High-Throughput Transcriptomics (HTTr) for Chemical Safety Screening

Logan J. Everett, Ph.D.

Bioinformatics Scientist

Biomolecular and Computational Toxicology Division

Center for Computational Toxicology & Exposure

Office of Research and Development, U.S. EPA

Research Triangle Park, North Carolina

Disclaimer

The views expressed in this presentation are those of the presenter and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. Company or product names do not constitute endorsement by US EPA.

Acknowledgements

HTTr Team

Josh Harrill
Richard Judson
Imran Shah
Woody Setzer
Katie Paul Friedman
Derik Haggard
Beena Vallanat
Joseph Bundy
Bryant Chambers
Jesse Rogers
Laura Taylor
Clinton Willis
Thomas Sheffield
Johanna Nyffeler
Mark Higuchi
Adam Speen

CCTE Leadership

Rusty Thomas
Sid Hunter
Drew Watkins
John Cowden
Kimberly Slentz-Kesler

Outline

- Tiered testing strategy using New Approach Methods (NAMs) to fill gaps in environmental chemical safety data
- Workflow for transcriptomic profiling of chemical effects *in vitro*
- Data analysis tools for high-throughput transcriptomics (HTTr) data
- Validation of HTTr data & ongoing research

Chemical Safety Testing Strategy Rationale

Rationale: Too Many Chemicals, Too Little Data!

- 1,000s of chemicals used in USA for non-food/drug applications
- Many chemicals lack safety data for human health & ecological impacts
- Traditional toxicity testing is costly and slow
  - *2-year rodent cancer bioassay costs over $1 million per substance*
- Fast, flexible, cost-effective method needed to fill gaps in safety data
- New Approach Methods (NAMs) aim to provide toxicity data without the use of animal testing (e.g. *in vitro* screening)

Tiered Chemical Safety Testing Strategy

Tier 1: Broad coverage, high content assays

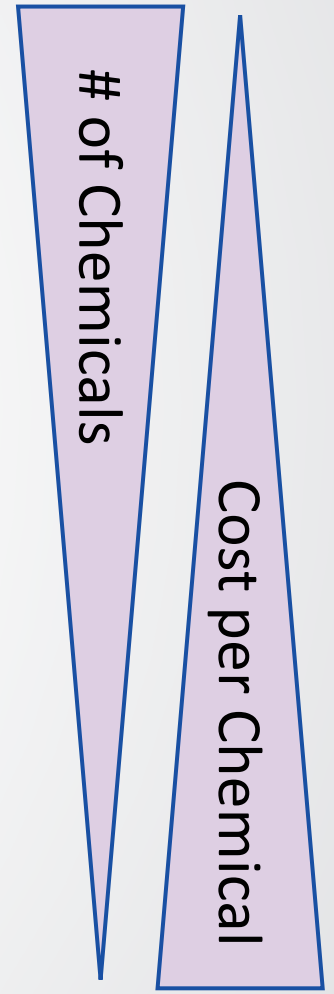
- Must be cost-effective enough to rapidly screen 1000s of chemicals
  - e.g. Transcriptomics and/or cell imaging applied *in vitro*
  - Acute exposure: 6 - 24 hours
  - Multiple cell types with different metabolic profiles
- Goals: Prioritize chemicals by bioactivity & potency for further testing

Tier 2: Targeted *in vitro* assays

- Goals: confirm bioactivity & potency of chemicals flagged for potential safety issues

Tier 3: Organotypic assays, systems modeling, and more

- Goals: identify likely tissue, organ, or organism effect of chemical



See also: Thomas, et al. Toxicol Sci 2019

Outline

- Tiered testing strategy using New Approach Methods (NAMs) to fill gaps in environmental chemical safety data

➢**Workflow for transcriptomic profiling of chemical effects *in vitro***

- Data analysis tools for high-throughput transcriptomics (HTTr) data

- Validation of HTTr data & ongoing research

Automated *in vitro* Chemical Screening Strategy

Cryopreserved Cell Stocks

Cell Expansion

Cell Line Examples:	
MCF-7	Breast cancer
U-2 OS	Bone cancer
HepaRG	Liver metabolism

