## Supplementary Methods

### RNA used in the ABRF-NGS Study

Standardized, commercial RNAs were sent to multiple sites for RNA-seq library preparation using different methods.  Data were generated on five NGS platforms: Illumina HiSeq 2000/2500, Roche 454 GS FLX+, Life Technologies Ion Personal Genome Machine (PGM) and Proton, and the Pacific Biosciences (PacBio) RS.  The HiSeq 2500 was used for the libraries from site W; all other Illumina libraries were sequenced on a HiSeq 2000.

Universal Human Reference RNA (740000, Agilent Technologies) and Ambion FirstChoice Human Brain Reference RNA (AM6000, Life Technologies) were used as the primary input RNAs for this study.  These samples were labeled as MAQC samples A and B, respectively, in the MicroArray Quality Control (MAQC) experiments initiated in 2005 and summarized in Nature Biotechnology, September 2006[9].  The A and B naming convention is maintained here.  These RNA samples were selected because they are well characterized and have been used for many benchmarking studies, including a concurrent complementary RNA-seq study led by the FDA[11].

External RNA Control Consortium (ERCC) "spike-in" synthetic transcripts were added at manufacturer recommended amounts (4456739, Life Technologies) to A and B standards.  These RNAs, also developed for the 2006 MAQC study, consist of different ratios of artificially generated, poly-adenylated RNA transcripts of various lengths and combined in differing known concentrations[29]. Analyzing the ratios of these synthetic transcripts following library construction enables detection of sample preparation and platform-based biases.

The quality of the RNA samples was assessed prior to distribution to the participating laboratories, using the Agilent Bioanalyzer 2100, Nanodrop ND-1000 Spectrophotometer (Thermo Scientific), and fluorometry. All shipments were on dry ice. The samples distributed to the participating core laboratories and the libraries produced from these samples are summarized in Table 2.

### Platform protocols

All platform-specific RNA processing, library preparation and sequencing methods for Illumina, Life Technologies, Pacific Biosciences, and Roche are described in the Supplemental Online Material.

### PrimePCR RT-qPCR

PrimePCR RT-qPCR reactions were run in 384-well plates according to the manufacturer's instructions (Bio-Rad). In short, 5 µl reactions contained 1x final SsoAdvanced SYBR Green Supermix (Bio-Rad), 1x final PrimePCR assay components, and 25 ng of cDNA, and were run in a CFX384 Touch real-time PCR detection system (Bio-Rad) using standard cycling parameters. Quantification cycle (Cq) value determination was done using CFX Manager software (Bio-Rad) with autocalculated baseline and fixed threshold settings (300 relative fluorescence units). A Cq value of 35 corresponds to a single molecule of cDNA input (see below).

Two µg of each MAQC RNA sample (MAQC A and B, and 1:3 mixture samples MAQC C and D) was reverse transcribed using the iScript advanced cDNA synthesis kit for RT-qPCR (Bio-Rad) in a 20 µl reaction. Prior to reverse transcription, MAQC B RNA was DNase treated using the Heat&Run gDNA removal kit according to the manufacturer's instructions (ArcticZymes). Absence of gDNA contamination in both MAQC A and DNase treated MAQC B was verified by qPCR using 25 ng of RNA as input and DNA specific assays[46]. MAQC samples A, B, C and D were measured in parallel in the same 384-well plate, each time for 96 different assays (n=1) (according to the sample maximization run layout strategy as described[47]).

All PrimePCR assays have been extensively wet-lab validated (see Bio-Rad tech note 6262 for more details on PrimePCR assay validation and performance characteristics). In accordance with the MIQE guidelines[48], assays were evaluated for specificity, efficiency, linear dynamic range, and background signal in negative controls. At least ten qPCR reactions were performed for each assay: cDNA from reference RNA, no-template control, gDNA, and seven points from a tenfold dilution series of synthetic templates (from 20 million to 20 copies). Amplification efficiencies were calculated from the results of the dilution series. Only assays that displayed good linear performance in the 20 to 20 million copy number range (r2 >0.99) with efficiencies between 90 and 110% were considered to be of sufficient quality. The average y-intercept of the standard curves was 35, indicating that a single template molecule results in a Cq value of 35 when amplified with SsoAdvanced SYBR Green Supermix in a qPCR reaction.

