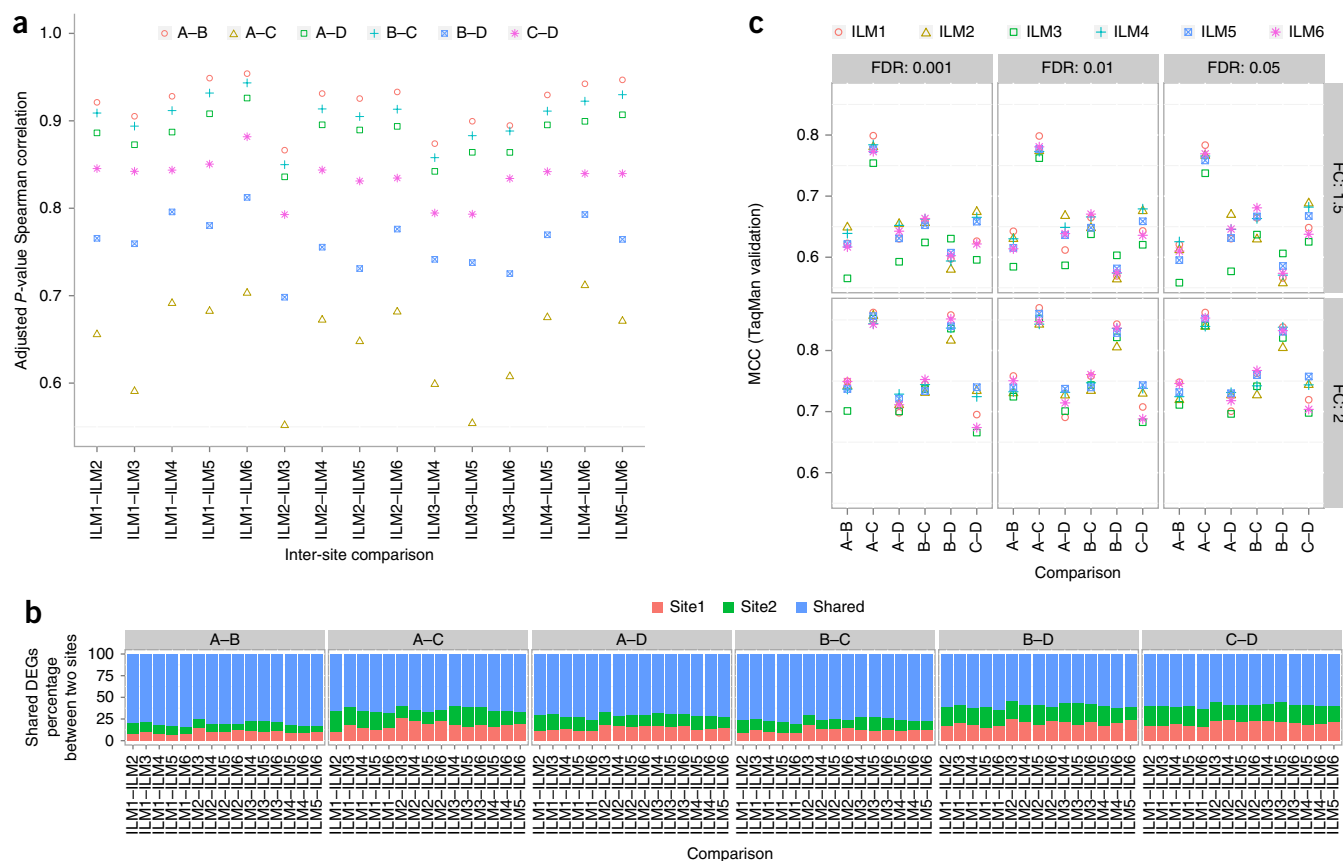
Figure 2 Evaluation of inter-site DEG reproducibility. For each of the six sites, all possible pairwise differential expression analyses were performed for all samples A–D, giving a total of six comparisons. We then assessed agreement across sites using different measures. (**a**) The Spearman rank correlation of the *q*-values from any two of the six sites, with color and shape indicating the samples compared. (**b**) Percentage of DEGs agreeing between two sites out of the union of DEGs detected at the two sites. Blue indicates the fraction of DEGs shared, the other colors represent the DEGs seen at only one of the sites. (**a**,**b**) Along the *x* axis we plot all 15 pairwise combinations of the six sites (ILM1 versus ILM2, etc.). (**c**) External validation by TaqMan using MCC as measurement. Along the *x* axis we plot all six possible pairwise combinations of the four samples. Different color and shape combinations represent the six sites.

across sites, and DEG detection vis-à-vis independent Taqman data from 779 genes querying the exact same RNA samples.

First, the intra-site and inter-site Pearson correlation coefficients (R^2) were all above 0.95 (**Supplementary Fig. 4**), and Q-Q plots of the gene expression values from different sites indicated that all sites had similar distributions that clustered together (**Supplementary Figs. 5–8**). Thus, simply calculating R^2 values of genes' expression measures and showing that samples cluster together merely shows the tendency of expression values to track each other; these high correlation coefficients mask the 8–20% false-positive rate described above[33,40].

Second, we examined, at each site, the differentially expressed genes for every possible pairwise comparison of samples. All six sites found similar numbers of DEGs (**Supplementary Fig. 9**), and the Spearman rank correlation of *P*-values showed that the inter-site rank agreement was very high for the common DEGs shared by all six sites, with a median correlation greater than 0.96 (**Supplementary Fig. 10a–c**). However, when we examined the complete list of DEGs found at each site (instead of just those DEGs common across sites), we found much lower correlations of 0.55–0.95 (**Fig. 2a**). As expected, one site (ILM3) always showed the lowest Spearman correlation of *P*-values (**Fig. 2a**), co-incident with an increase in site-specific DEGs (**Fig. 2**).

