**Supplementary Information**

**Comparison of whole transcriptome and targeted RNA sequencing for ecological high throughput transcriptomics**

Daniel L. Villeneuve1, Mackenzie Nash2, Adam Biales3, Kendra Bush2, Garrett Evensen4, Logan Everett5, Jonathan Haselman1, Monique Hazemi2, 5, Michelle Le2, Helen Poynton4, Bruce Seligmann7, Leah Wehmas6, Joanne Yeakley7, Kevin Flynn1

1 United States Environmental Protection Agency, Great Lakes Toxicology and Ecology Division, 6201 Congdon Blvd, Duluth, MN, USA, 55804.

2 Oak Ridge Institute for Science and Education Research Participant at U.S. EPA, Great Lakes Toxicology and Ecology Division, 6201 Congdon Blvd, Duluth, MN, USA, 55804.

3 United States Environmental Protection Agency, Great Lakes Toxicology and Ecology Division, 26 W. Martin Luther King Drive, Cincinnati, OH, USA 45520.

4 University of Massachusetts, Boston, School for the Environment, 100 Morrissey Blvd, Boston, MA, USA 02125.

5 United States Environmental Protection Agency, Biomolecular and Computational Toxicology Division, 109 T.W. Alexander Drive, Research Triangle Park, NC, USA 27711.

6 United States Environmental Protection Agency, Chemical Characterization and Exposure Division, 109 T.W. Alexander Drive, Research Triangle Park, NC, USA 27711.

7 BioSpyder Technologies Inc., Carlsbad, CA 92008.

**Supplementary Methods:**

A series of analyses were conducted on the data submitted by the Solvers and results provided to all judges to help facilitate evaluation of several criteria including accuracy, precision, and relative coverage of the transcriptome.

SM. 1: Analyses to aid Accuracy Evaluation:

To assess accuracy, the ratio of (average of high and control mRNA counts)/(average of middle mRNA counts) was calculated for each dataset to compare how similar the middle treatment samples were to the average of the high and control treatments. The middle treatment samples consisted of 50% control RNA and 50% high treatment RNA, so theoretically the middle treatment mRNA counts should be identical to the average of the high and control conditions.

First, all 3 technical replicates of the control and all 3 technical replicates of the high treatment mRNA counts were averaged together. Then the 3 technical replicates of the middle treatment mRNA counts were averaged together. The ratio of (mean high & control) / (mean middle) was then calculated for each feature of each count matrix (i.e., one ratio per row). All NA and 0 ratio values were removed, given that these values are not readily interpretable for the purposes of assessing accuracy in this context. Then, any ratio values that were < 1 were transformed to their inverse (i.e., 1/x), so that the magnitude of change relative to a value of 1 could be observed.

For instance, for a given feature, if the mean of high and control counts = 10 and the mean middle counts = 5, then the ratio would be (10/5) = 2. Namely, there was a 2-fold change in expression between the two values (i.e., one value was 2x the other). If the mean of high and control counts = 2.5 and the mean middle counts = 5 for a different feature, then the ratio would be (2.5/5) = 0.5. These values also possess a 2-fold difference in expression (albeit in a different direction of change). Taking the inverse of this ratio (since it is < 1) gives (1/0.5) = 2, which more clearly depicts that one value is 2x the other. Additionally, if the mean of high and control counts and mean middle counts were the same, the ratio would 1. This indicates no change in expression between the two values and that they were identical (as would be expected according to the experimental design). Boxplots of the ratio values for each solver-submitted dataset were constructed (Supplementary Figure S.1).

SM. 2: Analyses to aid Precision Evaluation:

To evaluate precision, the coefficients of variation of gene expression across the three technical replicates of each treatment were calculated. Coefficient of variation (CV) was calculated as:

.

Treatments were maintained while performing these calculations, such that there was a CV\_ctrl, CV\_mid, and CV\_high for each feature (i.e., row). Features with 0 counts across all 3 technical replicates had a mean of 0 counts and were removed as it is not possible to divide by 0. Boxplots of the CV values for each solver-submitted dataset were constructed, with plots separated by treatment condition (Supplementary Figures S.2A – control; S.2B – middle; S.2C – high).

Principal component analyses (PCA) were also conducted for each of the 24 count matrices submitted by the Solvers (Supplementary Figures S.3A-S.3F). In theory, since all three samples within each treatment group for a given species were technical replicates, samples within each treatment should cluster tightly together on a PCA plot (particularly with regard to position on the X-axis (PC1)).