**Data analysis and bioinformatics protocols**

All data, with analysis methods, are freely available at the Gene Expression Omnibus (GEO), http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46876. Additional study materials, code, scripts, and methods are available at:
http://physiology.med.cornell.edu/faculty/mason/lab/data3/sac2026/ABRF/index.html

***Sequence data preprocessing***

Sequences were aligned to the hg19 genome assembly (GRCh37). Read counts were calculated using the Rmake pipeline (http://physiology.med.cornell.edu/faculty/mason/lab/r-make/) with GENCODE v12 annotation. Read-level quantification for genes is achieved using Boost Interval

Container Library split_interval_map and a reference transcriptome in BED format following the union mode (illustrated in http://www-huber.embl.de/users/anders/HTSeq/doc/count.html). For gene expression analysis, sequences from HiSeq reads were aligned with STAR (https://code.google.com/p/rna-star/; parameters see Supplementary Table 5; v2.2.1d) and with ELAND (http://www.illumina.com/software/genome_analyzer_software.ilmn; ELAND_standalone.pl -if $input.R1.fastq.gz -if input.R2.fastq.gz -ref hg19/ --bam -it FASTQ -od . -op output -rt -l output.log; casava 1.8.2), PGM and Proton reads were aligned with TMAP (https://github.com/iontorrent/TMAP; command line: tmap map all -f hg19.fa -r <(zcat input.fastq.gz) -i fastq -Y -a 0  -o 1 -g 0 -n 5 -s output.bam stage1 map1 map2 map3; tmap.3.0.1), PacBio reads with GMAP (http://research-pub.gene.com/gmap/; gmap -D gmap_db/ -d hg19 -t 24 -f samse -n 0 input.fastq.gz; version 2013-10-04) , ILMN reads with ELAND (http://www.illumina.com/software/genome_analyzer_software.ilmn; ELAND_standalone.pl -if $input.R1.fastq.gz -if input.R2.fastq.gz -ref hg19/ --bam -it FASTQ -od . -op output -rt -l output.log; casava 1.8.2), and 454 reads with GS Reference Mapper (http://www.454.com/products/analysis-software/; GUI with "cDNA" settings). Only uniquely mapped reads were used for gene expression quantification. RNA expression levels were calculated as reads per kilobase of transcript per million mapped reads (RPKM) or, for paired-end sequencing, fragments per kilobase of transcript per million mapped reads (FPKM). Splice junction detections were generated by STAR RNA-seq aligner (parameters see Supplementary Table 5; v2.3.0e for 454, PGM, Proton, PacBio). Total junctions from 454, PGM, Proton, PacBio and HiSeq (ribo-depleted RNA and Illumina v2 kits from sites L, R, V) were used for comparison. Junction detection efficiency comparisons were normalized for read depth by using all PacBio data and subsets of data from other platforms (454: site I, HiSeq: site L-replicate1-Lane 5, PGM: site S-replicates 1-3, Proton: site S-replicate1). The resulting number of bases per platform used for this calculation ranged from 630 million to 5.451 billion bases.

***RNA-seq differential gene expression analysis***

The raw read counts were normalized by the trimmed mean of the M-values normalization method, which uses a weighted trimmed mean of the log expression ratios[49-51]. The mean-variance relationship of the counts was estimated, and the appropriate weights for each observation were computed based on their predicted variance, using *voom* from the limma package[38]. By applying the lmFit(), contrasts.fit() and eBayes() functions, also from the limma package, the log2 fold differences and standard errors were estimated by fitting a linear model for each gene, and empirical Bayes smoothing was applied for the standard errors. Benjamini and Hochberg adjustment for multiple hypothesis testing was applied at a variety of false discovery rates (FDR 0.05 or 0.01 or 0.001). Differentially expressed genes were evaluated at log2 fold change (FC) cutoffs (FC 1.5 or 2). Data from HiSeq site W, which used both ribo-depletion and poly-A library methods, was used for the comparison of different protocols from Illumina. Other platforms' data from 454 (sites C, P, I), PGM (sites H, S, P), and Proton (sites B, S) were used in the same fashion for cross-platform comparisons.

