High-Throughput Transcriptomics for Chemical Bioactivity Screening and Tiered Hazard Evaluation

Joshua A. Harrill, PhD
US EPA, Center for Computational Toxicology and Exposure (CCTE)
109 TW Alexander Drive
Research Triangle Park, NC 27709
E-mail: harrill.joshua@epa.gov
ORCID: 0000-0003-4317-6391
Conflict of Interest Statement

- The views expressed in this presentation are those of the presenter and do not necessarily represent the views or policies of the US Environmental Protection Agency, nor does mention of trade names or products represent endorsement for use.
- The presenter has no conflict of interest regarding the materials in this presentation.
Objectives

- Broad overview of the Next Generation Blueprint of Computational Toxicology at US EPA → emphasis on the role of **transcriptomics**.

- Provide information on technological and analytical innovations that support **high-throughput transcriptomics (HTTr)** chemical screening.
  - Targeted RNA-Seq technology.
  - Novel bioinformatics workflows and associated open-source tools.
  - Transcriptomic reference materials.
  - International effort to develop omics reporting frameworks.
Computational Toxicology Research Areas at EPA

The NexGen Blueprint of CompTox at US EPA

Thomas et al. (2019) DOI: 10.1093/toxsci/kfz058

ToxCast: Uses targeted high-throughput screening (HTS) assays to expose living cells or isolated proteins to chemicals and assess bioactivity and potential toxic effects.

Richard et al. (2016) DOI: 10.1021/acs.chemrestox.6b00135

Mostly targeted assays (chemical X $\rightarrow$ target Y). Incomplete coverage of human biological space.

New Strategy for Hazard Evaluation: Improve efficiency and increase biological coverage by using non-targeted profiling assays that cast the broadest net possible for capturing the potential molecular and phenotypic responses of human cells to chemical exposures.
High throughput profiling (HTP) assays are proposed as the first tier in a NAMs-based hazard evaluation approach.

**HTP Assay Criteria:**
1. Yield bioactivity profiles that can be used for potency estimation, mechanistic prediction and evaluation of chemical similarity.
2. Compatible with multiple human-derived culture models.
3. Concentration-response screening mode.

To date, EPA has identified and implemented two HTP assays that meet this criteria.

- High-Throughput Transcriptomics [HTTr]
- High-Throughput Phenotypic Profiling [HTPP]

The NexGen Blueprint of CompTox at US EPA
Thomas et al. (2019) DOI: 10.1093/toxsci/kfz058
Templated Oligo with Sequencing Readout (TempO-Seq)

The TempO-Seq human whole transcriptome assay measures the expression of greater than 20,000 transcripts.

Requires only picogram amounts of total RNA per sample.

Compatible with purified RNA samples or cell lysates.

Lysates are barcoded according to sample identity and combined in a single library for sequencing using industry standard instruments.

Scalable, targeted assay:
• 1) specifically measures transcripts of interest
• 2) ~50-bp reads for all targeted genes
• 3) requires less flow cell capacity than RNA-Seq

Yeakley et al. (2017) DOI: 10.1371/journal.pone.0178302
Chemical Screening in MCF7 Cells Using HTTr

### High-Throughput Transcriptomics Platform for Screening Environmental Chemicals


TOXICOLOGICAL SCIENCES, 2021, 1–22

doi: 10.1093/toxsci/kfab009
Advance Access Publication Date: 4 February 2021
Research Article

<table>
<thead>
<tr>
<th>Parameter</th>
<th>MCF7 Pilot</th>
<th>MCF7 Screen</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Type(s)</td>
<td>1</td>
<td>1</td>
<td>MCF7</td>
</tr>
<tr>
<td>Assay Formats:</td>
<td>2</td>
<td>2</td>
<td>High-Throughput Transcriptomics Cell Viability</td>
</tr>
<tr>
<td>Culture Condition</td>
<td>1</td>
<td>1</td>
<td>DMEM + 10% HI-FBS</td>
</tr>
<tr>
<td>Chemicals</td>
<td>44</td>
<td>1784</td>
<td>ToxCast chemicals</td>
</tr>
<tr>
<td>Time Points:</td>
<td>1</td>
<td>1</td>
<td>6 hours</td>
</tr>
<tr>
<td>Concentrations:</td>
<td>8</td>
<td>8</td>
<td>3.5 log10 units; semi log10 spacing</td>
</tr>
<tr>
<td>Biological Replicates:</td>
<td>3</td>
<td>3</td>
<td>Independent cultures</td>
</tr>
</tbody>
</table>
Experimental Design for HTTr

