A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium

SEQC/MAQC-III Consortium*

We present primary results from the Sequencing Quality Control (SEQC) project, coordinated by the US Food and Drug Administration. Examining Illumina HiSeq, Life Technologies SOLiD and Roche 454 platforms at multiple laboratory sites using reference RNA samples with built-in controls, we assess RNA sequencing (RNA-seq) performance for junction discovery and differential expression profiling and compare it to microarray and quantitative PCR (qPCR) data using complementary metrics. At all sequencing depths, we discover unannotated exon-exon junctions, with >80% validated by qPCR. We find that measurements of relative expression are accurate and reproducible across sites and platforms if specific filters are used. In contrast, RNA-seq and microarrays do not provide accurate absolute measurements, and gene-specific biases are observed for all examined platforms, including qPCR. Measurement performance depends on the platform and data analysis pipeline, and variation is large for transcript-level profiling. The complete SEQC data sets, comprising >100 billion reads (10Tb), provide unique resources for evaluating RNA-seq analyses for clinical and regulatory settings.

Technological advances have made deep RNA sequencing feasible, expanding our view of the transcriptome1 and promising to permit quantitative profiling with large dynamic range2. Recent comparisons of RNA-seq with established technologies for differential expression analysis have found good overall agreement between RNA-seq, qPCR and microarrays. In general, RNA-seq has provided increased detection sensitivity and opened new avenues of research in transcriptome analyses, such as the study of gene fusions, allele-specific expression and novel alternative transcripts. However, it has been shown that RNA-seq data have measurement noise which is a direct consequence of the random sampling process inherent to the assay. Assessments of RNA-seq have been limited to individual sequencing platforms and experiments, which may explain the variation of conclusions by study (see Supplementary Table 1 and Supplementary Notes, section 2.4 for discussion)3–6. Moreover, new platforms and protocols for RNA-seq have emerged in recent years. With the widespread adoption of RNA-seq, including the completion of large projects, such as the Encyclopedia of DNA Elements (ENCODE)7, The Cancer Genome Atlas (TCGA)8 and projects of the International Cancer Genome Consortium (ICGC)9, a comprehensive multisite, cross-platform analysis of RNA-seq performance is timely. Reproducibility across laboratories, in particular, is a crucial requirement for any new experimental method in research and clinical applications, and can only be tested in extensive comparisons of different sites and platforms. As in phase I of the MicroArray Quality Control project (MAQC-I)10, which tested agreement across sites and platforms for gene-expression microarrays, the US Food and Drug Administration (FDA) has coordinated the Sequencing Quality Control project (SEQC/MAQC-III), a large-scale community effort to assess the performance of RNA-seq across laboratories and to test different sequencing platforms and data analysis pipelines.

Here we report a multisite, cross-platform analysis of RNA-seq measurement performance in a controlled setting. We sequenced commercially available reference RNA samples spiked with synthetic RNA from the External RNA Control Consortium. Two distinct samples were assessed individually and also combined in known ratios. This allowed us to examine how well truths built into the study design, such as known relationships between samples within and across sites, could be recovered from measurements. With no independent ‘gold standard’ feasible, these ‘known truths’ support an objective assessment of performance. To this end, we examined a multitude of properties, including complementary metrics of reproducibility, accuracy and information content. Such a multidimensional characterization is critical for the development of more powerful analyses of the underlying biological mechanisms in complex data sets because often there is a trade-off between one desirable property and another, such as accuracy versus precision. Analyses focusing on measurement quality metrics11, spike-in controls and limits of detection12 and the effects of analytic pipeline choice are presented in separate studies (unpublished data).

The SEQC project also involved studies assessing RNA-seq in several research applications (Fig. 1a), including a performance analysis of neuroblastoma outcome prediction (unpublished data), a comparative investigation of toxicogenomic samples testing chemicals with different modes of action12 and a comprehensive survey of tissue-specific gene expression in rat13. In total, >10 billion reads (10 terabases) of RNA-seq data were produced and studied, which to our knowledge represents the largest effort to date to generate and

* A full list of authors and their affiliations appears at the end of the paper. Correspondence should be addressed to L.S. (leming.shi@gmail.com), C.E.M. (chm2042@med.cornell.edu) or D.P.K. (david.kreil@boku.ac.at or d.kreil@warwick.ac.uk).

Received 13 June 2013; accepted 11 May 2014; published online 24 August 2014; corrected after print 9 September 2014; doi:10.1038/nbt.2957
analyze comprehensive reference data sets. A rigorous dissection of sources of noise and signal indicates that, given appropriate data treatment and analysis, RNA-seq can be highly reproducible, particularly in differential gene-expression analysis.

RESULTS

Study design

We used the well-characterized reference RNA samples A (Universal Human Reference RNA) and B (Human Brain Reference RNA) from the MAQC consortium\(^{10}\), adding spike-ins of synthetic RNA from the External RNA Control Consortium (ERCC)\(^{14,15}\). We then mixed A and B in known ratios, 3:1 and 1:3, to construct samples C and D, respectively (Fig. 1 and Supplementary Fig. 1). All samples were distributed to several independent sites for RNA-seq library construction and profiling by Illumina’s HiSeq 2000 and Life Technologies’ SOLiD 5500 instruments. In addition, vendors created their own cDNA libraries, which were distributed to each test site to examine ‘site effects’ independent of the library preparation process (Fig. 1b). To support an assessment of gene models, three sites also independently sequenced samples A and B using the Roche 454 GS FLX platform, providing longer reads. In total, for samples A to D, 108 libraries were sequenced on a HiSeq 2000, another 68 libraries on SOLiD, and 6 libraries on a Roche 454.

To compare technologies, we also examined expression profiles of the same reference samples generated from Affymetrix HGU133Plus2.0 microarrays in the MAQC-I study, and profiled and analyzed these samples on several current microarray platforms (Online Methods). Besides RNA-seq and microarrays, we also considered qPCR-based protocols. We examined 843 TaqMan assays from the MAQC-I study and, in addition, performed 20,801 PrimePCR reactions.

Figure 1 The SEQC (MAQC-III) project and experimental design. (a) Overview of projects. We report on a group of studies assessing different sequencing platforms in real-world use cases, including transcriptome annotation and other research applications, as well as clinical settings. This paper focuses on the results of a multicenter experiment with built-in ground truths. (b-d) Main study design. Similar to the MAQC-I benchmarks, we analyzed RNA samples A to D. Samples C and D were created by mixing the well-characterized samples A and B in 3:1 and 1:3 ratios, respectively. This allows tests for titration consistency (c) and the correct recovery of the known mixing ratios (d). Synthetic RNAs from ERCC were both added to samples A and B before mixing and also sequenced separately to assess dynamic range (samples E and F). Samples were distributed to independent sites for RNA-seq library construction and profiling by Illumina’s HiSeq 2000 (three official + three unofficial sites) and Life Technologies’ SOLiD 5500 (three official sites + one unofficial site). Unless mentioned otherwise, data show results from the three official sites (italics). In addition to the four replicate libraries each for samples A to D per site, for each platform, one vendor-prepared library A5...D5 was being sequenced at the official sites, giving a total of 120 libraries. At each site, every library has a unique bar-code sequence, and all libraries were pooled before sequencing, so each lane was sequencing the same material, allowing a study of lane-specific effects. To support a later assessment of gene models, we sequenced samples A and B by Roche 454 (3x, no replicates, see Supplementary Notes, section 2.5). (c) Schema illustrating tests for titration order consistency. Four examples are shown. The dashed lines represent the ideal mixture of samples A and B expected for samples D and C. (d) Schema illustrating a consistency test for recovering the expected sample mixing ratio. The yellow lines mark a 10% deviation from the expected response (black) for a perfect mixing ratio. Both tests (c) and (d) will reflect both systemic distortions (bias) and random variation (noise).

Read depth dependency of gene detection and junction discovery

Because efficient quantitative expression profiling takes advantage of known gene models\(^3\) the choice of a reference annotation can considerably affect results, including performance assessments. Our data showed that the number of reads mapped to known genes depends on the accuracy and completeness of the gene models. Among all 23.2 billion reads that could be mapped to genes other than those encoding mitochondrial or ribosomal RNAs, 85.9% were mapped to RefSeq\(^{15}\), whereas 92.9% mapped to GENCODE\(^{16}\), and 97.1% to NCBI AceView\(^{17}\) (Fig. 2a). This is not a property of the samples examined (A, B, C and D), as a similar trend is seen when adding all reads from the SEQC neuroblastoma project (Supplementary Notes, section 1.2). The higher read fraction unique to AceView is genuinely due to the higher accuracy of its gene models (Fig. 2a,b), as AceView annotated exons cover fewer bases than those from GENCODE (191 Mb versus 203 Mb; Supplementary Notes, section 1.2).