### *Expression level CV calculations*

The inter-site coefficients of variation for normalized gene expression levels were calculated on the matrix of the same sample with the same platforms from all test sites for each gene. Only genes detected by all replicates for each platform were used for CV calculation.

### *Surrogate variable analysis*

Normalized gene expression values in log2 scale were used to detect latent variables using the sva package[39]. Using the twostepsva.build() function based on the two-step algorithm of Leek and Storey[52], three latent variables were constructed. Latent variables were removed in the DEG analysis by adding each latent variable to the design matrix for limma pipeline[37].

### *RNA-seq quality metrics*

Quality metric definitions were as follows: (1) *sequencing depth:* total number of reads sequenced; (2) *mapping rate:* percentage of reads which mapped uniquely to the reference genome; (3) *sequence directionality:* the number of reads which mapped to the forward and reverse strands compared to those of the AceView gene model; (4) *nucleotide composition:* the total number of A/G/C/T sequenced at each position across the length of the read; (5) *guanine-cytosine (GC) distribution:* the number of reads with a particular %GC content; (6) *read distribution:* the fraction of the reads which mapped to either exons, 3'UTRs, 5'UTRs, introns, or intergenic regions (or the intersection of any of the categories) as defined by the AceView gene models; (7) *coverage uniformity:* the number of reads covering each nucleotide position of all genes scaled to 100 nt; (8) *base quality scores (QV):* Phred-quality scores as calculated by Illumina's HiSeq Control Software for each nucleotide position across all reads; (9) *duplication rate:* the number of reads with exactly the same sequence content. See Wang *et al*.[53] for a more thorough description of metrics; and (10) *Mismatch rate* calculation: total number of mismatches was calculated by parsing and summing up the number in NM tag in bam files; total bases of indel mismatches comes from cigar field by parsing and summing up the number in front of "I" or "D". The total number of single base mismatches was total number of mismatches subtracted by total bases of indel mismatches. The mismatches rates were calculated using the number of mismatches divided by number of mapped bases from cigar field by samtools bam check. 454 GSRM alignment's cigar field and PAC GMAP alignment were using samtools calmd to calculate the MD tag before mismatches rates calculation. PGM and PRO were using TMAP alignment without softclipping.

**TaqMan gene expression analysis**

TaqMan data for samples A, B, C, and D were obtained through the Gene Expression Omnibus database (accession number GSE5350)[54]. Data for four replicates of each sample were analyzed. Undetectable $C_T$ values ($C_T$>35 or $C_T$=0) were removed prior to normalization. The data were normalized using the HTqPCR package[55] to the average $C_T$ of POLR2A (lowest standard deviation of

$C_T$ value) by subtracting the average $C_T$ of POLR2A from each TaqMan target to give the $\log_2$ difference between endogenous control and target genes. TaqMan differential gene analysis was performed as for RNA-seq data, without the trimmed mean and *voom* transformations.

