Evaluation of Sequencing Approaches for High-Throughput Toxicogenomics

R.S. Thomas[1], A.L. Karmaus[1,2], P. Kothiya[1,2], M.T. Martin[1]

[1]US EPA National Center for Computational Toxicology, [2]ORISE Fellow

Matt Martin

martin.matt@epa.gov | 919-541-4104

www.epa.gov

Abstract

Whole-genome *in vitro* transcriptomics has shown the capability to identify mechanisms of action and estimates of potency for chemical-mediated effects in a toxicological framework, but with limited throughput and high cost. We present the evaluation of three toxicogenomics platforms for potential application to high-throughput screening: 1. TempO-Seq utilizing custom designed paired probes per gene; 2. Targeted sequencing (TSQ) utilizing Illumina's TruSeq RNA Access Library Prep Kit containing tiled exon-specific probe sets; 3. Low coverage whole transcriptome sequencing (LSQ) using Illumina's TruSeq Stranded mRNA Kit. Each platform was required to cover the ~20,000 genes of the full transcriptome, operate directly with cell lysates, and be automatable with 384-well plates. Technical reproducibility was assessed using MAQC control RNA samples A and B, while functional utility for chemical screening was evaluated using six treatments at a single concentration after 6 hr in MCF7 breast cancer cells. All RNA samples and chemical treatments were run with 5 technical replicates. The three platforms achieved different read depths, with the TempO-Seq having ~34M mapped reads per sample, while TSQ and LSQ averaged 20M and 11M aligned reads per sample, respectively. Inter-replicate correlation averaged ≥0.95 for raw log2 expression values in all three platforms across all samples. When the ratio of MAQC samples A:B was correlated between the technologies and the reference MAQC-III Illumina results, r2 values of 0.83 for LSQ, 0.74 for TSQ, and 0.75 for TempO-Seq were observed, suggesting good technical reproducibility for each sequencing platform. When chemically-treated samples were evaluated, the inter-replicate and cross-technology correlations of fold-change values were significantly reduced. Bland-Altman plots revealed that genes with low read counts accounted for the greatest variability in fold-change space. Application of a minimum read-count cutoff was necessary to achieve good concordance. Finally, connectivity map (CMAP) analysis was conducted to evaluate the ability of each platform to identify modes-of-action in the chemically-treated samples. TempO-Seq showed the best concordance with mechanistically similar chemical treatments; however, this may be due to the increased read depth associated with the platform. In summary, the three sequencing platforms were able to measure whole-genome transcript levels with good technical reproducibility and show promise for the integration of toxicogenomics into high-throughput screening.

Objectives and Study Design

The generation of high-throughput global gene expression profiles using RNA-sequencing technologies for the evaluation of chemically-mediated effects could greatly advance the current toxicogenomics knowledgebase. Three high-throughput sequencing (HTS-Seq) approaches were evaluated to assess technical and functional performance in order to characterize the limitations and possible applications of HTS-Seq technologies.

Low Coverage Sequencing (Omega Bioservices)

- Omega Bio-Tek Mag-Bind Total RNA kit to **isolate total RNA**
- Library prep with **Illumina Stranded mRNA Sample Prep :** isolate mRNA using poly-T oligo attached to magnetic beads, fragment purified mRNA, copy first strand cDNA with random primers, purify, enrich with PCR
- Pooled libraries sequenced on Illumina HiSeq 2500

~ 11 million aligned reads per sample

Targeted Sequencing (Omega Bioservices)

- Omega Bio-Tek Mag-Bind Total RNA kit to **isolate total RNA**
- Library prep with **Illumina TruSeq RNA Access Library Prep Kit:** fragmentation, cDNA generation by random priming, ligate polyA sequencing adaptors, coding regions captured using optimized probe set (>425,000 probes covering 96.3% of RefSeq exome)
- Pooled libraries sequenced on Illumina HiSeq 2500

~20 million aligned reads per sample

TempO-Seq (BioSpyder)