As the data constitute the deepest sequencing of any set of samples yet reported and include a total of 12 billion mapped HiSeq 2000 RNA-seq fragments, we examined how well the known genes could be detected as a function of aggregate read depth, taking all replicate libraries, sites and samples together. We report read depth as the number of sequenced fragments because the mapping and counting of paired ends are highly correlated; single-ended reads can thus be used when the additional long-range information from read pairs is not required. At a sequencing depth of 10 million aligned fragments, about 35,000 of the 55,674 genes annotated in AceView\(^{17}\) were found by at least one read. Some of these reads were due to background noise, for instance, from genomic DNA contamination (see Supplementary Notes, section 1.2). For a comparison of alternative pipelines, annotations and the effect of read depth, we next focused on...
Figure 2 Gene detection and junction discovery. (a) The fraction of all reads aligned to gene models from different annotations, RefSeq, Encode and NCBI AceView. Reads aligning only to specific annotations are shown in dark green. (b) Known genes (left) and exon junctions (right) supported by at least 16 HiSeq 2000 or SOLiD reads are in green; genes or junctions annotated but not observed at this threshold are shown in gray. (c-e) Sensitivity as a function of read depth. (c) The number and percentage of all AceView-annotated genes detected for three RNA-seq analysis pipelines. The x axis marks cumulative aligned fragments from all replicates and sites. Vertical dotted lines indicate boundaries between samples A through D. (d) Known junctions detected. The numbers and percentages of all exon-exon junctions (supported by eight or more reads) are shown for different gene model databases (line style). Horizontal lines show the respective total numbers of annotated junctions. (e) Unannotated junctions supported by multiple platforms and pipelines. Subsets of unannotated junctions have expression levels with correct titration orders and mixing ratios (cf. Figs. 1b-d and 4a,b). (f) Distribution of junction expression levels. Unannotated junctions, then unannotated junctions supported by multiple platforms and pipelines, and known junctions show increasing expression levels (colors). Subsets expressed with mutual information about the samples and correct titration order and mixing ratio display a further shift toward higher expression levels (dashed lines). (g,h) Intra- (blue) and intersite reproducibility (orange) of detected known genes (g) and junctions (h). Pairwise agreement is shown by box plots, where the second set of box plots (upper group) indicates percentages.

We examined the detection of exon-exon junctions as a function of read depth for RefSeq, GENCODE and AceView annotation (Fig. 2d). In general, with each doubling of the read depth, many additional known junctions were detected for each doubling of read depth, additional genes were still being detected even at high read depths of >1 billion fragments, indicative of low expression levels per cell or small numbers of cells expressing these genes (Fig. 2c) for HiSeq 2000 and Supplementary Fig. 2 for SOLiD).

The number of additional known junctions decreased fastest for RefSeq, which provides the least complex annotation (Supplementary Fig. 3), and so practically all annotated junctions were observed at the highest read depth. In contrast, the AceView database is the most comprehensive and has the highest number of junctions supported by reads from this study, reaching over 300,000 junctions at the maximum read depth, more than three times the number detected at 10 million reads. Although GENCODE and AceView have similar total numbers of genes and similar footprints on the genome, considerably fewer annotated genes and junctions in GENCODE were supported by the observed reads (Fig. 2b and Supplementary Fig. 4). Therefore, all subsequent analyses presented in this manuscript are based on AceView, unless stated otherwise.

We next analyzed the reproducibility of detecting genes and junctions across measurement sites, platforms and analysis pipelines, as a key strength of RNA-seq is its inherent ability to identify splice sites de novo. To test this ability of RNA-seq to discover junctions, we first examined the HiSeq 2000 data because of the greater read length and depth. We considered three independent pipelines for de novo discovery of junctions independent of existing gene models: NCBI Magic17, r-make (which uses STAR18) and Subread19. All pipelines reported millions of junctions, with r-make predicting about 50% more than Subread and Magic, although almost all junctions found by Subread or Magic were also found by r-make. We also observed substantial but 12% lower agreement with TopHat2, regardless of whether it was run with or without gene model-guided
alignment^{(2)} (Supplementary Fig. 5a), giving a total of 1,110,550 junctions consistently found by all five analysis variants. In total, 2.6 million previously unannotated splice junctions were called by at least one of the five analysis pipelines, yet only 820,727 (32%) were consistently predicted by all the methods (Supplementary Fig. 5b), illustrating the considerable difficulty of reliably detecting splice junctions de novo with current analysis tools.

We then examined whether unannotated junctions were independently discovered in both HiSeq 2000 and SOLiD data, as junctions found only by a single platform or library preparation protocol could be technical artifacts (Supplementary Fig. 6). Junction discoveries from the SOLiD data reflected the lower read length and read depth as expected (simulation results in Supplementary Table 2). In particular, we discovered 87,117 unannotated potential junctions in the SOLiD data, of which 74,561 (86%) were also independently discovered from HiSeq 2000 reads using the Subread aligner. The number of these new junctions found in each of the four samples (Fig. 2e) followed the order expected corresponding to sample complexity: B < A < D < C.

We then used the built-in truths of the benchmark measurements to examine the accuracy of the sample-specific levels of support of the unannotated junctions in terms of their ability to capture the expected A/B sample mixing ratio and yield titration order consistency. For example (Fig. 1c), if a gene is more strongly expressed in sample A than in sample B (A > B) then we expect A > C > D > B because C = ¾ A + ¼ B and D = ¼ A + ¾ B, and we expect the inverse order if B > A. This consistency test is affected both by systemic distortions reducing accuracy and random variations reducing reproducibility. Another complementary test assesses the A/B mixing ratio recovery in samples C (3:1) and D (1:3), which can be examined in a plot of log2(C/D) versus log2(A/B). Deviations from the ideal line (Fig. 1d) are also affected by systemic distortions, which reduce accuracy, and by random variations, which reduce reproducibility. Because both tests only reflect reproducibility for genes with similar expression levels in samples A and B, we also require a clear differential signal as assessed by the 'mutual information', a measure of information content (Online Methods).

We observed that requiring a consistent titration order and the correct A/B mixing ratio clearly enriched for junctions with higher expression levels (Fig. 2f), which were the easiest to measure and quantify. A comparison with junctions detected in Roche 454 data confirmed that the more abundant junctions could reliably be detected across different sequencing platforms (Supplementary Fig. 7).

For an examination of how well junctions discovered by RNA-seq could be independently confirmed by a different technology, we performed qPCR with primer pairs designed to specifically validate 173 detected junctions. Of these, 136 were randomly-selected well-supported junctions that had been discovered de novo by all three RNA-seq analysis pipelines in both HiSeq 2000 and SOLiD data.

Figure 3 Sensitivity, specificity and reproducibility of differential expression calls. Robust cross-site analyses depend on pipeline choice and appropriate filter rules. Results are shown for five MAQC-III RNA-seq pipelines and the MAQC-I Affymetrix microarray platform (color). (a,c) Number of standardized differential expression calls. All possible pairwise intersite A versus A comparisons (solid) are shown next to all intra-site A versus B comparisons (x) as indicators for specificity and sensitivity. (b,d) Ratios of A versus B calls and A versus B calls give an estimate of the eFDR. For all platforms and pipelines, differential expression calls identify thousands of differences in inter-site comparisons of identical samples. These can be controlled for microarray by additional filters for effect size. In addition, RNA-seq also requires filters for expression strength due to the high sampling fluctuations at lower read counts. These were set to give similar numbers of A versus B expression calls (b), improving the eFDR to <1.5% except for several outliers. (e) Intersite reproducibility of differential expression calls. Comparing the identities and the directions of change for DEGs across sites, agreement is plotted for lists including the top-ranked genes as sorted by effect size (x axis). The observed response curves depend on pipeline and filter choice, showing more variation for shorter lists. The performance of several RNA-seq pipelines was comparable to that of differential expression profiling using the microarray measurements from the MAQC-I study, with the microarrays showing higher intersite reproducibility when considering only the differentially expressed genes with the strongest fold change (left side of e).
These junctions were chosen so that they log-uniformly covered a range of ~10–3,000 supporting reads, and so that half of them met all consistency tests. In addition, we also tested 13 known AceView junctions as positive controls and 24 unannotated junctions that had been discovered only by a single analysis pipeline in the HiSeq 2000 or SOLiD data, despite having support by many reads (~300–3,000). Only five assays were noninformative (Online Methods). Notably, in the remaining assays, all of the 13 positive controls and all of the 133 well-supported junctions were reliably identified by qPCR, with the numbers of supporting RNA-seq reads largely reflecting estimates of expression levels by qPCR (slope 0.95, Pearson (Spearman) correlation 0.74 (0.77), N = 146), even for junctions with low read numbers or not meeting all consistency tests (Supplementary Fig. 8). Moreover, 18 of the 22 pipeline-specific junctions (>80%) could be confirmed at least qualitatively. For the most comprehensive surveys of potential new junctions, one may therefore want to consider all discovered junctions, although it makes sense to prioritize well-supported junctions found consistently (Fisher, P < 4 × 10−4).