**PrimePCR RT-qPCR gene expression analysis**

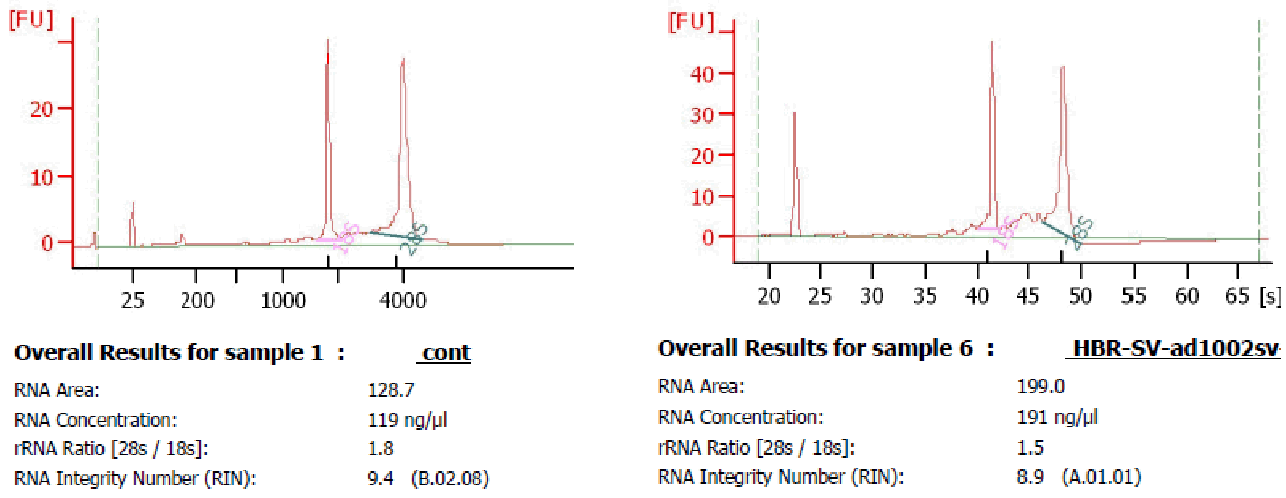
Undetectable Cq values (Cq>35 or Cq=0) were removed from data for samples A, B, C, and D. The standard deviations of the Cq values for each gene were calculated, and the gene MYSM1 exhibited the lowest standard deviation. The data were normalized by subtracting the average Cq of MYSM1 from each PrimePCR target to give the log2 difference between the endogenous control and the target genes. The normalized Cq values were then used to calculate the $R^2$ correlation to the RNA-seq data using lm() from the R stats package and summary function from the R base package.

**Comparison between RNA-seq and TaqMan data**

DEGs were validated from the Illumina RNA-seq data from each site, for six comparisons (A-B, A-C, A-D, B-C, B-D, C-D), using the DEGs from the TaqMan data. MCC (Matthews Correlation Coefficient)[40, 41] was used to compare performance metrics for external validations. Each DEG from the RNA-seq data was predicted based on its adjusted p-value and its fold difference. The determination of truth in the performances metric analysis was the detection of DEGs by the TaqMan data. Here, the true positive rate was the probability of a positive DEG result from RNA-seq given that TaqMan called the same gene differentially expressed, and the false positive rate is the probability of a positive DEG result from RNA-seq given that TaqMan did not call the gene as differentially expressed. MCC was calculated to measure how accurately RNA-seq can distinguish between DEGs and non-DEGs.

**Supplemental Methods**

RNA samples were evaluated for quality prior to distribution to the participating laboratories using the Agilent Bioanalyzer 2100 (Agilent) (Supp. Methods Fig. 1), Nanodrop ND-1000 Spectrophotometer (Thermo Scientific), and fluorometry. All shipments were on dry ice. The study coordinated shipments of library preparation and sequencing reagents so that all laboratory sites received reagents from similar manufacturing lots as determined by the vendor.



**Overall Results for sample 1 : cont**

| | |
|---|---|
| RNA Area: | 128.7 |
| RNA Concentration: | 119 ng/µl |
| rRNA Ratio [28s / 18s]: | 1.8 |
| RNA Integrity Number (RIN): | 9.4 (B.02.08) |

**Overall Results for sample 6 : HBR-SV-ad1002sv**

| | |
|---|---|
| RNA Area: | 199.0 |
| RNA Concentration: | 191 ng/µl |
| rRNA Ratio [28s / 18s]: | 1.5 |
| RNA Integrity Number (RIN): | 8.9 (A.01.01) |

**Supp. Methods Figure 1. RNA sample quality.** Agilent Bioanalyzer traces are shown for the Agilent Universal Human Reference Cell Culture RNA (MAQC A, left) and Ambion FirstChoice Human Brain Reference Tissue RNA (MAQC B, right).