- Test chemicals in 8-point dilution series
- Vehicle controls (DMSO)
- No treatment controls
- Reference chemical #1 (ex. Genistein, 10 µM)
- Reference chemical #2 (ex. Sirolimus, 0.1 µM)
- Reference chemical #3 (ex. Trichostatin A, 1 µM)

Used to track assay performance inclusive of cellular response.

Reference RNAs
Reference Lysates
Lysis Buffer Blanks
Reserved for Sequencing Vendor

Used to track assay performance independent of chemical treatments and responsivity of culture.
Use of Reference Samples in HTTr Screening

- Reference samples are intended to provide objective evaluation(s) of the technical performance of an ‘omics assay…NOT the biological response of an in vitro test system.
  - Use reference treatments for this latter purpose.

- Processed in parallel with test samples → they should be subject to the same manipulations and assay conditions as test samples.

- Implemented in a manner that facilitates monitoring of consistency of transcriptomics assay results generated within studies, across studies, across laboratories and over time.
Reference Samples: History of Use for HTTr

“Early days” (2017-2020) at US EPA:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Observations</th>
</tr>
</thead>
</table>
| Reference Pair #1             | Takara UHRR (636690) Takara HBRR (636530)        | • Comparable to Microarray Quality Control Consortium (MAQC) reference samples (doi: 10.1038/nbt1239).  
• Finite resource sourced from distinct individuals.  
• Not optimal for evaluating performance of cell-lysate compatible transcriptomics assays.                                                                                                                                                                                                                                                                                                                      |
| Reference Pair #2             | MCF7 Cells DMSO (0.5%) Treated TSA (1 µM) Treated | • Generated at US EPA  
• Fewer genes detected compared to Reference Pair #1.  
• Range of FC values smaller than Reference Pair #1.                                                                                                                                                                                                                                                                                                                                                               |

US EPA perceived a need to develop **replenishable** human-derived transcriptomics reference samples that are:

- Compatible with multiple assay technologies.
- Available as both purified RNA and cell lysates.
- Yield reproducible fold-change profiles across production batches.
Engineering of Transcriptomics Reference Samples

- Paired reference samples were prepared by combining the genetic material from different human-derived cell lines cultured under different conditions.
- Formulated to mimic the performance characteristics of MAQC samples.
- Prepared as both purified RNAs and cell lysates (BioSpyder, Inc.)

<table>
<thead>
<tr>
<th>Sample</th>
<th># of Genes Detected $^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSP_RNA_A</td>
<td>13,962</td>
</tr>
<tr>
<td>BSP_RNA_B</td>
<td>13,779</td>
</tr>
<tr>
<td>BSP_LYSATE_A</td>
<td>14,919</td>
</tr>
<tr>
<td>BSP_LYSATE_B</td>
<td>14,565</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample Pair</th>
<th># of Genes in Common $^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA_A &amp; RNA_B</td>
<td>12,881</td>
</tr>
<tr>
<td>LYSATE_A &amp; LYSATE_B</td>
<td>13,546</td>
</tr>
</tbody>
</table>

$^a$ Whole transcriptome TempO-Seq @ 8M mapped reads. Genes with count > 5 considered “detected”

Similar numbers of detected genes in engineered reference samples compared to MAQC or Takara samples.