In the complete data set comprising SEQC samples A, B, C and D, we consistently detected ~44,000 known genes (Fig. 2g) and ~310,000 known exons across pairs of replicate sites (Supplementary Fig. 9), constituting about 79% and 47% of all known genes and exons, respectively. Nearly 200,000 splice junctions were seen consistently (Fig. 2h), making up about 50% of all known junctions. This corresponds to about 90% of all detected known genes, 87% of detected known exons and 83% of detected known junctions, consistent with the explanation that larger features aggregating more reads are easier to measure reproducibly. The fluctuations in the detection of sequence features stemmed largely from sequencing depth–dependent sampling noise, which was reflected in the very similar intra- and inter-site agreements in the detection of known genes (Fig. 2g), exons (Supplementary Fig. 9) and junctions (Fig. 2h). Considering the low technical variance in addition to the unavoidable sampling noise, these results emphasize the value of biological replicates.

Improving differential expression analysis reliability

Studies on microarrays have shown that results of typical statistical differential expression tests thresholded by P-value need to be filtered and sorted by effect strength (fold-change) in order to attain robust comparisons across platforms and sites10. We thus sought to identify corresponding requirements for RNA-seq, examining the reproducibility of rank ordered lists of differentially expressed genes (DEGs), as well as the false discovery rate (FDR) reflecting the information content of the measurements. Because the same samples were profiled at all sites, the number of true DEGs is zero when comparing the same sample between any two sites. Any DEGs found for self–self comparisons thus represent technical differences and can be considered ‘false positives’ (Fig. 3a). We examined the number of inter-site A versus A ‘false positives’ relative to the number of DEGs in A versus B comparisons, giving an empirical estimate of the FDR (eFDR). We tested several RNA-seq data analysis pipelines for the six HiSeq 2000 sites, focusing on the set of 23,437 genes present on the Affymetrix HGU133Plus2.0 microarrays for comparison.

We found that unfiltered data for both RNA-seq and microarrays show many DEGs, both for false positives (A versus A) and the likely ‘real’ DEGs in the A versus B comparison (Fig. 3a), with the ratio of false positives versus true positives unacceptably high for both platforms (Fig. 3b). Also, we observed that different analysis pipelines vary in these measures of performance (Fig. 3a–e). Next, we applied the |log₂ fold-change|>1 filter advocated in the MAQC study on microarrays10, and for microarrays observed a reduction in the eFDR to below 1.5%, save for one outlier site (Fig. 3c,d). Notably, we found that applying pipeline-dependent filters for P-value, fold-change and expression-level (lowest third of all examined AceView genes, Supplementary Tables 3 and 4) successfully reduced the RNA-seq eFDR across sites without sacrificing sensitivity compared to the arrays (Fig. 3d). We note that results for SOLiD were very similar (Supplementary Figs. 10 and 11), with expression level thresholds reflecting the lower read depth for that platform.

Figure 4: Built-in truths for assessing RNA-seq. (a) Titration order A, C, D, B. Log₂ fold-change is related to cross-platform titration consistency. At sufficiently strong log₂ fold-change, reliable titration is also found across platforms. The dark blue line represents the 22,074 ‘unmissable’ genes showing the correct titration order with no contradiction in at least 14 HiSeq 2000 and 6 SOLiD samples. Most genes with high differential expression are in this class. (b) Known A/B mixing ratios in samples C and D. The yellow solid line traces the expected values after mRNA/total-RNA shift correction. The 1%, 10% and 25% most highly expressed genes are shown in red, cyan and magenta, respectively. On average, the most strongly expressed genes recover the expected mixing ratio best. Genes with inconsistent titration (cf. a) are colored gray. Black and gray symbols intermixing indicates that consistent titration (black) does not guarantee reliable recovery of the mixing ratio (and vice versa). (c) ERCC spike-in ratios can be recovered increasingly well at higher expression levels. From the response curves, one can calculate signal thresholds for the detection of a change46. (d) Variation of the total amounts of detected ERCC spikes. The lack of reliable titration indicates that the considerable differences between libraries of a given site and protocol are random, implying limits for absolute expression level estimates, in general, and using spike-ins for the calibration of absolute quantification, in particular. The observed variations likely arise in library construction, as the vendor-prepared libraries (colored cyan or gray) gave constant results across different sites. For a and b, all 55,674 AceView genes were tested.
After applying these filters, we found that most (but not all) RNA-seq pipelines achieved high inter-site reproducibility of differential expression calls with up to 95% concordance in DEGs (Fig. 3e and Supplementary Fig. 12). This concordance between sites was highest for the most strongly expressed genes. Moreover, the filters resulted in a good agreement of differential expression calls across platforms (e.g., A versus B on HiSeq 2000 compared with A versus B on SOLiD, Supplementary Figs. 13 and 14), suggesting that differential expression analyses from different platforms can be combined—for example, to extend existing studies with additional samples.

**Relative but not absolute expression measures satisfy tests**

After an examination of the reliability of differential expression analysis of genes, we next examined the quantification of RNA using four consistency tests exploiting ground truths built into the study design (Fig. 4). First, we considered titration-order consistency as introduced in Figure 1c. This metric is affected both by systemic distortions reducing accuracy and random variations reducing reproducibility. The majority of genes (59%) titrated correctly (Fig. 4a), with little disagreement between platforms (Supplementary Table 5). Genes with large differential expression performed best, with all genes showing consistent titration in several HiSeq 2000 and SOLiD sites, and no contradiction regarding the direction of change (blue curve). For the second built-in truth, we examined the A/B mixing ratio recovery (Fig. 1d) as another test reflecting accuracy and reproducibility. We observed the correct ratio for the majority of genes (Fig. 4b), with better agreement at higher expression levels (top 25%). Notably, the scatter of genes marked as titrating in this plot indicates that consistent titration does not guarantee a reliable recovery of the mixing ratio (and vice versa).

The third and fourth built-in truths leveraged the ERCC spike-ins.14 These analyses complement work examining fold-change recovery for these synthetic RNAs.47 Across platforms, we observed that with sufficiently high expression levels (log2,[conc] > 3), the expected ratios of 1:2, 2:3, 1:1 and 4:1 were accurately recovered using about 90 million mapped fragments (Fig. 4c), with high precision indicating good reproducibility. Finally, we examined the ERCC absolute titration levels, as the ERCC RNAs were spiked into samples A and B before samples C and D were created (Supplementary Fig. 1). We observed, however, that the fraction of reads aligning to ERCC spike-ins for a given sample varied widely between libraries and platforms, with measured ERCC ranges of 1–2.5% for HiSeq 2000 and 2.5–4.7% for SOLiD, with a clear ‘library effect’ observed for all sites and platforms, affecting reproducibility. (Fig. 4d and Supplementary Fig. 15). Indeed, when using the vendor-prepared library as the cross-site control, we observed very consistent measurements of the percentages of reads mapping to ERCCs, which indicates a large degree of variation from the preparation of libraries even at the same site. The resulting lack of meaningful absolute expression level measurements is moreover not specific to the ERCC spike-ins, as similar variation and substantial platform-specific differences could also be observed for human genes (Supplementary Fig. 16).
In addition, when the concentrations of the 92 ERCC spike-in RNAs are measured and compared to their nominal concentrations, some of them (e.g., ERCC-116) are consistently reported as being up to ten times below or above their expected concentrations, perturbing even the order of the ERCC scale. These discrepancies are highly reproducible, suggesting that the bias is sequence dependent. Although marginal trends could be observed as functions of GC content and average probe lengths for the unfolding of sequence regions (Supplementary Figs. 17 and 18), these did not pass tests for statistical significance. No consistent trend could be observed as a function of mRNA length (Supplementary Fig. 19), and the majority of the deviations is not explained by any of these co-variates, indicating a need for further investigation of such distortions and their possible sources. We observed, however, that the effect is also protocol dependent and is reduced in the absence of poly-A selection (Supplementary Fig. 20). This is in line with results in other studies, further underscoring the impact of protocol choice on quantification\textsuperscript{11,21,47}, where fragmentation time, poly-A enrichment by columns, beads, or ribo-depletion, hexamer or oligo-dT priming, library isolation by gel or beads, different ligation efficiency, and RNA quality at the start of library preparation have all been shown to have an effect. Consequently, just as for microarrays, absolute measurements by RNA-seq using a particular protocol are reproducible but not very accurate. This observation implies that the use of external spike-in controls to accurately infer absolute expression levels of a gene of biological interest will remain challenging as long as these deviations are not better understood and can thus be avoided or quantitatively modeled.

