RNA-Seq Data Comparison with Gene Expression Microarrays

A cross-platform comparison of differential gene expression analysis

Introduction

RNA-Seq is a powerful sequencing-based method that enables researchers to discover, profile, and quantify RNA transcripts across the entire transcriptome. Because the method does not require probes or primers, the generated data are completely unbiased, allowing for hypothesis-free experimental design. The ability to perform this type of analysis provides researchers a powerful tool for transcript discovery applications that are not possible using traditional microarray-based methods[1]. Beyond gene expression analysis, RNA-Seq can identify novel transcripts, novel isoforms, alternative splice sites, allele-specific expression, and rare transcripts in a single experiment.

Unlike microarrays, which measure continuous probe intensities, RNA-Seq quantifies discreet, digital sequencing read counts aligned to a particular sequence. The digital nature of this process supports an unlimited dynamic range, which enables researchers to quantify RNA activity at much higher resolution, important for capturing subtle gene expression changes associated with biological processes.

RNA-Seq offers several other advantages over microarrays (Table 1). While standard microarray probes only cover ~20% of a gene on average, capturing only a portion of the biologically relevant data, RNA-Seq can profile the entire transcript. The sequencing data can also be reanalyzed as novel exons are discovered, whereas the sample would have to be rerun on a microarray with updated probes.

The study presented in this white paper examines microarray and RNA-Seq data, comparing the ability of each platform to detect and quantify differential gene expression across two well-annotated samples. The results demonstrated that RNA-Seq and the microarray detected the same differentially expressed genes with high correlation. The strong data correlation between platforms is important, as it enables researchers to leverage legacy data when transitioning from a microarray to a sequencing platform. However, in addition to the detecting most of the same genes as the array, RNA-Seq also identified significantly more genes as being differentially expressed genes that were not identified by the array, exemplifying the superior sensitivity of sequencing technology.

A follow-up analysis of down-sampled sequencing reads showed that RNA-Seq sensitivity can be tuned down, using lower sequencing read depths, to an equivalent level as the microarray. This allows researchers to reduce per-sample costs, while still maintaining equivalent performance as gene expression microarrays.

Table 1: Comparison of RNA-Seq technology with expression microarrays

Application	RNA-Seq	Microarray
High run-to-run reproducibility	Yes	Yes
Dynamic range comparable to actual transcript abundances within cells	Yes	No
Able to detect alternative splice sites and novel isoforms	Yes	No
De novo analysis of samples without a reference genome	Yes	No
Re-analyzable data	Yes	No

Methods

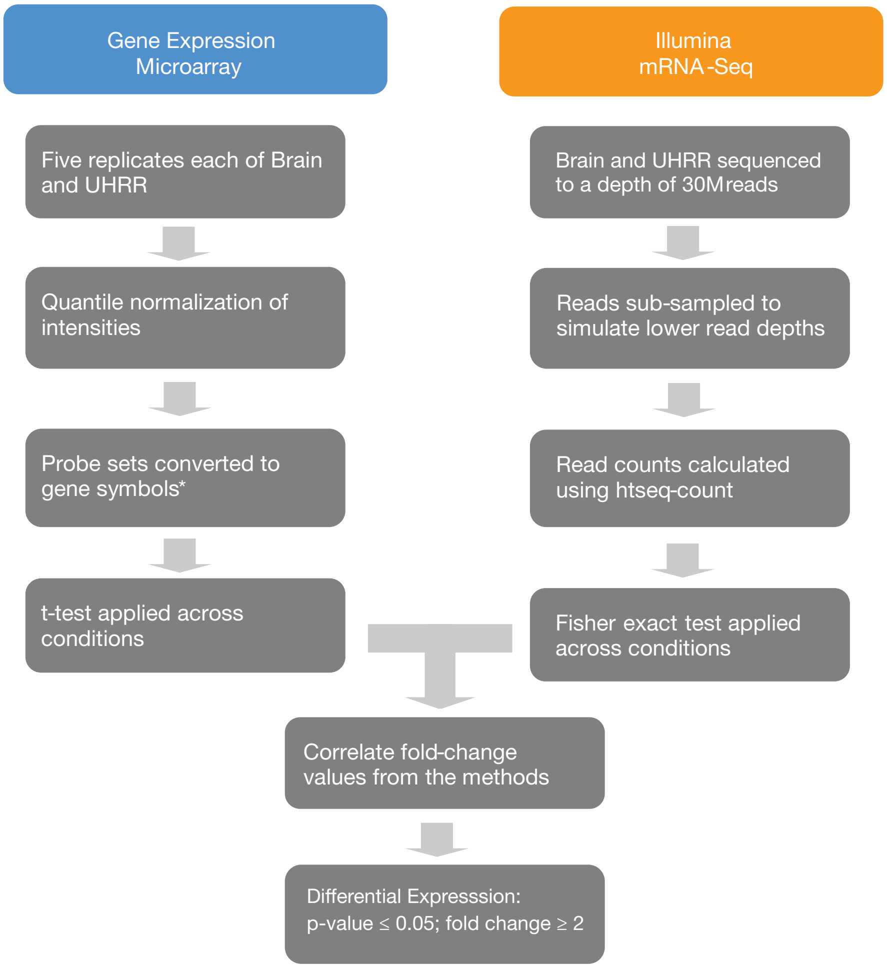
Statistical Considerations for Comparative Analysis