Cell Plating

BioTek MultiFlo ™ FX

Dispensing Test Chemicals

LabCyte Echo® 550 Liquid Handler

6 - 24 Hour Exposures

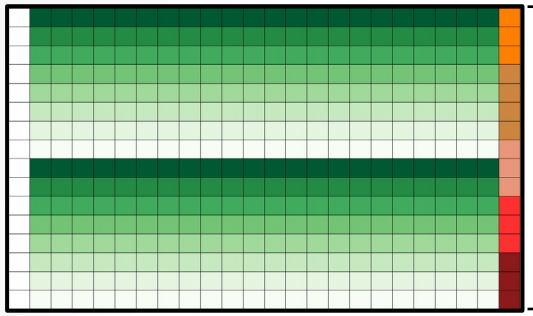
- **Cell imaging identifies cytotoxic exposures**
- **Targeted sequencing for changes in gene regulation (TempO-seq)**

Adapted from Joshua Harrill

High-Throughput *In Vitro* Chemical Screen Design

Test Chemicals:
- 8 Concentrations
- ½ Log_{10} Spacing

Chemical Dose Plate



Reference Chemicals & Vehicle Controls

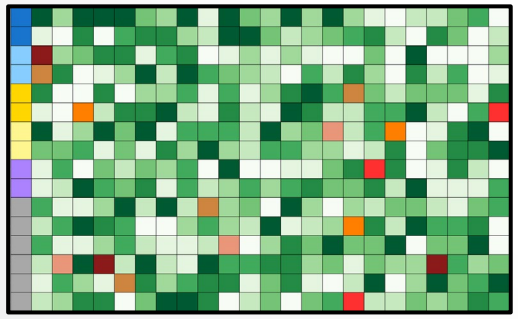
(Same on all plates)

Cell Stocks

Cells evenly plated

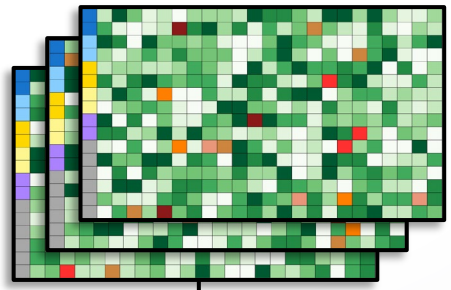
Chemical treatments randomized to test plate

Quality Control Samples
- Reference Mixtures
- Bulk Lysate from Single Exposure Experiment



