University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2009 *Paper* 247

Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments

James H. Bullard[*] Elizabeth A. Purdom[†]

Kasper D. Hansen[‡] Sandrine Dudoit[**]

[*] Division of Biostatistics, University of California, Berkeley, bullard@stat.berkeley.edu

[†] Division of Biostatistics and Department of Statistics, University of California, Berkeley, epurdom@stat.berkeley.edu

[‡] Division of Biostatistics, University of California, Berkeley, khansen@stat.berkeley.edu

[**] Division of Biostatistics and Department of Statistics, University of California, Berkeley, sandrine@stat.berkeley.edu

Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments

James H. Bullard, Elizabeth A. Purdom, Kasper D. Hansen, and Sandrine Dudoit

Abstract

The focus of this article is on the design and analysis of mRNA-Seq experiments, with the aim of inferring transcript levels and identifying differentially expressed genes. We investigate two mRNA-Seq datasets obtained using Illumina's Genome Analyzer platform to measure transcript levels in reference samples considered in the MicroArray Quality Control (MAQC) Project. We address the following four main issues: (1) exploratory data analysis for mapped reads, relating read counts to variables describing input samples and genomic regions of interest; (2) assessment and quantitation of biological effects (e.g., expression levels in Brain vs. UHR) and nuisance experimental effects (e.g., library preparation, flow-cell, and lane effects); (3) evaluation and comparison of methods for the identification of differentially expressed genes; (4) impact of base-calling calibration method (phi X vs. auto-calibration).

Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments

James H. Bullard*, Elizabeth A. Purdom*,
Kasper D. Hansen, and Sandrine Dudoit
Division of Biostatistics and Department of Statistics
University of California, Berkeley

* These authors contributed equally to this work.

Contents

List of Figures

Abstract

Background: High-throughput sequencing technologies, such as the Illumina Genome Analyzer, are powerful new tools for investigating a wide range of biological and medical questions. Statistical and computational methods are key for drawing meaningful and accurate conclusions from the massive and complex datasets generated by the sequencers. We provide a detailed evaluation of statistical methods for normalization and differential expression (DE) analysis of Illumina transcriptome sequencing (mRNA-Seq) data. **Results:** We compare statistical methods for detecting genes that are significantly DE between two types of biological samples and find that there are substantial differences in how the test statistics handle low-count genes. We evaluate how DE results are affected by features of the sequencing platform, such as, varying gene lengths, base-calling calibration method (with and without phi X control lane), and flow-cell/library preparation effects. We investigate the impact of the read count normalization method on DE results and show that the standard approach of scaling by total lane counts (e.g., RPKM) can bias estimates of DE. We propose more general quantile-based normalization procedures and demonstrate an improvement in DE detection. **Conclusions:** Our results have significant practical and methodological implications for the design and analysis of mRNA-Seq experiments. They highlight the importance of appropriate statistical methods for normalization and DE inference, to account for features of the sequencing platform that could impact the accuracy of results. They also reveal the need for further research in the development of statistical and computational methods for mRNA-Seq.

1 Background

For the past decade, microarrays have been the assays of choice for high-throughput studies of gene expression. Recent improvements in the efficiency, quality, and cost of genome-wide sequencing have prompted biologists to rapidly abandon microarrays in favor of ultra high-throughput sequencing, a.k.a., second-generation or next-generation sequencing: e.g., Applied Biosystems' SOLiD, Helicos BioSciences' HeliScope, Illumina's Genome Analyzer, and Roche's 454 Life Sciences sequencing systems. These high-throughput sequencing technologies have already been applied to monitor genome-wide transcription levels (mRNA-Seq), DNA-protein interactions (ChIP-Seq), chromatin structure, and DNA methylation (Chiang et al., 2009; Dohm et al., 2008; Hoen et al., 2008; Lee et al., 2008; Li et al., 2008; Marioni et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Wang et al., 2008).

We evaluate statistical methods for the inference of differential expression (DE) with mRNA-Seq, using reference samples from the MicroArray Quality Control (MAQC) Project (MAQC Consortium, 2006). With corresponding quantitative real-time polymerase chain reaction (qRT-PCR) data on roughly one thousand genes, we compare different normalization and DE procedures and assess possible biases related to the sequencing technology. For genes that are well-expressed in both samples being compared, the examined tests (Fisher's exact test and GLM-based tests) are indistinguishable. However, substantial differences exist in their ability to give reliable DE estimates when even just one of the samples yields low read counts (e.g. ≤ 10). One inherent bias of the Illumina platform is the preferential sequencing of longer genes (Oshlack and Wakefield, 2009). With the tests considered here, longer genes are more likely declared DE. We demonstrate that weighting the DE statistics by gene length can mitigate this effect.

While small "nuisance" technical effects can be observed due to differences in flow-cells/library preparations, we show that these do not impact substantially the differential expression calls for the MAQC dataset. We also find that not using the standard phi X control lane in each flow-cell, as in the base-calling calibration procedure recommended by Illumina, does not negatively impact DE detection. Moreover, auto-calibration without the phi X lane increases both quantity and quality of mapped reads. In this regard, there is no obvious benefit in using a phi X lane; doing away with such a control lane leads to more balanced and cost-effective designs.

We demonstrate that the greatest impact on DE detection is the choice of normalization procedure. As different lanes have different total read counts, i.e., *sequencing depths*, the usual approach is to scale gene counts within each lane by the total lane count: e.g., the now standard reads per kilobase of exon model per million mapped reads (RPKM) of Mortazavi et al. (2008) or the hypergeometric model of Marioni et al. (2008). We show that this form of global normalization is heavily affected by a relatively small proportion of highly-expressed genes and, as such, can give biased estimates of DE if these few genes are differentially expressed across the conditions under comparison. We propose alternative more robust quantile-based normalization procedures that remove the bias without introducing additional noise.

2 Methods

2.1 MAQC Datasets

This article considers two mRNA-Seq datasets related to the MicroArray Quality Control Project (MAQC Consortium, 2006) and obtained using Illumina's Genome Analyzer II high-throughput sequencing system (Illumina, 2008). The experiments analyze two biological samples: Ambion's human brain reference RNA and Stratagene's human universal reference RNA, herein referred to as Brain and UHR, respectively.

In the first experiment (MAQC-2), two types of biological samples (Brain and UHR) were assayed, each using seven lanes distributed across two flow-cells (Figure 11a). One library preparation was used for each of the two types of biological samples. In the second experiment (MAQC-3), four different UHR library preparations were assayed using 14 lanes from two flow-cells; each library preparation was assayed on only one of the flow-cells (Figure 11b).

As part of the original MAQC Project, around one thousand genes were also chosen to be assayed by qRT-PCR (Canales et al., 2006). We use these qRT-PCR data as a gold-standard to benchmark the gene expression values determined by mRNA-Seq. Additionally, a large number of microarray experiments were conducted. We compare the mRNA-Seq measures to those derived from a set of Affymetrix Human Genome U133 Plus 2.0 arrays (GSE5350, samples AFX_1_[A—B][1-5]; see Supplementary Text for details on array and qRT-PCR analysis).

2.2 Overview of the Illumina Sequencing Platform

We give a brief, non-technical overview of the steps involved in an Illumina mRNA-Seq experiment (Illumina, 2008). A sample of interest undergoes library preparation, a series of steps to convert the input RNA into small fragments of DNA that can be sequenced by the Illumina machine. Specifically, starting with any total RNA sample, Illumina's mRNA-Seq library preparation protocol includes poly-A RNA isolation, RNA fragmentation, reverse transcription to cDNA using random primers, adapter ligation, size-selection from a gel, and PCR enrichment (Illumina, 2009, Figure 6). The resulting cDNA *library* is placed in one of the eight *lanes* of a *flow-cell*. Individual cDNA fragments attach to the surface of the lane and subsequently undergo an amplification step, whereby they are converted into *clusters* of double-stranded DNA. The flow-cell is then placed in the sequencing machine, where each cluster is sequenced in parallel. Specifically, at each *cycle*, the four fluorescently labeled nucleotides are added and the signals emitted at each cluster recorded. For each flow-cell, this process is repeated for a given number of cycles, e.g., 35 cycles in the MAQC experiments. The fluorescence intensities are then converted into *base-calls*. The number of cycles determines the length of the *reads*; the number of clusters determines the number of reads.

2.3 Pre-processing of Sequencing Data

For the two MAQC experiments, 35 base-pair-long reads were obtained using Illumina's standard Genome Analyzer pre-processing pipeline, Version 1.3 (Bentley et al., 2008; Illumina, 2008). We used Bowtie to map reads to the genome (GRCh37 assembly) (Langmead et al., 2009).

Illumina's default base-calling algorithm, Bustard, can be calibrated in two ways. The method recommended by Illumina is to reserve one lane per flow-cell for sequencing DNA (typically phi X DNA) and use data from this control lane to determine base-calls and quality scores for the other seven lanes (Bentley et al., 2008, Supplementary Information, p. 7). Bustard can also be run using the auto-calibration method, which scores base-calls in a manner similar to the phred base-caller (Ewing and Green, 1998). In both MAQC experiments, one lane of each flow-cell was reserved for sequencing phi X genomic DNA. For one experiment (MAQC-2), we obtained both auto-calibrated and phi X-calibrated reads.

Except for Section 3.2.2, we focus on phi X-calibrated, purity-filtered reads that map uniquely to the genome, with up to two mismatches. The restriction to reads mapping to the genome implies that exon-exon junction reads are excluded ($\sim 10\%$ of the reads). Additionally, the library preparation protocol does not allow consideration of strand-specific counts, so reads mapping to the forward and reverse strands are pooled.

2.4 Definition of Union-intersection Genes

In our evaluation of DE, we focus on overall expression of a gene, rather than isoform-specific expression. There is no standard technique for summarizing expression levels of genes with several isoforms (see, for example, Marioni et al. (2008) and Mortazavi et al. (2008) for different approaches). For a given gene, we first define a *constitutive exon* as a set of consecutive exonic bases (i.e., portion of or entire exon) that belong to each isoform of the gene. We then define a *union-intersection (UI) gene* as a composite gene-level region of interest consisting of the union of constitutive exons that do not overlap with coding exons of other genes (based on Ensembl, Version 55; see Supplementary Text). We retain all genes identified with chromosomes 1–22, X, and Y. In addition to including protein-coding genes, the UI genes represent a number of other classes of Ensembl annotation, such as pseudogenes and small RNAs.

2.5 Normalization

In order to derive gene expression measures and compare these measures between (groups of) lanes, one first needs to normalize read counts to adjust for varying lane sequencing depths and potentially other technical effects. All but one of the normalization methods considered here are *global* procedures, in the sense that only a single factor d_i is used to scale the counts (per-lane).

We evaluate three types of global normalizations: (1) total lane counts, as in RPKM of Mortazavi et al. (2008), (2) per-lane counts for a "housekeeping" gene expected to be constantly expressed across biological conditions, e.g., POLR2A, (3) per-lane upper-quartile of gene counts for genes with reads in at least one lane. In order to make the normalized expression measures comparable, the scaling factors are themselves scaled so that their sum across all lanes is equal to the sum of the total counts across all 14 lanes (see Supplementary Text).

The expression quantitation problem can be framed in terms of generalized linear models (GLM),

$$\log(\mathrm{E}[X_{i,j}|d_i]) = \log d_i + \lambda_{a(i),j} + \theta_{i,j}, \tag{1}$$

where the natural logarithm of the expected value of the read count $X_{i,j}$ for the jth gene in the ith lane is modeled as a linear function of the gene's expression level $\lambda_{a(i),j}$ for the biological condition $a(i)$ assayed in lane i plus an offset ($\log d_i$) and possibly other technical effects ($\theta_{i,j}$).

Finally, we propose a quantile normalization procedure, inspired from the microarray normalization approach of Irizarry et al. (2003a) and its implementation in the R package aroma.light. Specifically, for each lane, the distribution of read counts is matched to a reference distribution defined in terms of median counts across sorted lane. The normalized data are rounded to produce integer values that can be used with the DE statistics described in Section 2.6, below.

2.6 Differential Expression

We compare three types of methods for inferring DE, each of which yields one test statistic per gene: Fisher's exact test statistic, likelihood ratio statistics based on a generalized linear model as in Equation (1), and t-statistics based on estimated parameters of the same GLM. Two different t-statistics are evaluated, which use different techniques for estimating the variance of the estimated parameters. We also assess the impact of flow-cell effects, either through the addition of parameters $\theta_{i,j}$ in the GLM or through a Mantel-Haenszel test, an extension of Fisher's exact test (see Supplementary Text).

All of the considered DE statistics can accommodate global normalization via an offset d_i. For the GLM-based statistics, the offset is handled as in Equation (1). Fisher's exact test and the Mantel-Haenszel test compare the distribution of the counts of the jth gene to that of d.

The likelihood ratio statistics are the most general, as they can be used for comparisons of any number of biological sample types and adjust for general experimental effects as well as sample covariates, e.g., RNA quality. The t-statistics are only applicable for testing differences between two groups. The t-statistics and likelihood ratio statistics are based on maximum likelihood estimators from the same GLM, but have different performance in certain cases. Distributional properties of all of the GLM-based statistics are derived under asymptotic theory; therefore, they may have poor behavior for small numbers of input samples or low counts (though this is not what we experience). In contrast, Fisher's exact test makes no assumption about sample size; however, it only adjusts for global experimental effects and even the Mantel-Haenszel extension allows only a single gene-level experimental effect.

Likelihood ratio statistics have been used in Marioni et al. (2008) for the special case of only a global lane effect (i.e., $\theta_{i,j} = 0$ in Equation (1)); these authors also mentioned applying an arcsine-root transformation for variance stabilization of the per-gene read proportions within each lane. Bayesian statistics with Gamma prior for the Poisson parameter have been found to yield similar results as the above GLM-based test statistics (Taub, 2009). Other test statistics considered in the recent mRNA-Seq literature include t-statistics with square root-transformed standard errors and Bayesian statistics based on the Beta-Binomial distribution (Hoen et al., 2008).

2.7 Receiver Operator Characteristic Curves using qRT-PCR Gold-standard

The qRT-PCR data of Canales et al. (2006) are used as gold-standard to determine "true" differential expression and derive receiver operator characteristic (ROC) curves for various mRNA-Seq and microarray DE methods. The qRT-PCR estimate of UHR to Brain expression log-fold-change is the difference of average expression measures for UHR and Brain across replicates (see Supplementary Text).

We divide the genes assayed by qRT-PCR into three sets, "non-DE", "DE", and "no-call", based on whether their absolute expression log-fold-change is less than a, greater than b, or falls within the interval $[a, b]$, respectively. We ignore the "no-call" genes when determining true/false positives/negatives. True positives (TP) are reported when the sequencing (or microarray) platform not only correctly declares a gene DE, but also agrees with qRT-PCR regarding the direction of DE. The true positive rate (TPR) is then defined as the total number of TPs divided by the total number of DE genes according to qRT-PCR; the false positive rate (FPR) is computed as usual. See Table 1 for a summary.

2.8 Software

In order to facilitate analysis and visualization of mRNA-Seq data, we developed the R packages Genominator and GenomeGraphs (Durinck et al., 2009). In addition to the analysis methods implemented in these packages, all R programs used in this study are available at: `www.stat.berkeley.edu/~bullard/mRNA-Seq`;

3 Results and Discussion

3.1 Comparison of mRNA-Seq Differential Expression Statistics

Lists of differentially expressed genes are typically produced by computing, for each gene, a test statistic comparing expression levels between the two types of biological samples and ranking the genes based on p-values assessing the statistical significance of the observed differences.

We evaluate various statistics for differential expression (see Section 2.6) and find that the main difference between test statistics is their ability to handle low counts, an issue of great importance when investigating differential expression in context of mRNA-Seq. When both samples have zero reads, clearly nothing can be said about differential expression. The more pertinent zero-count or low-count scenario occurs when a gene has zero reads for one sample and a reasonable number for the other. Around 700 genes ($\sim 1.8\%$) have zero reads in either Brain or UHR and 10 or more reads in the other tissue. Presumably, this represents an interesting biological phenomenon, where a gene in one tissue is completely non-expressed.