**Relative gene expression measurements agree across platforms**

Because we observed good performance for RNA-seq in consistency tests of relative expression levels, we then sought to compare alternative measurement platforms. We first examined the differential expression of 843 genes measured by TaqMan for samples A and B in the MAQC-I study. Although more strongly expressed than typical AceView genes, these genes nevertheless cover a wide range of expression levels (Supplementary Fig. 21). We found good and comparable agreement among different platforms (Fig. 5a, where Pearson and Spearman correlation coefficients are given; Supplementary Fig. 22). The HiSeq 2000 and SOLiD sequencing platforms performed similarly well for comparable effective read depths (Supplementary Fig. 33a). The read depth for which average RNA-seq performance met or exceeded that of the array is marked by a cyan bar. The corresponding read depths varied widely from 5 million (HGU133plus2 with MAS5) to about 50 million fragments (PrimeView with gcRMA/affyPLM), showing the strong effect of the reference gene set implied by the probes on the respective arrays and the employed microarray data-processing methods. Results are shown for the Subread pipeline. Alternative RNA-seq data analysis pipelines, however, can require up to double the number of fragments (TopHat2+Cufflinks, Subread). Alternative RNA-seq data analysis pipelines, however, can require up to double the number of fragments (TopHat2+Cufflinks, Subread, and BitSeq), where fragmentation time, poly-A enrichment by columns, beads, or ribo-depletion, hexamer or oligo-dT priming, library isolation by gel or beads, different ligation efficiency, and RNA quality at the start of library preparation have all been shown to have an effect. Consequently, just as for microarrays, absolute measurements by RNA-seq using a particular protocol are reproducible but not very accurate. This observation implies that the use of external spike-in controls to accurately infer absolute expression levels of a gene of biological interest will remain challenging as long as these deviations are not better understood and can thus be avoided or quantitatively modeled.
shape (Supplementary Fig. 23b). Although we observed, on average, agreement of absolute expression levels between different platforms, there were substantial deviations across the entire dynamic range for large numbers of individual genes. These deviations are systematic—that is, they are not a question of reproducibility, but rather reflect the accuracy of absolute expression measures. In particular, by comparing expression level estimates from HiSeq 2000 and SOLiD RNA-seq (Fig. 5b), we observed 5,056 genes that were expressed according to one platform but not the other (9%). This effect is only partly due to the nonstranded nature of the Illumina protocol used here, and the presence of 11,066 genes antisense to genes annotated on the opposite strand (Fig. 5c).

As an independent additional test, we generated 20,801 PrimePCR measurements of the SEQC samples A, B, C and D. We again observed that more than a thousand genes (5%) were not considered expressed by one platform but were clearly expressed according to the other (Supplementary Figs. 24 and 25). Although qPCR-based methods have traditionally been used as a reference ‘gold’ standard owing to their high sensitivity and dynamic range, it is noteworthy in this context that specific primer selection and protocol calibration are challenging in their own right25. PCR is affected by GC bias25, and considerable differences in expression level measurements from different PCR-based assays can be observed (Fig. 5d).

Performance assessment is metric dependent

A major promise of RNA-seq is the extension of expression profiling to the discovery and quantification of alternative transcripts. For transcript-specific profiling, however, no large-scale expression data from other technologies are available as an external reference point. The SEQC data represent an opportunity for the multiparametric comparison of transcript-specific measurements.

To support a balanced performance study of gene-level and transcript-specific expression profiling, we combined multiple metrics for a robust characterization of platforms, sites and data-processing options: (i) average measurement precision5, which directly assesses reproducibility; (ii) titration order consistency26 and (iii) recovery of the expected A/B mixing ratio, providing two complementary assessments reflecting both measurement accuracy and reproducibility, as well as (iv) differential expression and (v) the mutual information of sample titration, which capture different aspects of information content (Online Methods). For a summary view, we first focused on genes with a clear directional signal—that is, those that allowed an ordered discrimination of the samples A to D, as indicated by the mutual information metric (v). We then counted how many of these genes also satisfied a second requirement, for each of the metrics (i)–(iv). Such an integration of tests through counting genes that fulfill multiple assay criteria allows a comprehensive consideration of all the genes instead of restricting comparisons to a common subset of genes always identified as expressed. This is necessary for a meaningful comparison of pipelines and platforms with varying degrees of sensitivity (Supplementary Fig. 26a,b). The resulting four combined assays for the respective metrics (i)–(iv) are complementary—that is, a gene satisfying one does not generally satisfy the others (Supplementary Fig. 27). The average of the four assays then provides a consistency score for robust characterization of measurement performance (Supplementary Fig. 28).

For gene-level profiling, pipelines showed similar performances on average (Fig. 6a). Providing known gene models always considerably improved results (cf. our results for standard and gene model-guided TopHat2). Lower scores for transcript-level profiling indicate that the discrimination of alternative transcripts is more difficult, which is also reflected in stronger effects of pipeline choice (Fig. 6b).

RNA-seq has sparked an interest in transcript-specific profiling and the development of advanced algorithms for estimating alternative transcript abundances. With known gene models, similar approaches can now also be applied for microarrays. We thus next focused on a test set of 782 genes with multiple alternative transcripts of varying complexity and specifically selected to represent the full subset of spliced genes in AceView (Fig. 6c and Online Methods). Covering 5,691 alternative transcripts, this test set allows a first comparison of transcript-specific expression level estimates from RNA-seq and high-resolution transcript-level microarray data. We found that efficient transcript-specific measurements with good precision on microarrays for quantitative expression profiling (Fig. 6d, Supplementary Figs. 28 and 29) could complement the power of RNA-seq in the discovery and identification of new alternative transcripts (Fig. 2). In other words, the novel transcripts found by RNA-seq can lead to efficient measurements with good precision on microarrays, which can in turn aid in the confirmation and functional study of new transcript variants.

Finally, each metric showed a different and platform-specific response to signal strength (Supplementary Figs. 30 and 31), which for RNA-seq increases with transcript expression level and read depth. The read depth at which average RNA-seq performance meets or exceeds that of another platform thus directly depends on the chosen metric and the distribution of expression strength and differential signal in the samples measured. As a result, it also depends on the set of tested genes, over which the average performance is being computed. We show results here for the mutual information metric (Supplementary Fig. 32), which is of direct relevance for classifier performance. As expected, RNA-seq performance improved with increasing numbers of mapped fragments (Fig. 6e). In particular, Life Technologies’ SOLiD and Illumina’s HiSeq 2000 performed similarly well for comparable effective read depths (Supplementary Fig. 33a). The choice of reference platform considerably affects the number of RNA-seq reads required for obtaining comparable mutual information per gene (Supplementary Fig. 34). For some of the microarrays and data-processing methods tested, as few as 5 million mapped RNA-seq fragments were more than sufficient (HG113plus2 with MAS5), whereas ~50 million mapped fragments were required for others (PrimeView with gcRNA/affyPLM). The choice of RNA-seq pipeline also had an effect, with some tools requiring up to twice as many aligned fragments (cf. TopHat2+ Cufflinks27, Supplementary Fig. 35).

DISCUSSION

In a multisite, cross-platform study led by the FDA, four well-characterized reference RNA sample mixtures with built-in truths were profiled to test RNA-seq reproducibility, accuracy and information content in a detailed analysis of >30 billion reads on the reference samples alone. To our knowledge, the data presented here provide the deepest molecular characterization of any RNA samples to date.

We leveraged this deep data set to test the reliability and power of RNA-seq in exploring the complexity of the transcriptome. We studied the detection of known splice junctions and the discovery of unannotated junctions. De novo junction discovery was robust across sites both at low and high sequencing depths, even beyond 10 billion aligned fragments. Many unannotated splice junctions were detected by multiple platforms and pipelines, with concordance directly reflecting the junction expression level. Similar to observations by ENCODE at the gene level7, we observed three distinct classes of expression levels for splice junctions: highly expressed known junctions, known
and unannotated junctions at medium levels, and many unannotated junctions found only at low expression levels. Although it has been proposed that an abundance of weakly expressed transcripts may reflect biological noise, the lower expression of these junctions may alternatively be the reason that they have so far received less attention in traditional experiments. With alternative transcript expression being dependent on cell type and experimental conditions, deep RNA-seq will continue to play a key role in fully exploring the transcriptomic repertoire, including the construction of extensive maps of alternative transcripts highlighting splicing variants as well as alternative start and polyadenylation sites. De novo discovery constitutes a key strength of RNA-seq and that is reflected in the expansive transcriptional landscapes observed from different cells and tissues in the transcriptome re-annotation projects for human and rat highlighting the filtering of spurious junctions that cannot be explained by expression levels of alternative transcripts consistent with the distribution of reads that map to exons.