Finally, to evaluate the sensitivity and specificity of DEG detection from RNA-seq data at each test site, we calculated the

Matthews Correlation Coefficient (MCC)[41,42], with the true-positive rate (TPR) and false-positive rate (FPR) based on the Taqman data set (**Supplementary Fig. 11**). Scatter plots for pairwise comparisons across all sites and samples revealed good overall correlation between RNA-seq data and TaqMan data at the gene level (**Supplementary Fig. 12a**, with mean $R^2 = 0.729$). However, the similarity of the TaqMan and RNA-seq data was improved for all comparisons when using the exact TaqMan primer's coordinates on the transcriptome to quantify RNA-seq expression rather than the combined read count across the entire gene (**Supplementary Fig. 12b**, mean R^2 increase of 0.14). Nonetheless, in all cases, the site detected as an outlier by our analysis of false positives (ILM3) showed the lowest R^2 and MCC with the TaqMan data (**Fig. 2b**). The DEGs detected from TaqMan were then compared to the DEGs obtained from RNA-seq using the limma-voom method. Each of the six cross-sample comparisons had very similar MCC, TPR and FPR (**Fig. 2c** and **Supplementary Fig. 13**), and these measures also indicated lower agreement as the samples became more similar, as expected, with the biggest differences expected by design in comparisons of samples A and D, whereas conversely the mixture samples C and D were similar by design. Indeed, when applying a variety of information theoretic metrics (such as mutual information[33], **Supplementary Fig. 14**), we observed a similar loss of reproducibility among samples.

Figure 3 Inter-site DEG detection and validation. (**a**) Schematic plot of the comparison between intra-site DEGs and inter-site DEGs. We show site ILM1 and the comparison of sample A versus B as an example. Analogously, the analysis has been applied to all six sites and possible pairwise sample comparisons. (**b**) Spearman rank correlation of the adjusted *P*-value (*q*-value) for inter-site DEGs and intra-site DEGs. (**c**) Inter-site DEG validation by TaqMan, assessed by MCC for all six pairwise sample comparisons (A-B, A-C, A-D, B-C, B-D, C-D). For **b** and **c**, we compare six normalization methods: original (standard limma-voom processing only), and with additional processing by EDASeq, cqn, RUV2, sva, PEER. Thresholds for DEG calls: FDR: 0.05, FC: 2.0.


Cross-site data normalization improves RNA-seq quality

Because false positives and true positives were both affected by site-specific noise, we next compared DEG detection performance across sites (**Fig. 3a**) using five methods for RNA-seq normalization (EDAseq, cqn, RUV2, sva and PEER). We observed that EDASeq and PEER were the two top methods with the highest adjusted Spearman rank correlation of *P*-values between inter-site and intra-site DEG analysis (**Fig. 3b**). The more similar samples were, the lower inter-site reproducibility was for all methods. Using the common intra-site DEGs to validate inter-site DEGs showed that PEER consistently performed better, especially for the site with the largest bias (ILM3), where PEER successfully identified and compensated for this bias, for every comparison (**Supplementary Fig. 15**). This was also true when measured by MCC (**Fig. 3c** and **Supplementary Fig. 16a**).

We then further measured the impact of these normalization methods on the intra-site and inter-site quantification of differential gene expression. We compared the RNA-seq intra-site DEGs with the independent TaqMan data, using MCC as the evaluation measure. Although most methods did not improve the accuracy of intra-site DEGs detection, we found that EDAseq gave the highest similarity to Taqman expression measures (with mean MCC = 0.939 and s.d. = 0.019, **Fig. 4** and **Supplementary Fig. 16b**). This improvement was consistent across all pairwise comparisons and all test sites.

Quality control metrics flag sources of error and poor data

These results indicated a need to further investigate the underlying sources of variance that lead to so many false positives or irreproducible DEGs. Sample QC metrics (**Fig. 5**, **Supplementary Figs. 17** and **18**) indicated that data from a single site (ILM3) were distinct compared to the other sites. First, a nonrandom nucleotide composition bias was seen at the beginning of the sequencing reads, concomitant with a distinct, narrow bell curve of GC-content for the IML3 site (**Fig. 5a**). Also, site ILM3 had an overall higher sequencing error rate compared to the other sites (**Fig. 5c**). We saw that both
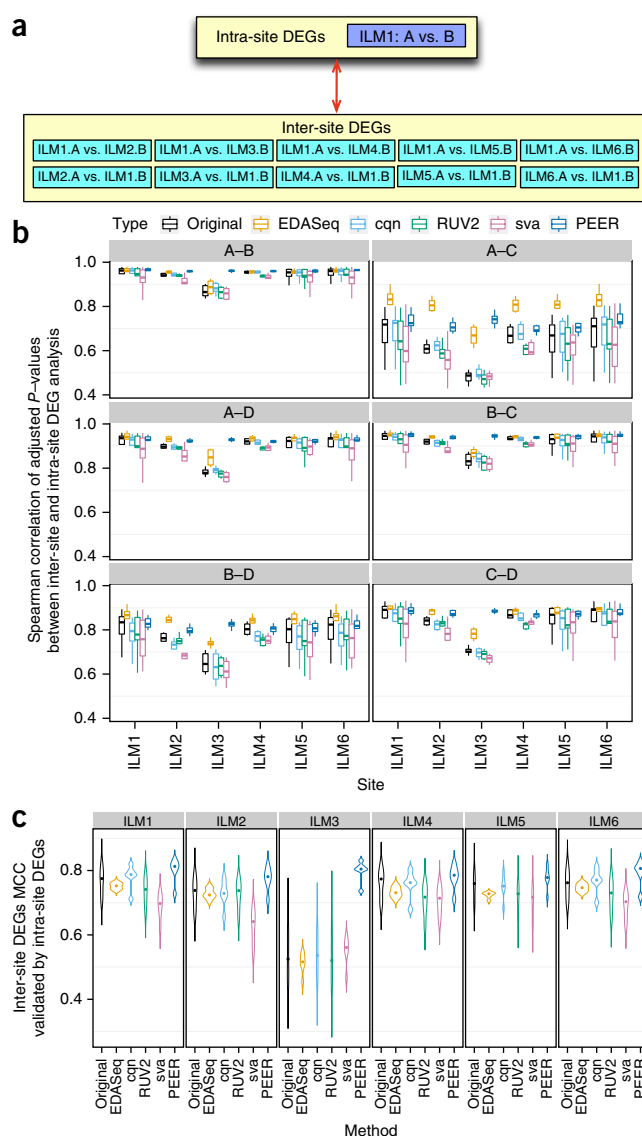
sample B (as a type) and ILM3 (as a site) had more reads near the 3′ end than the 5′ end of genes, indicating a shift in the coverage of the genes (**Supplementary Fig. 18c**). Coverage across the gene body was assessed using the coefficient of variation of the coverage across the length of the genes, and we saw that overall the ILM3 site had higher coefficients of variation (**Fig. 5d** and **Supplementary Fig. 18c**), thus demonstrating the value of a 'nucleotide composition metric' (described below) for QC in RNA-seq for identifying unusual and potentially problematic measurements.