For genes with zero counts in either sample, the t-statistics fail: the estimated standard errors become extremely large (or infinite in the case of the delta method t-statistic) and the nominal p-values cluster around one, regardless of the number of reads in the other sample. For Fisher's exact test and the GLM-based likelihood ratio test, however, we see a continuum of p-values as desired. For genes with reasonable counts in both samples, the choice of test statistic makes little difference in the nominal p-values (Figure 9). Because they cannot stably handle low-count genes, the t-statistics are failing to detect many "easy" cases of DE (i.e., genes with large differences in expression between the two conditions) and, as a result, have very low sensitivity. The poor performance of the t-statistics is reflected in ROC curves of the DE tests using qRT-PCR as gold-standard. Removal of genes with fewer than 20 reads in both samples completely accounts for the poor sensitivity of the t-statistics and results in equivalent ROCs for the various DE statistics, all of which are dramatically improved (Figure 1).

As the different mRNA-Seq DE tests show similar behavior, we will from here on focus only on the results from the GLM-based likelihood ratio tests. The results do not change when different test statistics are used, except for the already noted poor performance of the t-statistics for low-count genes.

3.2 Impact of Technical Effects on Differential Expression

3.2.1 Gene-length Biases in Differential Expression

It is expected from the mRNA-Seq assay that longer transcripts contribute more "sequencible" fragments than shorter ones expressed at the same level. There is clearly a positive association between gene counts and length, an association that is not entirely removed via scaling by gene length, as in the RPKM of Mortazavi et al. (2008) (Figure 13). This suggests either higher expression among longer genes or non-linear dependence of gene counts on length.

As noted by Oshlack and Wakefield (2009), the dependence of gene counts on length creates "gene length-related biases" in mRNA-Seq DE results: longer genes tend to have more significant DE statistics (Figure 2). All of the mRNA-Seq DE statistics evaluated here have an inherent dependence of their estimated standard errors on read counts. This is a serious shortcoming in terms of creating "gene-lists" for differential expression, as the resulting lists could favor long genes with small underlying effects as compared to short genes with large effects. Considering only estimated fold-changes is inadequate, as this ignores the fairly large range of standard errors for a given fold-change and gene length.

One can possibly remedy the length dependence of DE statistics using a fixed number of bases from each gene; repeating the DE analysis by randomly selecting 250 bp from each gene removes the association between DE significance and length (Figure 14). This also indicates that the cause of the association is the length of the gene and not differences in the underlying expression levels of longer genes. However, a fixed-length analysis is unsatisfactory, as it discards large amounts of data and there is no natural choice of common length.

A weighted analysis based on gene length might constitute a reasonable compromise towards a length-independent DE filter. Indeed, scaling each t-statistic by the inverse of the square root of length provides a length-independent ranking (Figure 2). However, the problem of choosing a cutoff still remains. Under the assumptions presented in Oshlack and Wakefield (2009), with the unweighted t-statistics and using the same cutoff across genes, power increases with gene length for a given level of DE. Under the same scenario, for the weighted t-statistics, both Type I error rate and power decrease with length.

3.2.2 Impact of Base-calling Calibration Method

The practice of reserving one lane out of eight, in each flow-cell, for sequencing bacteriophage phi X genomic DNA has important implications for experimental design, in terms of sample size and balance. We find that more reads are mapped to the genome with auto-calibration than with the standard phi X calibration, at each of three mapping stringency levels (Figure 3). Purity-filtered perfectly matching (FPM) reads are unlikely to contain sequencing errors and can serve as proxies for perfectly accurate reads. Similarly, purity-filtered reads with either 0, 1, or 2 mismatches (FMM) are comprised of both FPM reads as well as reads that represent sequencing errors. Then, the ratio (FMM-FPM)/FMM can be viewed as a rough estimate of the sequencing error rate, assuming no SNPs. For all lanes, the auto-calibration method produces slightly lower error rates (by $\sim 5\%$).

The increased number of reads is spread unevenly throughout the transcriptome. A majority of the UI genes have no change in read counts between calibration methods, whereas around 25% of the genes have 4 or more additional reads when using auto-calibration. When computing an (FMM-FPM)/FMM ratio for each gene for both phi X and auto-calibration, the auto-calibration has a lower error rate by about 3.8% on average.

The significance of differences in expression measures between the two calibration methods was evaluated by comparing observed differences to a permutation distribution of differences obtained by randomly swapping the auto-calibrated and phi X-calibrated sets of read counts for each of the 14 lanes. We find that in terms of absolute expression measures there are small, but statistically significant differences between the two calibration methods. However, relative expression measures, as used in DE analyses, do not appear to be significantly different (see Supplementary Text).

Although our assessment is based on only two flow-cells, it seems quite clear that auto-calibration is advantageous, as it yields more balanced designs, frees up one lane per flow-cell, and produces a larger number of higher quality reads per lane.

3.2.3 Lane, Flow-cell, and Library Preparation Effects

The Poisson distribution has been shown to provide a good fit to the distribution of gene-level counts across replicate lanes, after normalization by total lane counts (Lee et al., 2008; Marioni et al., 2008); our experience with both the MAQC data and unpublished datasets for *Drosophila melanogaster* supports this conclusion. The goodness-of-fit of the Poisson model across three different organisms and four different sequencing facilities strongly supports its validity as a model for lane variation and justifies the pooling of read counts across lanes by summation. Note, however, that the applicability of the Poisson distribution is questionable when analyzing *biological replicates* (i.e., samples from different individuals within a given biological group, such as, patients with the same type of cancer). The use of negative binomial or empirical Bayes methods, as described in the SAGE literature (Lu et al., 2005; Robinson and Smyth, 2007), may be sensible in such settings of increased variability.

Our analyses also confirm the previously noted small technical differences between flow-cells (Marioni et al., 2008), though there is evidence of slightly more variation between flow-cells than between replicate lanes (Figure 15c). Regardless of their statistical significance, estimated flow-cell effects are small and thus have a minor impact only in detecting extremely small biological effects; almost none for genes with more than 3 reads/lane.

To the best of our knowledge, there has been no published examination of the technical variation introduced during library preparation; replication of the library preparation is both expensive and time-consuming. There are clear library preparation effects on the total number of reads (Figure 11). After adjusting for differences in total lane counts, there is evidence for increased variation across replicate library preparations as compared to flow-cells and lanes (Figure 15d); however, this increased variability is mainly due to high-count genes for which there is high power to detect small differences. A direct comparison of library preparation effects to flow-cell and biological effects is not possible due to the experimental design, but comparison of the magnitude of the estimated differences suggests that library preparation effects are much smaller than the biological effects between Brain and UHR (Figure 4) and slightly larger for some genes than flow-cell effects (Figure 4 and Figure 15).

The biological differences between Brain and UHR samples may be much larger than those typically observed; therefore, technical sources of variation need not always be irrelevant. Finally, we note that the MAQC data are somewhat "ideal", in the sense that: (1) commercial-grade RNA was sequenced and (2) the sequencing was performed in-house by Illumina. A typical mRNA-Seq experiment begins with the extraction of RNA from biological specimens and variability induced during extraction may be much larger than the technical variability seen here.

3.3 Normalization of mRNA-Seq data

Because the total number of reads varies between lanes, read counts must be normalized to allow comparison of expression measures across lanes or samples. This subject has received relatively little attention in the mRNA-Seq literature. The common practice is to scale the gene counts by lane totals (Marioni et al., 2008; Mortazavi et al., 2008). We find, however, that more general quantile-based procedures yield much better concordance with qRT-PCR and are hopefully more robust than normalization by a single housekeeping gene.

Here, we evaluate a variety of normalization procedures and focus on two main questions: (1) Does the normalization improve DE detection (sensitivity)? (2) Does the normalization result in low technical variability across replicates (specificity)? To assess DE detection, we rely on the qRT-PCR data of Canales et al. (2006) as a gold-standard for determining true and false positives. Because there are a limited number of non-DE genes in the qRT-PCR data, we also assess goodness-of-fit to the Poisson model for replicate lanes (GLM 1 in Table 4).

The simplest form of normalization is achieved by scaling gene counts, in lane i, by a single lane-specific factor d_i. In essence, these *global* scaling factors define the null hypothesis of no differential expression: if a gene has the same proportions of counts across lanes as the proportions determined by the vector of d_i's, then it is deemed non-differentially expressed.

The standard total-count normalization results in low variation across lanes, flowcells, and library preparations, as discussed above. What has not been understood previously, is that this normalization technique reflects the behavior of a relatively small number of extremely high-count genes: 5% of the genes account for approximately 50% of the total counts in both Brain and UHR. These genes are not guaranteed to have similar levels of expression across different biological conditions and, in the case of the MAQC-2 dataset, they are noticeably over-expressed in Brain, as compared to the majority of the genes (Figure 5).

Accordingly, the performance of total-count normalization is not particularly impressive for detecting DE (Figure 6): sensitivity is only slightly higher as compared to the microarray data, even for genes with relatively large differences in expression (> 2 absolute log-ratio). When including genes with lower levels of differential expression (> 0.5 absolute log-ratio), performance is no better (and perhaps slightly worse) than that of microarrays. This contradicts general expectation given that the mRNA-Seq data are less noisy and thus better at detecting small expression differences. For small levels of DE, the bias in estimated log-ratios using total-count normalization makes the sequencing estimates less accurate.

We evaluate two alternatives for normalization of mRNA-Seq data. One approach relies on a single housekeeping gene like POLR2A, a standard technique for normalizing qRT-PCR expression measures. However, this is not a feasible solution in general, since it is not known *a priori* which genes have stable expression levels (in Canales et al. (2006), POLR2A was chosen only after examining many replicates for UHR and Brain across a number of plates).

In analogy with standard techniques for normalizing microarray data, we propose to match the between-lane distributions of gene counts in terms of parameters such as quantiles. For instance, one could simply scale counts within lanes by their median. In our case, due to the preponderance of zero and low-count genes, the median is uninformative for the different levels of sequencing effort. Instead, we use the per-lane upper-quartile (75th percentile), after excluding genes with zero reads across all lanes (see Supplementary Text).

Compared to total-count normalization, both POLR2A and upper-quartile normalization significantly reduce the bias of DE relative to qRT-PCR (Figure 7), with upper-quartile having bias near zero. ROC curves illustrate that both upper-quartile and POLR2A normalization are unequivocally better than total-count normalization at detecting DE and result in improved sensitivity of sequencing relative to microarray data.

A closer look at technical variation for the different normalization procedures shows that upper-quartile normalization does not noticeably increase the level of variability as compared to total-count normalization; POLR2A normalization is slightly more variable but still comparable (Figure 8).

Finally, it is also feasible to perform quantile normalization across lanes, as is often done in microarray experiments (Irizarry et al., 2003b). However, there does not seem to be added benefit to this more complicated normalization strategy. Quantile normalization performs similarly in the ROC analyses (Figure 16a) and induces comparable, or even slightly more, variability than upper-quartile normalization (Figure 8). We again recall the somewhat artificial nature of the MAQC data, which were obtained at essentially the same time, by one lab, using ideal RNA samples. As more data become available, there may be larger variations in gene count distributions necessitating more aggressive normalization.

4 Conclusions

Our main novel finding is the extent to which normalization affects differential expression results: sensitivity varies more between normalization procedures, than between test statistics. Although the standard total-count normalization results in Poisson variation across replicate lanes, it has poor detection sensitivity when benchmarked against qRT-PCR. Instead, we propose scaling gene counts by a quantile of the gene-count distribution (the upper-quartile) and show that such normalization improves sensitivity without loss of specificity.

It is possible that the improvement of POLR2A over total-count normalization is due to more closely matching the qRT-PCR data, which were normalized by POLR2A, rather than proper reflection of actual biological differences. Indeed, additional scaling of the microarray data by POLR2A slightly improves the ROC compared to the standard microarray quantile normalization. It is more likely, however, that total-count normalization, with its reliance on high-count genes, poorly reflects biological differences. This can be seen by taking a closer look at the POLR2A gene, which was chosen because of its very similar expression in UHR and Brain across many qRT-PCR replicates (Canales et al., 2006): the UHR to Brain fold-change of POLR2A is estimated as 1.3 for total-count normalization in contrast to 0.97 for upper-quartile normalization and 0.90 for microarray data.

In regards to DE test statistics, the GLM-based likelihood ratio statistics and Fisher's exact statistics perform equally well in terms of sensitivity and handling of low-count genes. We find likelihood ratio tests appealing because of their generality. Indeed, using the GLM framework, one can adjust for potential confounding variables, including quantitative covariates, e.g., age of sample, as well as accommodate different count distributions (negative binomial in cases of over-dispersion).

A serious concern with all the DE methods considered here is the inherent dependence of power on read count, which in turn is related to both gene expression level and length. As most DE studies produce gene-lists, which are often then related to functional annotation (e.g. GO), it is undesirable for significance values to be driven by features such as length. A weighted analysis based on gene length might lead to a reasonable length-independent ranking of genes, that would allow short genes with large effects to gain in significance compared to long genes with small effects.

We find that technical variation is quite low across lanes and flow-cells and slightly larger across library preparations. In all cases, however, the effect on differential expression results is minimal. We note that the MAQC datasets are unusual, in that we expect extremely large differences in expression between Brain and UHR and only small library preparation effects because of the high quality of the RNA. In practice, library preparation effects may be closer in magnitude to biological effects.

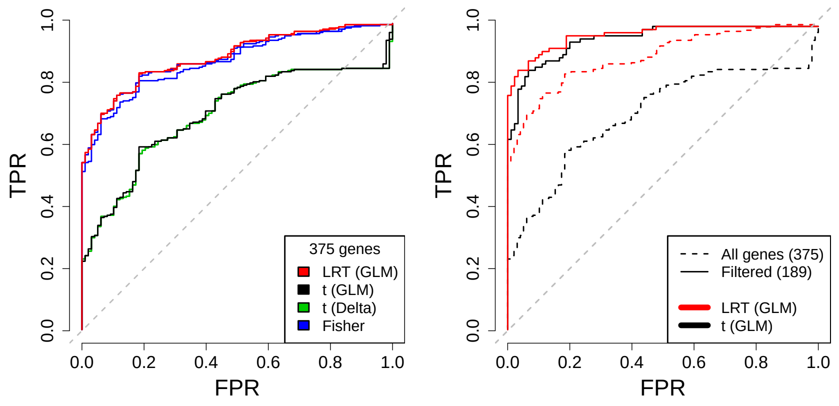
We have demonstrated that while there are some differences between phi X and auto-calibration in the early stages of the analysis pipeline, the differences in terms of differential expression are small. Overall, auto-calibration seems advantageous, as it yields more balanced designs, frees up one lane per flow-cell, and produces a larger number of higher quality reads per lane.

The analysis conducted in this work, as well as others, is predicated on a "whole-gene" view of expression profiling. We evaluated technical effects, phi X calibration, and normalization methods using a very constrained UI gene definition. We limited ourselves to such a strict definition in order to ensure that the evaluation was not biased by alternative splicing or overlapping genes. Our UI gene definition is a gross over-simplification, as a large amount of biologically relevant information is lost; we exclude more than 50% of the reads which fall within Ensembl genes.

As high-throughput sequencing becomes more prevalent, our ability to precisely characterize the transcriptome of a sample will dramatically increase. More refined analyses, such as isoform-level expression, allele-specific expression, and genome annotation (segmentation), involve comparing distinct regions within a sample as opposed to the same region across samples. Such analyses will require an understanding of the effect of sequence composition on base coverage to account for the heterogeneity of base-level count distributions within a gene.

		mRNA-Seq			
		Non-DE	DE +	DE −	
qRT-PCR	Non-DE	TN	FP	FP	N
	DE +	FN	TP	FP	P
	DE −	FN	FP	TP	

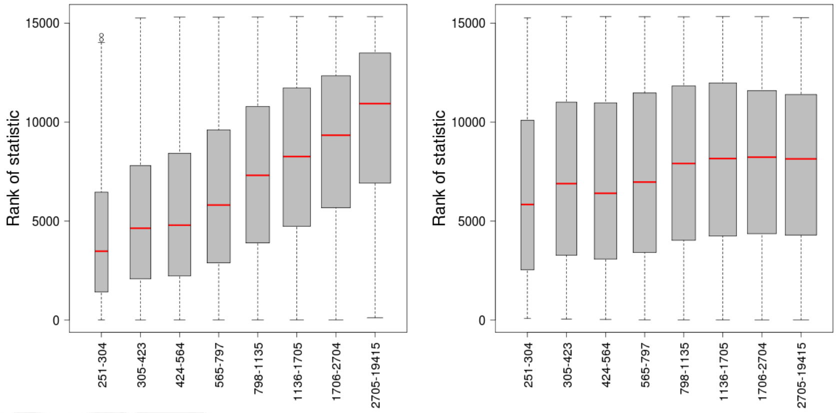
Table 1: Definition of true and false positive rates Synopsis of the rules for calling true/false positives and negatives, which take into account the sign of the direction of differential expression: "+" for over-expression in UHR, "−" for over-expression in Brain. The true positive rate (TPR) is estimated as TP/P and the false positive rate (FPR) as FP/N.