Our read-mapping results underscore how crucial comprehensive gene model annotations are to accurate expression profiling. The human genome now has >55,000 well-validated genes, and the majority of them do not encode proteins. Almost all human multixon genes exhibit alternative splicing, and spliced human genes have on average over nine alternative transcribed forms. Additional genes and transcripts are still being discovered, even beyond the already expansive gene annotation from ENCODE. Notably, the NCBI AceView database, which has >50,000 genes annotated from cDNA evidence, holds by far the largest and most extensively validated set of splice junctions, with >300,000 well supported by the RNA-seq data reported in this study alone. Transcripts that may explain a particular phenotype may be missed by less-extensive annotations, stressing that the most comprehensive annotation for expression profiling is vital to accurate clinical research. The characterization and, particularly, the quantification of alternative transcripts, however, still require further research. Although expression profiling of alternative transcripts is feasible, reattributing measurements to a set of alternative transcripts requires knowledge of all the alternative transcript forms of a gene, and involves combining information across the transcripts. For genome-scale RNA-seq, this is particularly difficult because of the sampling noise from low read counts for many transcripts, with recent work observing 300 million sequenced fragments to be required for the detection of a specific human alternative splicing event with 80% power.

This also highlights the value of targeted RNA-seq. Although simpler organisms such as Caenorhabditis elegans may be less affected by the difficulties of reliably attributing reads to alternative transcripts, analogous considerations will apply to research on mouse, rat and other complex transcriptionomes. The need for longer-range information is a consequence of the fact that certain complex gene models cannot be resolved by the local information provided by either microarray probes or individual short RNA-seq reads alone. Although read pairs of size-controlled fragments can improve on this, they also limit the recovery of shorter transcripts. Full alternative transcript profiling will thus greatly benefit from longer RNA-seq reads, which may eventually approach the full length of complex cDNAs. Combining deep RNA-seq for alternative transcript discovery with modern high-resolution microarrays for genome-scale quantification may provide an efficient approach for systematic transcript-level expression profiling.

Although none of the technologies we tested could provide reliable absolute quantification, relative expression measures agreed well across platforms, including RNA-seq, qPCR and microarrays. The majority of genes satisfied constraints based on the truths built into the study design. Going beyond earlier platform comparisons that considered individual performance metrics, we combined complementary metrics for a robust characterization of measurement performance that can be combined with further assays such as tests for strandedness, considerations based on the ERCC spike-in response, and tests for the exclusion of nonspecific background. Notably, the important but difficult task of estimating and removing background noise (Supplementary Fig. 36b) typically improves accuracy at the expense of precision (Supplementary Fig. 29).

Considering the substantial disagreements even between different types of qPCR-based assays, we conclude that there is no single ‘gold standard.’ Although our cross-platform comparisons reveal common trends, drastic systemic differences remain. Reference data sets such as the compendium presented here are invaluable for a systematic characterization of measurements, which is critical for reliable conclusions from large-scale experiments. Specifically, a closer examination of the varying amount of detected ERCCs per sample indicated substantial differences and inconsistencies even across libraries prepared from the same sample at the same site and sequenced by the same machine. This implies inherent limitations for the read-out of absolute expression level estimates and absolute quantification.

As the vendor-prepared libraries gave very uniform results across sites (Fig. 4d), the observed variations likely arose in library construction and may be partially explained by platform-specific differences in kit chemistry and varying degrees of sample polyadenylation. Therefore, in RNA-seq experiments, multiple libraries per examined condition or sample should be profiled.

We show that filters can improve robustness of differential expression calls and consistency across sites and platforms. For RNA-seq, removing small fold-changes as well as excluding low-expression measurements reduced the FDR considerably and, in general, gave an improvement over microarrays just as methodological differences of qPCR-based assays, we conclude that there is no single ‘gold standard.’ Although our cross-platform comparisons reveal common trends, drastic systemic differences remain. Reference data sets such as the compendium presented here are invaluable for a systematic characterization of measurements, which is critical for reliable conclusions from large-scale experiments. Specifically, a closer examination of the varying amount of detected ERCCs per sample indicated substantial differences and inconsistencies even across libraries prepared from the same sample at the same site and sequenced by the same machine. This implies inherent limitations for the read-out of absolute expression level estimates and absolute quantification.

As the vendor-prepared libraries gave very uniform results across sites (Fig. 4d), the observed variations likely arose in library construction and may be partially explained by platform-specific differences in kit chemistry and varying degrees of sample polyadenylation. Therefore, in RNA-seq experiments, multiple libraries per examined condition or sample should be profiled.

We show that filters can improve robustness of differential expression calls and consistency across sites and platforms. For RNA-seq, removing small fold-changes as well as excluding low-expression measurements reduced the FDR considerably and, in general, gave an improvement over microarrays just as methodological differences of qPCR-based assays, we conclude that there is no single ‘gold standard.’ Although our cross-platform comparisons reveal common trends, drastic systemic differences remain. Reference data sets such as the compendium presented here are invaluable for a systematic characterization of measurements, which is critical for reliable conclusions from large-scale experiments. Specifically, a closer examination of the varying amount of detected ERCCs per sample indicated substantial differences and inconsistencies even across libraries prepared from the same sample at the same site and sequenced by the same machine. This implies inherent limitations for the read-out of absolute expression level estimates and absolute quantification.

As the vendor-prepared libraries gave very uniform results across sites (Fig. 4d), the observed variations likely arose in library construction and may be partially explained by platform-specific differences in kit chemistry and varying degrees of sample polyadenylation. Therefore, in RNA-seq experiments, multiple libraries per examined condition or sample should be profiled.

We show that filters can improve robustness of differential expression calls and consistency across sites and platforms. For RNA-seq, removing small fold-changes as well as excluding low-expression measurements reduced the FDR considerably and, in general, gave an improvement over microarrays just as methodological differences of qPCR-based assays, we conclude that there is no single ‘gold standard.’ Although our cross-platform comparisons reveal common trends, drastic systemic differences remain. Reference data sets such as the compendium presented here are invaluable for a systematic characterization of measurements, which is critical for reliable conclusions from large-scale experiments. Specifically, a closer examination of the varying amount of detected ERCCs per sample indicated substantial differences and inconsistencies even across libraries prepared from the same sample at the same site and sequenced by the same machine. This implies inherent limitations for the read-out of absolute expression level estimates and absolute quantification.

As the vendor-prepared libraries gave very uniform results across sites (Fig. 4d), the observed variations likely arose in library construction and may be partially explained by platform-specific differences in kit chemistry and varying degrees of sample polyadenylation. Therefore, in RNA-seq experiments, multiple libraries per examined condition or sample should be profiled.
Already today, RNA-seq can be used as a versatile tool for relative expression profiling, with comparable or superior performance to microarrays in many applications given sufficient read depth and appropriate choice of analysis pipeline. An effective sequencing depth is clearly contingent on the experimental goals, with simple gene-level expression profiling only requiring 5–50 million single-ended reads for an appropriate analysis pipeline (cf. Supplementary Figs. 11 and 34 and Fig. 6e). A comprehensive characterization of alternative transcript expression benefits from the longer-range information of read pairs and requires considerably deeper sequencing. In our data set, at five million mapped fragments, >15,000 AceView genes could already be detected with strong support (16 reads), including ~10,000 RefSeq genes (Fig. 2c). Moreover, 10 million mapped fragments sufficed for differential expression analysis of the most strongly expressed genes in our study, and reliable results across sites (see Supplementary Fig. 11 for adapted filter parameters). Other applications may require deeper sequencing, as is reflected by different metrics responding differently to an increase in reads for the samples and genes studied. Classifier performance, for instance, is directly related to the mutual information metric. Although 5 million mapped fragments easily gave a mutual information per gene comparable to that of HGU133Plus2.0 microarrays with MAS5, performance comparable to newer arrays and processing methods required about 50 million mapped fragments in this study (Fig. 6e and Supplementary Fig. 34), or even required considerably more, depending on the RNA-seq analysis pipeline (cf. TopHat2+Cufflinks, Supplementary Fig. 35). In addition, the required read depth is also dependent on the genome size, transcriptional complexity, cellular distribution of stored versus active RNAs, biological noise and the panoply of all other factors of cell biology and RNA dynamics. Our comprehensive, multisite, cross-platform, benchmark measurements under controlled settings thus complement and extend comparisons for individual biological experiments (Supplementary Table 1 and Supplementary Notes, section 2.4 for further discussion).

In summary, the study and data collection presented here are a milestone in the development and dissection of RNA-seq as a method for transcriptome profiling. The results based on data sets of this size and complexity and an array of independent measures as introduced by this study will contribute to a better understanding of the power and limitations of RNA-seq. This work is complemented by SEQC companion studies analysing the application of RNA-seq to specific biological research and clinical questions, which are presented elsewhere. Together, the cumulative SEQC data sets with >100 billion reads (10 Tb) provide a key resource for testing future developments of RNA-seq, as required in clinical and regulatory settings.