To test whether these sources of bias were site-dependent, we examined the fifth library of each sample (replicate 5, for samples A,B,C and D), which was prepared at an independent seventh site and then sequenced at three of the test sites. With this experimental design we can distinguish between sources of variation arising from the library preparation (including RNA isolation) and sources arising from the
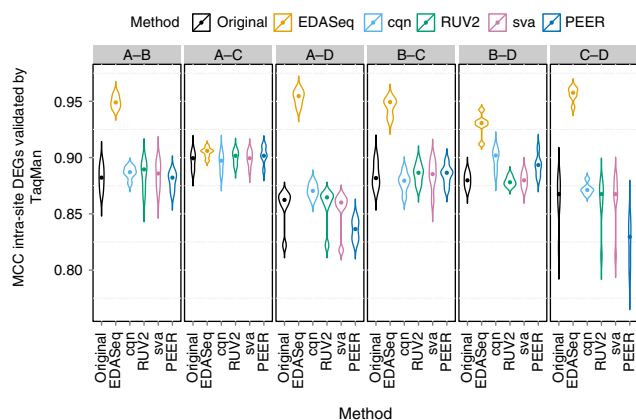
Figure 4 MCC evaluation of intra-site DEG detections using TaqMan data. Each violin plot summarizes data points from six sites. We compare six normalization methods: original (standard limma-voom processing only), and with additional processing by EDASeq, cqn, RUV2, sva, PEER. Thresholds for DEG calls: FDR: 0.05, FC: 2.0.
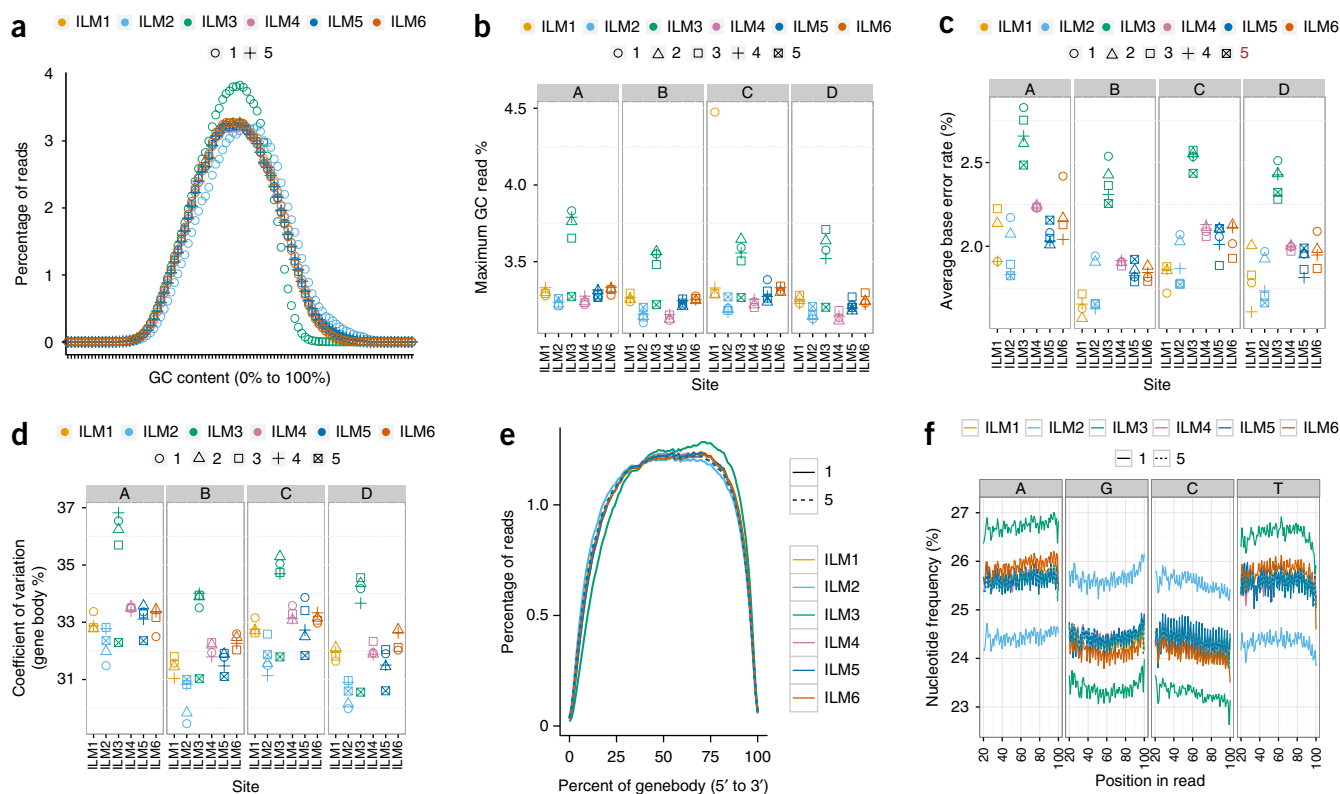
Figure 5 Examination of RNA-seq data quality identifies major sources of variation. (**a**) GC content distribution (sample A). Point shapes distinguish replicates (1: unfilled circle; 5: +). (**b**) The greatest percentage of reads contributing to some GC content bin (0–100%). A sample with more reads contributing to a particular GC content bin (%) indicates an abundance of reads with that particular GC content. (**c**) Average base error rate across all sequencing bases (y axis) across all sites (x axis). (**d**) Coefficient of variation of the percentage of gene body coverage (y axis), which is a measure of the evenness of coverage across all gene bodies for each site (x axis). (**e**) The percentage of reads that covers each nucleotide position of all of genes scaled to 100 bins, from 5′ UTR to 3′ UTR for sample A:1–5. Replicate 1 displayed site-dependent variation in gene body coverage for ILM3 (3′ bias), whereas replicate 5 showed similar gene body coverage regardless of where it was sequenced, suggesting that gene body coverage is influenced by library preparation. (**f**) Nucleotide frequency versus position for aligned reads. The percentage of each base was plotted as a function of the read length for each base (A, G, C, T) for two replicates (1, 5) for all sites. Replicate 1 displayed site-dependent base-composition frequencies, whereas replicate 5 showed similar base composition frequencies regardless of where it was sequenced, suggesting that base composition frequency is largely a result of library preparation. Only the 20th to the 100th bases are shown here; the full read range can be seen in **Supplementary Figure 4**. Vertical facets stand for sample A–D. Site information for ILM1–6 is color-coded. Replicates 1–4 were prepared and sequenced independently at each site, whereas replicate 5 was prepared at a single site and then sequenced at a subset of all sites. Point shapes distinguish replicates.

