ion torrent
by life technologies™

Transcriptome sequencing using the Ion Proton™ System

Key findings of transcriptome sequencing using the Ion Proton™ System and Microarray Quality Control (MAQC) samples

- **Highly correlated** — differentially expressed genes (DEGs) identified using the Ion Proton™ System were highly correlated with MAQC array and qPCR data
- **Improved discovery** — exceeds microarray sensitivity with enhanced discovery of DEGs
- **Flexible** — scalable chips facilitate adjustment of sequencing read depth to determine the sensitivity of transcript detection or allow multiplexing of samples for differential gene expression studies

Transcriptome sequencing provides fundamental insights into how genomes are organized and regulated

Understanding the structure, function, and organization of information within genomes is central to basic and translational research. With the advent of deep-sequencing technologies, it has become increasingly evident that transcription is an intricate and dynamic process involving a variety of RNA species, including short RNAs, polyadenylated RNA, and nonadenylated transcripts. Transcriptome sequencing is the most complete yet most cost-efficient method for identifying and quantifying RNA from multiple types of starting material. With the choice of two chips, the Ion PI™ Chip and the Ion PII™ Chip,* sequencing coverage depth can be adjusted to determine the sensitivity of detection for low-abundance transcripts and rare RNA types or to multiplex samples for differential gene expression studies (Table 1).

In contrast to microarray analyses that require *a priori* knowledge of transcripts in a cell for targets to be present on the microarray, transcriptome sequencing provides a hypothesis-neutral approach to interrogating all the transcriptional content of a genome. Compared to microarrays, transcriptome sequencing provides an unbiased digital readout with a greater quantitative linear dynamic range and improved detection at the extremes of the quantitation spectrum (Figures 1 and 2 in the supplementary information). Sequencing the transcriptome facilitates the identification of alternative splicing and transcript isoforms as well as aiding gene discovery. Further, quantification of differential gene expression enables the correlation of gene expression with phenotypic information.

Table 1. Transcriptome sequencing roadmap for the Ion Proton™ System using the Ion PI™ Chip and the Ion PII™ Chip.*

	Ion PI™ Chip	Ion PII™ Chip*
Aligned reads per run	60–80 million	240–320 million
Transcriptomes per run	1–4	3–16

Transcriptome sequencing results

Total RNA from the Ambion® FirstChoice® Human Brain Reference (HBRR) and Stratagene® Universal Human Reference (UHRR), spiked with the Ambion® ERCC RNA Spike-In Mix, were used to assess the performance of the Ion Proton™ System. The HBRR and UHRR samples were used in the Microarray Quality Control (MAQC) study, thus allowing comparison of Ion Proton™ transcriptome sequencing to all commercially available microarray platforms included in the MAQC publication (Shi et al., 2006, supplementary information).