Based on the way transcripts are sampled with RNA-Seq, the sequencing read counts for a transcript across experimental replicates arise from a Poisson distribution (see Appendixes A and B). In contrast, microarray intensities for a transcript across experimental replicates are well approximated by a normal distribution. This distribution difference between data types means that the statistical tests used to evaluate differential gene expression are different for each platform. RNA-Seq Poisson counts are typically analyzed using the Fisher Exact Test[2], while microarray normal intensities are usually analyzed by a t-test.

A thorough validation study was performed to prove that Fisher Exact Test data and t-test data can be compared in a fair manner (Appendix A). The results of this analysis showed that at a depth of 50 million mapped reads, RNA-Seq data analyzed by the Fisher Exact Test was as sensitive and specific as array intensities analyzed by a t-test in detecting a 1.25-fold change in transcript expression levels.

In practice, microarrays are not recommended for discriminating a fold change of 1.25, so much stricter fold-change requirements are commonly imposed for a transcript or gene to be called differentially expressed[3,4]. For this reason, the studies presented in the following sections impose the more realistic two-fold change requirement. With this higher fold change, a much lower sequencing read depth than 50 million mapped reads is required. The RNA-Seq data analysis in the following sections was performed using 10 million mapped reads or less.

Figure 1: Experimental design for data analysis


Gene Expression Microarray
Five replicates each of Brain and UHRR
Quantile normalization of intensities
Probe sets converted to gene symbols*
t-test applied across conditions
Illumina mRNA-Seq
Brain and UHRR sequenced to a depth of 30Mreads
Reads sub-sampled to simulate lower read depths
Read counts calculated using htseq-count
Fisher exact test applied across conditions
Correlate fold-change values from the methods
Differential Expresssion:
p-value ≤ 0.05; fold change ≥ 2

Experimental workflow showing how Brain and UHRR samples were processed independently on each platform for analysis

* Based on the curated mapping file, provided by planDBAffy

Data Generation

Two RNA sample types—MAQC brain (Brain) and Universal Human Reference RNA (UHRR)—were processed using five technical replicates of each on a popular version of a competitor microarray (referred to here as Microarray A) and RNA-Seq[3,5]. For the microarray-based analysis, the samples were processed by the Microarray Quality Control (MAQC) project according to the manufacturer's instructions [6,7]. For RNA-Seq, the sample cDNA libraries were prepared using the mRNA-Seq 8-Sample Prep Kit. Clonal DNA fragment clusters were amplified using the Cluster Station and DNA sequencing was carried out using the Genome Analyzer II system according to standard protocoll[8,9]. The samples were each sequenced to a depth of ~ 30 million mapped reads (only a portion of these reads were actually needed for the following analyses).

Data Analysis

For each platform, the data was processed and normalized to examine fold-change expression levels by the methods described in this section (Figure 1). The normalized values were correlated and compared between platforms.