sequencing itself. In the case of GC distribution, the fifth library from each sample did not exhibit an aberrant spike at 50% GC-content like replicates 1–4 sequenced at the ILM3 site; it is likely that the aberrant spike is a result arising purely from sample preparation (**Fig. 4a,b** and **Supplementary Fig. 18a**), which confirms the primary source of variation put forward by both the GEUVADIS and SEQC Consortium comparisons of RNA-seq measurements[33]. **Table 1** summarizes major sources of variation observed in our quality metrics as determined sequencing the fifth replicate libraries at sites ILM2, 3 and 5.

This control library was also able to reveal other features inherent to the sample preparation and sequencing. The fifth library replicate of each sample was always consistent in error rate with the other samples it was sequenced with, indicating that the sequencing error rate is indeed primarily a function of sequencing, and not affected by library preparation (**Fig. 5c** and **Supplementary Fig. 18b**). Plots of the uniformity of coverage across gene bodies showed that sample B, regardless of where it was prepared or sequenced, had more read coverage near the 3′ end than the 5′ end with respect to annotated gene models (**Supplementary Fig. 18c**), indicating that the stock of sample B, before its distribution to each site, may have been

contaminated with something that would have caused it to have depleted 5′ ends before poly(A)+ selection (e.g., RNase or cations). The fifth library of each remaining sample (A, C, D) had relatively

Table 1 Major sources of variation for quality metrics determined using fifth replicates

Quality metrics	Description	Major source of variation
GC content	Percentage of bases for each GC bin (1–100) for all aligned reads	Library preparation (including RNA isolation)
Gene body coverage evenness	Accumulative statistics for the read coverage of exonic regions from 5′ UTR to 3′ UTR for all genes Each gene is divided into 100 bins to calculate the gene body coverage	Library preparation (including RNA isolation)
Base error rate	The average base error rate for all aligned reads	Sequencing (inclusive of cluster generation)
Nucleotide composition	Nucleotide frequency versus position for aligned reads	Library preparation (including RNA isolation)

uniform coverage when sequenced at ILM3, whereas the corresponding samples prepared at ILM3 did not, notably demonstrating that library preparation can exacerbate poor gene body coverage uniformity (**Fig. 5d**,**e** and **Supplementary Fig. 18c**). Lastly, because the nucleotide composition metric (**Fig. 5f** and **Supplementary Fig. 17**) showed that the fifth library replicate had equal base composition regardless of sequencing site, these data demonstrate, for the first time to our knowledge, that the nucleotide composition bias of RNA-seq data likely arises from library preparation alone.

Finally, we observed that the latent experimental factors determined by PEER and sva are highly correlated with QC metrics and properties, and that these factors were responsible for the majority of false positives in inter-site DEG analysis. For sva, the first latent factor was significantly correlated with the GC content distribution quality metric of the sites ($P < 2 \times 10^{-7}$), the average error rate ($P < 6 \times 10^{-7}$) and the duplication by library (see **Supplementary Fig. 19**, $P < 2 \times 10^{-4}$). The second latent factor was significantly associated with the gene body coverage uniformity ($P < 3 \times 10^{-4}$). For PEER, the first latent factor was significantly correlated with the GC content distribution quality metric, the gene body coverage uniformity and the average error rate of the sites ($P < 2 \times 10^{-4}$). These additional metrics can be, and should be, used for tracking samples that may suffer from high false positives and inherent sample noise.