(a)

(b)

Figure 1: *Comparison of differential expression statistics: ROC curves.* (a) All DE statistics, no gene filtering. (b) GLM-based likelihood ratio statistics and t-statistics, before and after removing genes with fewer than 20 reads in either Brain or UHR. In both plots, a gene was declared non-DE if its qRT-PCR absolute log-ratio was less than 0.2 and DE if its absolute log-ratio was greater than 2.0. Note that the ROC curves do not reach the point (1,1), because of the sign condition in the definition of true positives.



(a) t-statistics

(b) Length-weighted t-statistics

Figure 2: *Differential expression statistics, by length.* Boxplots of the ranks of DE statistics vs. gene lengths for UI genes at least 250 bp-long and with non-zero counts in both Brain and UHR. (a) Delta method t-statistics. (b) Delta method t-statistics weighted by the inverse of the square root of gene length.

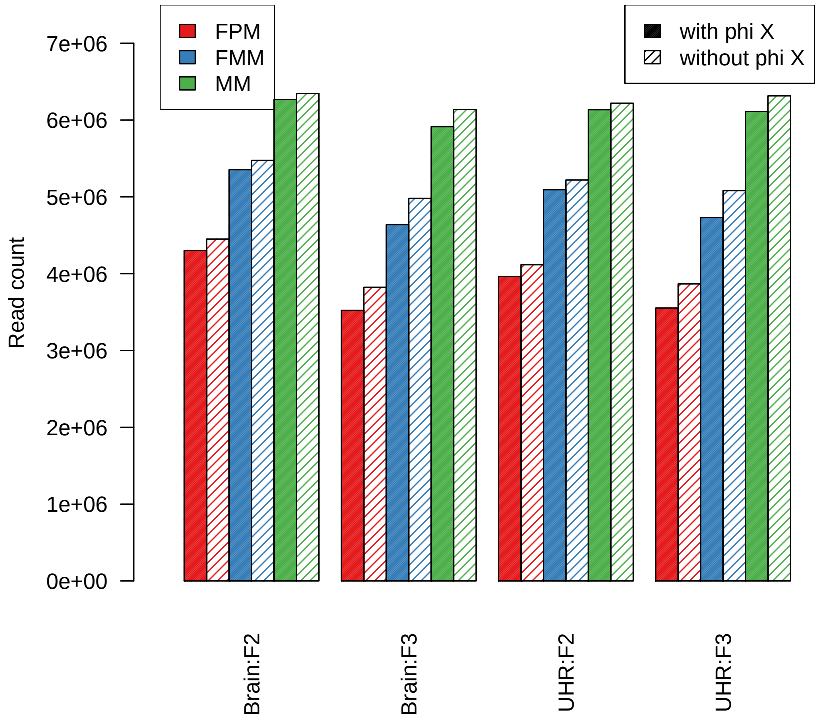


Figure 3: *Impact of base-calling calibration method on read-mapping.* Barplots of average read counts per lane with and without phi X calibration, for each of the four biological sample (Brain, UHR) and flow-cell (F2, F3) combinations. Reads are classified into three nested categories: purity-filtered perfectly matching reads (FPM); purity-filtered reads with either 0, 1, or 2 mismatches (FMM); unfiltered reads with either 0, 1, or 2 mismatches (MM).

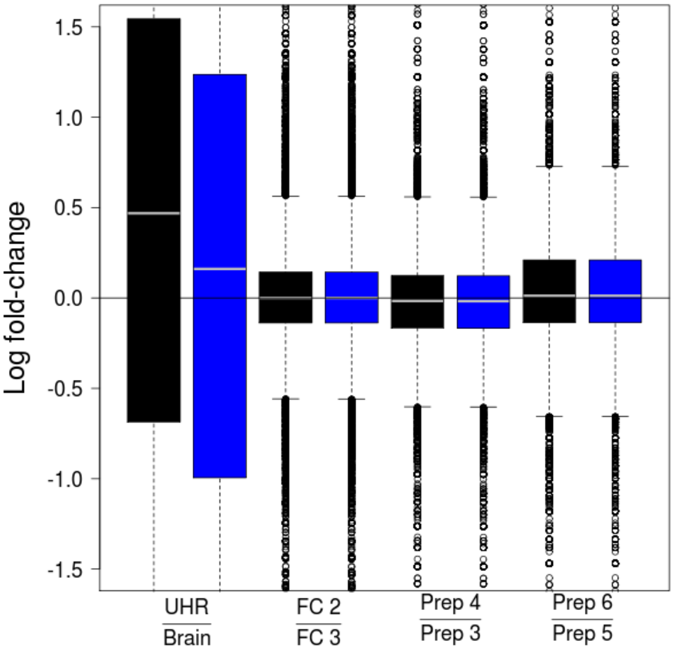
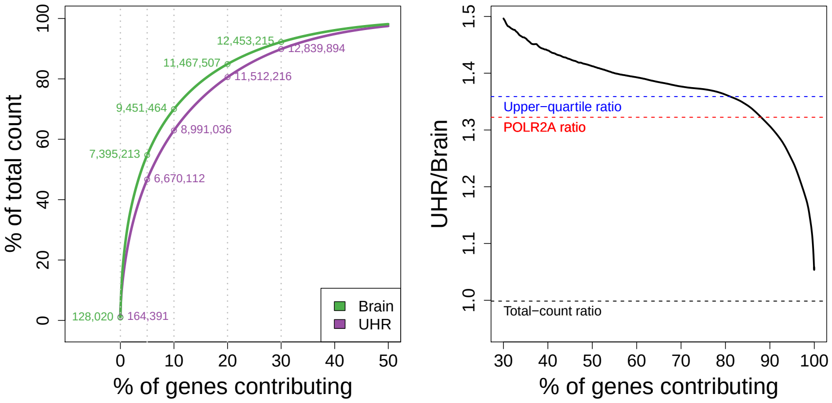
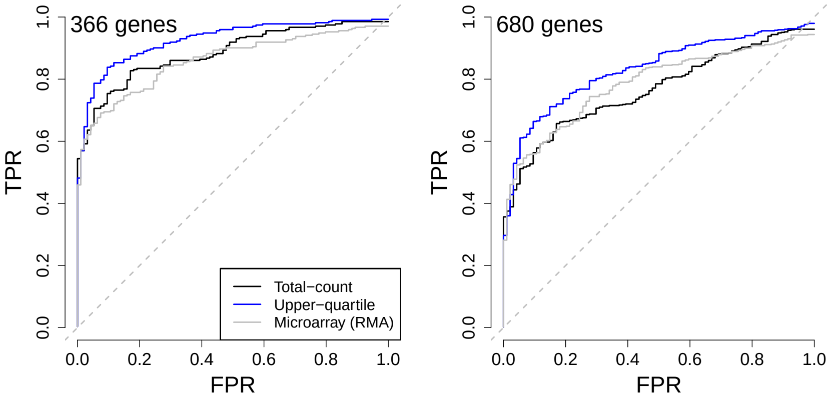


Figure 4: *Comparison of biological, library preparation, and flow-cell effects.* Boxplots of estimated log-fold-changes for UHR vs. Brain biological effects (GLM 2 in Table 4), flow-cell effects adjusting for biology (GLM 4), library preparation effects within flow-cell (GLM 7). Estimates are presented for total-count (black) and upper-quartile (blue) normalization.



(a) (b)

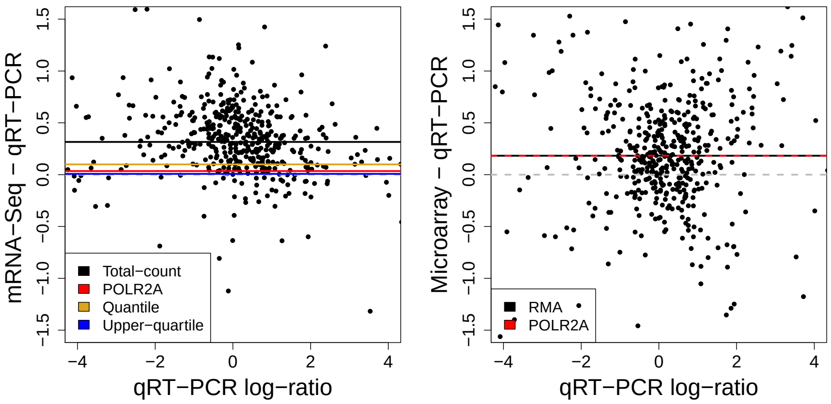
Figure 5: *Impact of highly-expressed genes.* (a) Cumulative percentage of total read count for Brain (green) and UHR (purple) samples, starting with the gene with the *highest* read count (across the seven Brain or UHR lanes). Cumulative read counts are marked for the 5, 10, 20, and 30 percent most highly expressed genes. (b) Running value of the UHR/Brain expression fold-change for unnormalized counts, starting with the gene with the *lowest* total count across all 14 lanes. Horizontal lines correspond to: the ratio of the counts for all genes (black), the ratio of the counts for the POLR2A gene (red), and the ratio of the per-lane upper-quartile of counts for genes with reads in at least one lane (blue).



(a) qRT-PCR positives: LR>2

(b) qRT-PCR positives: LR>0.5

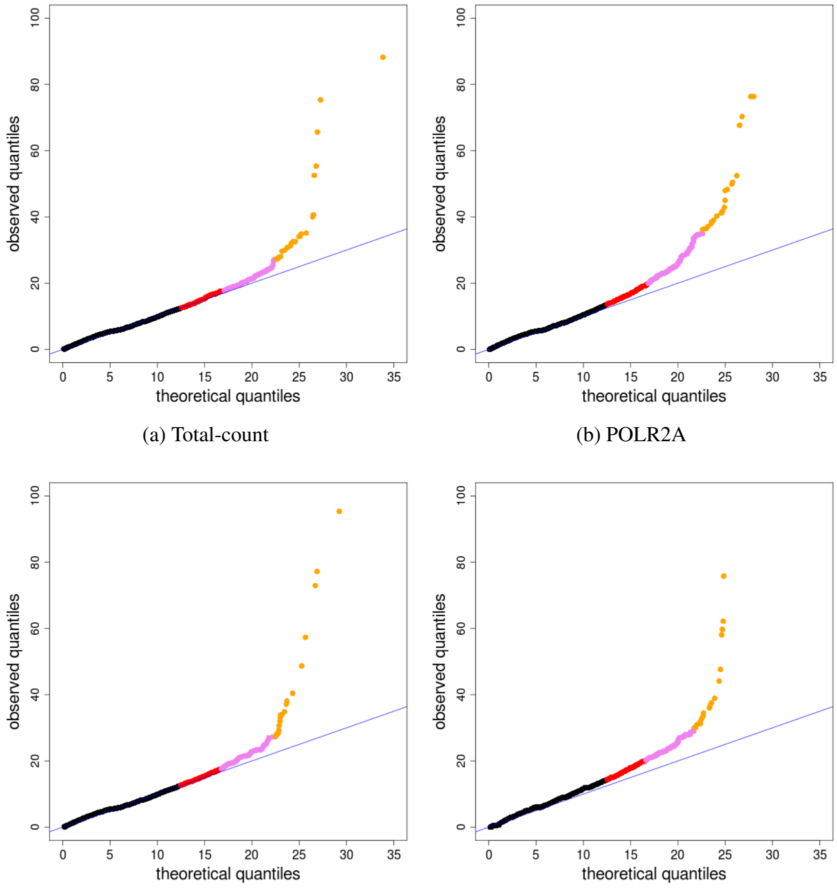
Figure 6: *Comparison of mRNA-Seq and microarray differential expression calls to qRT-PCR: ROC curves.* Genes common to all three platforms and present for both qRT-PCR and sequencing (see Supplementary Text) were evaluated and declared DE if their qRT-PCR absolute log-ratio was (a) greater than 2 or (b) greater than 0.5; genes were declared non-DE if their absolute log-ratio was less than 0.2. The GLM-based likelihood ratio test was used for the sequencing data. Two normalization procedures are presented for mRNA-Seq: total-count (black) and upper-quartile (blue) normalization. Microarray data were normalized using RMA (gray). Note that the ROC curves do not reach the point (1,1), because of the sign condition in the definition of true positives.



(a) Sequencing

(b) Microarray

Figure 7: *Comparison of mRNA-Seq and microarray differential expression calls to qRT-PCR: ROC curves.* Difference scatterplots comparing the estimates of UHR/Brain expression log-ratio from qRT-PCR to those from (a) mRNA-Seq, using the standard total-count normalization, and (b) microarrays, using the standard RMA normalization. Shown are the genes shared between all three platforms, present in both Brain and UHR according to both mRNA-Seq and qRT-PCR (see Supplementary Text), and having absolute qRT-PCR expression log-ratio less than 4. Horizontal lines in (a) represent the median UHR/Brain log-ratio for the sequencing data after the standard total-count normalization (black), POL2RA normalization (red), quantile normalization (yellow), upper-quartile normalization (blue); horizontal lines in (b) show the median UHR/Brain log-ratio for the microarray data after the standard RMA normalization (black) and POL2RA normalization (red).



(a) Total-count

(b) POLR2A

(c) Upper-quantile

(d) Quantile normalization

Figure 8: *Comparison of normalization procedures: Goodness-of-fit of Poisson model.* The multiplicative Poisson model (GLM 1 in Table 4) is fit to the seven Brain lanes in the MAQC-2 experiment after (a) total-count, (b) POLR2A, (c) upper-quartile, and (d) quantile normalization. Goodness-of-fit statistics are computed and displayed in χ^2 quantile-quantile plots. Genes with goodness-of-fit statistics in the top quantiles of the χ^2-distribution are displayed using colored plotting symbols: red $(1, 5]\%$, purple $(.1, 1]\%$, gold $[0, .1]\%$. Similar plots for UHR show the same patterns.

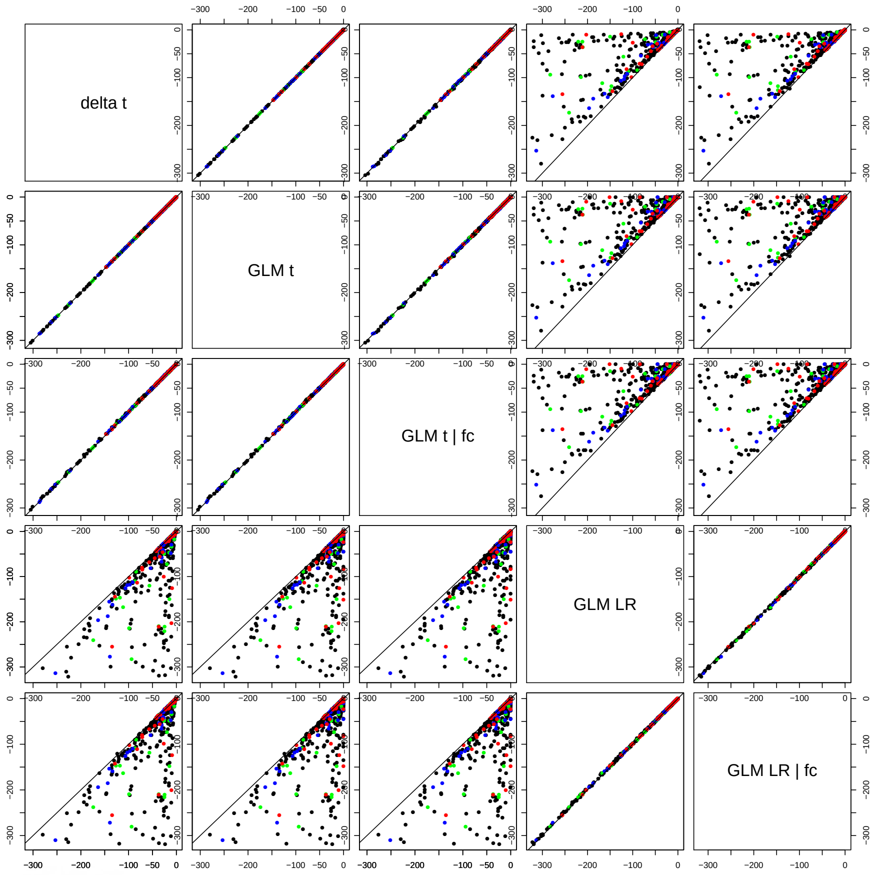


Figure 9: *Comparison of GLM-based DE statistics.* Scatterplot matrix of nominal p-values on the log-scale for differential expression statistics for genes assayed by both mRNA-Seq and qRT-PCR. Genes with zero p-values are not displayed. Plotting symbols are colored according to the read counts of the corresponding gene in the Brain and UHR samples. **Black**: ≥ 6 reads for both Brain and UHR; **Green**: ≥ 6 reads for Brain, < 6 reads for UHR; **Blue**: < 6 reads for Brain, ≥ 6 reads for UHR; **Red**:< 6 reads for both Brain and UHR.