Microarray Analysis

Intensity values for microarray probe sets were calculated and converted to gene-level intensity values. Fold-change ratios (in log space) were then constructed between samples and differential-gene expression was calculated from probe intensities using an unpaired two-sided t-test. Microarray probe sets were converted to gene IDs. In cases where multiple probe set IDs mapped to a single gene, the probe set whose fold change was closest to the mean fold change across all such probe sets was used for all subsequent analysis. Genes identified as having at least a two-fold change between conditions at a p-value threshold of 0.05 were considered differentially expressed between samples.

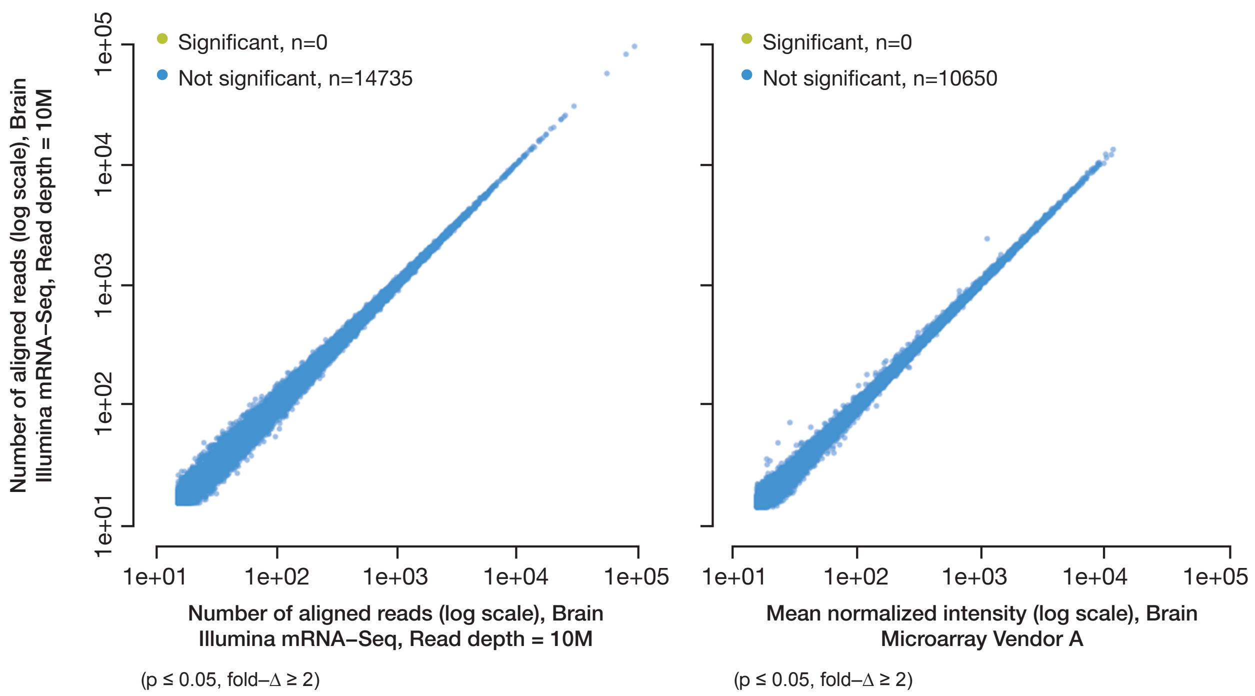
Illumina RNA-Seq Analysis

RNA-Seq sequencing reads were aligned to the reference genome by TopHat. Output SAM files were converted to gene-level read counts using htseq-count, an open-source tool available from EMBL[10]. Fold-change ratios (in log space) were constructed between samples and differential expression was quantified using a Fisher Exact Test on the total number of mapped reads per gene symbol. As with the microarray data, a fold-change cutoff of 2 and p-value threshold of 0.05 were used to determine differential gene expression. A threshold of 10 mapped reads was used to define detection at the gene level.