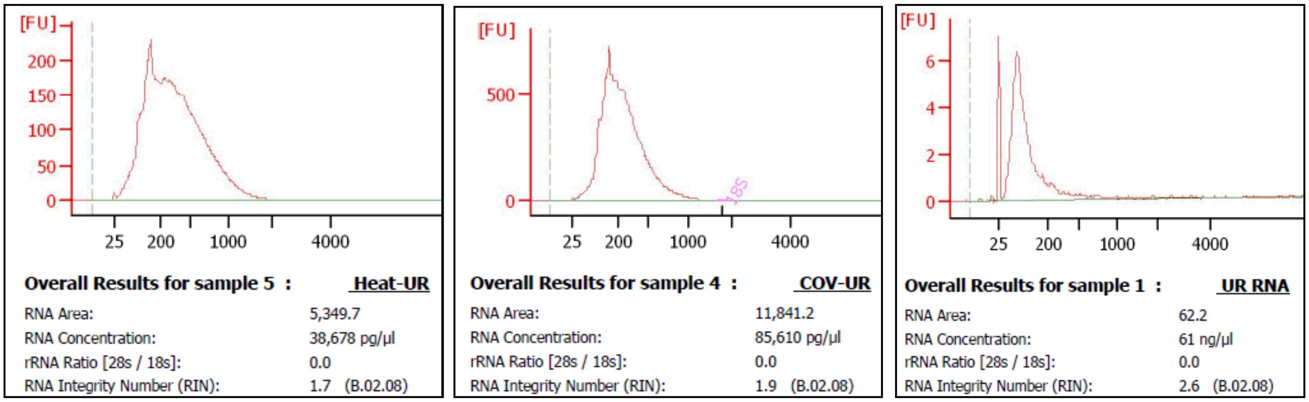
## ***Platform-Specific Protocols***

### Illumina HiSeq 2000/2500 and MiSeq

Starting Material and Enrichment

TruSeq libraries were synthesized from 2 ug of intact (RIN >7) and degraded (RIN <3) MAQC A and MAQC B RNA. Three replicate libraries of intact A and B were prepared at each of three core laboratory sites. Degraded RNA libraries were prepared at two additional core laboratory sites as listed in Table 2 (main text). Ribosomal RNA-depletion material was generated using the Ribo-Zero

Gold system (Epicentre Biotechnologies) according to the manufacturer's instructions. The preparation of three types of artificially degraded MAQC A and B RNA was performed at a single core laboratory site, using 75 ug of each RNA at a concentration of 1 ug/ul. Three techniques were used: (1) heat treatment in deionized water at 99 $^{o}$C for 10 min; (2) exposure to 1 ng/ul RNase A for a sufficient time period to result in a RIN of 3, with the RNase then neutralized with 10 ul RNase Inhibitor (RiboLock EO0381, Thermo Scientific); and (3) sonication within a Covaris S2 MicroVial for 6 x 55 sec at 5% DC, intensity 3, and 100 c/b. All resulting RNA sample degradations (i.e., RIN values) were analyzed using the Agilent 2100 Bioanalyzer (Supp. Methods Fig. 2).



| Overall Results for sample 5 : | Heat-UR |
|---|---|
| RNA Area: | 5,349.7 |
| RNA Concentration: | 38,678 pg/µl |
| rRNA Ratio [28s / 18s]: | 0.0 |
| RNA Integrity Number (RIN): | 1.7 (B.02.08) |

| Overall Results for sample 4 : | COV-UR |
|---|---|
| RNA Area: | 11,841.2 |
| RNA Concentration: | 85,610 pg/µl |
| rRNA Ratio [28s / 18s]: | 0.0 |
| RNA Integrity Number (RIN): | 1.9 (B.02.08) |

| Overall Results for sample 1 : | UR RNA |
|---|---|
| RNA Area: | 62.2 |
| RNA Concentration: | 61 ng/µl |
| rRNA Ratio [28s / 18s]: | 0.0 |
| RNA Integrity Number (RIN): | 2.6 (B.02.08) |

**Supp. Methods Figure 2. Degraded RNA sample quality.** RNA Bioanalyzer traces are shown for the three methods used to degrade the reference RNA. From left to right: heat treatment, sonication, and RNase A treatment.