x3

Triplicate Test Plates



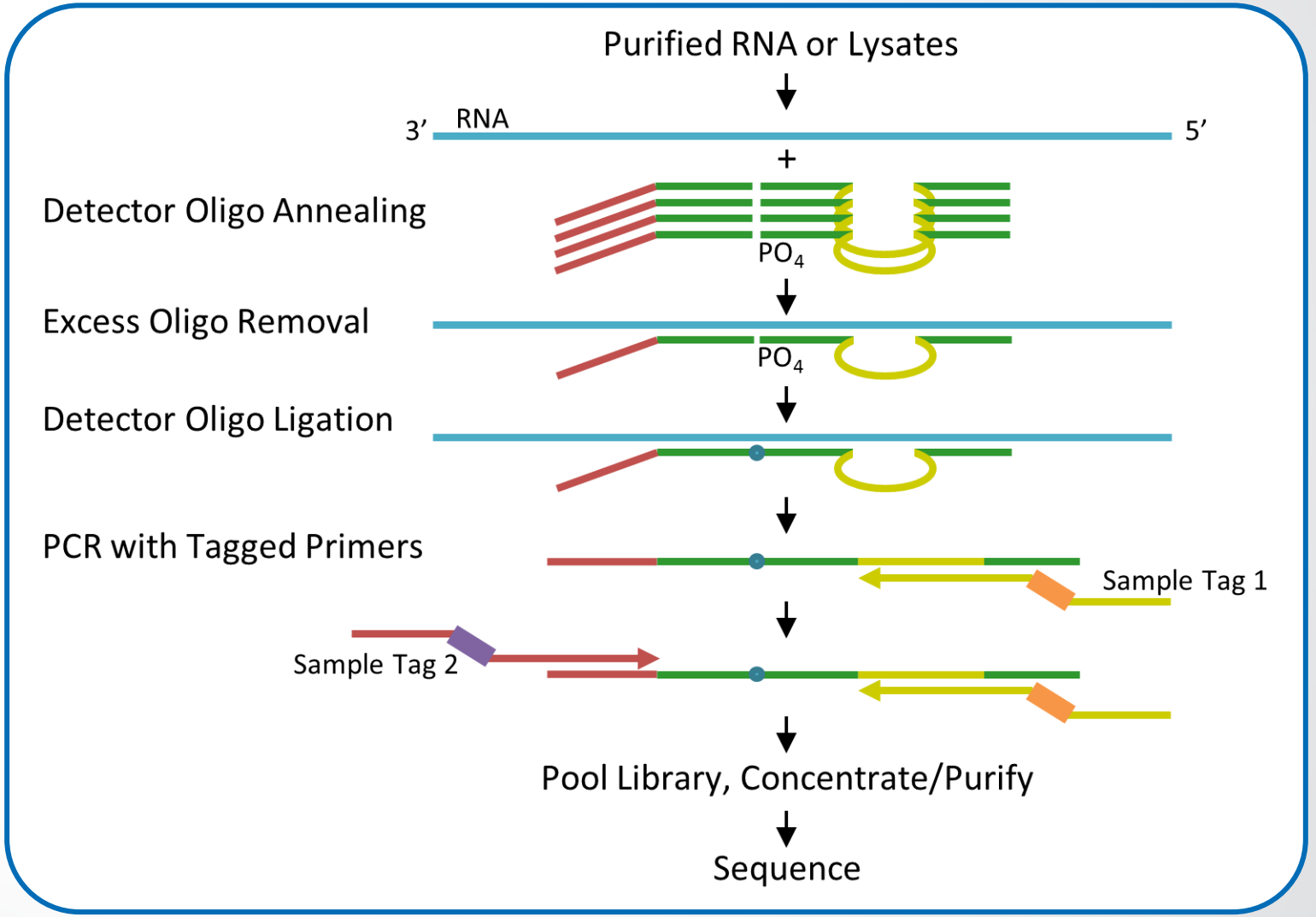
Sequencing/Imaging

Each 384 well test plate has:

- Cells from separately expanded batches
- Standardized dilution series for every chemical test sample, dispensed in an independently randomized manner
- Multiple quality control and reference chemicals to track assay performance

Targeted RNA-seq Assay (TempO-seq)

- Profiling of whole human transcriptome (~21,000 protein-coding genes)
- Captures sufficient biological signal at much lower cost than other methods
- Do **not** need to purify RNA



Yeakley, et al. PLoS ONE 2017

Outline

- Tiered testing strategy using New Approach Methods (NAMs) to fill gaps in environmental chemical safety data