Cross-platform applicability of normalization methods
Finally, we sought to gauge the utility of these inter-site normalization methods across multiple platforms. We used PGM RNA-seq data from the ABRF-NGS Consortium data (Online Methods), which used the same standardized RNA samples (A and B) as the SEQC Consortium, and were prepared using the Life Technologies RNA Sequencing kit at three independent sites (PGM1–3) with duplicate library preparations and sequenced using three Ion Torrent 318 chips. Sequencing reads were again aligned using the STAR[43] aligner and annotated using GenomicRanges[44] with AceView[45] genes.

We first examined the GC content of the mapped reads from PGM data, and found that some replicates showed abnormal GC content distributions (**Supplementary Fig. 20**). Two libraries in particular had a much higher maximum spike in their GC content (%GC) for their reads (>5.8%) in comparison to the rest of the libraries (mean 4.9%, **Supplementary Fig. 21**). The average base error rate (**Supplementary Fig. 21b**) was higher in PGM1 and PGM3 than in PGM2. We also observed for sample A, that replicate 4 from PGM1, replicate 2 from PGM2 and replicate 1 from PGM3, all had the lowest gene body coverage variation compared to other PGM data (**Supplementary Fig. 21c**). After the trimmed mean of M-values (TMM)[46] and limma-voom normalization, we found that samples A and B were well distinguished by multidimensional scaling (**Supplementary Fig. 21d**), and that the two replicates with abnormal GC content distributions (PGM1.A.4 and PGM2.A.2) were separated from the other replicates of sample A at dimension 2.

We then examined the inter-site false positive DEGs for the PGM data, each with two replicates for sample A and B. With the lowest stringency thresholds (FDR: 0.05; FC: 1.5), there were on average 114 false-positive DEGs (0.32%) using the original limma DEGs analysis (**Supplementary Fig. 22a**). Notably, applying PEER successfully removed almost all the false-positive DEGs (**Supplementary Fig. 22c**). The responsible hidden variable identified by PEER was significantly correlated with GC content ($P = 0.03$). Using the common intra-site DEGs to validate the called inter-site DEGs, the MCC[41,42] showed that PEER is also more accurate than the original limma-voom method (**Supplementary Fig. 22d**,**e**), indicating that

global data normalization analysis methods such as PEER can also be used to improve RNA-seq data across both Illumina and the PGM platforms.

DISCUSSION
Using the benchmark data sets created by the SEQC Consortium and the ABRF-NGS study on RNA-seq[32,33], we determined the relationship between the quality of a data set indicated by a wide range of quality metrics and the results of differential gene expression (DEG) analysis of samples both within a site and across sites. We then rigorously tested a variety of commonly used statistical tools for RNA-seq data normalization (sva, RUV2, cqn, EDASeq, PEER) using multiple samples and metrics. Overall, the reproducibility of intra- and inter-site DEGs across all sites showed a higher correlation for comparisons between more biologically different samples (A versus B), and a lower reproducibility for more similar samples (A versus C, B versus D, C versus D), reflecting the expected greater challenge of reliably identifying smaller differences. Indeed, the unique study design allowed a *reductio ad absurdum* experiment, comparing replicates of the exact same sample across sites, where we notably still observed thousands of DEGs that were deemed statistically significant but clearly reflected technical differences between sites and not differences between the compared RNA samples. The application of GC content bias correction packages including cqn[14] and EDASeq[15] could not remove these false positives, likely because GC content bias is not the only source that contributes to bias in gene expression data. Similarly, RUV factor analysis based on the ERCC control gene set was not sufficient.

However, the majority of RNA-seq false positives (>85%) could successfully be removed by subtracting the effects of latent variables identified by either sva[26,27] or PEER[30], which could be achieved by jointly analyzing the set of measurements of all genes across multiple sites, without a decrease in the sensitivity or specificity of DEG detection at each site or across sites. These latent variables were shown to be significantly associated with GC-content, gene body coverage uniformity, average base error rate and insert size. This confirms the impact of two already recognized RNA-seq latent variables, GC content and insert size[15,22], and it also identifies two more relevant contributions to technical variation, gene coverage variation and error rate. Furthermore, our use of the cross-site, internal control library (#5) has demonstrated that GC content is preparation-specific, not laboratory-specific, and we have introduced the coefficient of variation for gene body coverage as an important quality measure in RNA-seq (**Table 1**), which quantifies this 5′-3′ bias across both platforms.

Our results also indicate that a tradeoff is sometimes made between different goals of normalization. For example, although EDASeq did not effectively remove inter-site false positives (**Fig. 1**), it did consistently improve the detection of DEGs as compared to the TaqMan reference set (**Fig. 4**). Conversely, although PEER sometimes ranked lower on comparability to the Taqman reference set, it had the greatest impact on removing site-specific bias. Moreover, it worked best in making data from the HiSeq and PGM platforms comparable for cross-platform analyses. Notably, genes tested by TaqMan were (on average) more highly expressed, and this may affect normalization method performance for this reference set. Regardless, we have shown that RNA-seq quality metrics and bias removal can successfully be used on multiple platforms. Because many aspects of library preparation and normalization are universal aspects of working with RNA, including isolation, purification, priming, amplification, reagent batch and kit version, the recommendations and most of the observations presented here will likely be applicable to any sequencing platform used for RNA-seq[47,48].