METHODS

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

Accession codes. All SEQC (MAQC-III) data sets are available through GEO (series accession number: GSE47792)45.

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

ACKNOWLEDGMENTS

All SEQC (MAQC-III) participants freely donated their time and reagents for the completion and analysis of the project. Many participants contributed to the sometimes-heated discussions on the topic of this paper during numerous e-mail exchanges, teleconferences and face-to-face project meetings. The common conclusions and recommendations reported in this paper evolved from this extended discourse. The authors gratefully acknowledge support by the National Center for Biotechnology Information (NCBI)’s Supercomputing Center, the FDAs Supercomputing Center, China’s National Supercomputing Center of Shanghai, the Vienna Scientific Cluster High Performance Computing Facility (VSC), the Vienna Science and Technology Fund (WWTF), Baxter, the Austrian Institute of Technology, and the Austrian Centre of Biopharmaceutical Technology. This work was supported in part by China’s Program of Global Experts. This work was supported in part by the US National Institutes of Health (NIH) grants R01CA163256, R01HG006798, R01NS076465, R44HG005297, U54CA119338, PO1HG00205, R24GM102656 and the Intramural Research Program of the NIH, National Library of Medicine, National Institute of Environmental Health Sciences (NIEHS) Z01 ES102345-04, Shriners Research Grant 85500, an Australia National Health and Medical Research Council (NH&MRC) Project grant (1023454) and Victorian State Government Operational Infrastructure Support (Australia), the National 973 Key Medical Basic Research Program of China (2010CB945401), the National Natural Science Foundation of China (31240038 and 31071162), and the Science and Technology Commission of Shanghai Municipality (11DZ2260300). We greatly appreciate SAS Institute, Inc. for kindly hosting several face-to-face meetings of the SEQC (MAQC-III) project.

AUTHOR CONTRIBUTIONS

Project coordination: US Food and Drug Administration.

Project lead: Weida Tong & Leming Shi.

Manuscript lead: David P. Kreil.

Scientific management: David P. Kreil, Christopher E. Mason, Weida Tong & Leming Shi.

Next-generation sequencing technology lead: Christopher E. Mason.

The following authors contributed to project leadership: Zhengqiang Su, Pawel P. Labaj, Shen Li, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Shi, Charles Wang, Gary P. Schroth, Robert A. Settlerquist, John F. Thompson, Wendell D. Jones, Wenzhong Xiao, Weihong Xu, Roderick J. Jensen, Rebekah Kelly, Joshua Xu, Ana Conesa, Cesare Furlanello, Hanlin Gao, Huixiao Hong, Naderer Jafari, Stan Letovsky, Yang Liao, Fei Lu, Edward J. Oakley, Zhiyu Peng, Craig A. Praul, Javier Santoyo-Lopez, Andreas Scherer, Tielu Shi, Gordon K. Smyth, Frank Staudeler, Peter Sykacek, Xin-Xing Tan, E. Aubrey Thompson, Jo Vandesompele, May D. Wang, Jian Wang, Russell D. Wolfinger, Jiri Zavadil, Weida Tong, David P. Kreil, Christopher E. Mason & Leming Shi.

DISCLAIMER

The views presented in this article are those of their authors and do not necessarily reflect current or future opinion or policy of the authors’ institutions or agencies. Any mention of commercial products is for clarification and do not necessarily reflect current or future opinion or policy of the authors’ institutions or agencies. Any mention of commercial products is for clarification and do not necessarily reflect current or future opinion or policy of the authors’ institutions or agencies.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

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

1. Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470–476 (2008).
2. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat. Methods 5, 621–628 (2008).
3. Łabaj, P.P. et al. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. Bioinformatics 27, i383–i391 (2011).
4. Liu, S., Lin, L., Jiang, P., Wang, D. & Xing, Y. A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. Nucleic Acids Res. 39, 578–588 (2011).
5. McIntryre, L.M. et al. RNA-Seq: technical variability and sampling. BMC Genomics 12, 293 (2011).
6. Teung, J.M., Morley, L., Li, M. & Cheung, V.G. RNA-sequence analysis of human B-cells. Genome Res. 21, 991–998 (2011).
7. Djebali, S. et al. Landscape of transcription in human cells. Nature 489, 101–108 (2012).
8. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 1061–1068 (2008).
9. International Cancer Genome Consortium. International network of cancer genome projects. Nature 464, 993–998 (2010).
10. Shi, L. et al. The Microarray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat. Biotechnol. 24, 1151–1161 (2006).
11. Li, S. et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat. Biotechnol. doi:10.1038/nbt.3000 (24 August 2014).
12. Wang, C. et al. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat. Biotechnol. doi:10.1038/nbt.3001 (24 August 2014).
13. Yu, X. et al. A rat RNA-seq transcriptomic Bodymap across eleven organs and four developmental stages. Nat. Commun. 5, 3230 (2014).
14. Baker, S.C. et al. The External RNA Controls Consortium: a progress report. Nat. Methods 2, 731–734 (2005).
15. Pruitt, K.D., Tatusova, T., Brown, G.R. & Maglott, D.R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 40, D30–D35 (2012).
16. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774 (2012).
17. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive CDNA-supported gene and transcripts. Genome Biol. 7, S12 (2006).
18. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
19. Liao, Y., Smyth, G.K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 41, e108 (2013).
20. Kim, D. et al. TopHat2: accurate alignment of transcripts in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2013).
21. Li, S. et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat. Biotechnol. doi:10.1038/nbt.2972 (24 August 2014).
22. Xu, W. et al. Human transcriptome array for high-throughput clinical studies. Proc. Natl. Acad. Sci. USA 108, 3707–3712 (2011).
23. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18, 1509–1517 (2008).
24. VanGuider, H., Vrana, K. & Freeman, W. Twenty-five years of quantitative PCR for gene expression analysis. Biotechniques 44 (suppl.), 619–626 (2008).
25. Aird, D. et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12, R18 (2011).
26. Shippy, R. et al. Using RNA sample titrations to assess microarray platform performance and normalization techniques. Nat. Biotechnol. 24, 1123–1131 (2006).
27. Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-Seq. Nat. Biotechnol. 31, 46–53 (2013).
28. Pickrell, J.K., Pai, A.A., Gilad, Y. & Pritchard, J.K. Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet. 6, e1001236 (2010).
29. Dai, M. et al. Evolving gene transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. 33, e175 (2005).
30. Liu, Y. et al. Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. PLoS ONE 8, e66883 (2013).
31. Levin, J.I. et al. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. Genome Biol. 10, R115 (2009).
32. Agawal, A. et al. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics 11, 383 (2010).
33. Raghavachari, N. et al. A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. BMC Med. Genomics 5, 28 (2012).
34. Ting, T., Yu, Y., Du, T. & Shi, L. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-seq studies. Sci. China Life Sci. 56, 134–142 (2013).
35. Benjamini, Y. & Speed, T.P. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 40, e172 (2012).
36. Robinson, M.D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25 (2010).
37. Smyth, G.K. in Bioinformatics and Computational Biology Solutions Using R Biocomputing (eds) Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A. & Dudoit, S.) 397–420 (Springer, New York, 2005).
38. Huber, W., von Heydebreck, A., Siltmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18, S96–S104 (2002).
39. Wu, Z., Irizarry, R., Gentleman, R., Murillo, F.M. & Spencer, F. A model based background adjustment for oligonucleotide expression arrays. J. Amer. Stat. Assoc. 99, 909–917 (2004).
40. Hochreiter, S., Clevert, D.-A. & Obermayer, K. A new summarization method for affymetrix probe level data. Bioinformatics 22, 943–949 (2006).
41. Fasold, M., Stadler, P.F. & Binder, H. G-Stack modulated probe intensities in expression arrays—sequence corrections and signal calibration. BMC Bioinformatics 11, 207 (2010).
42. Mueckstein, U., Leparc, G.G., Posekany, A., Hofacker, I. & Kreil, D.P. Hybridization thermodynamics of NimbleGen Microarrays. BMC Bioinformatics 11, 35 (2010).
43. Sykacek, P. et al. The impact of quantitative optimization of hybridization conditions on gene expression analysis. BMC Bioinformatics 12, 73 (2011).
44. Rapaport, F. et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 14, R95 (2013).
45. Xu, J. et al. Cross-platform ultra-deep transcriptomic profiling of human reference RNA samples by RNA-Seq. Sci. Data (in the press).
These authors contributed equally to this work.
ONLINE METHODS

Study design and data The SEQC (MAQC-III) main study design is based on the well-characterized MAQC-I RNA samples: the Universal Human Reference RNA (UHRR, from ten pooled cancer cell lines, Agilent Technologies, Inc.) and the Human Brain Reference RNA (HBR, from multiple brain regions of 23 donors, Life Technologies, Inc.)39. To these, two different ERCC spike-in mixes were added34 (50 µl of ERCC mix was spiked into 2,500 µl of total RNA) to give: sample A—UHRR with ERCC spike-in mix E, and sample B—HBR with ERCC spike-in mix F. These were then combined in ratios of 3:1 and 1:3, respectively, to generate samples C and D (Fig. 1b and Supplementary Fig. 1).