Library Construction and Sequencing

Following ribo-depletion, all recovered RNA was processed using the Illumina TruSeq RNA Sample Preparation Kit v2 protocol at the "elute-fragment-prime" step, followed by the standard TruSeq protocol. Completed libraries were evaluated by DNA quantitation and Bioanalyzer analysis, and then submitted to a single core laboratory site for sequencing. Sequencing libraries were constructed with barcodes to allow multiplexing of 12 samples per lane, pooled to target 200 million clusters per channel and 100 million reads per library, and distributed over multiple channels of three flow cells to normalize for lane and run variability. Sequencing was carried out on Illumina HiSeq 2000 and 2500 instruments using protocols HCS 1.5.15.1 and RTA 1.13.48 and paired-end 50 bp reads. One of the replicate libraries for intact MAQC B was also sequenced on a MiSeq instrument using the 4nM protocol (v2.2.0.2) targeting 15 million clusters with paired-end 250 bp reads. The recently released TruSeq RNA Sample Preparation Kit v3 differs from v2 by including ribosomal RNA depletion and preserves cDNA strand orientation. Lab site W compared polyA enriched and ribo-depleted libraries constructed using the v3 kit.

**Life Technologies Ion Torrent PGM**

Starting Material and Enrichment

Four different Ion Torrent PGM libraries were constructed at three core laboratory sites using the MAQC A, MAQC B, ERCC 1, and ERCC 2 RNAs. Five micrograms of each MAQC RNA was enriched for polyA RNA (MRRK1010, MPG Kit, PureBiotech) using the recommended Life Technologies Ion protocol for Transcriptome Profiling of Low-Input RNA Samples (April 2011). The MPG-Streptavidin was prepared from 100 ug (10 ul) of the complex and resuspended in 5 ul of Release Solution instead of 20 ul. This process was repeated as a second round of enrichment to further deplete rRNA from the samples. The resulting RNA was assessed for yield and purity using an Agilent 2100 Bioanalyzer PicoChip. No enrichments were performed for the ERCC samples.

Library Construction

Whole transcriptome library preparation was performed using 5-10 ng of fragmented enriched polyA RNA according to the manufacturer’s protocol (Ion Total RNA-Seq Kit V2 protocol #4476286B Life Technologies). Size selection of a 315 bp product was performed using a standard Pippin prep protocol (Sage Science) followed by purification with AMPure beads (Beckman-Coulter Genomics).

The synthetic ERCC 1 and ERCC 2 control RNA library construction was performed directly from 30 ng of the non-polyA enriched sample.

Template Preparation and Sequencing

Emulsion PCR was performed using the One Touch system (Life Technologies). Beads were prepared from 70-100 million copies using the One Touch 200 Template Kit v2 #4471263. Some libraries were prepared for 70 to 100 million copies and others using the standard 210 million copies as stated in the RNA-SEQ protocol #4476286B. Sequencing was conducted using an Ion PGM 200 sequencing kit (#4474004) on the 318 Ion chip. Data were collected using the Torrent Suite v3.0 software.

**Life Technologies Proton**

Starting Material, Enrichment and Library Construction

Libraries were prepared from 1 ug of MAQC A, B, C, and D RNA containing ERCC controls using either polyA enrichment (as described for PGM). After enrichment, 8-9 ng of polyA mRNA was used for ligation reactions. The size selection step was adjusted to 220 bp using a standard Pippin Prep protocol, generating library competent molecules with a template insert size of approximately 150 bp. The Ion Total RNA-seq Kit v2 (4476286 Rev D, Life Technologies) was used to prepare the MAQC libraries for sequencing. The resulting material was quantified using the Agilent Bioanalyzer High Sensitivity Chip.