In general, given advanced data processing, even substantial bias could be corrected and value extracted from experiments combined from multiple laboratories, highlighting the need to archive and share the original sequencing reads from RNA-seq experiments. These best practices for quality control and analysis of RNA-seq data from different experiments or laboratories can readily be implemented, and they are of immediate relevance not just for large-scale RNA-seq studies, but also the analysis of smaller experiments in the context of other data, such as in-house data or those from public repositories. With the globalization of research collaborations and the emergence of an increasing number of large RNA-seq cohorts, obtaining sequencing data across different institutes and platforms is inevitable. The ENCODE project and GEUVADIS Consortium have provided extremely valuable guidelines and best practices for RNA-seq experiments and this work validates and extends their conclusions to other efforts such as GTEx[49], the Epigenomics Roadmap[50], the human Brainspan Project[51] and the Nonhuman Primate Reference Transcriptome Resource[52]. These metrics and internal controls complement those currently in use and create additional resolution insights into the quality of an RNA-seq data set, further establishing RNA-seq as a reliable, universal tool for differential expression profiling.

METHODS

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

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

AUTHOR CONTRIBUTIONS
P.Z. developed the first quality metric tools and algorithms for the SEQC and ABRF groups and lead the data production at the bench as well as the computational work. S.L., P.Z., L.S., C.W., D.T.-M., J.T.-M. and C.E.M. designed the experiments. S.L., P.P.Ł., P.Z., W.S., M.W., D.T.-M., J.T.-M., D.P.K. and C.E.M. contributed to the analysis. S.L., P.P.Ł., P.Z., D.P.K. and C.E.M. wrote the manuscript and made the figures. P.S., W.S., L.S., J.P., P.-Y.W., M.W., C.W., D.T.-M. and J.T.-M. edited the manuscript and helped with analysis.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Irizarry, R.A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–350 (2005).
2. Wang, H., He, X., Band, M., Wilson, C. & Liu, L. A study of inter-lab and inter-platform agreement of DNA microarray data. *BMC Genomics* **6**, 71 (2005).
3. MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24**, 1151–1161 (2006).
4. Casciano, D.A. & Woodcock, J. Empowering microarrays in the regulatory setting. *Nat. Biotechnol.* **24**, 1103 (2006).
5. Ball, C.A. & Brazma, A. MGED standards: work in progress. *OMICS* **10**, 138–144 (2006).
6. Hong, F., Wittner, B., Breitling, R., Smith, C. & Battke, F. RankProd: Rank Product method for identifying differentially expressed genes with application in meta-analysis. *R package version 2.28.0* (2011).
7. Dudley, J.T., Tibshirani, R., Deshpande, T. & Butte, A.J. Disease signatures are robust across tissues and experiments. *Mol. Syst. Biol.* **5**, 307 (2009).
8. Glenn, T.C. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* **11**, 759–769 (2011).
9. Loman, N.J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–439 (2012).
10. Editorial. Prepare for the deluge. *Nat. Biotechnol.* **26**, 1099 (2008).
11. Ji, H. & Davis, R.W. Data quality in genomics and microarrays. *Nat. Biotechnol.* **24**, 1112–1113 (2006).
12. Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
13. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
14. Hansen, K.D., Irizarry, R.A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).
15. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12**, 480 (2011).
16. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
17. Benjamini, Y. & Speed, T.P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).
18. van Heesch, S. *et al.* Systematic biases in DNA copy number originate from isolation procedures. *Genome Biol.* **14**, R33 (2013).
19. Hansen, K.D., Brenner, S.E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38**, e131 (2010).
20. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
21. Cheung, V.G. *et al.* Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol.* **8**, e1000480 (2010).
22. DeLuca, D.S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).
23. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011).
24. Goncalves, A., Tikhonov, A., Brazma, A. & Kapushesky, M. A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics* **27**, 867–869 (2011).
25. Schulze, S.K., Kanwar, R., Golzenleuchter, M., Therneau, T.M. & Beutler, A.S. SERE: single-parameter quality control and sample comparison for RNA-Seq. *BMC Genomics* **13**, 524 (2012).
26. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. & Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
27. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).
28. Mooney, M. *et al.* Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris. *PLoS ONE* **8**, e61088 (2013).
29. 't Hoen, P.A. *et al.* Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31**, 1015–1022 (2013).
30. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).
31. Gagnon-Bartsch, J.A. & Speed, T.P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539–552 (2012).
32. Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.* doi:10.1038/nbt.2972 (24 August 2014).
33. 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).
34. Li, Y., Terrell, A. & Patel, J.M. in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data* 445–456 (ACM, 2011).
35. Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, e178 (2010).
36. Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* **11**, 473–483 (2010).
37. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
38. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
39. Anders, S., Pyl, P.T. & Huber, W. HTSeq: A Python framework to work with high-throughput sequencing data. *bioRxiv* doi:10.1101/002824 (20 February (2014).
40. Łabaj, P.P. *et al.* Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics* **27**, i383–i391 (2011).
41. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**, 412–424 (2000).
42. Shi, L. *et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **28**, 827–838 (2010).

43. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
44. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).
45. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.* **7** (suppl. 1), S12 1–14 (2006).
46. 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).
47. Tripathi, A.K. *et al.* Transcriptomic dissection of myogenic differentiation signature in caprine by RNA-Seq. *Mech. Dev.* **132**, 79–92 (2014).
48. Bragg, L.M., Stone, G., Butler, M.K., Hugenholtz, P. & Tyson, G.W. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput. Biol.* **9**, e1003031 (2013).
49. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
50. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
51. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
52. Pipes, L. *et al.* The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic Acids Res.* **41**, D906–D914 (2013).

ONLINE METHODS

Sample definitions. Sample A was Universal Human Reference RNA (catalog no. 740000) and Sample B was Human Brain Reference RNA (catalog no. 6050) from Stratagene and Ambion, respectively. Sample C was a 3:1 mixture of A and B (vol/vol), and sample D was a 3:1 mixture of A and B (vol/vol).