life technologies™

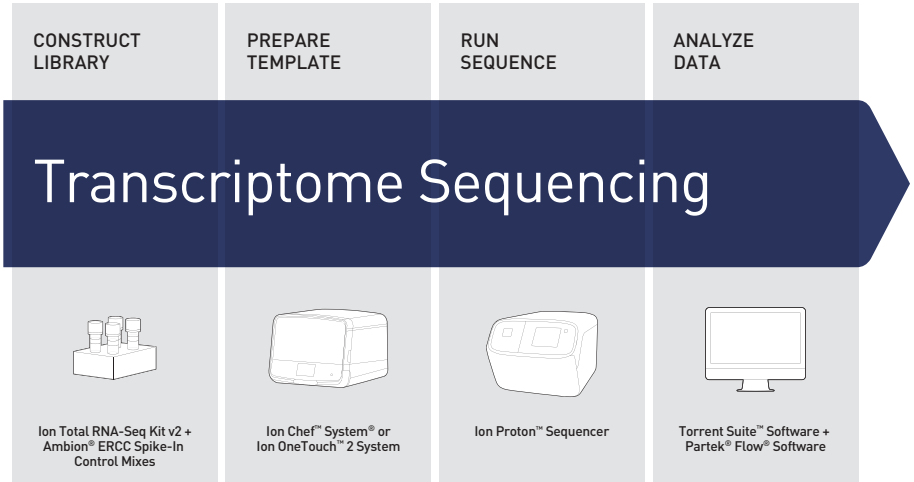

Figure 1. The transcriptome workflow for the Ion Proton™ System. Life Technologies supplies an easy-to-implement, cost-effective, scalable transcriptome sequencing workflow for the Ion Proton™ System with a rapid <24 hour workflow from library (which could include 2–8 barcoded samples) to primary data analysis results. Following library construction and template preparation, sequencing runs are completed in just 4 hours. Products for transcriptome sequencing include the Ion Total RNA-Seq Kit v2, RiboMinus™ Eukaryote Kit v2, and the Ambion® ERCC RNA Spike-In Mix. The Partek® Flow® software package provides a simple and optimized transcriptome workflow that minimizes the data analysis time (~5 hours) by providing comprehensive tools and auto-analysis directly from the Ion Proton™ System.

Highly expressed and uninformative rRNA species were removed with the Ambion® RiboMinus™ Eukaryote Kit v2 followed by the creation of strand-specific libraries using the Ion Total RNA-Seq Kit v2 (Figure 1). Maintenance of strandedness allows the discrimination of sense and antisense transcription. Sequencing was performed using the Ion Proton™ instrument and the Ion Proton™ I Sequencing Kit.

Using the Ion PI™ Chip, three technical replicates for UHRR resulted in 57.7–64.8 million preprocessed reads that were aligned using a two-step alignment method to reference sequence hg19, yielding 89–93% of reads mapped to the human genome (Table 2).

Ion Proton™ transcriptome data can be easily processed using either the open source methods described at the Ion Community (http://ioncommunity.lifetechnologies.com/docs/DOC-7062) or the Partek® Flow® v2.2 software package (Figure 3, supplementary information, and http://info.partek.com/IonTorrent_RNA-Seq).

Coordinates that correspond to annotated RefSeq regions were determined for the three UHRR technical replicates, with a very high proportion—19,618–19,742 of known RefSeq genes—detected at ≥3 read counts, with 36–37% of bases aligned to RefSeq exons and only 2% aligning to rRNA species, indicating successful depletion of uninformative transcripts (Table 2). All transcriptome technical replicates were highly correlated, with a Pearson correlation coefficient (R) of >0.99 for UHRR samples and a mean correlation coefficient of 0.997 for all pairwise comparisons (Figure 2, supplementary information).

Depth of transcript detection

With a random sampling strategy of transcriptome sequencing, most reads will identify abundant transcripts, with an increasing number of reads needed to identify lower-abundance transcripts and low levels of gene expression. For parity with microarrays using polyadenylated samples, it is estimated that transcriptome sequencing would require 20–25 million reads/sample for differential expression and gene-level quantification (The ENCODE Consortium, 2011, supplementary information), while for transcript-level quantification and alternate splicing the recommendation is 40 million reads/sample (Toung et al., 2011, supplementary information), and sensitive detection of low-abundance transcripts will require >100 million reads/sample (Toung et al., 2011, supplementary information).

The near saturation of the protein-coding transcripts can be confidently detected at moderate sequencing read depth, while extreme read depth benefits the detection of noncoding and low-expression transcripts of mainly putative regulatory function (Tarazona et al., 2011, supplementary information). The increased sequencing depth comes at the cost of increasing the noise and detection of off-target transcripts that will adversely influence differential expression analysis. Thus, a reasonable balance of 30–40 million aligned reads per sample is recommended for the detection of novel transcripts, to not unduly increase the noise level of a transcriptome sequencing dataset.