Ion Template Preparation and Sequencing

Emulsion PCR was performed using the One Touch 2 (OT2) system following the Ion P1 Template OT2 200 protocol (Life Tech 0007488 Rev2.0) by using 315-615 million DNA molecules post-library preparation. Enriched spheres were quantified and approximately 400 to 800 million spheres were recovered. P1 Chips were loaded according to the spinning protocol and sequencing was performed using the Proton 200 sequencing kit (MAN0007491 Rev 3.0). Base calls were collected with Torrent Suite using v3.4.1 software.

**Roche 454 GS FLX+**

Starting Material and Enrichment

The MAQC A and MAQC B RNAs were subjected to two rounds of polyA enrichment at a single laboratory site before distribution to the other 454 data generation core laboratories. Each RNA sample was enriched using the Oligotex mRNA Mini Kit (Qiagen), starting from 60 µg of RNA, according to manufacturer's instructions, using the spin column <0.25 mg method. A second enrichment step was performed following step 5 of the protocol, according to the manufacturer's instructions. Final elution was performed twice using 50 µl of 700C OEB elution buffer. The resulting enriched RNA was quantified by Nanodrop spectrophotometry and evaluated on the Agilent Bioanalyzer 2100 using an RNA PicoChip.

Library Construction

Library synthesis and sequencing was performed with MAQC A and MAQC B samples at three core laboratory sites. Each site constructed one cDNA library from each of the polyA-enriched RNA samples. Enriched RNA (200 ng) was reduced to 19 µl in a vacuum centrifuge at 60 $^{o}$C, followed by library construction using the Roche cDNA Rapid Library Preparation Method Manual XL+ (May 2011) with the following modifications: (1) RNA Fragmentation Reagent kit (AM#8740, Life Technologies) was used in place of the RNA Fragmentation Solution; (2) all magnetic particle concentrator (MPC) pelleting steps were held for 2 minutes; (3) Roche rapid library multiplex identifier (RL-MIDs) adaptors were used for the adaptor ligation step; (4) the final libraries were quantified using a Qubit fluorometer (Life Technologies) and average fragment sizes were determined by analyzing 1 µl of the libraries on the Agilent Bioanalyzer 2100 using a High-Sensitivity DNA LabChip; and (5) the library concentrations were determined using the average fragment size from the

Bioanalyzer analysis.  Final samples were diluted to 1x10$^8$ molecules/µl in Tris-HCl pH 8 buffer with 0.001% Tween-20.

Template Enrichment and Sequencing

Libraries were diluted to 1x10$^6$ molecules/µl for sequencing.  Emulsion-based clonal amplification and sequencing on the Roche 454 Genome Sequencer FLX+ was performed according to the manufacturer's instructions (454 Life Sciences).  Each library was sequenced on one full PicoTiterPlate (PTP) per laboratory site.  An additional PTP per library was sequenced at one of the laboratories for a total of four PTPs per MAQC sample.  Sequencing was done using the Roche XL+ sequencing kit with software version 2.6 or 2.8 with Flow Pattern A.  Signal processing and base calling were performed using the bundled 454 Data Analysis Software (v.2.6 and 2.8).

## Pacific Biosciences (PacBio) RS

The RNA-seq methods used in this study for full length cDNA sequencing on the Pacific Biosciences RS are an early access protocol provided to the ABRF NGS consortium; refinements to the protocol are under development by the vendor.

Starting Material

MAQC A and MAQC B RNA samples were used to generate PacBio libraries.  The polyA+ fraction was purified from total RNA using Invitrogen Dynabeads Oligo(dT)25, according to the manufacturer's protocol (61002, Life Technologies).  MAQC A RNA (100 µg) was mixed with 1.33 mg of washed Dynabeads, and 200 µg MAQC B RNA was mixed with 2.66 mg of washed Dynabeads, collected by magnetic precipitation, washed as directed, and eluted into 27 µl of 10 mM Tris-HCl.  The amount and purity of the polyA+ RNA was assessed using an Agilent 2100 Bioanalyzer RNA NanoChip.