SM. 3: Analyses to aid Evaluation of Transcriptome Coverage:

Transcriptome coverage was evaluated on a relative basis. Percent coverage of the transcriptome for each solver dataset was determined by taking the quotient of the number of features in the solver-submitted count matrix over the maximum number of features present in any solver-submitted count matrix for that given species.

For instance, from the six *C. dilutus* datasets submitted, the highest total annotated features belonged to Solution 2C with 24,509 features. So for the Solution 1 Technician 1 dataset, the percent coverage was determined to be (1536/24509)\*100= 6.3%. For the Solution 2C dataset, the percent coverage would be 100%.

SM.4: BioSpyder’s Selection of Sentinel Genes for the Challenge

BioSpyder’s selection of sentinel genes was based on the objective of implementing an assay that would demonstrate feasibility, and compete effectively against whole transcriptome RNAseq assays. However, for the purposes of the Challenge, the sentinel gene selection did not go through the standard validation with reference samples, optimization of the gene content or incorporation of attenuators, which are important features that increase sensitivity and robustness of commercial TempO-Seq assays. The approach used for the Challenge mirrored that used to design the human S1500 sentinel gene assay, which contains the content to measure 2700 genes. This content includes the L1000 gene set and additional genes selected, at the time, to have at least 3 genes representing every known molecular pathway identified in humans (Mav, D, et al. (2018) PLOSone. 13(2): e0191105. doi.org/10.1371/journal.pone.0191105).

For each targeted species an annotated reference genome was used with KEGG analysis to identify pathways and select 3 genes that identified each pathway, preferentially selecting genes that were more specific for the taxon, where possible. KEGG analysis was also used to identify orthologs of the human pathway genes used in the human S1500 assay, and those ortholog genes were added. For algae (*Raphidocelis subcapitata*), the well annotated species *Monoraphidum neglectum* was used to identify homologous genes that overlapped with the specified genome assembly (NCBI BioProject PRJDB5653, which was not well annotated itself), from which 114 pathways, and the genes within those pathways, were identified and targeted. The resulting algae *Raphidocelis subcapitata* S1500 assay employed for the Challenge measures all 114 pathways using content measuring 856 genes. For midge *Chironomous dilutus*, the NCBI BioProject PRJNA480192 was used as the reference. 132 unique pathways were identified in KEGG, and after selecting 3 genes for each pathway, and identifying and selecting additional gene orthologs to the Human, the final assay measured 1450 genes. Similarly, the *Daphnia magna* S1500 contained content measuring 2378 genes, and the *Pimephales promelas* S1500 contained content measuring 1832 genes. Thus, near-complete pathway analysis can be carried out using the data obtained from each species using the sentinel assays implemented for each.

Key assumptions for this approach were that the databases used contained all relevant pathways, and that the compounds tested only act by modulating those pathways. These assumptions may be reasonable for humans. However, because so little data were available for the species being tested in the Challenge, genes and pathways not in the KEGG database were unaccounted for. It is likely that compounds tested during the Challenge modulated pathways and processes not represented by the current sentinel gene sets. Long term, additional strategies can be employed to expand and improve the sentinel gene sets. However, for the Challenge the objective was to demonstrate competitive feasibility of using a sentinel array vs sequencing the whole transcriptome, not to provide a final sentinel assay for each species.

SM.5: In silico creation of Sentinel Gene Sets from Whole Transcriptome Data.

Whole transcriptome data from the larval fathead minnow and *Daphnia magna* studies (Villeneuve et al. 2022; Flynn et al. 2024; Villeneuve et al. 2024) were used to create subsets that matched the sentinel gene sets employed by one of the Solvers (BioSpyder Inc.; <https://www.biospyder.com/>, accessed 03-15-2025). Additionally, in the case of fathead minnow, 20 random subsets with the same number of features as the sentinel gene set employed by BioSpyder (1832 probes) were created (see Supplementary Information for details).

To create the fathead minnow subsets, the 1,829 probes from the BioSpyder probe manifest (Supplementary Information Table S.3) were matched to gene symbols or RefSeq IDs from the annotated whole transcriptome RNAseq data (Villeneuve et al. 2022; Flynn et al. 2024), in which the reads were mapped to either the transcriptome (Villeneuve et al. 2022) or genome (Flynn et al. 2024). Of the 1,829 BioSpyder probes, 1,594 matched to a transcript/gene name from the RNASeq annotations, and 235 genes were unassigned and not included in this analysis. However, because the same probe name/symbol may apply to more than one individual RNAseq gene/transcript feature, the 1,594 matched gene names corresponded to 3,697 transcripts or 1,832 gene IDs from the RNAseq file. Consequently, a total of 1832 gene level features were included in the *in silico* sentinel gene set.