Expression profiles of BioSpyder RNA and lysates are highly correlated.
Evaluating HTTr Assay Performance

Correlation of $\log_2(FC)$

Pearson Correlation

Flagged Plates

Plates with potential performance issues flagged for additional scrutiny

* modified from (House et al. 2017)
HTTr Bioinformatics Pipeline

Primary Goals:

- Reproducible & open source
  - github.com/USEPA/httrpl_pilot
  - github.com/USEPA/CompTox-htrpathway

- Automate and efficiently execute computationally intensive steps.

- Focus on concentration-response modeling and **molecular point-of-departure (mPOD)** determination.

- Store analysis results in a queryable database structure (MongoDB).
## HTTr Quality Control Metrics

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
<th>Threshold</th>
<th>Additional Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>FrVC</td>
<td>Fraction of viable cells (PI-negative or Casp3/7-negative)</td>
<td>Reject &lt; 50%</td>
<td>Highly cytotoxic conditions no longer represent molecular initiating event</td>
</tr>
<tr>
<td>NMR</td>
<td>Number of mapped reads, defined as sum of total read counts summed over all detected probes</td>
<td>Reject &lt; 300,000</td>
<td>Threshold =10% of target depth</td>
</tr>
<tr>
<td>FMR</td>
<td>Fraction of uniquely mapped reads</td>
<td>Reject &lt; 50%</td>
<td>Majority of reads must align to a single probe sequence</td>
</tr>
<tr>
<td>Ncov&lt;sub&gt;5&lt;/sub&gt;</td>
<td>The number of probes with at least 5 uniquely mapped reads</td>
<td>Reject &lt; 5,000</td>
<td>Based on Tukey’s Outer Fence (3*IQR) of all viable samples cultured on each plate (test samples, vehicle controls, and reference chemical treatments)</td>
</tr>
<tr>
<td>Nsig&lt;sub&gt;80&lt;/sub&gt;</td>
<td>The number of probes capturing the top 80% of signal in a sample</td>
<td>Reject &lt; 1,000</td>
<td></td>
</tr>
<tr>
<td>GiC</td>
<td>Gini coefficient computed for each sample based on the distribution of raw counts for all probes including those with 0 aligned reads</td>
<td>Reject &gt; 0.95</td>
<td></td>
</tr>
</tbody>
</table>

Harrill et al. (2021) DOI: [10.1093/toxsci/kfab009](https://doi.org/10.1093/toxsci/kfab009)
HTTr QC Results – MCF7 Screen

- The screen contained 32,886 TempO-Seq samples.
- None of the lysis buffer blank samples passed the QC criteria.
- >99% of test samples were of acceptable quality based on QC criteria.
- In some cases, samples flagged for viability did not fail other QC criteria.
Signature Concentration-Response Modeling

Method intended to address coordinated changes in expression in genes belonging to the same gene set / signaling pathway.

GitHub: github.com/USEPA/CompTox-httrpathway
(Richard Judson)
Signature Scoring Procedure

Count data per chemical

Estimate fold-changes for all genes

DESeq2

ssGSEA

Catalog of gene set signatures with toxicological relevance, annotated for known molecular targets

- Bioplanet (Huang, et al. Front Pharmacol 2019)
- DisGeNET (Pinero, et al. Database 2015)
- MSigDB (Liberzon, et al. Cell Syst 2015)


- Score coordinated responses at each concentration
- Test for multiple genes in a signature enriched among most extreme fold-changes
Signature Scoring of Reference Treatments

- Differential expression analysis of 3 reference chemical exposures repeated 37 times (MCF7)
- Computed distribution of correlations between each repeat analysis
- Signature scores have higher reproducibility than fold-changes, especially for weaker effect sizes
• Reference treatments produced higher absolute signature scores for signatures associated with primary mechanisms of action.

• The expected biology was identified!