To determine the optimal read depth, thresholds of 3, 5, and 10 read counts were used to evaluate transcript detection sensitivity as a function of mapped reads for a UHRR transcriptome sequencing run (Figure 4, supplementary information). The saturation curves indicate that, at all read counts, saturation in the discovery of new genes was not reached at 68 million reads. At a read count threshold of ≥10, the new detection rate (NDR) of genes (the number of additional transcripts per 1 million reads) was 53, at 40 million mapped reads. However, the NDR at 50 million reads was nearly half that observed at 40 million, with the addition of 18 million reads resulting in only a 4% gain in gene detection. The data indicate that the rate of new gene detection decreases considerably after 40 million reads, at which point >18,900 genes were detected at 3 read counts. Thus, the Ion Proton™ platform allows the flexibility to accommodate studies with sample size variability for differential expression analysis or the detection for low-abundance transcripts and rare RNA types, with the availability of the the Ion PII™ Chip.*

Strong correlation of Ion Proton™ transcriptome data with MAQC array and qPCR data

Differentially expressed genes (DEGs) between HBRR and UHRR samples were determined, and inter-platform concordance was used to illustrate the sensitivity of Ion Proton™ transcriptome

Table 2. Transcriptome sequencing results on the Ion Proton™ System using the Ion PI™ Chip. Reads, preprocessed with cutadapt software v0.9.5 to remove reads shorter than 16 bp, were mapped to the human genome (Hg19) and to ERCC reference sequences using a two-step alignment method with TopHat2 software (v2.0.5), and the unmapped reads were aligned with Bowtie2 software (v2.0.0.7 with a very sensitive local option). The results of TopHat2 and Bowtie2 alignments were merged into the final bam file using the Picard MergeSamFiles module (v1.75). Following mapping, the HTSeq package (v0.5.3p9) with the RefSeq gene model was used to assess coverage of known exon genomic coordinates of 22,335 RefSeq genes.

Sample	Replicate	Raw reads	Mapped genome reads	Percent of genome-mapped reads	Number of RefSeq genes covered at ≥3x	Percent of bases aligned to exons	Percent of bases aligned to rRNA species
UHRR	1	57,791,379	52,576,390	90.98	19,618	36.2	2.0
UHRR	2	64,770,307	60,210,866	92.96	19,575	36.5	2.0
UHRR	3	57,669,610	51,305,437	88.96	19,742	37.2	1.9

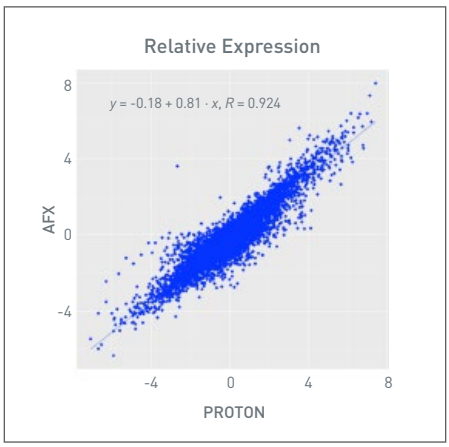

Figure 2A. Differentially expressed genes (DEGs) between HBRR and UHRR samples were determined with the Bioconductor DESeq package (v1.10.1). Scatter plot comparison of $\log_2$ (HBRR/UHRR) ratios from the Ion Proton™ System and MAQC microarray expression data. A total of 17,279 RefSeq genes were compared, with 9,127 genes having detectable expression on both platforms. Data plotted here for transcriptome sequencing used 35 million mapped reads, with a Pearson correlation coefficient (R) of 0.92, indicating the datasets are highly correlated.

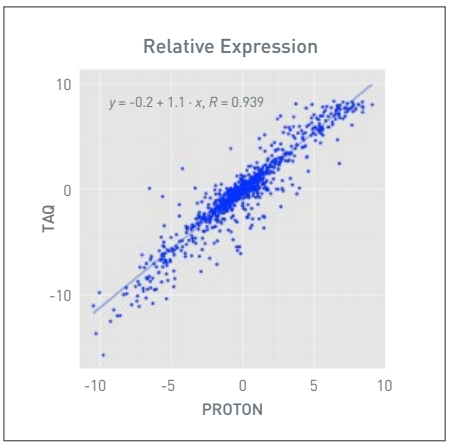

Figure 2B. Differentially expressed genes (DEGs) between HBRR and UHRR samples were determined with the Bioconductor DESeq package (v1.10.1). Scatter plot comparison of $\log_2$ (HBRR/UHRR) ratios from the Ion Proton™ System and MAQC qPCR data. Differential expression for 950 genes was compared for the two platforms with the transcriptome sequencing dataset of 35 million mapped reads. The Pearson correlation coefficient (R) was 0.94, demonstrating that the qPCR and transcriptome datasets are highly correlated.