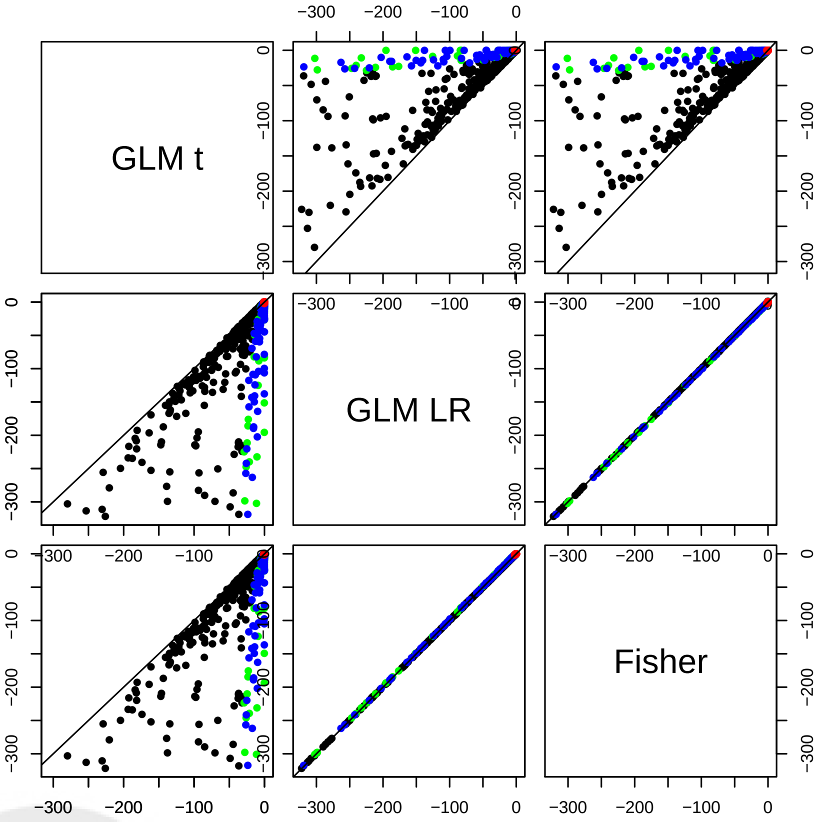
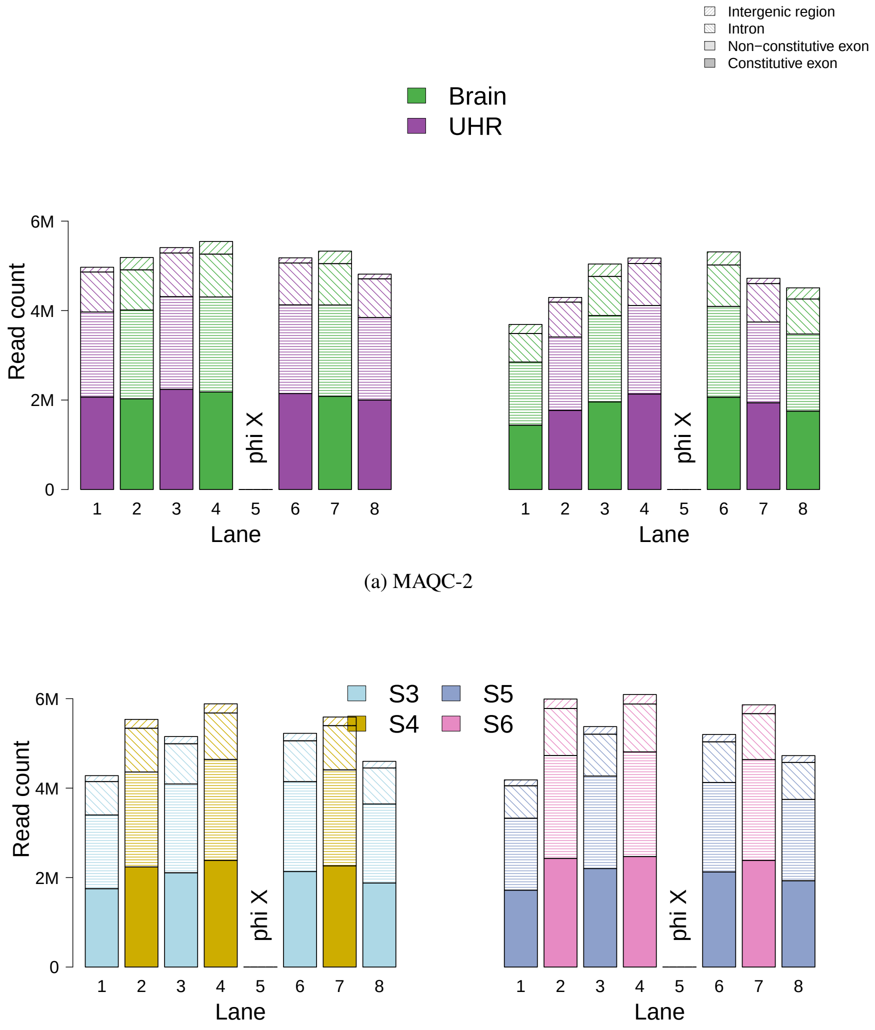


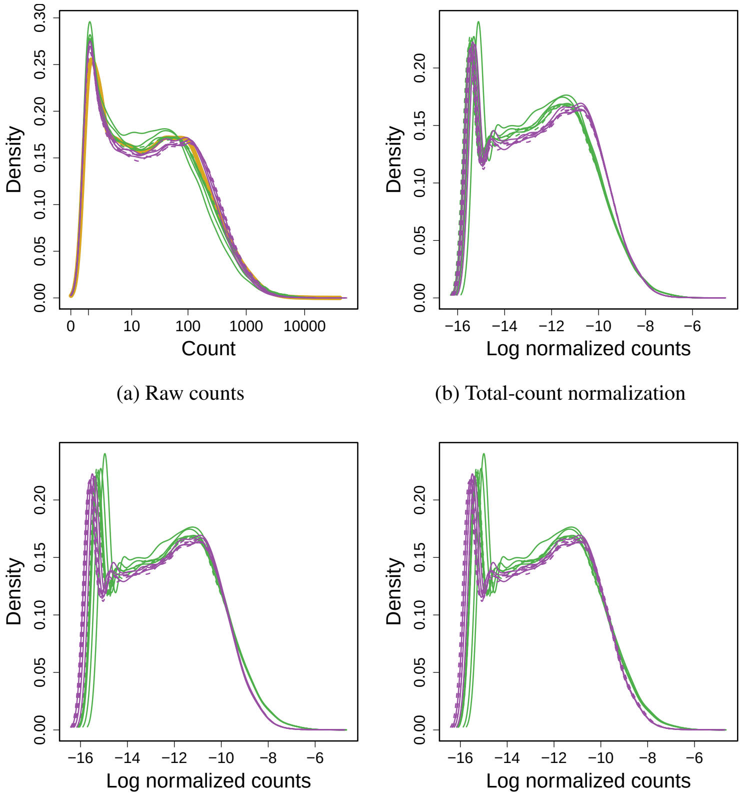
Figure 10: *Comparison of GLM-based and Fisher DE statistics.* Scatterplot matrix of nominal p-values on the log-scale for differential expression statistics for genes assayed by both mRNA-Seq and qRT-PCR. Plotting symbols are as described in the caption of Figure 9.



(a) MAQC-2

(b) MAQC-3

Figure 11: *Experimental design and per-lane read counts.* Barplots of per-lane read counts for: (a) MAQC-2 experiment and (b) MAQC-3 experiment. There are fourteen lanes in each experiment. The MAQC-2 experiment assayed single library preparations of Brain and UHR RNA, each using seven lanes distributed across two flow-cells. The MAQC-3 experiment assayed four different library preparations of UHR RNA, each across 3-4 lanes within a single flow-cell. The fifth lane in each flow-cell was reserved for phi X genomic DNA. Only purity-filtered reads that map uniquely to the genome with up to two mismatches in the first 35 bases (FMM) are retained. Total lane counts are partitioned into read counts for introns, constitutive exons, non-constitutive exons, and intergenic regions.



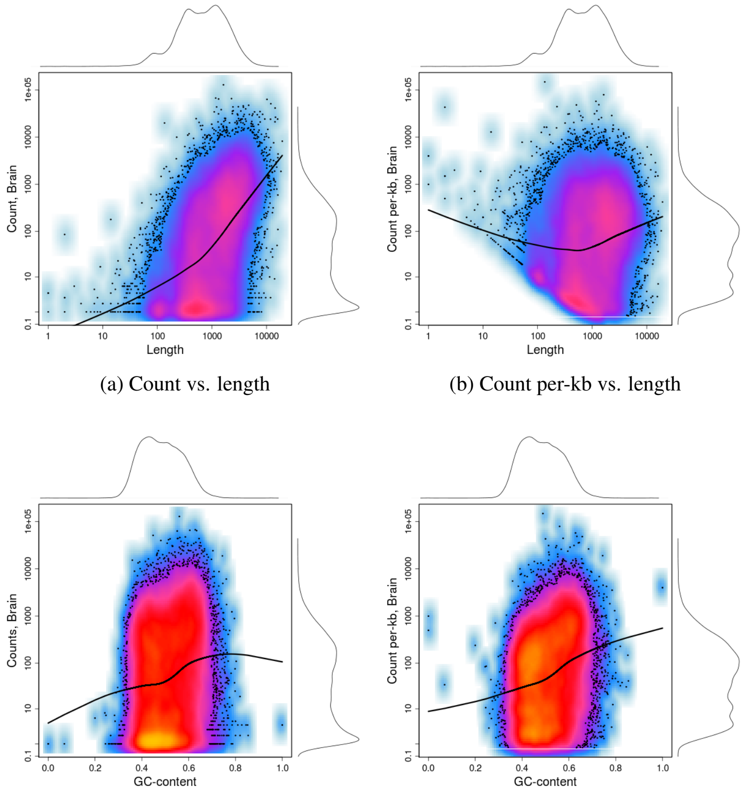
(a) Raw counts

(b) Total-count normalization

(c) Upper-quartile normalization

(d) POLR2A normalization

Figure 12: *Distribution of UI gene counts.* (a) Raw counts, (b) log counts after total-count normalization, (c) log counts after upper-quartile normalization, (d) log counts after POLR2A normalization. The density of quantile-normalized counts is the same for each lane and shown in yellow in panel (a). Brain samples are shown in green and UHR samples are shown in purple.



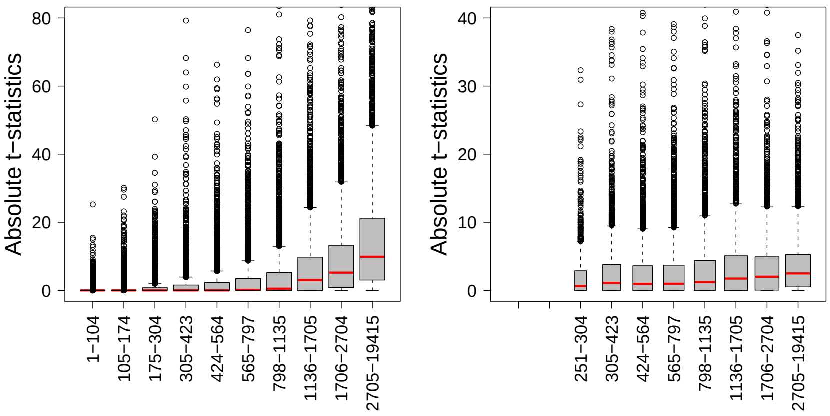
(a) Count vs. length

(b) Count per-kb vs. length

(c) Count vs. GC-content

(d) Count per-kb vs. GC-content

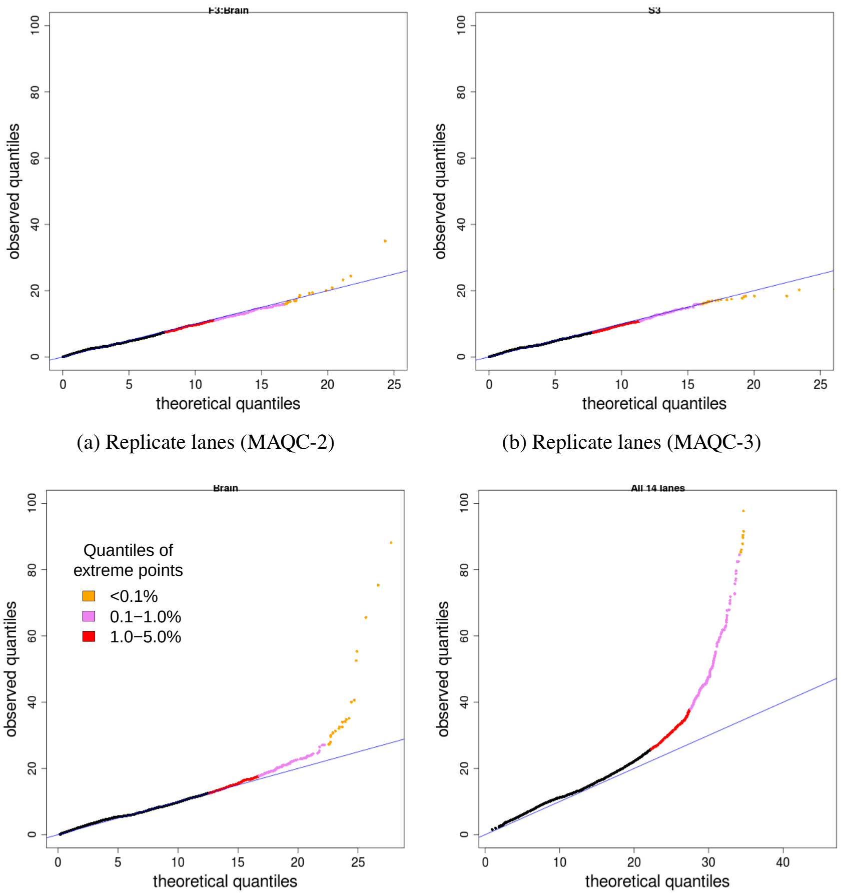
Figure 13: *Distribution of UI gene counts for Brain, by length and GC-content.* For each gene, total and per-kb read counts were summed over the seven Brain lanes (MAQC-2). Bivariate binned Gaussian kernel density smoothers are displayed for: (a) total gene count vs. gene length; (b) per-kb gene count vs. gene length; (c) total gene count vs. gene GC-content; (d) per-kb gene count vs. gene GC-content. Marginal Gaussian kernel density smoothers are displayed above and to the right of the plots for gene length/GC-content and gene count, respectively. The curves represent lowess fits. Only genes with non-zero read counts were included.



(a) Full-length UI genes

(b) 250-bp UI gene regions

Figure 14: *Differential expression statistics, by length.* Boxplots of absolute DE *t*-statistics (delta method) stratified by length for: (a) full-length genes and (b) a random sample of 250 base-pairs for each full-length gene longer than 250 base-pairs. The width of each boxplot is proportional to the number of genes within each length stratum.