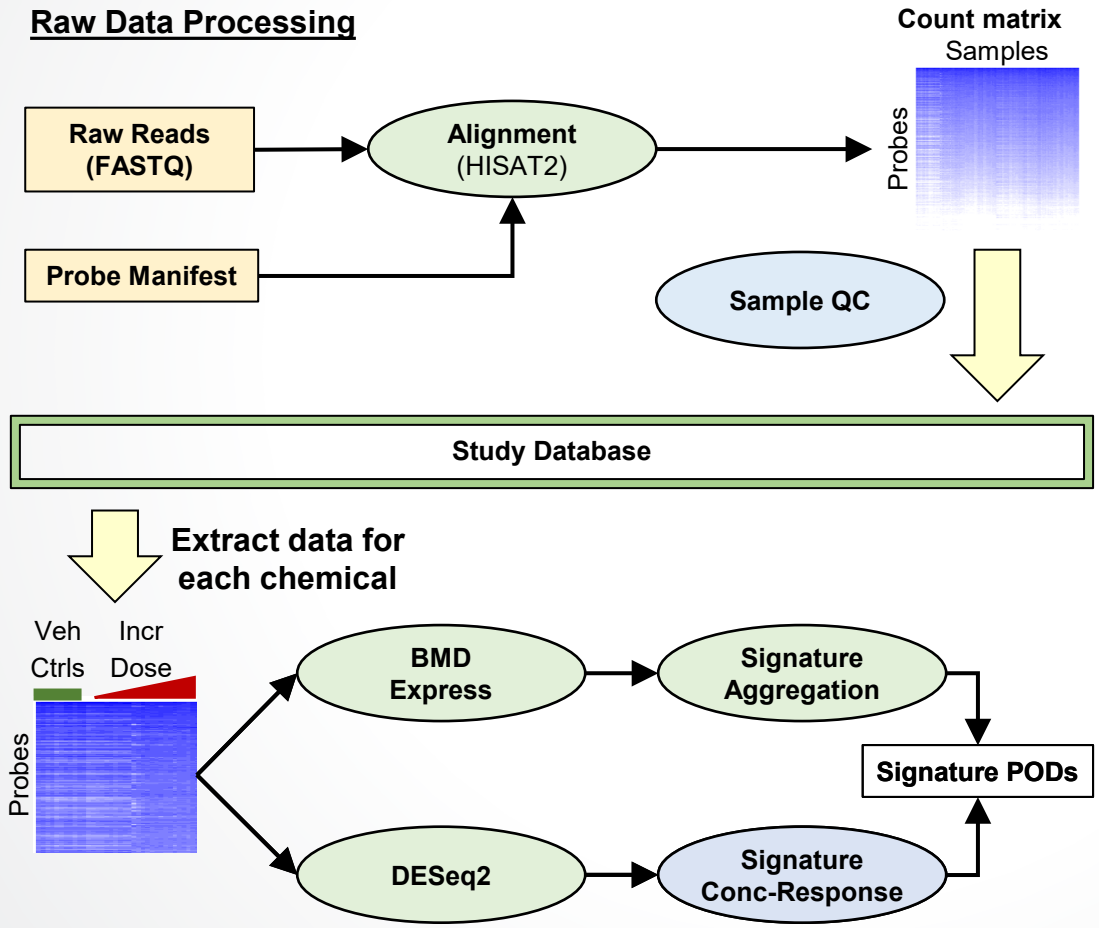
- Workflow for transcriptomic profiling of chemical effects *in vitro*

➢**Data analysis tools for high-throughput transcriptomics (HTTr) data**

- Validation of HTTr data & ongoing research

HTTr Bioinformatics Pipeline

10,000s of samples
Terabytes of data



Harrill, et al. Toxicol Sci 2021

Primary Goals:

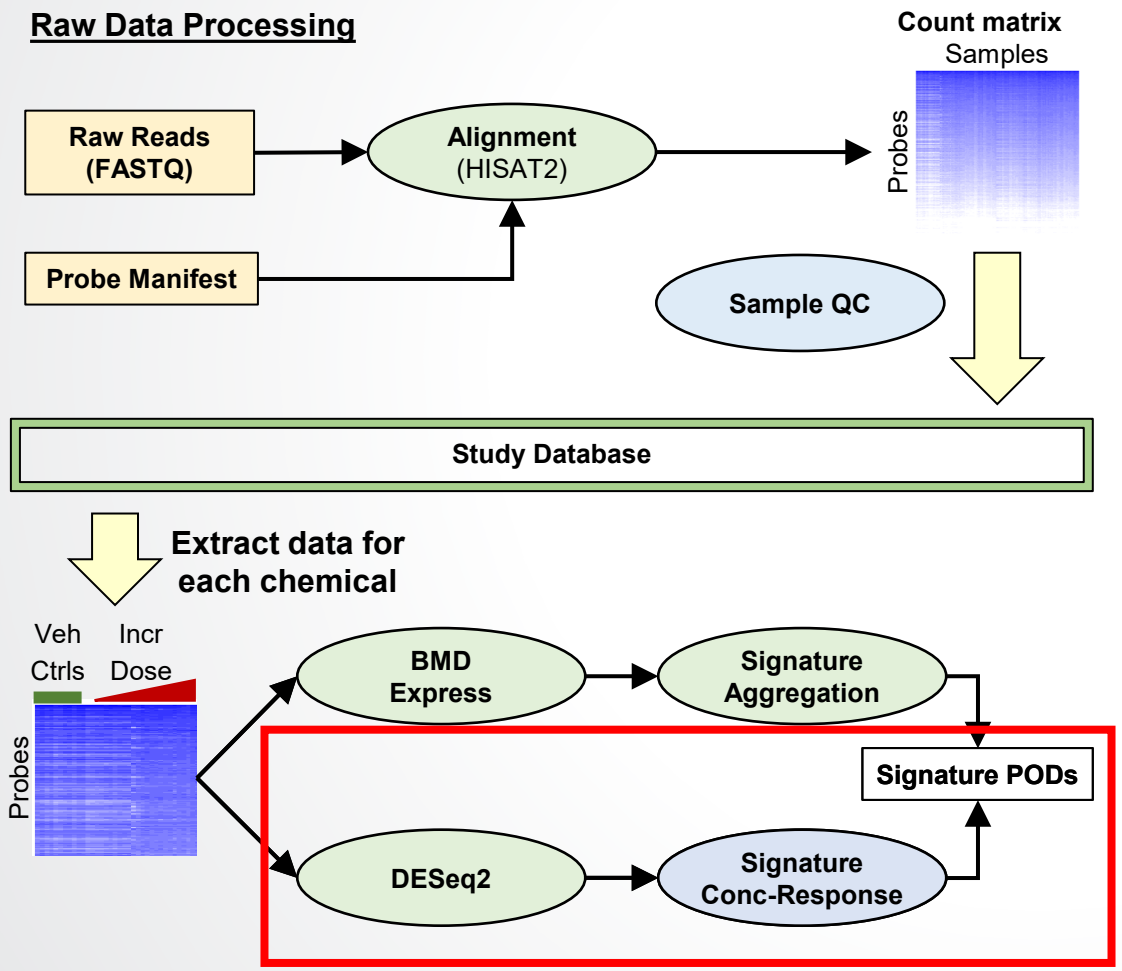
- Speed up & automate compute intensive steps
- Reproducible & open source