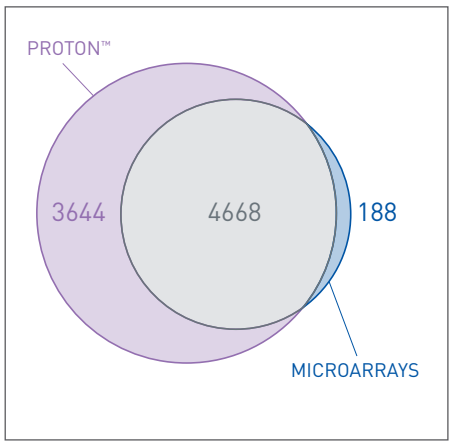

Figure 3. Detection of significantly differentially expressed genes (DEGs) by transcriptome sequencing. Venn diagram demonstrating the concordance of DEGs (≥2-fold change between HBRR and UHRR) between MAQC microarray data and transcriptome data at mapped reads of 35 million.

sequencing. By comparing DEGs, the relative gene expression as determined by transcriptome sequencing on the Ion Proton™ System could be compared to microarrays and qPCR. For optimal detection of DEGs by transcriptome sequencing, the fold change between HBRR and UHRR was calculated at 35 million mapped reads. For a gene to be considered significantly differentially expressed, the fold change between UHRR and HBRR samples had to be ≥2.

A total of 17,279 RefSeq genes were shared between the MAQC microarray and transcriptome sequencing datasets; of these, 9,127 showed detectable expression in both datasets. A scatter plot comparing the fold change demonstrates that the datasets were strongly correlated, with a Pearson correlation (R) of 0.92 between transcriptome count data and microarrays (Figure 2A). A fold change comparison of transcriptome data to qPCR data from the MAQC study (a set of 950 TaqMan® gene assays) also showed high correlation, with a Pearson correlation coefficient (R) of 0.94 (Figure 2B).

A Venn diagram illustrates that, out of 8,500 significant DEGs with a fold change of ≥2, 55% (n = 4,668) of transcriptome DEGs were concordant with microarray DEGs (Figure 3, while 43% (n = 3,644) of DEGs were not concordant between the two platforms and were detected only by transcriptome sequencing. In contrast, only 2% (n = 188 genes) showed differential expression by the use of arrays but not by transcriptome sequencing. To confirm the detection of additional DEGs by transcriptome sequencing, 128 TaqMan® gene assays were used to validate that a subset of DEGs not detected by microarray were true, significant DEGs. The fold change ($\log_2$ scaled) correlation was 0.85 between transcriptome sequencing and the 128 TaqMan® gene assays (Figure 5, supplementary information). The results indicate that transcriptome sequencing shows improved detection of DEGs compared to microarrays.

Figure 4. Identification of splice isoforms using transcriptome sequencing and the Ion Proton™ System. Shown is a 12.3 kb region of the *SHC1* gene showing alternate splicing in the UHRR sample.

Splice variant detection

Using the transcriptome-specific workflows in Partek® Flow® software package v2.2, the transcriptome data were assessed for alternate splicing in the UHRR samples. A number of alternatively spliced mRNAs have been identified in the UHRR control (Brosseau et al., 2010, supplementary information). An example of splicing isoforms is shown for the *SHC1* gene (Figure 4). The *SHC1* gene encodes three main isoforms—p46Shc, p52Shc, and p66Shc—that differ in length due to use of two alternative promoters (Luzi et al., 2000, supplementary information). The p66Shc form contains the entire p46Shc and p52Shc sequence and an additional amino-terminal proline-rich CH2 domain, and has been implicated in lifespan determination, cellular response to oxidative stress, and apoptosis. p46Shc and p52Shc are thought to be involved in Ras signaling pathway activation and differ in size due to a difference in initiation codon usage and in subcellular localization, with p46Shc targeted to the mitochondrial matrix.