• Reference treatments did not produce higher absolute signature scores in a set of synthetic “random” signatures.
Signature Score Concentration-Response Modeling

Concentration response modeling of signature scores using `tcplfit2` ([github/USEPA/CompTox-ToxCast-tcplFit2/](https://github/USEPA/CompTox-ToxCast-tcplFit2/))

 Ranked of Active Signatures

**Signed, Scaled Area Under the Curve (ssAUC)**

\[
\text{3-log}_{10}(\text{BMC}) \times | \text{Top / Cutoff} | \times \text{Sign (Top)}
\]

Used to discern mechanism

**mPOD**

Most sensitive signature

OR

Statistic based on distribution of active signatures (5\text{th} %ile)
Fulvestrant Signature (Top 100 Up & Down Genes)

The expression of fulvestrant signature “down” genes goes down following ER antagonist treatment.

These gene level data are noisy!

Signature level results display correct directionality!

The expression of fulvestrant signature “down” genes goes up following ER agonist treatment.

Harrill et al. (2021) DOI: 10.1093/toxsci/kfab009
Chemicals with known pharmacological targets in MCF7 cells show an “early wave” of biological activity. Other potent toxicants (organometallics, dyes, etc) cause many signatures to be affected near the onset of biological activity.
Mechanistic Clustering Using Signature ssAUC

Potent chemicals with known specificity for a molecular target expressed in MCF7 cells tend to cluster together when using signature scores (ssAUC) as the response metric.

Chemicals with BMC_{0.05} < 1 \mu M
Exploring Similarities in Chemical Response (1)

Data visualization tools (UMAP) help identify chemicals that produce similar responses.
Exploring Similarities in Chemical Response

Similar classification of chemicals as ER agonists or ER antagonists using HTS or HTTr signatures.
To develop frameworks for the standardisation of reporting of ‘omics data generation and analysis, to ensure that all of the information required to understand, interpret and reproduce an ‘omics experiment and its results are available.

**Purpose:** to ensure that sufficient information is available to enable an evaluation of the quality of the experimental data and interpretation, and support reproducibility.

**NOT** to stipulate the methods of data analysis or interpretation… **Rather,** provide guidance on reporting of information that fosters transparency and reproducibility.

<table>
<thead>
<tr>
<th>Project Name</th>
<th>Project Leads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolomics Reporting Framework (MRF)</td>
<td>Mark Viant (Univ. Birmingham, UK)</td>
</tr>
<tr>
<td>Transcriptomics Reporting Framework (TRF)</td>
<td>Joshua Harrill (US EPA)</td>
</tr>
<tr>
<td></td>
<td>Carole Yauk (University of Ottawa)</td>
</tr>
<tr>
<td></td>
<td>Matt Meier (Health Canada)</td>
</tr>
<tr>
<td>OECD Secretariat</td>
<td>Magda Sachana</td>
</tr>
</tbody>
</table>
Each module has:
1) a reporting template (Excel)
2) a narrative guidance

- Both TRF and MRF
- TRF only
- MRF only
OORF Reporting Templates

**8. Data Analysis Reporting Module (DARM) for Detection of Enriched Biological Pathways**

<table>
<thead>
<tr>
<th>Reporting Category</th>
<th>Reporting Element</th>
<th>Required / Optional</th>
<th>Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1. Software Documentation</td>
<td>Software and Documentation</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Software</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Operating System</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Additional Libraries used</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Software Availability</td>
<td>Optional</td>
<td></td>
</tr>
<tr>
<td>8.2. Description of Data Used as Data Description</td>
<td>Data used as input</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Methods used to produce input</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pre-filing of input data</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pre-processing and/or normalization</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Background set(s) used</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td>8.3. Contrasts for Which Contrasts</td>
<td>Required</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.4. Database of Pathways or Gene Biological Entity or Biological Set</td>
<td>Required</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Species Name</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Version or Date of Biological Set</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td>8.5. Enriched Pathways Statistical Test Performed to Identify</td>
<td>Required</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistical Threshold Applied</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multiple Testing Correction Method</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Additional Filtering</td>
<td>Required</td>
<td></td>
</tr>
<tr>
<td>8.6. Outputs Outputs and Supporting Files</td>
<td>Output and Supporting Files</td>
<td>Required</td>
<td></td>
</tr>
</tbody>
</table>