github.com/USEPA/httrpl_pilot

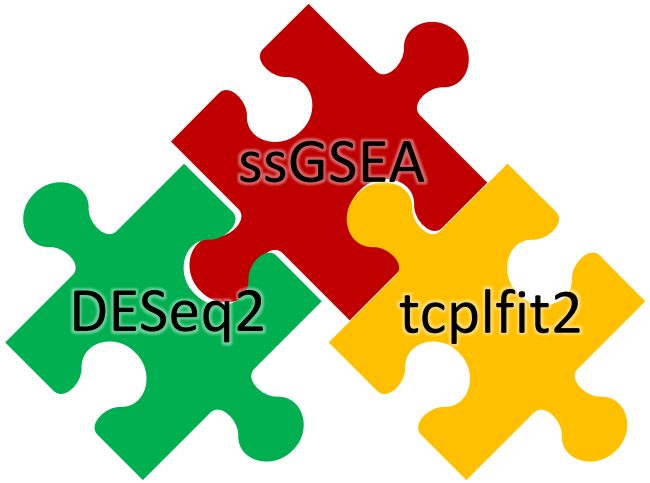
github.com/USEPA/CompTox-httrpathway

Currently exploring multiple analysis strategies for estimating & summarizing points of departure (PODs)

HTTr Bioinformatics Pipeline



Signature Concentration-Response Modeling



github.com/USEPA/CompTox-httrpathway

(Richard Judson)

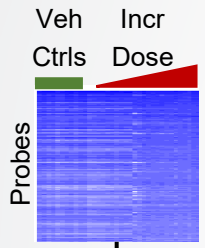
Primarily interested in transcriptional changes that:

- **Are coordinated across known pathways/gene sets**
- **Fit standard curve-models across all concentrations**

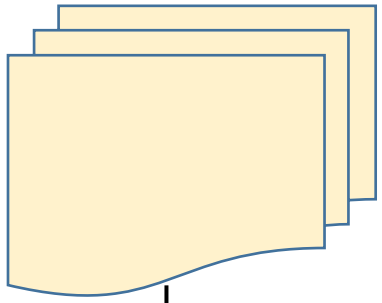
Harrill, et al. Toxicol Sci 2021

Signature Scoring

Count data per chemical



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



- **Bioplanet** *(Huang, et al. Front Pharmacol 2019)*
- **CMap** *(Subramanian, et al. Cell 2017)*
- **DisGeNET** *(Pinero, et al. Database 2015)*
- **MSigDB** *(Liberzon, et al. Cell Syst 2015)*

Estimate fold-changes for all genes

DESeq2

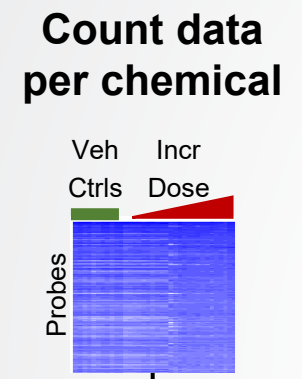
Compute signature scores from all gene expression changes