Conclusions

Ion Proton™ semiconductor transcriptome sequencing exceeded microarray sensitivity using MAQC samples for comparison, and detected expression levels of DEGs were highly correlated with results from microarrays and qPCR. The Ion Proton™ System, combined with Ambion® RNA kits, offers fast, flexible, and high-quality transcriptome sequencing that is scalable with research needs, at an affordable price. For instance, one to two transcriptomes per run are possible using the Ion PI™ Chip, and higher multiplexing or rare-transcript discovery will be possible using the Ion PII™ Chip.* The combination of affordable instrument pricing with scalable chips allows the easy implementation of transcriptome sequencing for studies with sample size variability or requiring experimental fine-tuning for the detection for low-abundance transcripts and rare RNA types.

Explore the transcriptome at lifetechnologies.com/iontranscriptome

Download supplementary information at lifetechnologies.com/iontranscriptomesuppinfo

For Research Use Only. Not for use in diagnostic procedures.

* The content provided herein may relate to products that have not been officially released and is subject to change without notice.

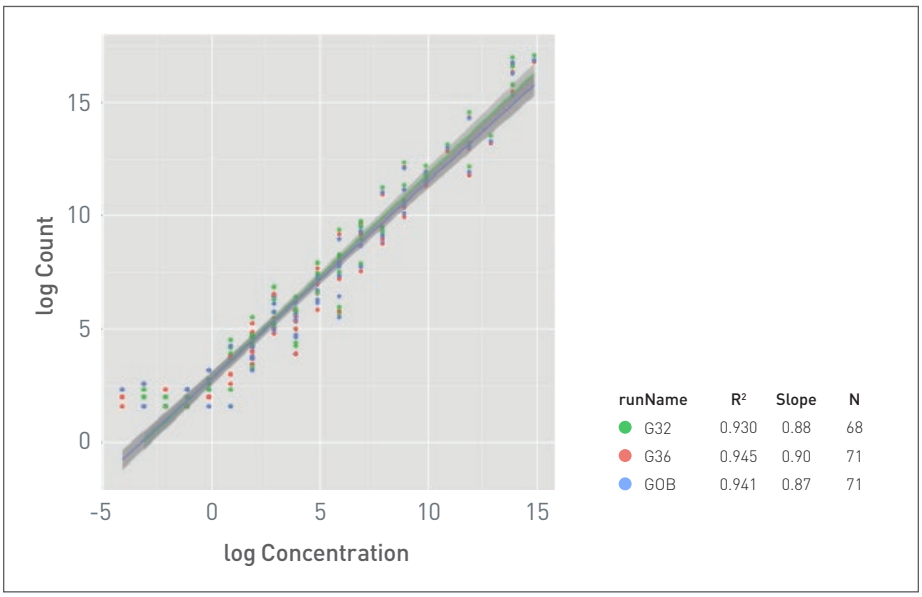

Figure 1. Scatter plot analyses of External RNA Controls Consortium (ERCC) spike-in controls. Read counts on the y-axis refer to total ERCC aligned reads; the x-axis denotes the relative concentration of each ERCC transcript in the pool that was spiked into each sample. Read counts for the 92 external ERCC spike-in transcripts were used to evaluate dynamic range, sensitivity, and variability of the transcriptome sequencing runs. The ERCC transcripts are polyadenylated, unlabeled RNAs that have been certified and tested by the National Institute of Standards and Technology (NIST) as controls to test sample RNA for performance and to assess sources of variability. ERCC transcript lengths range between 250 and 2,000 nucleotides and have been balanced for GC content to closely represent characteristics of endogenous eukaryotic mRNAs. The ERCC pool of transcripts is configured in known titrations designed to represent a large dynamic range of expression levels. Using this information, counts of reads mapping to each ERCC transcript were compared using a linear regression analysis. Excellent sensitivity was observed with a strong dose-response correlation, as assessed by an R^2 metric ranging from 0.930 to 0.945 with good variability as determined by slope (m) values of 0.87 to 0.90.