Each platform vendor designated three ‘official sites’ before samples were distributed, and these are marked with a star (*) below. Data produced by the official sites were used in all of the analyses performed. In addition, data produced by the unofficial sites were incorporated in some analyses, for example, the analysis of gene detection and junction discovery as a function of read depth (Fig. 2, Supplementary Figs. 2–4, 6, 7 and 37) and the study of sensitivity, specificity and reproducibility of differential expression calls (Fig. 3, Supplementary Figs. 10–14 and 38–40).

Illumina HiSeq 2000 data were provided by six sites (Supplementary Tables 6 and 7): (i) Australian Genome Research Facility; (ii) Beijing Genomics Institute*; (iii) City of Hope; (iv) Weill Cornell Medical College*; (v) Mayo Clinic*; and (vi) Novartis, generating 100–100 nt read-pairs.

Life Technologies SOLID 5500 data were provided by four sites (Supplementary Tables 8 and 9): (i) University of Liverpool; (ii) Northwestern University*; (iii) Penn State University*; and (iv) SeqWright Inc.*, generating 51+36 nt read-pairs, except for Liverpool which applied a protocol variant giving single 76-nt reads.

All official sites created four replicate measurements of each sample A to D, and also sequenced a vendor-prepared fifth replicate (Fig. 1b). The other HiSeq 2000 sites sequenced four replicate libraries of each sample A to D. In Liverpool, one site-prepared library and one vendor-provided library of each of the samples A to D were sequenced.

For comparisons of gene-level expression profiling, samples A to D were also hybridized to a variety of commercial microarray platforms: (i) Affymetrix HuGene2.0 (one site: Stanford); (ii) Affymetrix PrimeView (one site: Stanford); (iii) Agilent 60K (one site: Boku University Vienna); and (iv) Illumina Bead arrays (two sites: City of Hope, and University of Texas Southwestern Medical Center). In addition, MAQC-1 Affymetrix HGU133Plus2.0 data from six sites were reanalyzed. Providing another independent platform, 20,801 PrimePCR Center*; and (v) City of Hope; (iv) Weill Cornell Medical College*; (v) Mayo Clinic*; and (vi) Novartis, generating 100–100 nt read-pairs.

Life Technologies SOLID 5500 data were provided by four sites (Supplementary Tables 8 and 9): (i) University of Liverpool; (ii) Northwestern University*; (iii) Penn State University*; and (iv) SeqWright Inc.*, generating 51+36 nt read-pairs, except for Liverpool which applied a protocol variant giving single 76-nt reads.

All official sites created four replicate measurements of each sample A to D, and also sequenced a vendor-prepared fifth replicate (Fig. 1b). The other HiSeq 2000 sites sequenced four replicate libraries of each sample A to D. In Liverpool, one site-prepared library and one vendor-provided library of each of the samples A to D were sequenced.

For comparisons of gene-level expression profiling, samples A to D were also hybridized to a variety of commercial microarray platforms: (i) Affymetrix HuGene2.0 (one site: Stanford); (ii) Affymetrix PrimeView (one site: Stanford); (iii) Agilent 60K (one site: Boku University Vienna); and (iv) Illumina Bead arrays (two sites: City of Hope, and University of Texas Southwestern Medical Center). In addition, MAQC-1 Affymetrix HGU133Plus2.0 data from six sites were reanalyzed. Providing another independent platform, 20,801 PrimePCR measurements were also performed, with at least ten qPCR reactions per assay for all samples A…D (see Supplementary Notes, section 3.6 for more detail). Specificity was confirmed by analyses of qPCR product lengths. This allowed the identification of nonspecific assays, identifying the target but also picking up additional unintended targets due to unexpected or unavoidable cross-reactivity (giving a qualitative validation but no meaningful quantitative read-out) and of noninformative assays, failing to pick up the target but picking up unintended targets. We provide information on RNA-seq read coverage flanking all 250 candidate junctions considered for validation in file Supplementary Data 1. Supplementary Data 2 provides the qPCR primer sequences employed, qPCR results and expression level estimates, as well as the corresponding RNA-seq expression level estimates for the 173 assays performed. Further data and results are collected as archive file Supplementary Data 3.

Data processing—assessing expression estimates. A variety of tools/pipelines to process RNA-seq data were compared (see Supplementary Protocols for pipeline parameters used):

TopHat2 std: TopHat v2.0.0 (ref. 20) + CuffDiff v2.0.0 (ref. 27).
TopHat2 G: TopHat v2.0.0 with -G parameter (providing the reference GTF file) + CuffDiff v2.0.0.
Magic: NCBI AceView MAGIC37.
BitSeq: SHRIMP2 v2.2.2 (ref. 48) + BitSeq v0.4.2 (ref. 49).
Subread: Subread 1.3.0 (ref. 19). The Subread pipeline uses the sub-junc function to identify exon-exon junctions and the featureCounts function to obtain count summaries for each gene and spike-in transcript (see Supplementary Notes, section 3.1 for more detail).

e: r-make: Cornell’s r-make pipeline incorporating STAR (http://physiology.med.cornell.edu/faculty/mason/lab/r-make/).

For LifeTech reads, all alignments were processed in color space.

Applying consistency tests based on truths built into the study design to expression levels of individual junctions, we considered the number of reads hitting a specific exon-exon junction as an indicator of expression level. Except for r-make, which provides raw read counts, each pipeline already has a built-in approach to normalization. To analyze Agilent 1M microarray data, a variance-stabilizing normalization (vsn) was used. Probe sequences-specific signals have been modeled using established methods, saturation effects detrended and outlier probes downweighted. Transcript variant expression levels have been estimated using a hierarchical Bayesian approach similar to modern methods applied for RNA-seq data analysis (see ‘Transcript quantification for Agilent high-density microarrays’ section).

CustomCDFs (v16, re-mapped to the latest AceView) were used for an analysis of the Affymetrix data (HGU133Plus2.0, PrimeView, and HuGene2.0)28, respectively covering 24,623, 17,984 and 29,879 genes. PrimeView and HuGene2.0 data were analyzed using established methods (correction for probe sequence-specific effects by gcRNA29, conservative normalization across arrays by vrn30, robust probe set summarization by affyPLM), whereas for the HGU133Plus2.0 microarrays, a combination of more recent tools that appeared to be more efficient were used (correction for probe-specific saturation effects by Hook41, conservative normalization across arrays by vsn38 and factor-based probe set summarization by FARMS46).

For Illumina Bead microarrays and Agilent 60K microarrays variance stabilization normalization (vsn) was applied.
Discovery of transcriptome complexity at high read depth. The detection and discovery of junctions was performed by Subread using data from all six HiSeq 2000 sites as well as all four SOLID sites, and compared to results by r-make using data from all six HiSeq 2000 sites.

We applied consistency tests for the truths built into the study design to junction expression levels. The sample expression levels of almost one-third of all known AceView junctions (and 38% of all detected) follow the expected titration order while also correctly yielding the expected A/B mixing ratio and show a clear differential signal as assessed by the mutual information, a measure of information content (Supplementary Fig. 37 and Supplementary Table 11). Of the well-supported new junctions, 5,189 passed these rigorous filters. Conservatively assuming a 1:3 ratio as in the AceView junctions, there may easily be three times as many new junctions, thereby adding over 15,000 likely new junctions to the already extensive AceView annotation. Furthermore, considering that we essentially required junctions to be independently detected in by SOLID, where 98% of junctions were detected by a single site (LIV) and thus corresponding to just about 1/65 of the total HiSeq 2000 sequencing volume, and detection power being about 2× lower (Supplementary Table 2), there may well be up to 5,000 × 3 × 65 × 2 = 2 million new junctions to be discovered in samples A and B alone. Even more junctions than in samples A and B were discovered in the SEQC neuroblastoma study.