Normalized Comparisons between Microarray and RNA-Seq

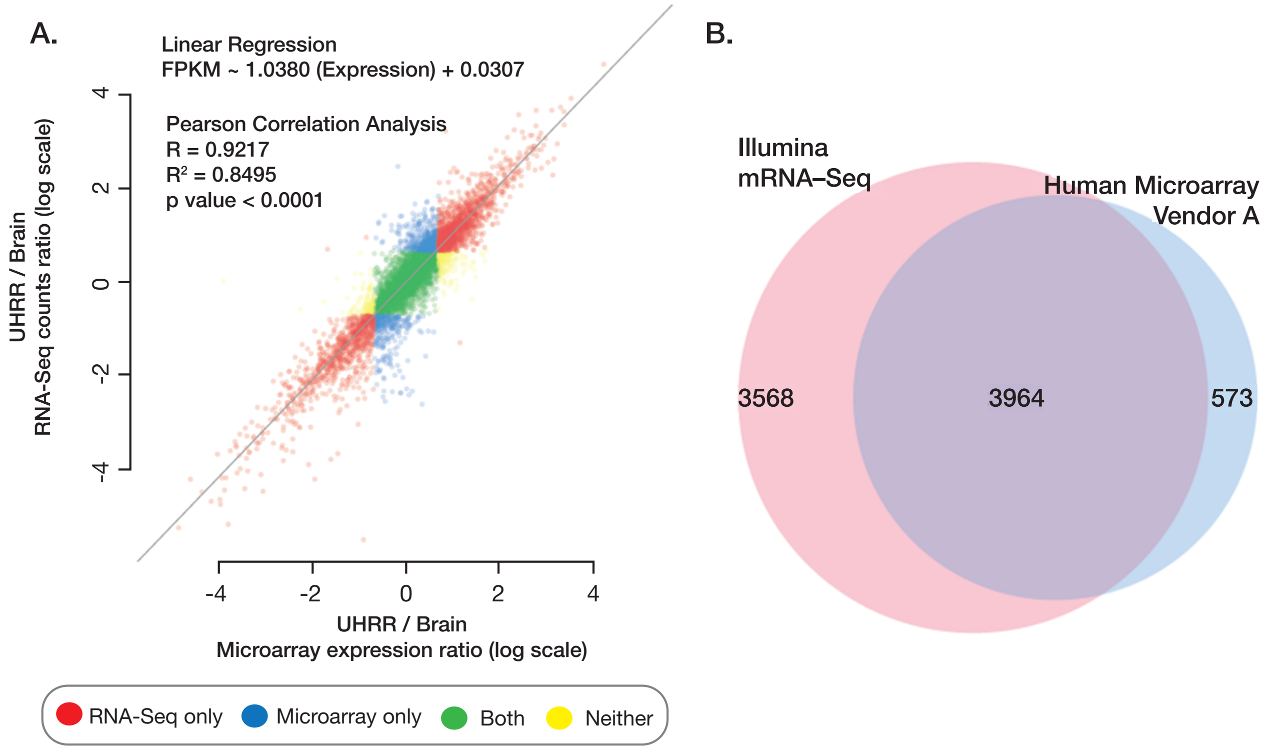
Normalized fold-change expression ratios per gene were compared and correlated to examine the data reproducibility within and between platforms. This data comparison was visualized by creating scatterplots and quantified by calculating the Pearson correlation coefficients (R^2) and corresponding p-values.

Figure 2: Scatterplots of Technical Replicates of Brain Sample



Scatterplots of technical replicates of Brain sample analyzed by Illumina RNA-Seq (left) and human genome microarray (right). The false-positive rates are comparable between the two methods and both methods have extremely high correlation between replicates (Pearson $R^2 > 0.99$ for both RNA-Seq and microarray). The plots demonstrate that RNA-Seq identifies more genes and spans a wider dynamic range compared to the microarray.

Figure 3: Fold-change comparison of gene expression between platforms



A) Scatterplot of fold change per gene as measured by Illumina RNA-Seq and human genome microarray. Genes identified as differentially expressed by both methods are plotted in red; genes identified as differentially expressed by either human genome microarray or Illumina RNA-Seq are plotted in yellow and blue, respectively; genes not identified as differentially expressed by either method are plotted in green. B) Corresponding Venn diagram demonstrating the increased sensitivity of RNA-Seq compared to microarray to detect differentially-expressed genes at a set specificity.

False-positive rates were also estimated and compared across platforms. For the microarray platform, the false-positive rate was determined by comparing two versus three replicate arrays of the same sample. The RNA-Seq false-positive rate was estimated by comparing two independent sets of reads from the same sample over a range of read depths (1–10 million). The sensitivity of differential gene expression detection was assessed using the same read depths per sample.