- Prompts data providers to report details of their experiment.
- Links to narrative guidance with descriptions of what type of information to enter in each field.
Refining the OORF with Paired Trials

<table>
<thead>
<tr>
<th>Participant</th>
<th>Step</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data provider:</td>
<td>1</td>
<td>I. Identify omics dataset</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II. Process and analyze dataset</td>
</tr>
<tr>
<td></td>
<td></td>
<td>III. Populate reporting fields in one or more modules of the TRF/MRF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IV. Write an ease-of-use commentary</td>
</tr>
<tr>
<td>Trialling coordinator:</td>
<td>2</td>
<td>I. Review report for completeness (return to data provider if incomplete)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II. Truncate and blind results and send to ‘end user’</td>
</tr>
<tr>
<td>End user:</td>
<td>3</td>
<td>I. Use blinded reporting template to reprocess and reproduce the original analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II. Write a second ease-of-use commentary</td>
</tr>
<tr>
<td>Trialling coordinator:</td>
<td>4</td>
<td>I. Review report for completeness (return to end user if incomplete)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II. Concordance analysis of two completed TRF/MRF reporting templates</td>
</tr>
<tr>
<td>TRF/MRF expert group:</td>
<td>5</td>
<td>I. Review concordance analysis and two ease-of-use commentaries</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II. Update TRF/MRF reporting templates as required</td>
</tr>
</tbody>
</table>
OECD Omics Reporting Framework Project

- Genesis and progress towards the OORF detailed in this publication.
- Early draft available at OECD omics website.
- Review and approval by OECD Extended Advisory Group for Molecular Screening and Toxicogenomics (EAGMST) in Q3 2022!

https://www.oecd.org/chemicalsafety/testing/omics.htm
Summary

- Innovations in High-Throughput Transcriptomics Screening:
  - Targeted RNA-Seq assay.
  - Scalable laboratory workflows for chemical exposure and lysate generation.
  - Reproducible, open-source data analysis pipeline(s).
  - Improvements to ToxCast pipeline (tcpl) concentration-response modeling software.
  - Novel approach for signature level concentration-response analysis.
  - Data visualization techniques for exploring chemical / biological response similarity.

- Development of transcriptomics reference materials.

- Reporting frameworks for toxicology studies involving omics technologies.


Acknowledgements

EPA Office of Research and Development Center for Computational Toxicology and Exposure (CCTE)

Logan Everett
Imran Shah
Richard Judson
Katie Paul Friedman
Derik Haggard
Joseph Bundy
Laura Taylor
Bryant Chambers
Jesse Rogers
Thomas Sheffield
Woody Setzer

Johanna Nyffeler
Clinton Willis
Megan Culbreth
Gabrielle Byrd
Felix Harris
Dan Hallinger
Terri Fairley

Logos Reporters Framework Contributors

BioSpyder
Garrett McComb
Jo Yeakley
Bruce Seligmann
Joel McComb
Pete Shepherd
Milos Babic
Dalia Gonzalez
Kyle LeBlanc

Leadership Team
Joshua Harrill
Carole Yauk
Mark Viant
Matthew Meier
Magda Sachana

EAGMST Co-chairs
Rusty Thomas
Maurice Whelan

Tim Gant
Scott Auerbach
Rick Beger
Mounir Bouhifd
Jason O-Brien
Lyle Burgoon
Florian Caient
Donatella Carpi
Tao Chen
Brian Chorley
John Colbourne
Raffaella Corvi
Claire O'Donovan
Laurent Debrauwer
Timothy Ebbels
Drew Ekman
Frank Faulhammer
Laura Gribaldo

Gina Hilton
Stephanie Jones
Aniko Kende
Thomas Lawson
Sofia Batista-Leite
Pim Leonards
Mirjam Luijten
Alberto Martin
Laura Moussa
Serge Rudaz
Oliver Schmitz
Tomasz Sobanski
Volker Strauss
Monica Vaccari
Vikrant Vijay
Ralf Weber
Antony Williams
Andrew Williams