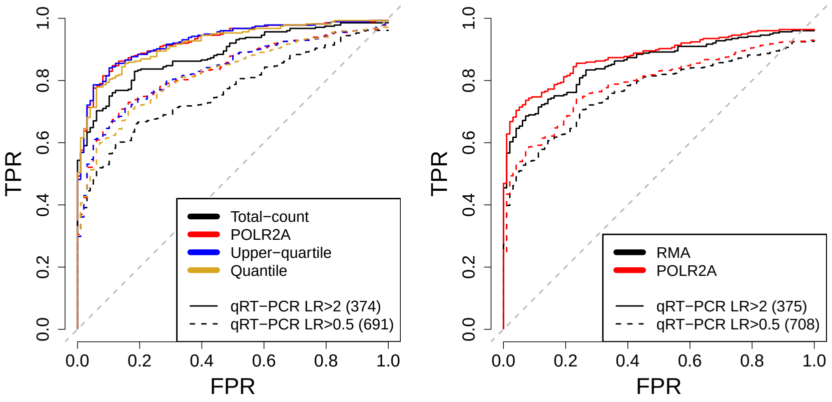
(a) Replicate lanes (MAQC-2)

(b) Replicate lanes (MAQC-3)

(c) Lanes across flow-cells (MAQC-2)

(d) Lanes across library prep and flow-cells (MAQC-3)

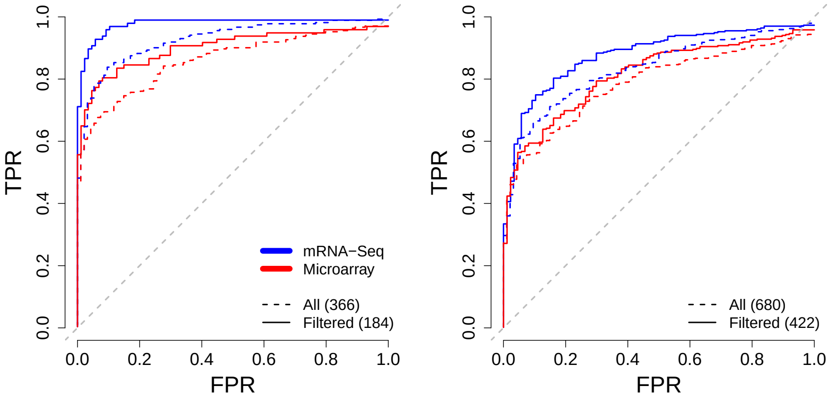
Figure 15: *Goodness-of-fit of gene-level multiplicative Poisson model across lanes, flow-cells, and library preparations.* The multiplicative Poisson model (GLM 1 in Table 4) is fit to the following sets of lanes representing different combinations of biological samples, library preparations, and flow-cells. Panel (a): Four replicate Brain lanes in flow-cell F3. Panel (b): Four replicate UHR lanes of library preparation S3 in flow-cell F4. Panel (c): Seven Brain lanes across flow-cells F2 and F3. Panel (d): Fourteen UHR lanes of four library preparations across flow-cells F4 and F5. Goodness-of-fit statistics are computed and displayed in χ^2 quantile-quantile plots. Genes with goodness-of-fit statistics in the top quantiles of the χ^2-distribution are displayed using colored plotting symbols as indicated in legend.



(a) mRNA-Seq

(b) Microarray

Figure 16: *Comparison of mRNA-Seq and microarray normalization procedures: ROC curves.* Panel (a): ROC curves comparing mRNA-Seq DE calls for total-count, POLR2A, upper-quartile, and quantile normalization. Panel (b): ROC curves comparing microarray DE calls for RMA and POLR2A normalization. Genes were declared DE if their qRT-PCR absolute log-ratio was greater than 2.0 (solid) or greater than 0.5 (dashed); genes were declared non-DE if their absolute log-ratio was less than 0.2.



(a) LR ≥ 2

(b) LR ≥ 0.5

Figure 17: *Comparison of mRNA-Seq and microarray DE calls: ROC curves.* ROC curves comparing upper-quartile normalized mRNA-Seq and RMA normalized microarray DE calls. Genes were declared DE if their qRT-PCR absolute log-ratio was (a) greater than 2 or (b) greater than 0.5; genes were declared non-DE if their absolute log-ratio was less 0.2. Considered were those genes common to both platforms and present in both qRT-PCR and sequencing (solid). Filtering is as before: genes with fewer than 20 reads in either Brain or UHR were excluded.

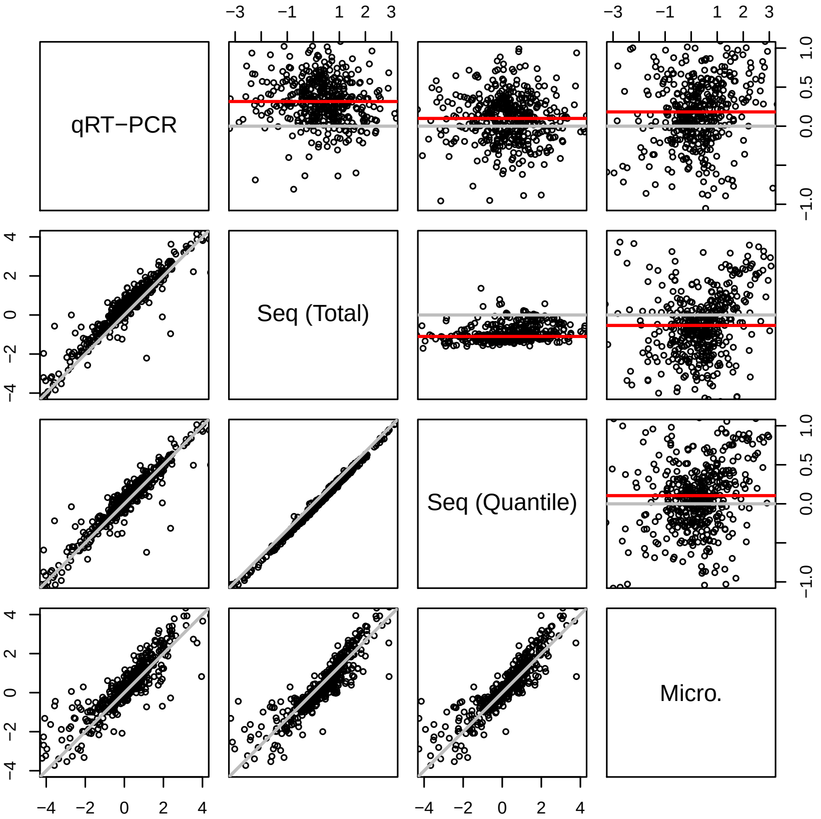


Figure 18: *Pairs plot, comparison of normalization.* In the lower diagonal are standard x-y plots, while in the upper-diagonal are MA plots of the difference x-y plotted against the average of x and y. No difference between the platforms is shown by a grey line; in the MA plots, the value of the median difference is shown as a red line. Note that difference between total counts and median normalization is just a shift of the log-ratio values and thus we show only the total-counts normalization.

5 Supplementary Text

5.1 Data

5.1.1 mRNA-Seq Data

The calibration method used in Bustard for quality-scoring of base-calls is highly relevant in terms of experimental design. In the auto-calibration method, base-calls are scored in a manner that is similar to the phred base-caller (Ewing and Green, 1998). An alternative, recommended by Illumina, is to reserve one control lane per flow-cell for sequencing DNA, typically bacteriophage phi X genomic DNA (Bentley et al., 2008).

Bustard also provides a variety of read quality measures. For a given cluster, the *chastity* c_k at cycle k is defined as the highest of the four fluorescence intensities divided by the sum of the highest two intensities. The *purity filter* (PF) discards any read for which the chastity at any of the first 12 sequencing cycles is less than 60%, i.e., $\min_{1 \leq k \leq 12} c_k < 0.60$ (Bentley et al., 2008, Supplementary Information, p. 6). For the MAQC-2 and MAQC-3 datasets, the percentage of reads passing the purity filter (out of the total number of clusters) varies between 50% and 76% per lane. Summaries of the Genome Analyzer output are provided in Tables 2 and 3.

We used Bowtie (Langmead et al., 2009), Version 0.10.1, to align reads to the genome (*H. sapiens*, NCBI 37.1 assembly). We used a strict alignment policy, which enforces a strong definition of uniqueness: a perfect match is a read that perfectly matches a position and does not match elsewhere, even when allowing up to two mismatches. In this regard, we minimize the chance that a perfect match read is a read with an error that happens to perfectly match elsewhere. The Bowtie command for implementing this mapping strategy is:

```
-r -v 2 -a -m 1 -p 8 --quiet h_sapiens_37_asm
```

Mapped reads were classified into the following three nested categories: (1) *purity-filtered perfect match* (FPM) reads, that passed the purity filter and mapped uniquely as described above; (2) *purity-filtered mismatch* (FMM) reads, that passed the purity filter and mapped with either 0, 1, or 2 mismatches; (3) *mismatch* (MM) reads, that mapped with either 0, 1, or 2 mismatches, regardless of purity filtering.

As a result of the above pre-processing steps, we therefore have six sets of mapped reads, corresponding to two calibration methods (auto-calibration and phi X calibration) and three mapping stringencies (FPM, FMM, and MM). In our main analysis, we focus on phi X-calibrated, purity-filtered reads that map uniquely to the genome, with up to two mismatches (FMM).

Note that, by mapping to the genome, we do not capture exon-exon junction reads, which would be relevant in studies of alternative splicing. In any given lane, around 10% of the reads mapped to exon-exon junctions. Additionally, the library preparation protocol does not allow consideration of strand-specific counts, i.e., reads mapping to the forward and reverse strands are pooled.

Table 2: *MAQC-2: Pre-processing summary.* The table reports summaries from Illumina's standard Genome Analyzer pre-processing pipeline: Firecrest image analysis and Bustard base-calling (Bentley et al., 2008). "Yield (kb)": Product of number of purity-filtered clusters and number of bases per cluster (per lane). "Raw clusters": Average ± standard deviation of per-tile number of clusters detected by the image analysis module of the pipeline. "PF clusters": Average ± standard deviation of per-tile number of detected clusters that meet the purity filtering criterion. Note that the fifth lane in each flow-cell was reserved for sequencing phi X genomic DNA.

Flow-cell	Lane	Biology	Yield (kb)	Raw clusters	PF clusters
F2	L1	UHR	296866	128513 ± 8346	88353 ± 9043
F2	L2	Brain	277172	113931 ± 13407	81641 ± 13201
F2	L3	UHR	324216	134627 ± 9441	92633 ± 8700
F2	L4	Brain	294120	112663 ± 6475	84883 ± 5541
F2	L6	UHR	310230	131166 ± 8986	88637 ± 9422
F2	L7	Brain	283315	113651 ± 7401	80947 ± 8981
F2	L8	UHR	287474	122293 ± 12000	82135 ± 10298
F3	L1	Brain	203301	117128 ± 6695	58086 ± 16486
F3	L2	UHR	260693	135475 ± 7102	74483 ± 14256
F3	L3	Brain	273610	118160 ± 6825	78174 ± 10553
F3	L4	UHR	313353	136806 ± 7869	89529 ± 9365
F3	L6	Brain	288766	120813 ± 7309	82504 ± 9424
F3	L7	UHR	288312	136649 ± 7037	82374 ± 11148
F3	L8	Brain	243072	116163 ± 6596	69449 ± 10014

Table 3: *MAQC-3: Pre-processing summary.* Cf. Table 2 caption.

Flow-cell	Lane	Lib. Prep.	Yield (kb)	Raw clusters	PF clusters
F4	L1	S3	258032	114634 ± 10132	79272 ± 13356
F4	L2	S4	334535	144311 ± 11288	96547 ± 10952
F4	L3	S3	311489	120708 ± 9614	88997 ± 9017
F4	L4	S4	354932	141855 ± 11771	101409 ± 8263
F4	L6	S3	316489	119383 ± 8989	90425 ± 7053
F4	L7	S4	336469	140959 ± 10373	97105 ± 7050
F4	L8	S3	278196	113575 ± 8529	79484 ± 8161
F5	L1	S5	251885	118150 ± 7715	81780 ± 8955
F5	L2	S6	364813	162401 ± 11096	105285 ± 4996
F5	L3	S5	324904	124526 ± 7446	92829 ± 3300
F5	L4	S6	371195	158799 ± 11354	106055 ± 3718
F5	L6	S5	314057	118976 ± 6562	89730 ± 2854
F5	L7	S6	357717	157457 ± 9475	102204 ± 4237
F5	L8	S5	288629	122585 ± 10022	82465 ± 8166

5.1.2 qRT-PCR Data

For benchmarking purposes, we use the *quantitative real-time polymerase chain reaction* (qRT-PCR) data of Canales et al. (2006) to obtain distinct measures of gene expression (Gene Expression Omnibus (GEO), Series GSE5350, `www.ncbi.nlm.nih.gov/geo`). In this TaqMan assay, a quantitative measure of template abundance is provided by the *threshold cycle* (C_T), i.e., the number of PCR cycles at which one detects a significant exponential increase in the fluorescence of a labeled TaqMan probe. The greater the threshold cycle, the less abundant the template.

As described in Canales et al. (2006, p. 1120–1121), for each of 997 protein-coding genes, between four (994 genes) and eight (3 genes) C_T measures were obtained for each of Brain and UHR. Due to annotation differences, of the 997 genes assayed by qRT-PCR, 965 matched a unique UI gene. We find there is no systematic relationship between gene expression measures and mapping status.

Following Canales et al. (2006), a detection limit of 35 was set on the raw C_T values. For each type of biological sample (Brain and UHR), genes were further classified as present (P) if they were detectable in at least three fourths of the qRT-PCR assays and absent (A) otherwise. According to this criterion, 797 genes were declared present in both Brain and UHR samples, 26 present in only Brain samples, 76 present in only UHR samples, and 40 absent in both types of samples. The C_T measures available from GEO were normalized as in Canales et al. (2006), separately for the Brain and UHR samples, using the POLR2A gene as a reference.

In what follows, the qRT-PCR expression measures are represented as

$$Y_{i,j} \equiv \Delta C_{i,j} \times \log 2, \tag{2}$$

where $\Delta C_{i,j} = C_{i,POLR2A} - C_{i,j}$ are POLR2A-normalized threshold cycles C_T for protein-coding genes $j = 1, \ldots, 939$, in TaqMan assays $i = 1, \ldots, n_j$ ($n_j = 8$ for all but three genes that have $n_j = 16$). The qRT-PCR measures are originally on a log base-2 scale. Multiplication by $\log 2$ transforms these measures to the natural logarithmic scale used throughout. The qRT-PCR estimate of UHR to Brain expression log-fold-change is the difference of averages: $\bar{Y}_{UHR,j} - \bar{Y}_{Brain,j}$.

5.1.3 Affymetrix Microarray Data

Affymetrix microarray data were downloaded from GEO:
(GSE5350, MAQC_AFX_123456_120CELs.zip). To minimize variation across labs, we used data only from lab 1, i.e., AFX_1_[A—B][1-5].CEL$. Arrays were preprocessed using RMA (Irizarry et al., 2003b) and then differential expression was determined by the R/Bioconductor package limma (Smyth, 2004), using the standard pipeline of *lmFit* and *eBayes*.

In order to match Affymetrix probesets with our UI genes, we used the R/Bioconductor package biomaRt, which retrieves data from Ensembl, Version 55. In cases where multiple probesets matched to a single UI gene, we took the median measurement for the log-ratio, standard errors, and p-values, so as not exclude a large fraction of the microarray data.

5.2 Defining Genomic Regions of Interest

Using Ensembl, Version 55, annotation, we define a *union-intersection (UI) gene* as a composite gene consisting of unions of constitutive exons that do not overlap a coding region of another gene. Specifically, for a given gene, a *constitutive exon* is defined as a set of consecutive exonic bases (i.e., portion of or entire exon) that belong to each isoform of the gene of interest. We further exclude any portion of such region that overlaps the coding region of any other gene, either constitutive or alternative, on either strand (Figure 19a). A gene model defined according to this union-intersection principle can be viewed as representing all isoforms of a given gene. Reads are assigned to a given gene if their 5'-end falls within the region, as depicted in Figure 19b.

Figure 20 examines basic features of the set of UI genes considered in the present article and built using gene annotation from Ensembl (`www.ensembl.org`, Version 55). Figure 21 displays an example gene and its base-level read counts.

We also define an Ensembl gene as the union of all exons from a given gene, excluding regions which overlap any other gene on either strand. Our definition of UI genes is clearly more restrictive than that of Ensembl genes, as it retains only constitutive exons. The genome coverage of UI genes is 42,708,318 base-pairs, whereas the coverage of Ensembl genes is 82,020,267 base-pairs.

We call an Ensembl gene or a UI gene present, if it has at least one read in both Brain and UHR samples. Filtered genes are defined as having at least 20 reads in both samples.

5.3 Generalized Linear Models for Gene-level Counts

Consider J genes and let $X_{i,j}$ denote the number of reads mapping to gene j in lane i. Sums of counts over all lanes or genes are represented with the standard "·" symbol, e.g., $X_{i,\cdot}$ denotes total counts in lane i.

Generalized linear models (GLM) provide a flexible and extensible statistical inference framework for mRNA-Seq. Though we focus on models with the *log/Poisson link* function (McCullagh and Nelder, 1989), what follows may be applied using alternative link functions and distributions, such as the negative binomial.