Results and Discussion

Within-Platform Reproducibility

To assess within-platform reproducibility and examine false-positive rates, repeated runs of the same Brain sample were analyzed by each platform (Figure 2). Both methods showed high reproducibility and equivalent false-positive rates for each sample, with R^2 values > 0.99. However, at a read depth of 10 million mapped reads, RNA-Seq identified over 4,000 more genes than the microarray-based analysis, demonstrating much higher sensitivity.

Cross-Platform Expression Correlation

To evaluate data correlation between microarray intensities and RNA-Seq counts, fold-change ratios of differentially expressed genes between Brain and UHRR were plotted and compared (Figure 3A). The analysis was performed on subsets of genes, broken out into those that were significantly differentially expressed by both platforms ($R^2 = 0.90$), a single platform ($R^2 = 0.54$ and 0.49 for genes identified by microarray or RNA-Seq, respectively), or neither platform ($R^2 = 0.46$). Compared to the within-platform reproducibility, the Pearson correlation coefficients across all genes are on the order of 0.85, as opposed to 0.99 or greater. While the data correlation is still significant, the lower R^2 values indicate a discrepancy between the platforms in the ability to identify genes as differentially expressed. The gene subset segmentation revealed that again, RNA-Seq counts identified significantly more differentially expressed genes (Figure 3b).

To validate the findings from Figure 3B, the analysis was repeated on a subset of the data constrained to 1,044 genes for which PCR data were available (Table 2). Of these genes, 613 were identified as differentially expressed by PCR, based on a two-fold change cutoff. RNA-Seq identified 148 differentially expressed genes that were not identified as such by the microarray. Of these 148 genes, 119 or 80.4%

Table 2: PCR Validation of Differential Gene Expression

	Detected Differentially Expressed Genes	Detected Differentially Expressed Genes Validated by PCR	Concordance with PCR
RNA-SEQ only	148	119	80.4%
Microarray only	28	18	64.3%
Both platforms	312	296	94.9%
Neither platform	556	180	67.6%

Table showing the breakdown of differentially expression genes identified by RNA-Seq and Microarray A from a subset of 1,044 genes for which there is known PCR data. RNA-Seq identified more differentially expressed genes than Microarray A, and had higher concordance with PCR data.

were also identified as differentially expressed by PCR. Comparatively, the microarray identified 28 differentially expressed genes that were not identified as such by RNA-Seq, only 18 (64.3%) of which were confirmed by PCR. Overall, RNA-Seq (at 10 million mapped reads) detected 460 of the 613 differentially expressed genes identified by PCR. Comparatively, Microarray A only identified 360 of these genes. These results demonstrate that RNA-Seq consistently detects more differentially expressed genes with a lower false-positive rate than Microarray A.

Analysis of Gene Detection Using RNA-Seq at Lower Read Depths

As shown in the previous sections, at 10M mapped reads per sample, RNA-Seq demonstrated much higher sensitivity by identifying 44% more differentially expressed genes (Figure 3B). To determine the read depth at which the platforms offer equivalent sensitivity at comparable false-positive rates, a random down sampling was performed on the total number of RNA-Seq reads for each sample.

In this analysis, the RNA-Seq false-positive rate was first assessed by comparing two independent sets of reads from the same sample at each down sampled read depth. Those genes identified as differentially expressed between sets are false positives by definition, so they were used to calculate the overall false-positive rate. Imposing the two-fold change requirement with a p-value threshold of 0.05 resulted in zero false-positive identifications at all read depths, which is equal to the false-positive rate achieved by the microarray using the same criterion.

The RNA-Seq fold-change ratios were then re-calculated for each down-sampled read depth and correlated to the original microarray data. Differential gene expression for RNA-Seq was again determined by the Fisher exact test. Sensitivity was then estimated by calculating the percentage of the detected genes identified as differentially expressed between Brain and UHRR samples for each platform. The point at which the two platforms identified approximately the same percentage of detected genes as differentially expressed was assumed to be the read depth at which RNA-Seq offers equivalent sensitivity to microarray. This point was identified at 2 million aligned reads (highlighted row in Table 3).