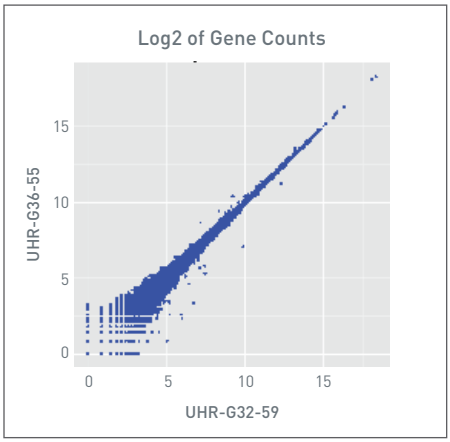

Figure 2. Correlation of technical replicates from transcriptome-mapped reads using the Ion Proton™ System. Scatter plot comparisons of $\log_2$ gene counts for two UHRR replicates from the RefSeq set are shown, with all pairwise comparisons between technical replicates demonstrating a Pearson correlation coefficient (R) of >0.99 and with a mean correlation coefficient of 0.997 for all pairwise comparisons. The scatter plot illustrates a wide dynamic range spanning 6 $\log_{10}$ units (2^{20}), more than necessary to measure the full range of expression in cells.

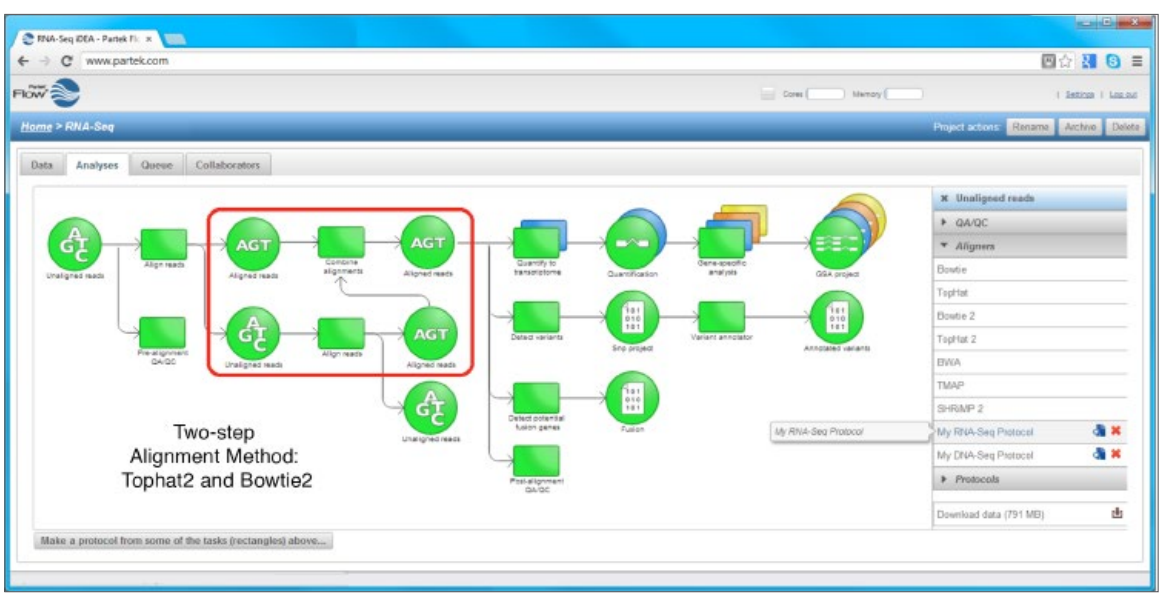

Figure 3. Partek® Flow® software supports two-step alignment of Ion Proton™ transcriptome data for an integrated data analysis workflow.