We formulate a gene-level GLM for read counts $X_{i,j}$, such that $\log(\mathrm{E}[X_{i,j}]) = \lambda_{a(i),j} + \theta_{i,j}$, where $a(i) \in \{1, \ldots, A\}$ is the biological group (e.g., Brain or UHR) corresponding to lane i, $\lambda_{a(i),j}$ is the parameter of interest representing the expression level of gene j in biological group $a(i)$, and $\theta_{i,j}$ is a nuisance parameter representing experimental effects, such as library preparation, flow-cell, and lane effects. Suitable identifiability constraints need to be specified for each experimental design under consideration, e.g., $\sum_j \exp(\lambda_{a,j}) = 1$ for each biological group a.

Ultimately, the parameter of interest is the ratio of transcript counts in biological group a_2 vs. a_1, i.e., a transcript expression fold-change. In the mRNA-Seq assay, each transcript is divided into a number of fragments. As a result, the parameter $\exp(\lambda_{a_2,j} - \lambda_{a_1,j})$ represents the ratio of the number of fragments for biological group a_2 vs. a_1. Under certain assumptions for the library preparation process (concerning fragmentation, in particular), it can be argued that transcript and fragment fold-changes are proportional, with a single proportionality constant across all genes.

It is clear any reasonable model must normalize read counts to adjust for the large differences in sequencing depths between lanes (or samples). This can be achieved by introducing a lane-level parameter, δ_i, in the GLM,

$$\log(\mathrm{E}[X_{i,j}]) = \delta_i + \lambda_{a(i),j} + \theta_{i,j}. \tag{3}$$

Instead of fitting the above GLM jointly to all J genes (in the tens of thousands), it is equivalent to fit the following log-linear regression model per gene,

$$\log(\mathrm{E}[X_{i,j}|d_i]) = \log d_i + \lambda_{a(i),j} + \theta_{i,j}, \tag{4}$$

where d_i is a lane-level random variable, such as the total lane count $X_{i,\cdot}$, and the offset $\log d_i$ is to be treated as a quantitative covariate whose regression coefficient is set to one.

To evaluate the presence of experimental effects, we fit the model of Equation (1) with different choices of $\theta_{i,j}$ that account for groupings of lanes into flow-cells (fc) or library preparations (prep), as well as interactions of these effects with biological (bio) effects, where appropriate (see Table 4). We use likelihood ratio statistics per gene to compare the fits of models with various combinations of effects, e.g., (1+bio+fc) vs. (1+bio). We also use χ^2 goodness-of-fit statistics to assess deviation from a particular null model – again per gene (such an approach was also applied in Marioni et al. (2008), to assess goodness-of-fit for a particular model of inter-lane variation).

We note that by estimating a large number of parameters we risk overfitting and introducing noise into our estimators. In this instance, however, our goal is an overall assessment of experimental effects. To correct for such effects in practice and obtain reliable estimators of $\lambda_{a,j}$ for specific genes, more sophisticated approaches may be appropriate, such as pooling data across genes using empirical Bayes methods.

5.4 Normalization

The total-count, upper-quartile, and POLR2A normalization procedures involve a choice of a global scaling factor d, which is a vector of length equal to the number of lanes (14 in each of the two MAQC datasets). The offset d is incorporated in GLM-based tests (i.e., the LLR and the t-statistics) as described below. We note that the total-count offset corresponds to the maximum likelihood estimator of the parameter δ_i in the GLM of Equation (3).

We rescale the offset vector d so that the sum of its elements is equal to the total count across all lanes (roughly 67 million). This is done solely for the purpose of comparing normalizing factors and only affects Fisher's exact test, for which the distribution of the test statistic depends on the actual magnitude of d. The GLM-based tests are unaffected, because differences in the overall magnitude of the normalizing scale factor can be absorbed into an intercept term.[1] In our implementation of Fisher's exact test, d is acting as observed data even though it is not (except in the case of total-count normalization). GLM clearly give a more logical framework for allowing different choices of global normalization.

[1] The GLM of Equations (3) and (4) do not include an intercept term, but this is just a matter of reparameterization.

The final normalization considered is quantile normalization (Irizarry et al., 2003a), as implemented in the R package aroma.light (Bengtsson, 2009); the median across sorted lanes was chosen as the reference distribution. The normalized data are rounded to produce integer values that can be used with each of the DE statistics described below.

5.5 Differential Expression Statistics

Identifying genes that are differentially expressed between A conditions corresponds to testing the following J per-gene null hypotheses: $H_0(j) : \lambda_{1,j} = \cdots = \lambda_{A,j}$, where $\lambda_{a,j}$ is the expression level of gene j in samples of type a. We evaluate three main types of DE tests.

- **Log-likelihood ratio (LLR) statistics for GLM**:

  $$T_j^{LLR} = 2(\mathrm{l}_j(\hat{\lambda}, \hat{\theta}) - \mathrm{l}_j(\hat{\lambda}^0, \hat{\theta}^0)) \dot{\sim} \chi^2(A-1), \tag{5}$$

  where l_j denotes the log-likelihood function for the jth gene and $(\hat{\lambda}, \hat{\theta})$ and $(\hat{\lambda}^0, \hat{\theta}^0)$ denote, respectively, the maximum likelihood estimators (MLE) of the biological and experimental effect parameters under the full model and null model.

- **t-statistics for GLM (2 sample comparisons,** $A = 2$):

  $$T_j^t = \frac{(\hat{\lambda}_{2,j} - \hat{\lambda}_{1,j}) - 0}{\sqrt{\widehat{\mathrm{Var}}[\hat{\lambda}_{1,j}] + \widehat{\mathrm{Var}}[\hat{\lambda}_{2,j}]}} \dot{\sim} N(0,1), \tag{6}$$

  where the variances of $\hat{\lambda}$ may be estimated from (1) the standard GLM fitting procedure glm in R (**?**), e.g., based on an estimator of the information matrix obtained from the Hessian of the log-likelihood function, or (2) the delta method, where $\widehat{\mathrm{Var}}[\hat{\lambda}_{a,j}] = 1/\sum_i \mathrm{I}(a(i) = a) X_{i,j}$ (assuming $\theta_{i,j}$ constant across samples).

- **Fisher's exact test** is based on the $2 \times A$ contingency table created by cross-tabulating genes with biological sample type (Brain and UHR). The Mantel-Haenszel test of conditional independence within stratum extends Fisher's exact test to account for a single additional experimental effect (e.g., flow-cell). In all cases, except quantile normalization, row 1 corresponds to the total number of reads observed in the jth gene; in quantile normalization, it corresponds to the rounded quantile-normalized value. For total-count normalization, row 2 corresponds to the lane totals less the number of reads in the jth gene. In the case of POLR2A, upper-quartile, and quantile normalization, pseudo-counts are generated to match the total number of reads (see Normalization section, above). The tests are implemented using the fisher.test and mantelhaen.test functions in R.

5.6 Receiver Operator Characteristic Curves

Definition of true and false positive rates Given a “DE” (positive, P) or “non-DE” (negative, N) call from qRT-PCR, define a true positive (TP) as the event that the test of interest (based on either sequencing or microarray data) calls a gene DE that qRT-PCR called DE and that the direction of DE agrees between the two assays. Let a false positive (FP) event occur when the test calls a gene DE that qRT-PCR called non-DE (Table 1). We consider a true positive rate (TPR) defined as

$$\Pr(\text{TP}|\text{qRT-PCR is DE}) = \frac{\Pr(\text{TP}, \text{qRT-PCR is DE})}{\Pr(\text{qRT-PCR is DE})}$$

and estimated with

$$\frac{(\text{\# TP and qRT-PCR is DE})/(\text{total \# genes})}{(\text{\# qRT-PCR is DE})/(\text{total \# genes})} = \frac{\text{TP}}{\text{P}}.$$

Note that this is not the standard definition of TPR, usually expressed in terms of TP, FP, TN, and FN. We consider the standard definition of false positive rate (FPR),

$$\Pr(\text{FP}|\text{qRT-PCR is non-DE}),$$

estimated with

$$\frac{(\text{\# FP and qRT-PCR is non-DE})/(\text{total \# genes})}{(\text{\# qRT-PCR is non-DE})/(\text{total \# genes})} = \frac{\text{FP}}{\text{N}}.$$

Analysis of qRT-PCR data Conceivably, every gene could be declared differentially expressed at some cutoff, which means any “false positive” could be due either to noise or errors or to extremely high sensitivity of the (sequencing or microarray) platform. Furthermore, the qRT-PCR measures of DE are themselves imperfect, though generally accepted as the best available such measures – they have very low levels of variation and the variation is extremely uniform across genes. Rather than rely on the p-values from a test statistic for differential expression in qRT-PCR, we instead remove the 12 genes with standard errors greater than .25. In this manner, we focus on the more biologically relevant fold-change rather than the standard errors.

5.7 Experimental Effects: Lane, Flow-cell, and Library Preparation

We investigate various experimental effects for gene-level counts, including lane, flow-cell, and library preparation effects. For this, we rely on the total-count normalization, which gives the best results in terms of goodness-of-fit of the Poisson model for replicate lanes. Figure 22 displays mean-difference scatterplots of expression fold-changes vs. overall expression measures for lanes representing different combinations of flow-cells, library preparations, and biological groups (Brain and UHR). It is immediately clear that the magnitude of the differences between biological groups dwarfs any of the experimental effects. Mean-difference scatterplots of log-fold-change vs. overall expression are preferable to scatterplots of expression measures, as the latter often give a misleading impression of concordance between samples.

Replicate lanes Figures 15a and 15b show quantile-quantile (QQ) plots of χ^2 goodness-of-fit statistics for the multiplicative Poisson model fit within sets of replicate lanes for each UI gene (GLM 1, Table 4). Note that zero-read genes have undefined χ^2-statistics and are not plotted. Each QQ-plot is very close to the 0, 1 line; in particular, at worst only the top 0.1% of genes (and less than 10 genes for many of the sets of replicates) do not closely follow the null distribution – a remarkably good fit for non-simulated data. When goodness-of-fit is assessed without correcting for differences in total number of reads, the results unsurprisingly show lack-of-fit. Analogous QQ-plots stratified by read count for MAQC-3 (Figure 23) indicate that genes with a reasonable number of reads (average of 3 or more reads per lane) show excellent fit; genes with fewer reads exhibit poor fit. This discrepancy most likely results from the breakdown of the asymptotic χ^2 approximation.

Flow-cell and library preparation effects We assess whether different aspects of the experimental design (flow-cell, library preparation) influence our ability to estimate the biological effects of interest. In Figures 24c and 24d, we see that when we ignore flow-cell or library preparation designation, the QQ-plots demonstrate lack-of-fit as compared to similar plots for replicate lanes. In particular, flow-cell and library preparation QQ-plots show deviation for the top 1% and 5% of genes, respectively, whereas analogous plots for replicate lanes only show deviation in the top 0.1% (if at all). Explicitly adjusting for flow-cell or library preparation effects results in near linear QQ-plots.

Next, to assess the significance of technical effects compared to biological effects, we compare various parameterizations of the log-linear regression model using likelihood ratio statistics (Table 4). The count-stratified QQ-plots of Figure 25a demonstrate that globally, the most significant differences between models are related to biology, as opposed to flow-cell.

Figure 4 demonstrates that flow-cell effects are much smaller in magnitude than biological effects. Although a direct comparison of library preparation effects to flow-cell and biological effects is not possible (due to confounding and nesting in MAQC-2 and MAQC-3, respectively), the boxplots suggest that both technical effects are much smaller than biological effects.

In summary, the above analysis suggests that there are both flow-cell and library preparation effects, but of less significance and of smaller magnitude than biological effects. Ignoring flow-cell has only a minor impact in detecting extremely small biological differences; almost none when genes have greater than 3 reads/lane.

5.8 Phi X Calibration Analysis

In each flow-cell, one lane out of eight was reserved for sequencing bacteriophage phi X genomic DNA and used by Genome Analyzer's base-caller Bustard for base-calling and quality-scoring (Bentley et al., 2008, Supplementary Information, p. 7). This practice has important experimental design implications, in terms of sample size and balance. We used the MAQC-2 dataset to investigate the impact of the calibration method (phi X calibration vs. auto-calibration) at various levels of the analysis pipeline,

Table 4: *Log-linear regression models.* The following class of log-linear regression models are considered separately for each gene j: $\log(\mathrm{E}[X_{i,j}|X_{i,\cdot}]) = \log X_{i,\cdot} + \lambda_{a(i),j} + \theta_{i,j}$. Each row in the table corresponds to a different parameterization of the biological effect λ (bio) and experimental effect θ, to represent different combinations of biological, library preparation, and flow-cell effects. Specifically, library preparation (prep) and flow-cell (fc) effects are denoted, respectively, by $\beta_{b(i)}$ and $\gamma_{c(i)}$, where $a(i)$, $b(i)$, and $c(i)$ map lane i to its corresponding biological, library preparation, or flow-cell group, respectively. Recall that in MAQC-2, biological effects (λ) are confounded with library preparation effects (β), and in MAQC-3, library preparation effects (β) are nested within flow-cell effects (γ). The gene index j is omitted to simplify notation.

Dataset	Model	Formula	$\lambda_{a(i)}$	θ_i	# parameters	Constraints
MAQC-2	1	1	0	α	1	$\lambda_{Brain} = 0, \gamma_{F2} = 0$
	2	1 + bio	$\lambda_{a(i)}$	α	2	
	3	1 + fc	0	$\alpha + \gamma_{c(i)}$	2	
	4	1 + bio + fc	$\lambda_{a(i)}$	$\alpha + \gamma_{c(i)}$	3	
MAQC-3	5	1	0	α	1	$\beta_{S3} = 0, \gamma_{F4} = 0$
	6	1 + fc	0	$\alpha + \gamma_{c(i)}$	2	
	7	1 + fc:prep	0	$\alpha + \beta_{b(i)}$	4	

including base-calling, read-mapping, and (differential) expression inference.

5.8.1 Base-calling and Quality-scoring

We first examine the effect of the calibration method on base-calls by cycle and by lane (in base-calling, a cycle refers to a position in a read, here, from 1 to 35).

The pseudo-color image in Figure 27 illustrates that there is good overall agreement between phi X and non-phi X-calibrated reads (less than 3% discrepancy). However, the discrepancy rate between the base-calls for the two calibration methods varies between cycles (higher for later cycles) and between lanes and flow-cells (higher for flow-cell F3). Furthermore, Figure 28 shows that not all base substitutions are equally likely, with phi X calls of ‘C’ being more frequently assigned another base by auto-calibration and the ‘C’ to ‘G’ transversion being the most common substitution.

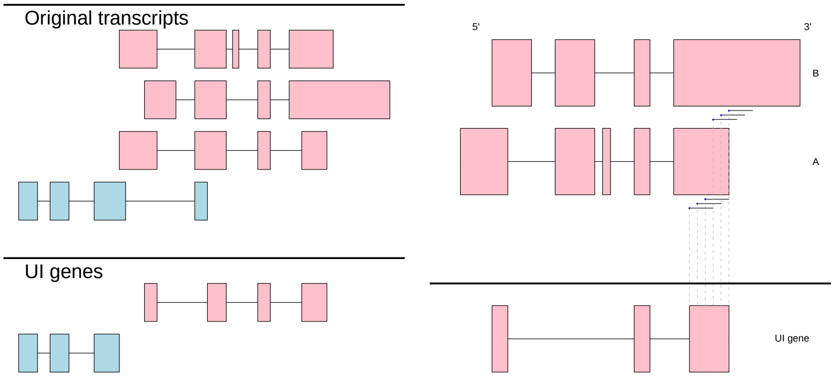
Overall, quality scores assessing the base-calls tend to be higher with auto-calibration. Figure 30a shows per-cycle quality scores for phi X and non-phi X-calibrated reads averaged across the seven lanes of each flow-cell. The quality scores for auto-calibration are generally higher at each cycle and, as previously noted, quality degrades through cycle (Bentley et al., 2008). Flow-cell F3 generally has lower quality scores and much steeper drops in quality for higher cycles. Additionally, Figure 30b, which shows the difference in quality scores by lane, demonstrates substantial variation in differences of quality scores between flow-cells and between lanes within flow-cells. The differences in base-calling quality scores between flow-cells F2 and F3 may explain the flow-cell effects reported earlier on downstream gene expression measures.

5.8.2 Absolute and Relative Expression Measures

Next, we consider the impact of the calibration method on (differential) expression statistics, based on purity-filtered perfect match (FPM) reads.