At the time of this study, the Illumina HiSeq™ 2000 system produced up to one billion reads passing filter in a single read experiment. Making the conservative assumption that 70% of reads align to the reference genome, a sequencing run would produce 700 million mapped single reads. With 2 million mapped single reads offering comparable sensitivity to a gene expression microarray, then > 300 samples could theoretically be multiplexed on a single HiSeq 2000 run. For such an experiment, the retail cost of all required sequencing reagents would amount to less than $100 (USD) per sample. Since the average price of the competitor microarray is $250–400 per sample, the cost of using RNA-Seq is quite favorable by comparison.

Table 3: Analysis of gene detection using RNA-Seq at lower read depths

Number of Mapped Reads	Total number of detected genes	Number (%) of differentially-expressed genes identified	
		Illumina RNA-Seq	Microarray Vendor A
10,000,000	16203	7532 (46.49%)	4537 (28.00%)
5,000,000	15396	6339 (41.17%)	4537 (29.47%)
4,000,000	15109	6016 (39.82%)	4537 (30.03%)
3,000,000	14795	5504 (37.20%)	4537 (30.67%)
2,500,000	14566	5233 (35.93%)	4537 (31.15%)
2,000,000	14332	4777 (33.33%)	4537 (31.66%)
1,000,000	13513	3496 (25.87%)	4537 (33.58%)

Summary of down sampling analysis, reporting the percentage of differentially expressed genes at each sub-sampling level. With approximately 2.0 million mapped reads (highlighted in gray), the sensitivity of RNA-Seq is approximately equal to that of the microarray (33.2% for each platform).

Summary

Illumina RNA-Seq is a powerful tool for whole-transcriptome analysis. Because there is no need to design probes or primers, the technology provides unbiased data across the entire transcriptome of any species, enabling a broad range of transcript discovery applications not possible with microarray-based analysis. The digital nature of RNA-Seq allows for much higher resolution and an unlimited dynamic range, providing very high sensitivity differential expression analysis. A comparative analysis of data from a competitor microarray and RNA-Seq using well-studied Brain and UHRR samples showed that RNA-Seq offers superior performance. Within-platform variability analysis showed that RNA-Seq and the microarray each produced equivalently high reproducibility between replicates, but RNA-Seq identified 4,000 additional differentially expressed genes. A similar result was revealed in a cross-platform data comparison. While RNA-Seq largely detected the same differentially expressed genes as the array (demonstrated by high correlation coefficients between the data), it also identified a significant number of differentially expressed genes missed by the array. PCR validation of the cross-platform analysis showed that RNA-Seq did indeed detect more differentially expressed genes than the array, and those detections were in agreement with known PCR data a much higher percentage of the time.

A follow-up analysis of down sampled RNA-Seq reads showed that at the reduced read depth of 2 million mapped reads, RNA-Seq sensitivity can be lowered to an equivalent level as the microarray. This allows researchers to reduce per-sample costs by using lower sequencing read-depths, while still maintaining equivalent performance to gene expression microarrays.

References

1. Wang et al., 2009 RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57-63.
2. Fisher RA. (1922) On the interpretation of χ^2 from contingency tables, and the calculation of P. J R Stat Soc. 85 (1): 87–94.
3. MAQC Consortium, et al. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. Sep; 24 (9): 1151-61.
4. Guo L, Lobenhofer EK, Wang C, Shippy R, Harris SC, et al. (2006) Rat toxicology study reveals analytical consistency across microarray platforms. Nat Biotechnol. 2006 Sep; 24 (9): 1162-9.
5. Platform GPL570 (2009) Gene Expression Omnibus. National Center for Biotechnology Information. www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570
6. GeneChip Expression Analysis Technical Manual (P/N 702232 Rev. 3). http://media.affymetrix.com/support/downloads/manuals/expression_analysis_technical_manual.pdf
7. GeneChip Expression Platform: Comparison, Evolution, and Performance (P/N 701632 Rev. 2). 2004. http://media.affymetrix.com/support/technical/technotes/expression_comparison_technote.pdf
8. Illumina Genome Analyzer$_{IIx}$ User Guide (PN 15015143 Rev. A).
9. Illumina mRNA Sample Preparation Guide (PN 1004898 Rev. D).
10. Anders S. (2010) Counting reads in features with htseq-count. www-huber.embl.de/users/anders/HTSeq/doc/count.html

Appendix A: Simulated Comparison of the Fisher Exact Test and t-Test

To more easily compare results from RNA-Seq and gene expression microarrays, simulated data were used to evaluate how the sensitivity and false-positive rates of the Fisher exact test applied to Poisson-distributed data compares to a t-test applied to Poisson- and normal-distributed data.