RNA quantification, purity and intactness assessment. Concentrations were based on total RNA as measured by OD_{260} using a NanoDrop 2000 UV-Vis spectrophotometer. RNA was run on an Agilent Bioanalyzer 2100 to assess intactness. Acceptable values were defined as: $A_{260}/_{280}$ ratio in the range of 1.8–2.2, ribosomal RNA ratio (28S/18S) > 1.8 and RNA integrity number (RIN) > 8.0.

Library preparation and sequencing. All SEQC (MAQC-III) data sets are available through the Gene Expression Omnibus GEO site (series accession number: GSE47792). All ABRF-NGS RNA-seq data, with analysis methods, are also available at the GEO (series accession number: GSE46876).

For Illumina, 250 ng of total RNA from the identical MAQC samples from 2006 were used to create aliquots for all sites and all technologies. Libraries were prepared in quadruplicates as six different sites using reagents from Illumina's TruSeq RNA Sample Preparation Kit (v2) and following Illumina's Low Sample (LS) protocol in their TruSeq RNA Sample Preparation v2 Guide. At each site, each library was indexed with a unique barcode, pooled together and paired-end sequenced (100 × 100) on 16 lanes across two flow cells on Illumina's HiSeq2000 platform. Control cDNA libraries from the four control RNAs were made at a seventh site, which was then distributed to all sites for testing the 'machine effect'. For the PGM, libraries were constructed at three core laboratory sites using the MAQC A, MAQC B, ERCC 1 and ERCC 2 RNAs. Further details are provided in the ABRF-NGS[32] paper, but briefly, 5 μg of each RNA was enriched for poly-A RNA (MRRK1010, MPG Kit, PureBiotech) using the recommended Life Technologies Ion protocol for Transcriptome Profiling of Low-Input RNA Samples (April 2011 version). The resulting RNA was assessed for yield and purity using an Agilent 2100 Bioanalyzer PicoChip, all with RINS above 8. Site definitions are as follows: ILM1: Australian Genome Research Facility; ILM2: Beijing Genomics Institute; ILM3: Cornell; ILM4: City of Hope; ILM5: Mayo Clinic; ILM6: Novartis. We used a set of quality metrics (**Supplementary Figs. 1**–**4**) to gauge the variability of the RNA-seq data within and between six SEQC test sites.

Whole transcriptome library preparation for PGM was performed using 5–10 ng of fragmented enriched poly-A RNA according to the manufacturer's protocol (Ion Total RNA-Seq Kit V2 protocol #4476286B Life Technologies). Size selection of a 315- bp product was performed using a standard Pippin prep protocol (Sage Science) followed by purification with AMPure beads (Beckman-Coulter Genomics). Emulsion PCR was done using the One Touch system (Life Technologies). Beads were prepared from 70–100 million copies using the One Touch 200 Template Kit v2 #4471263. For each of the MAQC samples, PGM1 had four replicates, whereas PGM2 and PGM3 each had two replicates. Sequencing was conducted using an Ion PGM 200 sequencing kit (#4474004) on the Ion Torrent 318 chip.

RNA-seq data preprocessing. Image processing and base calling were accomplished in real time with Illumina's HiSeq Control Software (HCS). Demultiplexing was carried out using Illumina's CASAVA (v1.8) software. For the PGM, data were collected using the Torrent Suite v3.0 software. Sequences were aligned to the hg19 genome assembly (GRCh37) using STAR[43] RNA-seq aligner. Using the R packages GenomicRanges[44] and Rsamtools[53], expression values were calculated for each AceView[45] annotated gene as the number of reads that overlapped with that gene's exonic coordinates.

For any read, if a read overlapped exactly with one gene, the read was counted for that gene; otherwise, the read was counted as ambiguous and discarded. The lowest 30% of genes (n = 21,710), as determined by the sum of all inter-site and intra-site depth-normalized counts for each gene, were then removed from each sample. Genes with average low read counts of ≤2.7 mapped reads across the whole gene are extremely variable, and their removal is recommended by the SEQC Consortium in the SEQC main manuscript. Due to the lower sequencing depth of the PGM data, the read count for each gene is much less than in the ILM data set, we filtered out the lowest 50% of

AceView[45] genes to achieve an average read count across all replicates with at least two reads before gene count normalization. This ensured that we only examined consistently detected genes at all sites from all platforms.

Surrogate variable analysis. Normalized gene expression values for all samples were used to detect latent variables using the sva package[26]. Two latent variables were constructed using twostepsva.build() function based on the two-step algorithm of Leek and Storey[26,27]. Latent variables in the DEG analysis were removed by adding the latent variable in the design matrix for limma approach mentioned above.

PEER analysis. Normalized gene expression estimates for all samples were used to detect latent variables using PEER package[30]. The covariates associated with sample type were included for inference and the inferred hidden confounders were removed from the signal. The optimal number of hidden confounders was found to be two and three for ILM and PGM data sets respectively, as the robust analysis of higher numbers of confounders has indicated (data not shown) that influence of further confounders is negligible and thus these can be omitted.

GC bias correction. We applied two R packages cqn[17] and EDASeq[18] to correct the GC content bias and normalized the gene expression, respectively. Then the normalized expression matrix was fed in limma lmfit(), contrasts. fit() and eBayes() functions for differentially expression analysis.

Remove unwanted variables analysis. We applied RUV2 function[31] to remove the unwanted variables in the normalized expression values on the $\log_2$ scale. The 23 ERCC read counts were used as the control.

3′ UTR gene counting. Gene counts were created as previously described, except 3′UTR coordinates were used in place of exon coordinates.