The significance of differences in estimates between the two calibration methods can be assessed by comparing observed differences to a permutation distribution of differences obtained by randomly swapping the auto-calibrated and phi X-calibrated sets of read counts for each of the 14 lanes. Such a permutation scheme respects the joint distribution of gene counts within lane and the experimental design (lane/flow-cell/library preparation/biological sample structure). The empirical cumulative distribution function (ECDF) and scatterplots of permutation p-values in Figure 31 suggest that, although small in magnitude, the differences in absolute expression measures are significant, especially for ROI with large read counts (Figure 31, Panels (a) and (c)). However, differences in expression fold-changes between UHR and Brain do not appear to be significant (Figure 31, Panels (b) and (d)).

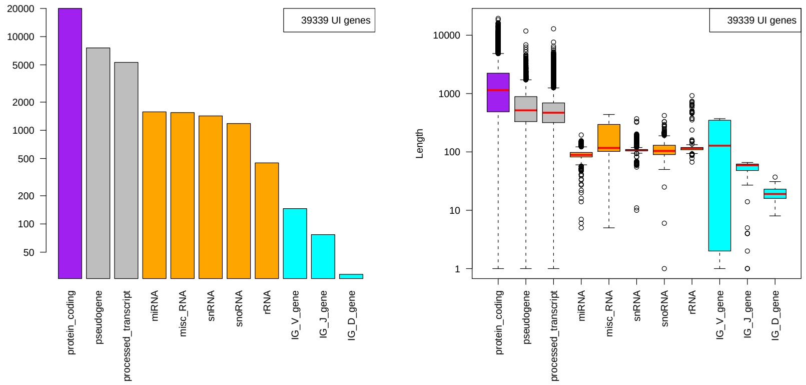
In summary, while there are some differences between phi X and auto-calibration in the early stages of the analysis pipeline, the differences in terms of differential expression are small. Unfortunately, we only have two flow-cells from which to assess the impact of auto-calibration vs. phi X calibration. However, it seems quite clear, using these two flow-cells, that auto-calibration is advantageous, as it yields more balanced designs, frees up one lane per flow-cell, and produces a larger number of higher quality reads per lane.



(a) UI vs. Ensembl genes

(b) UI genes and read-counting

Figure 19: *Union-intersection and Ensembl gene models.* Panel (a): Illustration of union-intersection (UI) and Ensembl gene definitions for two genes (pink and blue) with multiple isoforms (Section 2.4). The original transcripts, as would be reported by Ensembl, are displayed in the top panel. Below are the corresponding UI and Ensembl gene models. Note that because the genes overlap, the entire exon region is removed, not just the overlap. Panel (b): Illustration of read-counting for a gene with two isoforms. Isoform A has a shorter 3'-most exon as compared to Isoform B. The UI gene model includes the entire 3'-most exon for Isoform A. In addition to reads originating from the constitutive portion of the UI gene, reads emanating exclusively from Isoform B may also be counted.



(a) Ensembl annotation

(b) Length by Ensembl annotation

Figure 20: *UI gene Ensembl annotation.* Panel (a): Barplots of the distribution of UI genes by Ensembl annotation. Panel (b): Boxplots of UI gene lengths by Ensembl annotation. Ensembl annotation categories are sorted in decreasing order of their cardinalities; only categories comprising more than ten UI genes are displayed (22 out of 25 categories).

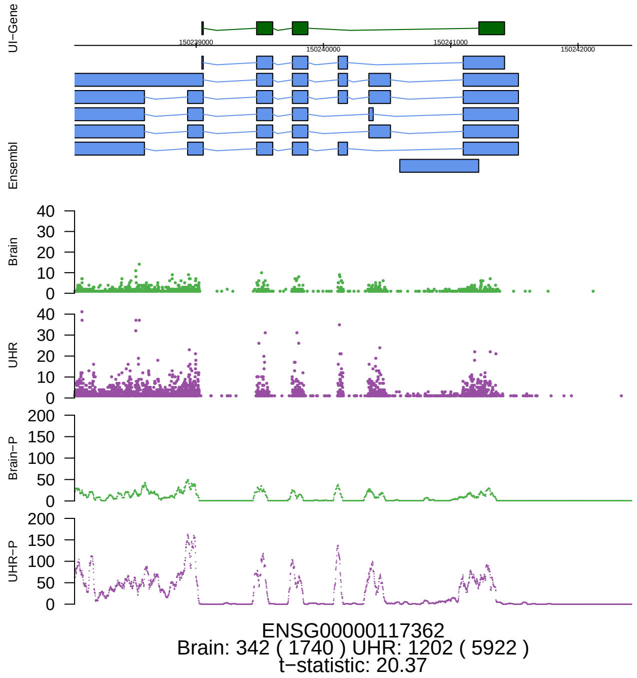
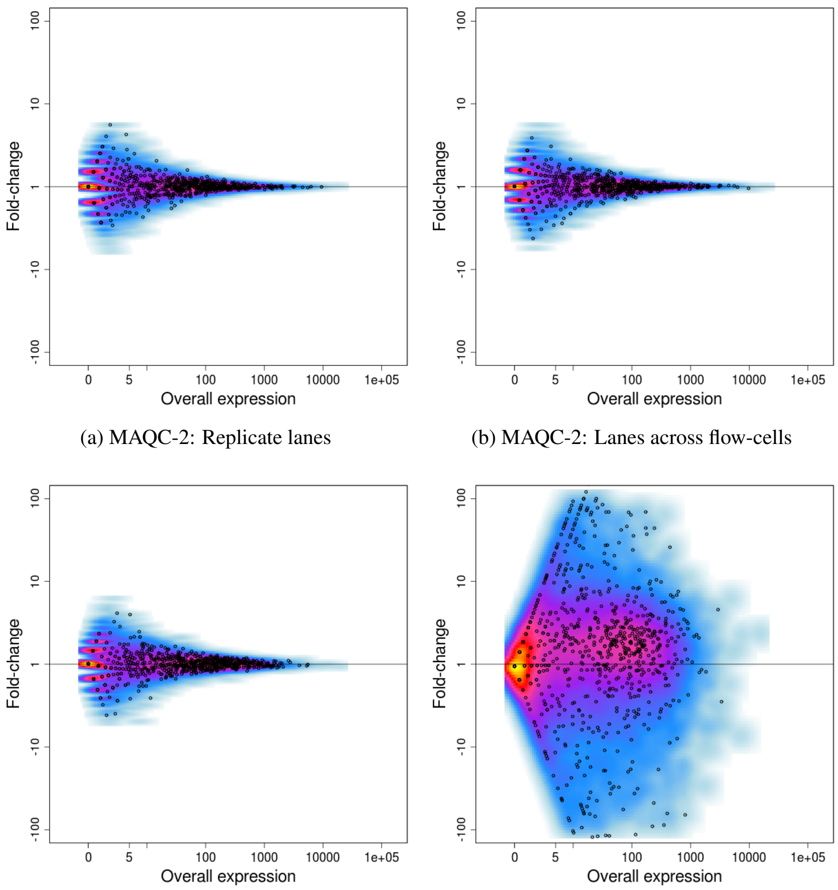


Figure 21: *Base-level read counts.* The plot provides two representations of base-level read counts summed across the seven Brain and seven UHR lanes for Ensembl Gene ENSG00000117362. Regions corresponding to the union-intersection gene model and Ensembl transcripts are indicated by dark green and light blue boxes, respectively. The top two read tracks (Brain, UHR) display numbers of reads with 5'-end at a given base (Section 2.4). UI gene counts for the Brain and UHR samples are reported below the tracks; Ensembl gene counts are in parentheses. The t-statistics for UHR vs. Brain differential expression are based on GLM adjusting for flow-cell effects (1+bio+fc, Table 4). The second set of tracks (Brain-P, UHR-P) correspond to a "pileup" representation of "overlap" counts, i.e., of numbers of reads overlapping a given base.



(a) MAQC-2: Replicate lanes

(b) MAQC-2: Lanes across flow-cells

(c) MAQC-3: Lanes across library prep.

(d) MAQC-2: Lanes across biological groups

Figure 22: *Mean-difference scatterplots of read counts across lanes, flow-cells, library preparations, and biological groups.* Scatterplots of expression fold-changes vs. overall expression measures for pairs of lanes representing different combinations of biological samples, library preparations, and flow-cells. Panel (a): Replicate Brain lanes in flow-cell F3. Panel (b): Brain lanes in flow-cell F3 vs. F2. Panel (c): UHR library preparation S4 vs. S3 lanes in flow-cell F4. Panel (d): UHR vs. Brain lanes in flow-cell F2. Only the qRT-PCR genes are individually plotted as a representative sample of genes; for comparison, these genes are plotted over the bivariate Gaussian kernel density smoothers of the MD-plots for all UI genes that contain reads in any lane of either the MAQC-2 or MAQC-3 datasets. Expression measures were normalized by total lane counts and then multiplied by 10^6 to make the scales commensurate when comparing different numbers of lanes.

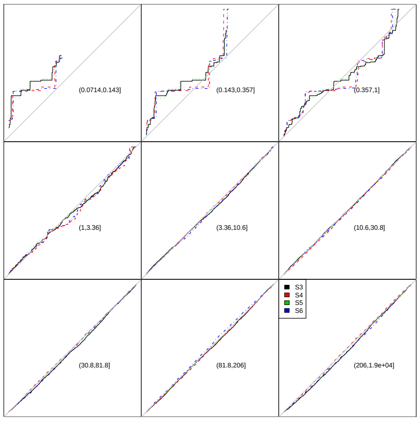
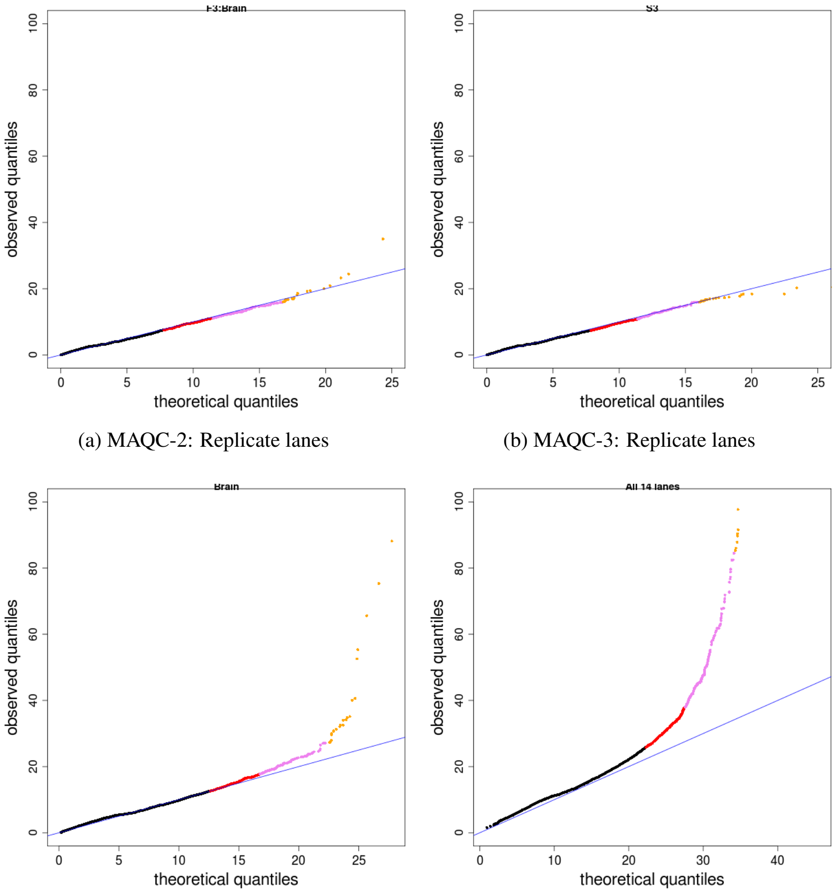


Figure 23: *MAQC-3: Goodness-of-fit of ROI-level Poisson model for replicate lanes, by count.* The multiplicative Poisson model of Equation (1) is fit to each UI gene within library preparation. Goodness-of-fit statistics are computed and displayed in uniform quantile-quantile plots for the corresponding nominal χ^2 p-values. The QQ-plots are stratified according to UI gene counts averaged over all fourteen lanes. The count strata partition the UI genes into nine groups of approximately the same cardinality, but vastly different count ranges.



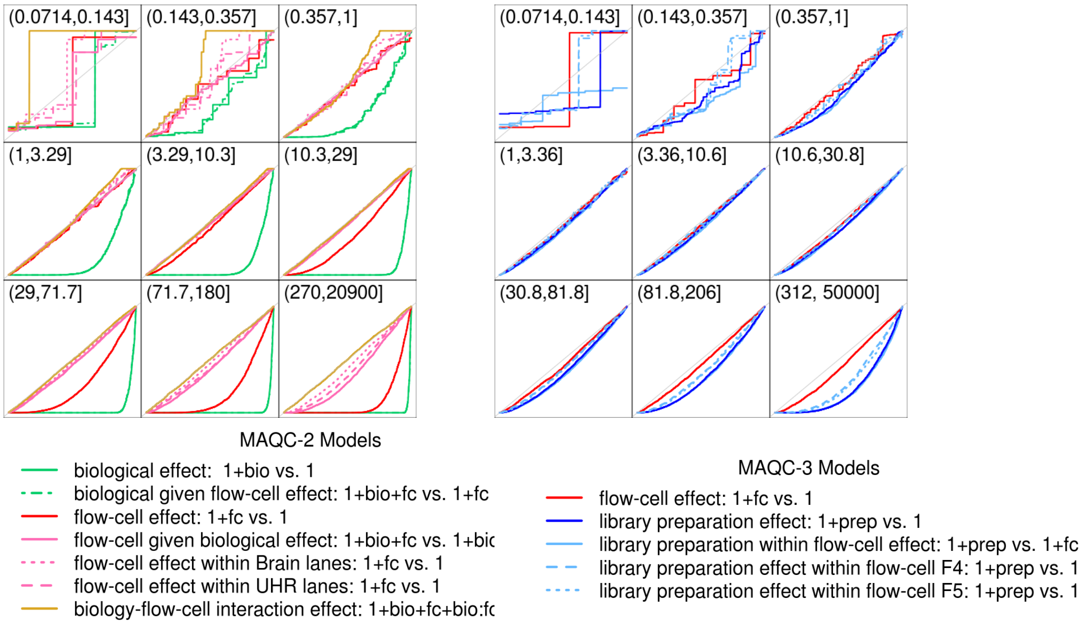
(a) MAQC-2: Replicate lanes

(b) MAQC-3: Replicate lanes

(c) MAQC-2: Lanes across flow-cells

(d) MAQC-3: Lanes across library prep.

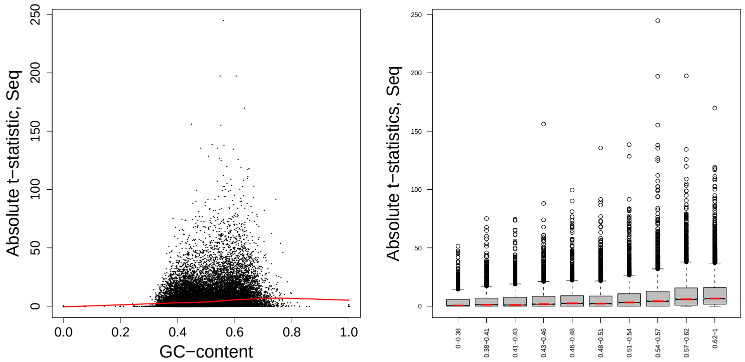
Figure 24: *MAQC-2 and MAQC-3: Goodness-of-fit of ROI-level multiplicative Poisson model across lanes, flow-cells, and library preparations.* The multiplicative Poisson model of Equation (1) is fit to the following sets of lanes representing different combinations of biological samples, library preparations, and flow-cells. Panel (a): Four replicate Brain lanes in flow-cell F3. Panel (b): Four replicate UHR lanes of library preparation S3 in flow-cell F4. Panel (c): Seven Brain lanes across flow-cells F2 and F3. Panel (d): Fourteen UHR lanes of four library preparations across flow-cells F4 and F5. Goodness-of-fit statistics are computed and displayed in χ^2 quantile-quantile plots. The top 5%, 1%, and 0.1% quantiles are indicated in red, violet, and orange, respectively.



(a) MAQC-2

(b) MAQC-3

Figure 25: *Count-stratified QQ-plots comparing the fit of log-linear regression models with various formulations of the biological and experimental effect parameters.* The log-linear regression model of Equation (1) is fit to each UI gene for various formulations of the biological and experimental effect parameters, λ and θ, respectively (Table 4). Models are compared with log-likelihood ratio statistics and the associated nominal χ^2 p-values are displayed in uniform quantile-quantile plots. The QQ-plots are stratified according to UI gene counts averaged over all fourteen lanes. The count strata partition the UI genes into nine groups of approximately the same cardinality, but vastly different count ranges.