cDNA Synthesis

Full-length, double-stranded cDNA was synthesized from the polyA+ mRNA using the first five steps of the Invitrogen SuperScript Full Length cDNA Library Construction Kit II (A13268, Life Technologies); which included: (1) first strand cDNA synthesis; (2) RNase I treatment; (3) 5'G Cap-Antibody selection; (4) second strand cDNA synthesis; and (5) cDNA size fractionation by Sephacryl column chromatography and precipitation. The cDNA was amplified with limited PCR using Phusion Hot Start Flex DNA Polymerase (M0535L, New England Biolabs), including PCR primers adapted from the 3' oligo-dT primer used for first-strand cDNA synthesis and the 5' adaptor used for second-strand cDNA synthesis, as described in the Invitrogen Superscript manual (primer sequences: pGGG ACA ACT TTG TCA AAG AAA and pTCG TCG GGG ACA ACT TTG TAC, respectively). The PCR amplification profile was 98 °C for 2 min, 14 cycles x (98 °C for 0.5 min, 64 °C for 0.5 min, and 72 °C for 4 min) and a final extension at 72 °C for 4 min. PCR-amplified cDNA was purified using 0.6x volume of Agencourt AMPure PB Beads (Beckman Coulter, Life Sciences Division), as specified by the supplier (Pacific Biosciences). The purified cDNA was recovered in 50 µl of TE buffer.

Total cDNA was divided into three MW size classes of 1-2 kb, 2-3 kb, and 3-8 kb using SYBR green and 0.8% agarose gel size fractionation, and recovered using Qiagen QIAquick Gel Extraction (Qiagen). Each cDNA size class was re-amplified using PCR conditions identical to those listed above. In order to prepare enough cDNA for multiple sequencing library constructions, a total of eight 100 µl PCR reactions were performed in parallel for each of the three cDNA size classes. The resulting PCR product was purified using 0.6x volume of AMPure PB beads as described above. Purified cDNA was quantified by a fluorometric Qubit assay and evaluated using an Agilent 2100 Bioanalyzer DNA 12000 chip.

## Library Construction and Sequencing

Each of the three size-fractionated cDNA pools from the MAQC A and MAQC B samples were distributed to three core laboratory sites for library preparation and data generation, resulting in a total of 18 libraries. SMRTbell libraries were prepared from 0.5 and 1.0 µg of each cDNA size class according to the PacBio Large Insert Template Library Prep Kit. Double stranded cDNA was subjected to DNA damage repair, end repair, and blunt-end ligation to hairpin adaptors. Incomplete SMRTbell templates were degraded with a combination of Exonuclease III and Exonuclease VII. Intact cDNA SMRTbells were purified by three sequential AMPure PB purifications. Sequencing primers were annealed to the SMRTbell templates and subsequently bound to the sequencing polymerase using the Pacific Biosciences DNA/Polymerase v 2.0 binding kit, following manufacturer's instructions. The samples were sequenced on the PacBio RS using "C2" chemistry with SMRTcell loading via diffusion. The data collection times were adjusted per cDNA size bin. For size bins less than 2 kb, the 2 x 45 minute movie protocol was used, while bins at or above 2 kb used a 1 x 90 minute movie protocol.

Platform specific library preparation parameters for all the participating sites are summarized in Table 1 (main text).

**Data used in each section**

*Figure 2:*

Samples A-D, intact and degraded, sequenced at all sites.

*Figure 3:*

3a-e,g. samples A-B for all sites, except for ILMN, sites L,R,V were used here.

3f. 454: site I all replicates; ILMN: site L, sample A-1_D1DJ4ACXX_CGATGT_L005_R1_001 and sample B-1_D1DJ4ACXX_GCCAAT_L005_R1_001; PAC: all data; PGM: site S, replicates 1-3; PRO: site S, replicate 1.

*Figure 4:*

4a. Sample B was used.

POLYA: site W replicate 1; RIBO: site W replicate 1; PAC: site H replicate 1; 454: site I replicate 1;

PGM: site S replicate 1; PRO: site S replicate 1; MISEQ: site W replicate 1.

4b. sample A and B for data in PRO, PGM, 454, ILMN site W all data.

*Figure 5:*

ILMN site W all data.