In the case of *Daphnia magna*, of the 2,378 probes in the BioSpyder probe manifest (Supplementary Information Table S.4), only 1,110 had a gene ID compatible to the annotated whole transcriptome used for RNASeq. To retrieve compatible information for the other 1,268 probes, BLAST (Altschul et al., 1990) searches were performed in R using the "rBLAST" package v0.99.4 (Hahsler & Anurag, 2024) against the *Daphnia magna* reference genome (GenBank accession no. GCF\_020631705.1) using the 50 bp BioSpyder probe sequences. For each probe, the BLAST result with the lowest E-value was retained, and the corresponding gene symbol was determined from the gene accession using the "rentrez" package v1.2.3. The BLAST search was performed for all 2,378 probes for consistency. This allowed for the identification of the most likely gene IDs for 1,263 of the 1,268 previously incompatible/unassigned probes based on sequence similarity. The probe assignments determined with BLAST were combined with the original 1,110 compatible assignments, resulting in a final subset with 2,158 unique features.

Following creation of the subsets, resulting count matrices represented all the original samples, concentrations, etc. from the whole transcriptome studies, but only included the sentinel genes (or random subsets for fathead minnow). The subset count matrices were analyzed with BMDExpress 3 (Phillips et al. 2019). The pre-filtering steps (e.g., application of Williams trend test) and quality filters applied to the subsets was identical to those applied in the whole transcriptome studies (Villeneuve et al. 2022; Flynn et al. 2024; Villeneuve et al. 2024). Transcriptomic points of departure were calculated as the 10th percentile of the distribution of BMCs for each of the concentration responsive genes (CRGs) identified for the sentinel genes or random subsets.



Figure S.1 Box plot depicting relative accuracy as evaluated based on the ratio of the mean of (High + control counts) divided by the counts for the middle treatment. A ratio of 1 indicates agreement between the observed and expected count values for the middle treatment. Ratios greater than 1 indicate the counts for the middle treatment either over- or under-estimated the counts expected based on (High + control / 2). Note, figure 1 excludes extreme outliers. Sol. = solution, T1 = technician 1, T2 = Technician 2, CDil = C. dilutus, DMag = D. magna, PProm = P. promelas, RSub. = R. subcapitata.



Figure S.2A Distribution of coefficients of variation of count values for individual genes, based on three technical replicates of control sample. Sol. = solution, T1 = technician 1, T2 = Technician 2, CDil = C. dilutus, DMag = D. magna, PProm = P. promelas, RSub. = R. subcapitata.



Figure S.2B Distribution of coefficients of variation of count values for individual genes, based on three technical replicates of middle treatment sample. Sol. = solution, T1 = technician 1, T2 = Technician 2, CDil = C. dilutus, DMag = D. magna, PProm = P. promelas, RSub. = R. subcapitata.



Figure –S.2C Distribution of coefficients of variation of count values for individual genes, based on three technical replicates of high treatment sample. Sol. = solution, T1 = technician 1, T2 = Technician 2, CDil = C. dilutus, DMag = D. magna, PProm = P. promelas, RSub. = R. subcapitata.



Figure S.3A. Principal component plots of counts for each species (*Chironomous dilutus, Daphnia magna, Pimephales promelas, Raphidocelis subcapitata*) based on Solution 1, Technician 1.



Figure S.3B Principal component plots of counts for each species (*Chironomous dilutus, Daphnia magna, Pimephales promelas, Raphidocelis subcapitata*) based on Solution 1, Technician 2.



Figure S.3C. Principal component plots of counts for each species (*Chironomous dilutus, Daphnia magna, Pimephales promelas, Raphidocelis subcapitata*) based on Solution 2A.



Figure S.3D. Principal component plots of counts for each species (*Chironomous dilutus, Daphnia magna, Pimephales promelas, Raphidocelis subcapitata*) based on Solution 2B.



Figure S.3E. Principal component plots of gene counts for each species (*Chironomous dilutus, Daphnia magna, Pimephales promelas, Raphidocelis subcapitata*) based on Solution 2C.



Figure S.3F. Principal component plots of gene counts for each species (*Chironomous dilutus, Daphnia magna, Pimephales promelas, Raphidocelis subcapitata*) based on Solution 3.