RNA-seq quality metrics. R-make (http://physiology.med.cornell.edu/faculty/mason/lab/r-make/) is an open-source package that we used for all quality metrics evaluation. R-make depends on BEDTools[54], samtools[55], BamTools[56], STAR[43] and interval container library[57]. In brief, quality metric definitions were as follows: sequencing depth: total number of reads sequenced; mapping rate: percentage of reads which mapped uniquely to the reference genome; sequence directionality: the number of reads which mapped to the forward and reverse strands compared to those of the AceView gene model; nucleotide composition: the total number of A/G/C/T sequenced at each position across the length of the read; guanine-cytosine (GC) distribution: the number of reads with a particular %GC content; read distribution: the fraction of the reads that mapped to either exons, 3′UTRs, 5′UTRs, introns, or intergenic regions (or the intersection of any of the aforementioned categories) as defined by the AceView gene models; coverage uniformity: the percentage of reads covering each nucleotide position of all genes scaled to 100 bins; error rate: the number of mismatches in each unique, aligned read with respect to the reference genome for each nucleotide position across all reads; base quality scores: Phred-quality scores as calculated by Illumina's HCS for each nucleotide position across all reads; insert size: the distance between two paired fragments as calculated by the start position of read-2 minus the end position of read1; and duplication rate: the number of reads with exactly the same sequence content.

RNA-seq differential gene expression analysis. Lists of differentially expressed genes were generated using the limma-voom pipeline[58,59] and compared to the total set AceView genes consistently observed at all sites (n = 45,656). All samples used four replicates, e. g., four of sample A at site 1 versus four replicates of sample A at site 2, etc. The limma package[58,59] has implemented RNA-seq differential gene expression analysis. In the current study, the differential gene expression analysis followed the limma package[58,59] user guide (http://www.bioconductor.org/packages/2.12/bioc/vignettes/limma/inst/doc/usersguide.pdf). Briefly, the trimmed mean of M-values normalization method, which uses a weighted trimmed mean of the log expression ratios, was applied to the raw gene counts[46,60,61]. Using voom() from the limma package[58,59], the mean-variance relationship of the counts was estimated, and the appropri-

ate weights for each observation were computed based on their predicted variance. By applying the lmFit(), contrasts.fit() and eBayes() functions, also from the limma package, the fold changes and standard errors were estimated by fitting a linear model for each gene, and empirical Bayes smoothing was applied to the standard errors. We used the Benjamini and Hochberg adjustment for multiple testing at a variety of FDR values (0.05 or 0.01 or 0.001). Differentially expressed genes were evaluated at $\log_2$ FC cutoffs (1.5 or 2).

TaqMan gene expression analysis. TaqMan data for samples A, B, C and D were obtained through GEO (accession number GSE5350)[3]. Each TaqMan assay was run in four replicates for each sample. Undetectable C_T values ($C_T > 35$) were removed before normalization. The data were normalized using the HTqPCR package[62] to the average C_T of POLR2A by subtracting the average C_T of POLR2A from each TaqMan target to give the $\log_2$ difference between endogenous control and target gene[3]. TaqMan differential gene analysis was performed as for RNA-seq data, minus the TMM and voom transformations.

Gene expression quantification correlation of TaqMan data and RNA-seq data. We obtained the TaqMan primer sequence from 2006 MAQC Consortium. We then mapped the sequence using blat to hg19 refseq transcriptome have 100% alignment (available at http://physiology.med.cornell.edu/faculty/mason/lab/data3/sac2026/ABRF/Data/SEQC/taqman_refseq_mapping.bed). We then converted the transcriptome alignment results to genome locations using in-house R script, considering three conditions: (i) single-exon genes; (ii) multi-exon genes (sense or anti-sense strand) with primer in one exon; (iii) multi-exon genes (sense or anti-sense strand) with primer spanning two exons. After double confirmation with the UCSC genome browser on the actual sequence on the genome, we annotated the read count for SEQC project using the genome locations of TaqMan's 863 primer sequences. We then compared the TaqMan-normalized gene expression level with the primer sequence–annotated RNAseq-normalized gene expression using scatter plot and calculated the Pearson correlation.

Validation of DEGs from RNA-seq data using TaqMan data. DEGs from RNA-seq data from each site for six comparisons (A-B, A-C, A-D, B-C, B-D, C-D) were validated using the DEGs from the TaqMan data. Based on our FDR and FC cutoff, for example, genes with adjusted P value smaller than 0.05 and absolute fold change greater than 1.5 and declare them to be differentially expressed, our findings might include both truly differentially expressed genes (true positives) and nondifferentially expressed genes (false positives). Given a list of declared DEGs from sequencing data and the information about which genes in TaqMan are truly differentially expressed and which genes are not, we can calculate the true-positive rate (TPR) and false-positive rate (FPR). TPR is defined as the proportion of true DEGs that are declared to be differentially expressed, whereas FPR is the proportion of non-DEGs that are also declared to be differentially expressed, which both range from 0 to 1. The Matthews Correlation Coefficient (MCC) was chosen as a measure of DEG detection accuracy[41,42], which combines test sensitivity and specificity.

53. Morgan, M., Pages, H. & Obenchain, V. Rsamtools: Binary alignment (BAM), variant call (BCF), or tabix file import. *R package version 1.14.3* (2014).
54. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
55. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
56. Barnett, D.W., Garrison, E.K., Quinlan, A.R., Stromberg, M.P. & Marth, G.T. BamTools: a C. API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692 (2011).
57. Faulhaber, J. Boost interval container library http://www.boost.org/doc/libs/1_55_0/libs/icl/doc/html/index.html (2010).
58. Law, C.W., Chen, Y., Shi, W. & Smyth, G.K. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
59. Smyth, G.K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds. Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S.) 397–420 (Springer, New York, 2005).
60. Robinson, M.D. & Smyth, G.K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9**, 321–332 (2008).
61. Robinson, M.D. & Smyth, G.K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881–2887 (2007).
62. Dvinge, H. & Bertone, P. HTqPCR: high-throughput analysis and visualization of quantitative real-time PCR data in R. *Bioinformatics* **25**, 3325–3326 (2009).