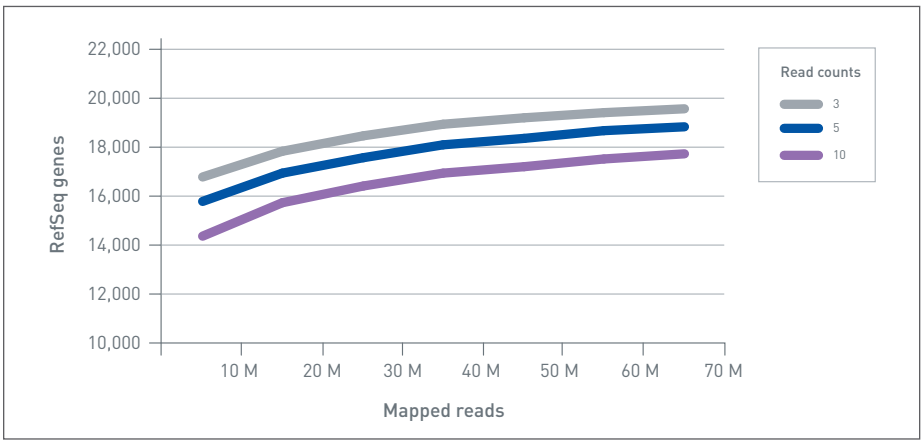

Figure 4. Gene detection in a UHRR transcriptome sequencing sample as a function of cumulative mapped reads. Solid lines indicate gene level detection as counts from mapped reads are accumulated, with detection level thresholds for transcriptome data shown at 3, 5, and 10 read counts per gene. At a threshold of ≥10 read counts, 40 million mapped reads resulted in a new discovery rate (NDR) of 53 new genes per million mapped reads, with an additional 18 million reads resulting in only a 4% gain in gene detection and an NDR of 25 new genes per million reads.

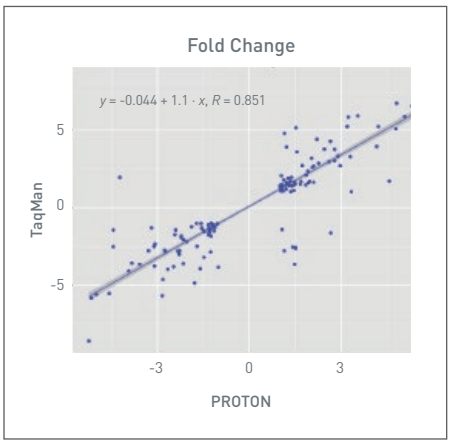

Figure 5. Scatter plot comparison of $\log_2$ (HBRR/UHRR) ratios from the Ion Proton™ System and MAQC qPCR data. Differential expression for a subset of 128 genes selected from the 3,644 DEGs detected by transcriptome sequencing only was compared between platforms. The Pearson correlation coefficient (R) was 0.85, demonstrating that the qPCR and transcriptome results are highly correlated.

References

Brosseau Lucier JF, Lapointe E et al. (2010) High-throughput quantification of splicing isoforms. *RNA* 16(2), 442–449.

Luzi L, Confalonieri S, Di Fiore PP et al. (2000) Evolution of Shc functions from nematode to human. *Curr Opin Genet Dev* 10(6), 668–674.

MAQC Consortium, Shi L, Reid LH et al. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 24(9), 1151–1161.

Tarazona S, Garcia-Alcalde F, Dopazo J et al. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Res* 21(12), 2213–2223.

The ENCODE Consortium. Standards, Guidelines and Best Practices for RNA-Seq V1.0. 1.0. 6-1-2011.

Toung JM, Morley M, Li M et al. (2011). RNA-sequence analysis of human B-cells. *Genome Res* 21(6), 991–998.

Explore the transcriptome at lifetechnologies.com/iontranscriptome

For Research Use Only. Not for use in diagnostic procedures.

* The content provided herein may relate to products that have not been officially released and is subject to change without notice.