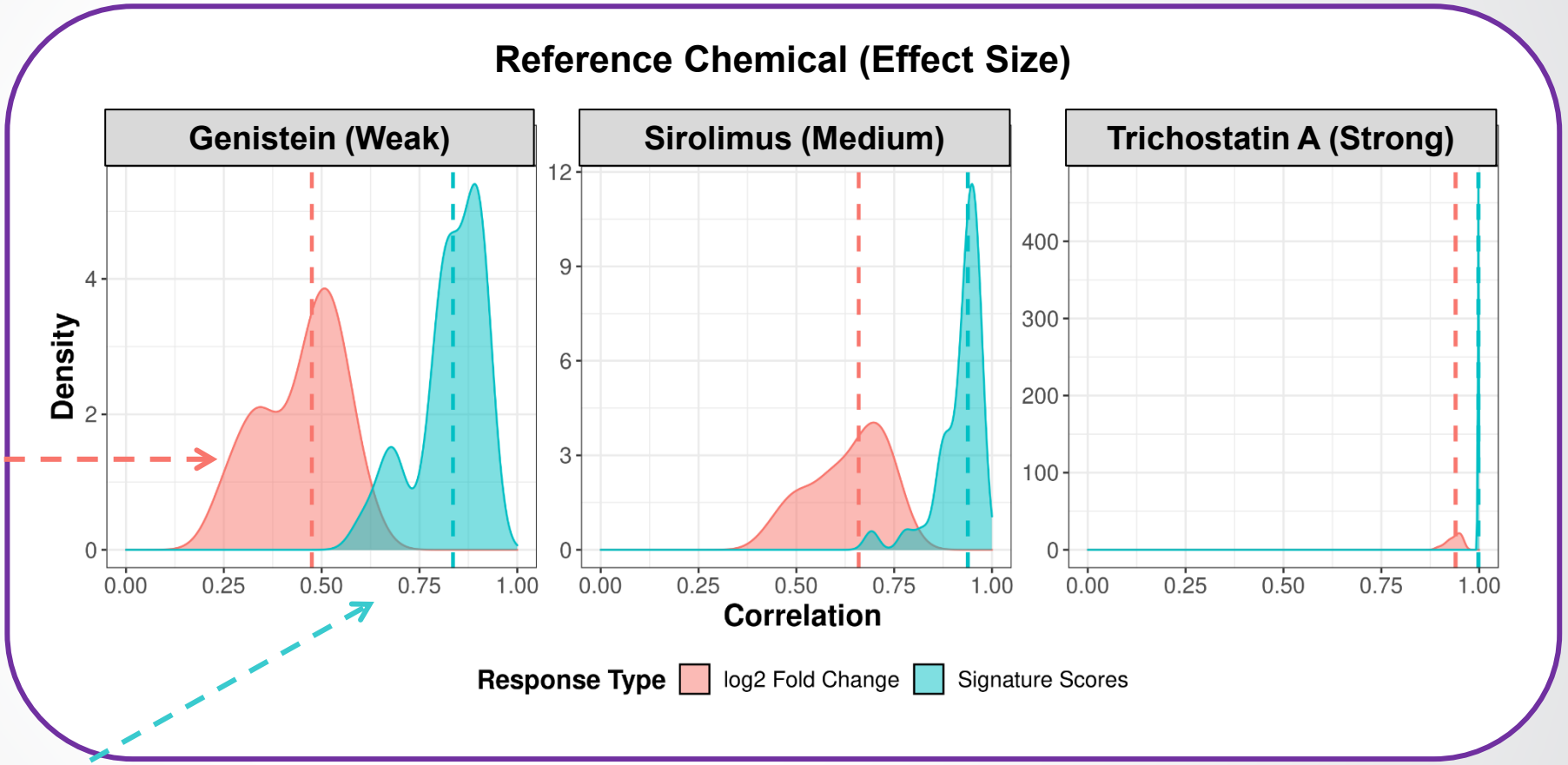
ssGSEA

Single-Sample Gene Set Enrichment Analysis (ssGSEA)
(Barbie, et al. Nature 2009)