- **Cell lysate** input (ie. capture-free method)
- **Detector oligos annealed** (25 base probes, two per gene, designed to target a gene-specific region; have adaptor sequences allowing sample-specific barcodes to be used)
- Illumina-compatible adaptors on ligated detector oligos ultimately enable standard dual index sequencing

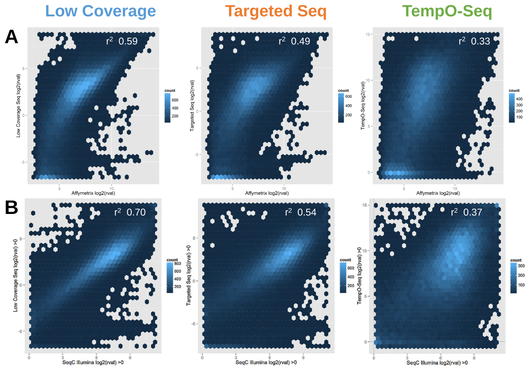
~34 million mapped reads per sample

Use all three sequencing technologies to quantify gene expression for MAQC control samples A and B as well as samples from MCF7 cells treated with a single concentration of five chemicals for 6 hrs.

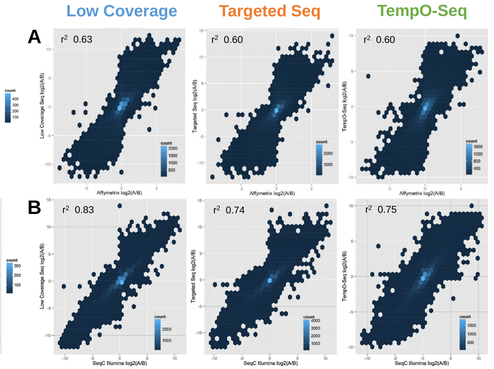
Objectives for the evaluation of three HTS-Seq platforms:

- ► Assess technical and inter-replicate reproducibility
- ► Identify chemical-mediated differential gene expression signatures
- ► Evaluate output from Connectivity Mapping to assess functional utility for toxicogenomics screening

Evaluating Technical Performance



▲ ***Figure 1. Correlation of normalized expression values (rval) to MAQC Affymetrix (A) and SEQC Illumina (B) datasets****: The mapped reads were normalized and log2 transformed to obtain "rval". For low coverage and targeted sequencing FPKM was used. For TempO-Seq each gene was normalized as total reads relative to the average of the sum of total reads for that gene across all replicates. The lower r2 values for Targeted Seq and TempO-Seq may be due to differing probe efficiencies across genes and may not be appropriate for measuring absolute transcript abundance.*

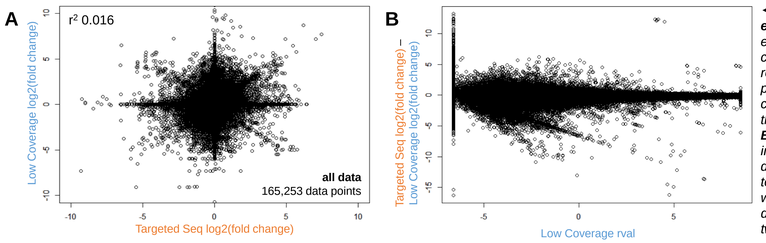


▲ ***Figure 2. Evaluation of MAQC control sample A:B ratio correlation between sequencing technologies and MAQC Affymetrix (A) and SEQC Illumina (B)****: As a surrogate for fold change, the ratio of control sample A vs. B was evaluated. Pearson's correlations (r^2) show better concordance with SEQC than microarray. The dynamic range achieved among platforms was more similar to the SEQC dataset compared with Affymetrix. All three platforms show similar performance for measuring fold-change gene expression changes.*

Table 1: Inter-replicate correlations for raw normalized values (rval)