Simulation of a 1.25-fold change using Poisson Data

100,000 random counts were generated from a Poisson distribution with a mean of 20, simulating an RNA-Seq experiment consisting of 10,000 transcripts, each with 10 replicate measurements. A Fisher exact test was performed on the sum of all counts from the first five replicates versus the sum of all counts from the second five replicates. The percentage of transcripts with a p-value less than or equal to 0.05 was taken to be the false positive rate. A t-test was also performed for the first five replicates versus the second five replicates of each simulated transcript, and the corresponding false positive rate was again calculated at the 0.05 p-value level. Next, five replicate values were generated from a Poisson distribution with mean 25 for each of the 10,000 transcripts, corresponding to a true fold change of 1.25. The differential expression analysis was repeated on this data set using the Fisher exact test and the t-test to evaluate the sensitivity of each to detect the relatively low 1.25 fold change at the 0.05 p-value level.

Simulation of a 1.25-fold change using Normal Data

To demonstrate the relative power of a t-test to detect a 1.25-fold change from five replicates of normally distributed data, a similar analysis was performed by generating 100,000 random values from a normal distribution with mean 20 and standard deviation 2, simulating a microarray experiment consisting of 10,000 probe sets and 10 technical replicates. This normal distribution corresponds to a coefficient of variation (CV) of 10%, which is common for replicate array data. False positives were assessed by performing a t-test comparing the first five replicates to the second five replicates. Sensitivity was assessed by performing a t-test comparing the first five replicates of this distribution to five replicates of data generated from a normal distribution with mean 25 and standard deviation 2, again simulating a true fold change of 1.25.

Comparison of False Positive Rates and Platform Sensitivity

A comparative analysis of the simulated data showed that the two tests' results are highly correlated for Poisson data, with R^2 values of 0.91 for both the false positive rate and sensitivity. This correlation demonstrates that the false positive rates and sensitivities can be compared across the two tests for Poisson data in a fair manner. However, a t-test applied to normally distributed data has over a two-fold increase in sensitivity to detect a 1.25-fold change compared to a Fisher exact test applied to Poisson-distributed data. (Table A1). This result can be attributed to the Poisson distribution with mean read count of 20 having a coefficient of variation of 22%, compared to the 10% associated with the normal distribution. This indicates that a higher number of counts is necessary to detect such a small fold difference. As a result, the simulation was repeated using a five-fold increase in the mean read count (Table A2). Here, the mean read count for the 10,000 transcripts was increased to 100, thus reducing the CV to the 10% level seen in microarrays, and sensitivity to detect a 1.25-fold change was assessed by comparing five replicates with mean 100 to five replicates with mean 125.

Table A1: Summary of simulation results for a mean expression value of 20

	False-Positive Rate	Sensitivity
Poisson data, Fisher exact test	4.72%	36.35%
Poisson data, t-test	4.42%	29.16%
Normal data, t-test	4.30%	91.71%

Table A2: Summary of simulation results for a mean expression value of 100

	False-Positive Rate	Sensitivity
Poisson data, Fisher exact test	4.48%	96.00%
Poisson data, t-test	4.06%	88.75%
Normal data, t-test	4.27%	91.83%

Using higher read counts per transcript (approximately 100), a Fisher exact test applied to Poisson data was equally sensitive to a t-test applied to normally-distributed data in detecting a 1.25-fold change in expression. At this higher read count, each platform also produced equivalent false positive rates.