Transcript quantification for Agilent high-density microarrays. Quantification of transcript expression from the probe level information was carried out using a linear mixed model independently for each gene and sample. Denoting the expression level of probe \( p \) as \( y_p \), we modeled the probe expression as the sum of effects from transcripts with a probe-match:

\[
y_p = \sum_{t=1}^{T} \delta_{t,p} x_t
\]

The Kronecker delta \( \delta_{t,p} \) is one exactly if the probe \( p \) is matching the transcript \( t \), and zero otherwise, whereas \( x_t \) denotes the unknown abundance for transcript \( t \). Further, we assume Gaussian additive and multiplicative error variances. Probe-level noise tends to exhibit a strong spatial correlation structure, which we account for by using a latent Gaussian process function \( \Phi \). We employ a squared exponential covariance function where the probe distance in transcript space is used to parameterize the covariance.

Inference is performed by maximizing the joint marginal likelihood of all considered probes given with respect to the hidden transcript abundances \( x_t \) and the noise covariance parameters. To mitigate the computational complexity of Gaussian process models (cubical scaling in the number of probes), we randomly chose probes for each gene selecting a subset of at most 700 probes, including probes falling onto junctions.

Discrete nature of RNA-seq data. With the discrete nature of RNA-seq data and considering that most analysis tools work on a log2 scale, consistent ways need to be found for dealing with unexpressed or not detected features, which are supported by zero reads. As the lowest positive expression is just a single read, a common approach is the addition of a pseudo-count (e.g., 0.5 in voom\(^2\)). An alternative well-established for microarray data analysis is the application of asinh as a variance-stabilizing transform assuming an additive-multiplicative error model. The transform is approximately linear for small values; for larger values it is well approximated by a logarithm. Another approach (natively applied, for example, in Magic\(^17\)) is to use a threshold below which measurements are considered below detection sensitivity. Here that threshold has been set to the highest minimum read count of all measurement samples adjusted for library size (the total number of reads), and this threshold is then applied consistently as a floor to all expression levels. We have applied this approach to improving consistency in our studies to the data from each pipeline and platform (and thus we were not adding the pseudo-count of 0.5 reads when using voom).

To identify genes, transcripts and junctions with clear support of sequence reads, thresholds were applied. Support was considered sufficient when at least 16, 16, or 8 at reads were observed, respectively. For Figure 2b, expression above background level as determined by the Magic pipeline was additionally required for genes.

Sensitivity, specificity and reproducibility of differential expression calls. In this part of the study the subset of 23,420 AceView genes (of the version frozen for the SEQC study) present on the MAQC-1 Affymetrix HG-U133Plus2.0 microarray was used.

As the array data were already processed and normalized with state-of-the-art methods (see ‘Data processing’ sections) no further processing was required. For RNA-seq data, weighted trimmed mean of log fold-change normalization\(^16\) improved results (data not shown), in agreement with a recent performance comparison of normalization methods\(^53\). For this normalization step, the TMM implementation provided in the Bioconductor R package edgeR\(^24\) was employed.

Several of the RNA-seq pipelines examined exploit multimapping reads. This increases power\(^11\) and, more generally, also allows the analysis of alternative transcripts. Those pipelines report expression-level estimates rather than read counts. For a uniform approach to differential expression analysis, RNA-seq data were therefore analyzed using an established approach supporting such pipelines. Precision-based weights were attached to normalized expression estimates on the log-scale to account for higher variability at low expression levels using voom\(^35\) of the limma package\(^37\). The voom function has been developed to account for different variances as a function of signal intensity. For count data, such a variation is expected by theory, whereas for expression level estimates, it is empirically justified. So, in general, we apply the voom model for expression level dependent variance to account for the different platform-specific noise characteristics as a function of the expression level (Supplementary Fig. 38).

Differential expression was then assessed for both microarray and sequencing platforms using the empirical Bayes moderated \( t \)-statistic of the limma package\(^37\). A \( P \)-value threshold of 0.01 unadjusted for multiple testing was used, as suggested in the MAQC-I study\(^10\). As the number of DEGs was similar for \( P < 0.01 \) and the \( q_{BY} \) is the Benjamini–Yekutieli adjusted false-discovery rate, downstream analysis is not qualitatively affected by this choice (Supplementary Fig. 39 versus Supplementary Fig. 40).

An estimate of the empirical eFDR was computed by comparing the number of DEGs for intrasite A versus B and intersite A versus A comparisons. For each A versus A analysis two eFDRs were calculated (using the A versus B comparison of the two sites considered in the matching A versus A comparison).

Further filters were applied in order to control the eFDR, with parameters chosen to give similar numbers of A versus B differential expression calls:

- AFX: \( \log2(\text{fold-change}) > 1 \)
- MAGIC: \( \log2(\text{fold-change}) > 1.7 \) and AveExp \( > 32\% \)
- r-make: \( \log2(\text{fold-change}) > 1.7 \) and AveExp \( > 33\% \)
- Subread: \( \log2(\text{fold-change}) > 1.7 \) and AveExp \( > 32\% \)
- BitSeq: \( \log2(\text{fold-change}) > 2 \) and AveExp \( > 19\% \)
- TopHat2 –G: \( \log2(\text{fold-change}) > 2 \) and AveExp \( > 23\% \)

When filtering for average log, expression level (AveExp), the stated fraction of weakly expressed genes to be removed also included in that percentage the genes that were not observed at all. This was to allow comparisons across different pipelines observing varying numbers of genes.

Metrics for a robust characterization of platforms, sites and data-processing options. The complementary metrics examined react differently to rescaling, shifts and other consequences of data processing. As a result, individual pipelines can show varying performance in specific assays. For a robust performance characterization, we combine the complementary metrics. For the SEQC pipelines and platforms, however, identify varying numbers of targets (see Supplementary Notes, section 3.4 for more details) With this number varying considerably, performance needs to be assessed in terms of actual counts or fractions of all genes, rather than fractions of observed genes. An increased sensitivity of some pipelines and platforms can be demonstrated by not limiting analysis only to genes observed by all pipelines and platforms.
1) We count a gene as preserving titration monotonicity, when $A > B$ and, as expected $A \geq C \geq D \geq B$ or, conversely, for $A < B$.
2) A gene is considered precise for a sample, if the standard error across technical replicates is $<10\%$.
3) A deviation $<10\%$ from the expected behavior of
   \[
   \log \frac{C}{D} = \log \left( k_1 \frac{A}{B} + (1 - k_1) \right) - \log \left( k_2 \frac{A}{B} + (1 - k_2) \right)
   \]
   where the correction $z$ of the known mixing coefficients $k_1 = 3z/(3z + 1)$ and $k_2 = z/(z + 3)$ arising out of different ratios of mRNA versus total RNA in the samples A and B has been determined by a nonlinear robust fit (nlrob) from an independent RNA-seq library (library #5). The obtained value, $1.45 \pm 0.01$ is very much in line with the experimental estimate of $1.43 \pm 0.10$. The plots and statistics shown in the paper give the same picture with either value. To ensure pipeline independence, the experimental value $1.43$ is being used.
4) For the purpose of this metric, we call a gene differentially expressed if it is significant at a Benjamini-Yekutieli corrected FDR of $5\%$ in an empirical Bayes moderated $t$-test across the expression level estimates of samples A and B (limma).
5) Finally, we calculate the mutual information of sample titration by extending the approach introduced for two state measurements\textsuperscript{43}. The mutual information between gene or transcript expression and titration requires modeling the probability of a measurement being from sample A, C, D or B, under the constraint that these labels are ordered. To avoid a dependency of our assessment on choosing mutual information of sample titration as evaluation measure, we complemented this assay with three other alternative measures. For that purpose, we evaluated the mutual information for discriminating A versus B, and the mutual information for discriminating C versus D using an established approach\textsuperscript{43}. In order to add a further measure that does not depend on modeling assumptions, for all genes and transcripts, we also calculated a nonparametric estimate of the probability that the respective measurement fulfills the order constraint which is implied by the titration experiment. All four measures are illustrated in Supplementary Figure 41 for the official HiSeq 2000 and SOLiD sites and a number of different quantification pipelines. Although the complementary measures suggest different numbers of ‘good’ transcripts and genes, they qualitatively agree with no exception on how they rank the different platforms and pipelines. This confirms that we can select the mutual information of sample titration to represent this class of information preserving measures to add an independent robust viewpoint for characterizing quantification performance in Figure 6.

Further details and information complementing the Online Methods as well as additional results are provided in Supplementary Tables 12–15 and Supplementary Figs. 42–46.

48. David, M., Dzamba, M., Lister, D., Ilie, L. & Brudno, M. SHRiMP2: Sensitive yet practical short read mapping. Bioinformatics 27, 1011–1012 (2011).
49. Glaus, P., Honkela, A. & Rattray, M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics 28, 1721–1728 (2012).
50. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014).
51. Rasmussen, C.E. Gaussian Processes for Machine Learning (MIT Press, 2006).
52. Law, C.W. et al. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 15, R29 (2014).
53. Dillies, M.-A. et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief. Bioinform. 14, 671–683 (2013).
54. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010).