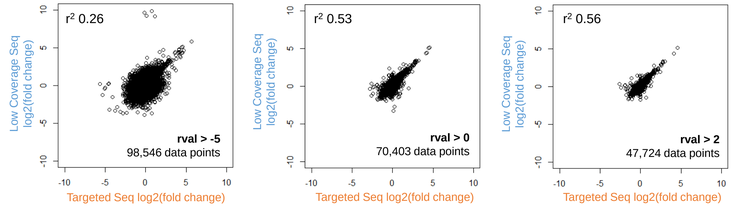
Technology	MAQC Control A	MAQC Control B	10 µM Chlorpromazine	10 µM Ciclopirox	10 µM Genistein	100 nM Sirolimus	1 µM Tanespimycin	DMSO
Affymetrix	0.99	0.99	-	-	-	-	-	-
SeqC Illumina	0.99	0.99	-	-	-	-	-	-
Low Coverage	0.95	0.96	0.96	0.96	0.95	0.96	0.96	0.96
Targeted	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.95
TempO-Seq	0.97	0.97	0.95	0.95	0.95	0.96	0.94	0.95

Note: Chemical treatments were chosen from the Connectivity Map, encompassing unique modes of action. The inter-replicate correlations reflect Pearson's correlation (r^2) across 5 replicates. These correlations were calculated for the normalized expression values (rval). All treatments were conducted independently for each technology for 6 hrs in MCF7 cells.



◄ ***Figure 3. Chemical-treated gene expression correlation****: **A.** Chemical-elicited gene expression (log2 fold change) was compared to evaluate reproducibility of gene expression profiles across platforms. Despite high correlation of rval (data not shown), the fold change correlations were low. **B.** A Bland-Altman plot was used to investigate the effect of rval on the difference in fold change between technologies, revealing that genes with lowest rval showed the greatest difference in fold change between the two platforms.*

Evaluating Technical Performance (Continued)



▲ ***Figure 4. Chemical-treated gene expression correlation among the sequencing technologies****: Chemical-elicited gene expression (log2 fold change) from Targeted Seq and Low Coverage Seq were compared, revealing low correlation (Figure 3A). To address the effect of low rval, as determined based on the Bland-Altman plot in Figure 3B, a filter was applied requiring rval to be greater than log2(-5), log2(0), or log2(2), respectively. This filtering resulted in increased correlation, reaching r^2 0.56.*

Evaluating Functional Performance

Table 2: Number of matching mechanisms in top 10 CMAP results

Chemical	Low Coverage	Targeted Seq	TempO-Seq
Genistein	0	0	3
Ciclopirox	0	0	3
Sirolimus	0	0	2
Tanespimycin	1	1	4
Chlorpromazine	1	1	1

Note: Differentially expressed genes for CMAP were identified using filtering criteria: |fold change| > 2 and t-test p < 0.01

◄ ***Connectivity Mapping (CMAP) Analysis****: Genes identified as differentially expressed were used as input for CMAP. The resulting output was ranked based on p-value and the top ten profiles were evaluated for mechanism of action (MOA) that match the reference chemicals. Overall, TempO-Seq resulted in the most matching MOAs among the top CMAP outputs.*

Summary

Gene expression profiles were successfully generated using three high-throughput sequencing technologies

- ► The technical reproducibility across replicates within a technology for all sequencing platforms was very high with Pearson's correlations of r^2 >0.95.
- ► The normalized expression values for MAQC control samples A and B were highly correlated to results from MAQC Affymetrix and SEQC Illumina datasets. Furthermore, the A:B gene expression demonstrated good dynamic range comparable to SEQC for all technologies, outperforming Affymetrix microarrays.
- ► Due to differing probe efficiencies across genes, Targeted Seq and TempO-Seq showed lower performance for measuring absolute transcript abundance; however, all three platforms showed high technical performance for measuring fold change gene expression changes.
- ► The TempO-Seq platform showed better functional performance for correctly identifying chemical MOA than the other two platforms; however, this may be due to differences in read depth or slight differences in cell and treatment protocols among vendors.
- ► Future work will seek to define a minimum mapped read requirement, refine how significant differential expression is identified, establish an automated in-house CMAP algorithm, and incorporate concentration-response modeling for chemical-mediated gene expression.

U.S. Environmental Protection Agency
Office of Research and Development