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

Signature Scoring

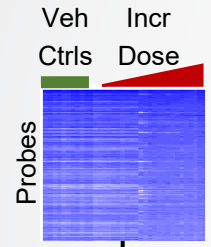




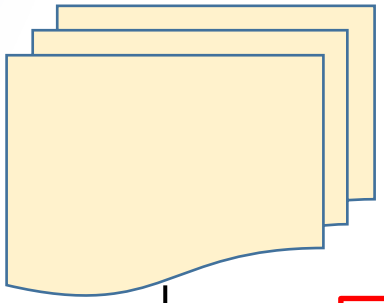
- Differential expression analysis of 3 reference chemical exposures repeated 37 times (MCF-7)
- Computed distribution of correlations between each repeat analysis
- **Signature scores have higher reproducibility than fold-changes, especially for weaker effect sizes**

Signature Scoring

Count data per chemical



Catalog of gene set signatures



DESeq2

ssGSEA

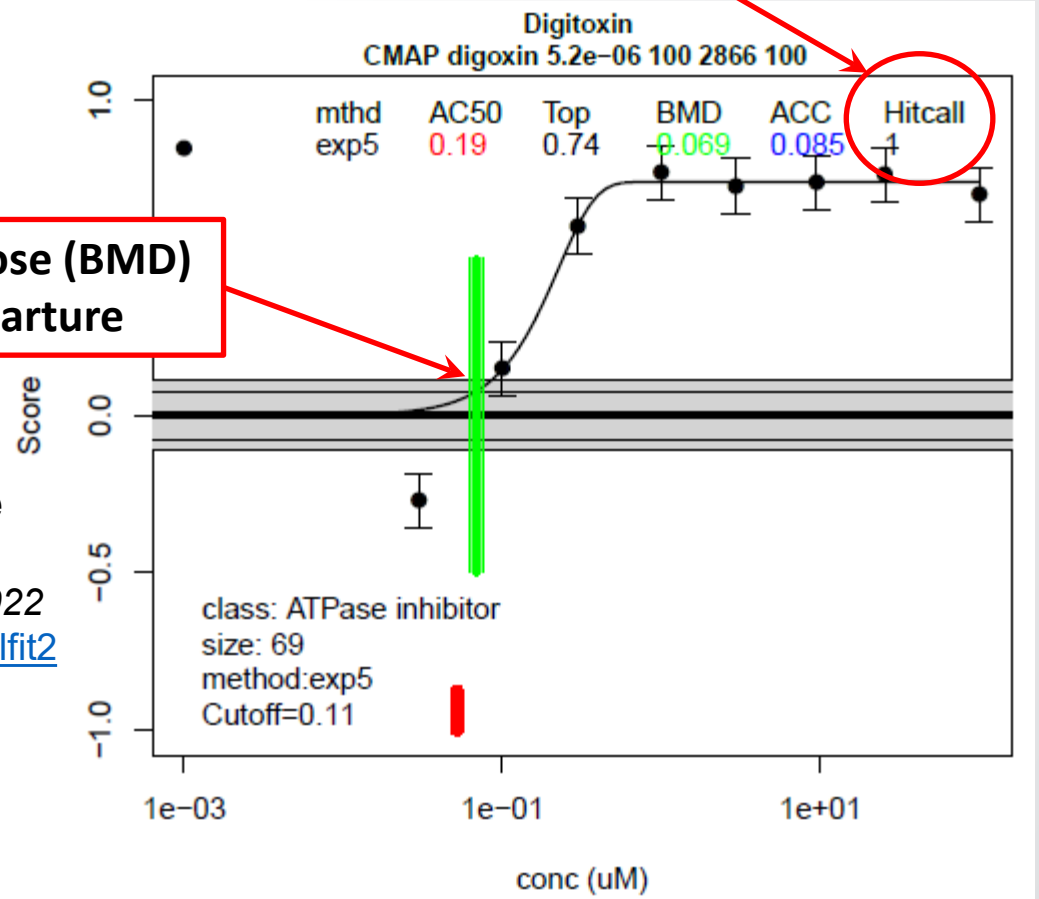
Concentration-Response Curve Fitting (*tcplfit2*)

Sheffield, et al. Bioinformatics 2022

CRAN.R-project.org/package=tcplfit2

Use Benchmark Dose (BMD) as Point of Departure

Hit Call (0-1) indicates confidence that signature is concentration-responsive at all