(a) mRNA-Seq, UI genes

(b) mRNA-Seq, UI genes

Figure 26: *MAQC-2: Differential expression statistics, by GC-content.* Panel (a): Scatterplot of absolute mRNA-Seq DE statistics vs. GC-content for all UI genes. Panel (b): GC-content-stratified boxplots of absolute mRNA-Seq DE statistics for UI genes partitioned by GC-content into ten groups of approximately the same cardinality. For mRNA-Seq, the DE statistics are t-statistics for differences of biological effects $\lambda_{UHR,j} - \lambda_{Brain,j}$, based on GLM adjusting for flow-cell effects (1+bio+fc, Equations (1) and (6), Table 4).

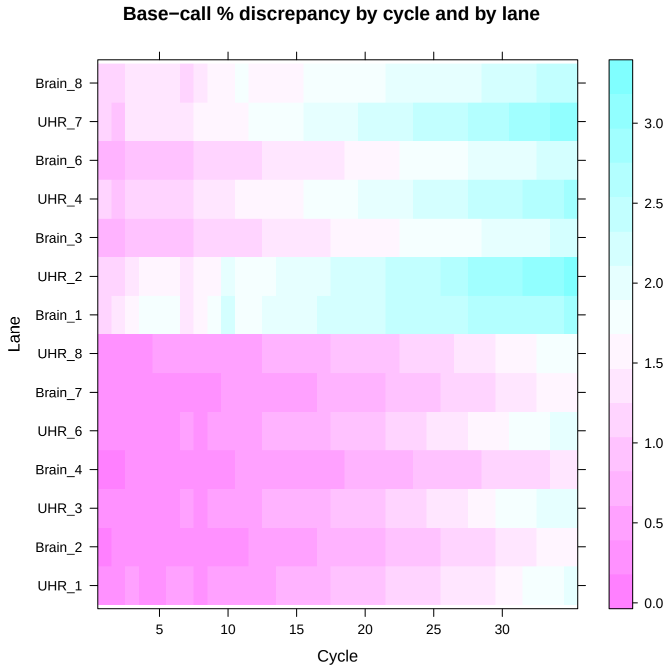


Figure 27: *MAQC-2: Impact of phi X calibration, base-calling.* Pseudo-color image of the per cycle and per lane percentage (out of 11,244,980–13,680,634 clusters per lane) of base-calling differences with and without phi X calibration.

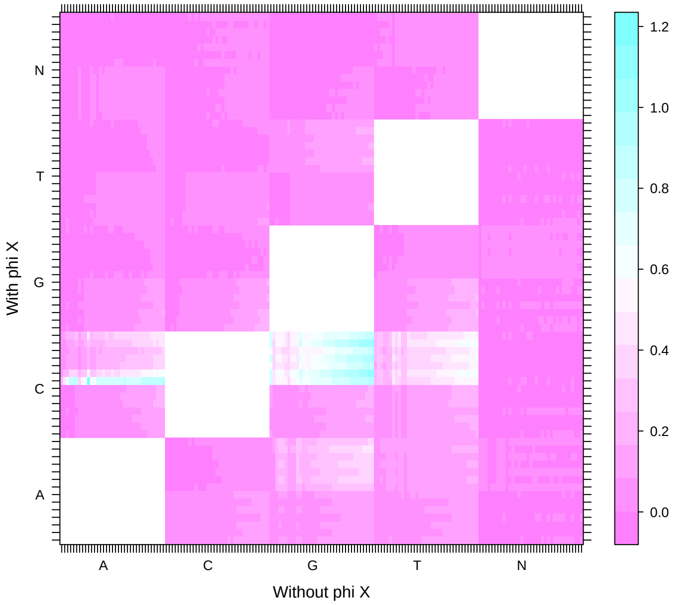


Figure 28: *MAQC-2: Impact of phi X calibration, base-calling.* Pseudo-color image of the per cycle and per lane joint distribution of base-calls with and without phi X calibration. Each cell in the image corresponds to the percentage (out of 11,244,980–13,680,634 clusters per lane) of base-call pairs of a given type, at a given cycle and in a given lane, e.g., an (A,C) pair corresponds to a base-call of 'A' with phi X and 'C' without phi X calibration. A base-call of 'N' is returned when all four fluorescence intensities are zero. Concordant base-calls are not displayed, as they dwarf discrepant calls.

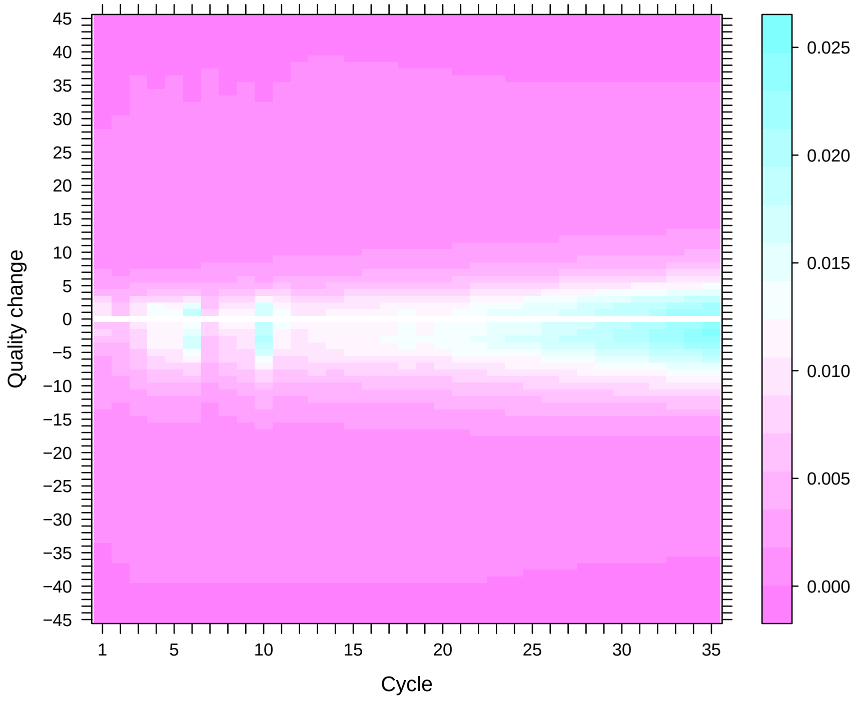
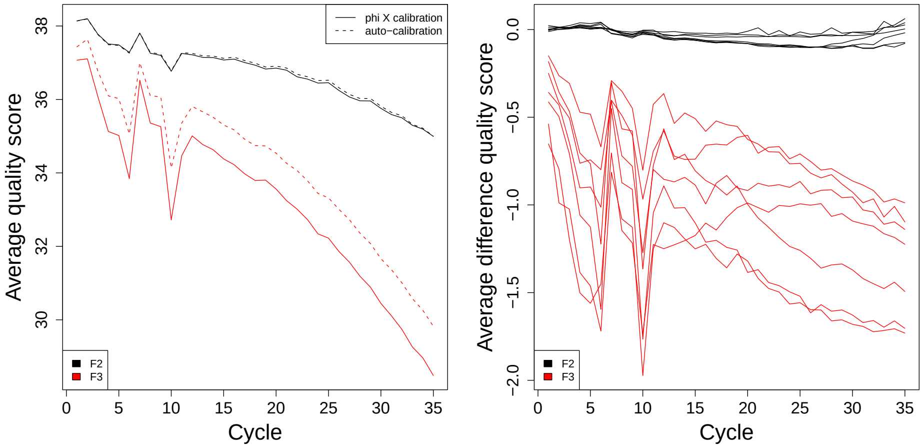


Figure 29: *MAQC-2: Impact of phi X calibration, quality-scoring.* Pseudo-color image of the per-cycle distribution of changes in quality scores with vs. without phi X calibration. The frequencies for equal quality scores are not displayed, as they consistently exceed 75%. Note that a quality change of 5 could correspond to an increase in quality from 35 with phi X calibration to 40 with auto-calibration, or an increase in quality from 5 to 10.



(a) All lanes combined

(b) By lane

Figure 30: *MAQC-2: Impact of phi X calibration, quality-scoring.* Plots of per-cycle average quality scores (out of 11,244,980–13,680,634 clusters per lane) with and without phi X calibration. Panel (a): Average quality scores are averaged across seven lanes for flow-cells F2 and F3. Panel (b): Average difference of quality scores between phi X calibration and auto-calibration for fourteen lanes.

References

Henrik Bengtsson. *aroma.light: Light-weight methods for normalization and visualization of microarray data using only basic R data types*, 2009. URL `http://www.braju.com/R/`. R package version 1.11.2.

D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

R. D. Canales, Y. Luo, J. C. Willey, B. Austermiller, C. C. Barbacioru, C. Boysen, K. Hunkapiller, R. V. Jensen, C. R. Knight, K. Y. Lee, Y. Ma, B. Maqsodi, A. Papallo, E. H. Peters, K. Poulter, P. L. Ruppel, R. R. Samaha, L. Shi, W. Yang, L. Zhang, and F. M. Goodsaid. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature Biotechnology*, 24(9):1115–1122, 2006.

D. Y. Chiang, G. Getz, D. B. Jaffe, M. J. T. O'Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, M. Meyerson, and E. S. Lander. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods*, 6(1):99–103, 2009.

J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105, 2008.

S. Durinck, J. Bullard, P. T. Spellman, and S. Dudoit. GenomeGraphs: integrated genomic data visualization with R. *BMC Bioinformatics*, 10(1):Article 2, 2009.

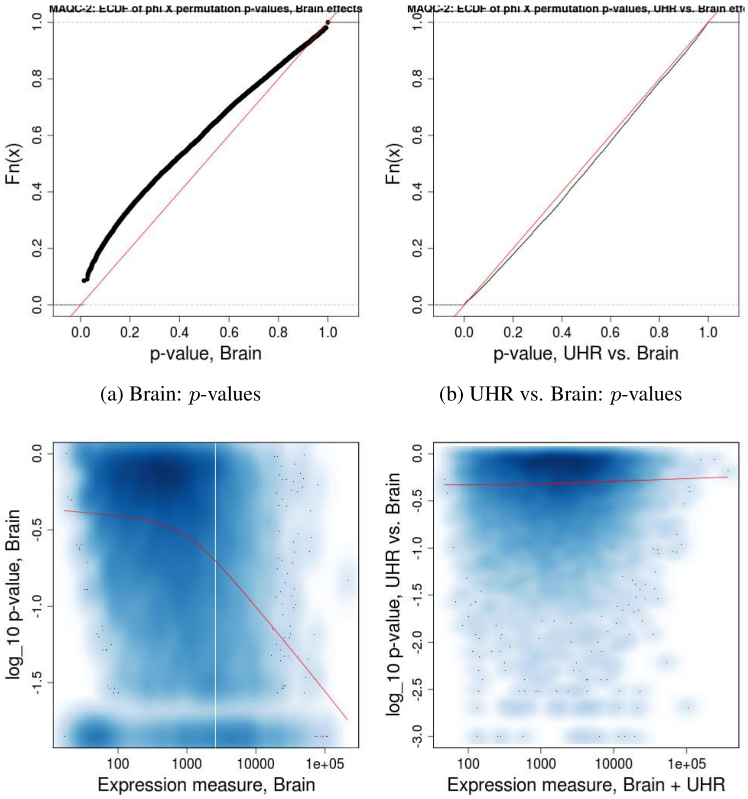
B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8(3):186–194, 1998.

P. A. C.'t Hoen, Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. A. M. Vossen, R. X. de Menezes, J. M. Boer, G.-J. B. van Ommen, and J. T. den Dunnen. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research*, 36(21):e141, 2008.

Illumina. *Sequencing Analysis Software User Guide For Pipeline Version 1.3 and CASAVA Version 1.0 T*. Illumina, Inc., December 2008. URL `icom.illumina.com/icom/software.ilmn?id=277`. Part # 1005359 Rev. A.

Illumina. *Preparing Samples for Sequencing mRNA*. Ilumina, Inc., 2009. URL `icom.illumina.com/icom/software.ilmn?id=277`. Part # 1004898 Rev. A.

R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003a.



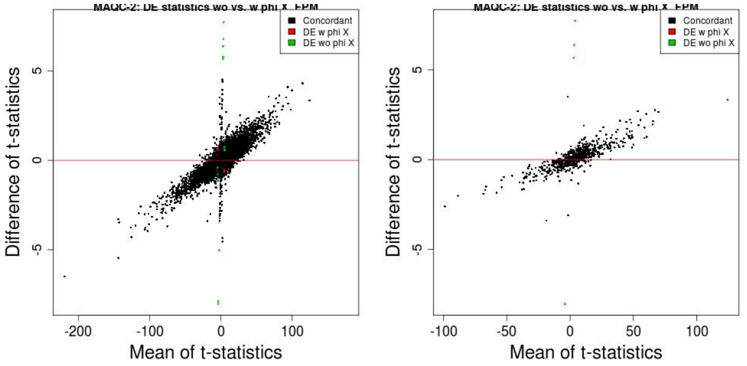
(a) Brain: p-values

(b) UHR vs. Brain: p-values

(c) Brain: p-values vs. counts

(d) UHR vs. Brain: p-values vs. counts

Figure 31: *MAQC-2: Impact of phi X calibration, biological effect estimation.* Panel (a): Empirical cumulative function of permutation p-values for differences in Brain effects $\hat{\lambda}_{Brain,j}$ without vs. with phi X calibration. Panel (b): Empirical cumulative distribution function of permutation p-values for differences in biology effects $\hat{\lambda}_{UHR,j} - \hat{\lambda}_{Brain,j}$, i.e., expression log-ratios, without vs. with phi X calibration. Panel (c) : Bivariate binned Gaussian kernel density smoother of permutation p-values for differences in Brain effects $\hat{\lambda}_{Brain,j}$ vs. read counts summed over the seven Brain lanes. Panel (d): Bivariate binned Gaussian kernel density smoother of permutation p-values for differences in biology effects $\hat{\lambda}_{UHR,j} - \hat{\lambda}_{Brain,j}$ vs. read counts summed over all fourteen lanes. Estimates of (absolute and relative) biological effects are based on GLM with only biological effects: $\hat{\lambda}_{a,j} = \log(X_{+a,j}/X_{+a,\cdot})$, $a \in \{Brain, UHR\}$, for the UI genes having non-zero counts with both types of calibration for each of the fourteen lanes. Two-sided p-values are computed based on $1,000$ random permutations of the phi X and non-phi X sets of read counts for each of the fourteen lanes (from the possible $2^{14} = 16,384$), with a floor of $2/1,000$.



(a) All UI genes

(b) qRT-PCR genes

Figure 32: *MAQC-2: Impact of phi X calibration, differential expression statistics – Purity-filtered perfectly matching reads (FPM).* Mean-difference scatterplots of DE statistics without vs. with phi X calibration. Panel (a): All UI genes. Panel (b): genes assayed by qRT-PCR. DE statistics are t-statistics for differences of biological effects $\lambda_{UHR,j} - \lambda_{Brain,j}$, based on GLM adjusting for flow-cell effects (1+bio+fc, Equations (1). Genes are declared differentially expressed if their nominal Bonferroni Gaussian adjusted p-values do not exceed 0.05. Discrepant DE calls are highlighted using red and green plotting symbols: red for DE according to phi X base-called lanes only and green for DE according to non-phi X base-called lanes only.

Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4 (1465-4644 (Print)):249–64, 2003b.

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 2009. (In press).

A. Lee, K. D. Hansen, J. Bullard, S. Dudoit, and G. Sherlock. Novel low abundance and transient RNAs in yeast revealed by tiling microarrays and ultra highthroughput sequencing are not conserved across closely related yeast species. *PLoS Genetics*, 4 (12):e1000299, 2008.

H. Li, M. T. Lovci, Y.-S. Kwon, M. G. Rosenfeld, X.-D. Fu, and G. W. Yeo. Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model. *PNAS*, 105(51):20179–20184, 2008.

Jun Lu, John K Tomfohr, and Thomas B Kepler. Identifying differential expression in multiple sage libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, 6:165, 2005.

MAQC Consortium. The MicroArray Quality Control (MAQC) project shows interandintraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, 2006.

J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 2008.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Number 37 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 2nd edition, 1989.

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008.

U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008.

A. Oshlack and M. J. Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4(14), 2009.

Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, Nov 2007.

Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3, 2004.

M. A. Taub. *Analysis of high-throughput biological data: some statistical problems in RNA-seq and mouse genotyping*. PhD thesis, Department of Statistics, UC Berkeley, 2009.

E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.