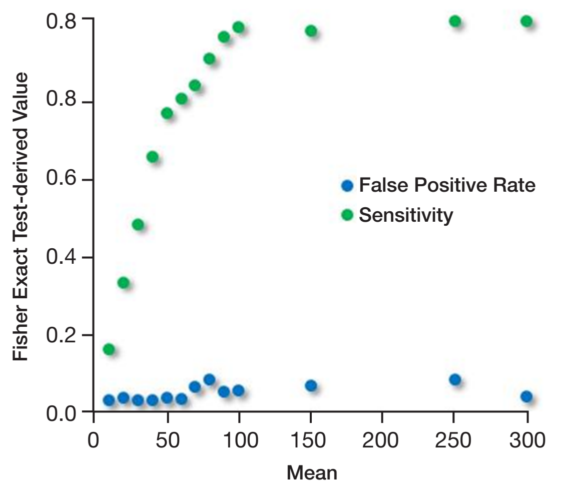
The overall read depth needed for RNA-Seq (Poisson data/Fisher Exact Test) to achieve equivalent performance to a microarray (normal data/t-test) can be extrapolated from RNA-Seq data sub-sampled to a range of read depths. For this extrapolation, the 10th percentile of read counts was calculated across all detected genes as overall read depth increased from 1 million to 30 million mapped reads. Figure A2 shows the linear relationship between the 10th percentile of read counts and overall experimental read depth. This relationship indicates that at a depth of 50 million mapped reads, RNA-Seq data analyzed by the Fisher exact test are equally sensitive and specific to array intensities analyzed by a t-test in detecting a 1.25-fold change in experiment-wide transcript expression levels.

Appendix B: Experimental Validation of Simulated Results

Validation that RNA-Seq Read Counts Arise from a Poisson Distribution

To show that read counts from RNA-Seq data can be approximated by a Poisson distribution, RNA-Seq data from the Human Body Map 2.0 Project were downloaded[1]. Human brain was analyzed on an Illumina HiSeq 2000 instrument after standard library preparation from poly(A)-selected mRNA. The SAM file, containing over 64 million 1 x 75 bp reads, was parsed into ten equally sized SAM files, approximating ten replicate analyses at lower read depth. The number of counts per gene symbol was determined as described previously in this document, and the overall distribution of the

Figure B1: Fisher exact-derived false-positive rates and sensitivities as a function of read count per gene



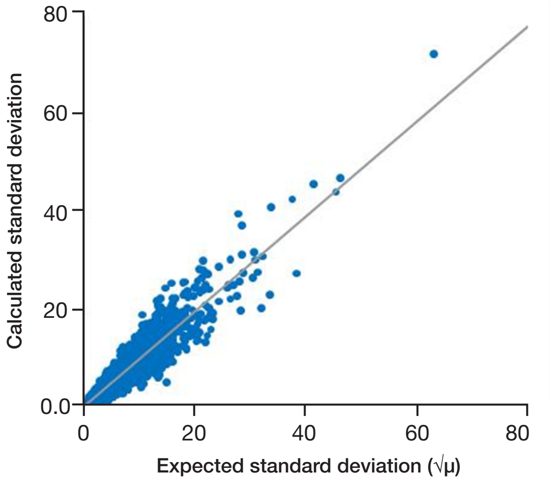
Experimental results showing that RNA-Seq provides equivalent sensitivity to microarrays at a mean read count of approximately 100.

replicates was analyzed. Figure B1 demonstrates the high correlation between the expected standard deviation from a Poisson distribution (where the standard deviation is equal to the square root of the mean) and the measured standard deviation across the 10 simulated replicates. The theoretical and measured values are highly correlated, with R^2 equal to 0.92.

Experimental Validation of Simulated Platform Performance

Using the Human Body Map data, the false-positive rate for the simulated RNA-Seq data was assessed across the range of read counts by performing a Fisher exact test on the first five replicates versus the second five. Sensitivity was assessed by performing the same analysis after multiplying the second replicates by a factor of 1.25, corresponding to the 1.25-fold change in the simulated study. The results support the theoretical conclusion that sensitivity equivalent to microarrays is reached at a mean read count of approximately 100 (Figure B2).

Figure B2: Correlation of RNA-Seq counts with the expected standard deviation from a Poisson distribution



Scatterplot demonstrating high correlation between the expected standard deviation from a Poisson distribution and the measured standard deviation across the ten simulated replicates.

Appendix References

1. Body Map 2.0 (Illumina HiSeq). Data downloaded on 28 February 2010. www.broadinstitute.org/igvdata/BodyMap/hg19/IlluminaHiSeq2000_BodySites

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

Pub. No. 470-2011-004 Current as of 12 April 2011