Outline

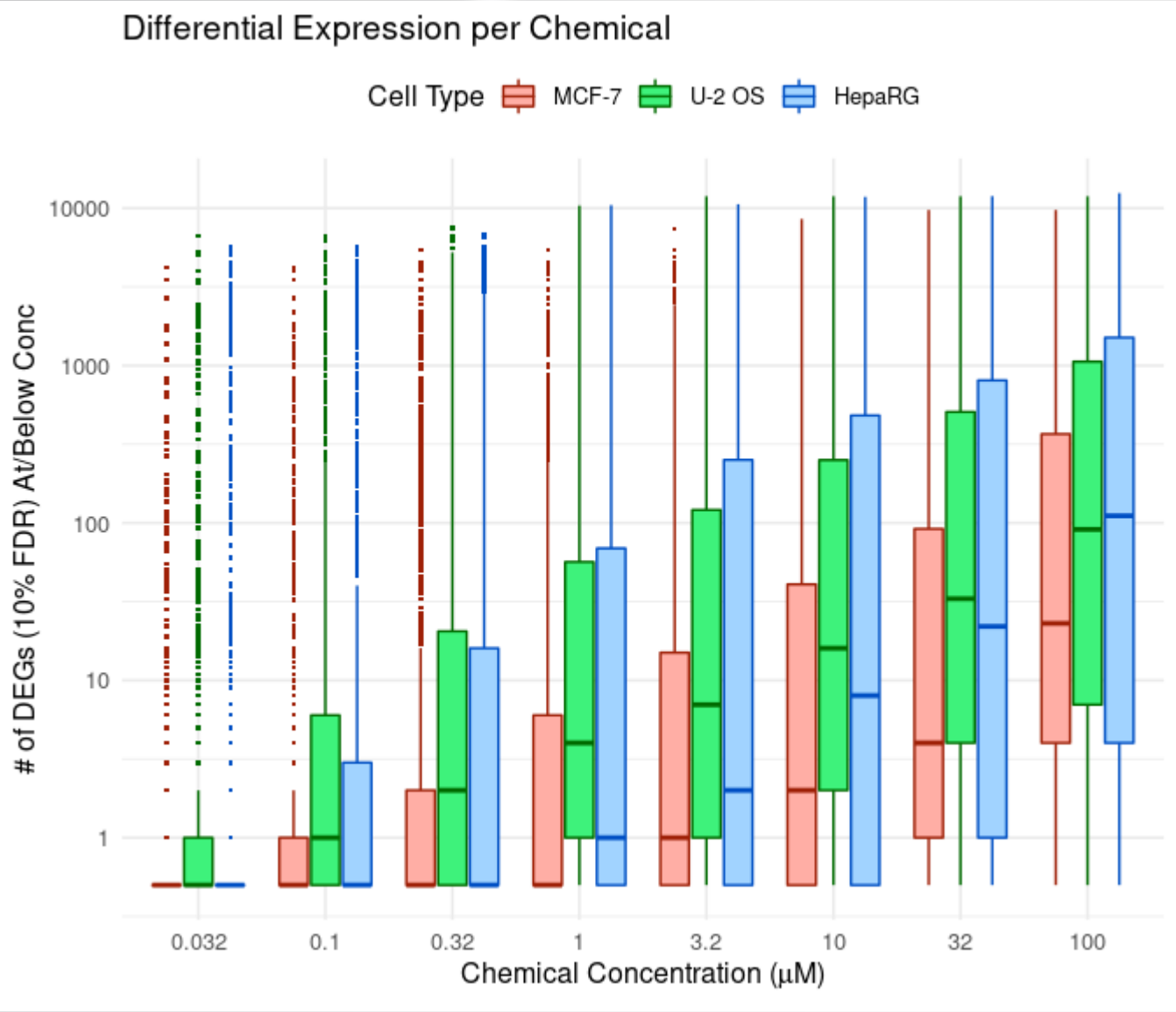
- Tiered testing strategy using New Approach Methods (NAMs) to fill gaps in environmental chemical safety data
- Workflow for transcriptomic profiling of chemical effects *in vitro*
- Data analysis tools for high-throughput transcriptomics (HTTr) data

➢**Validation of HTTr data & ongoing research**

HTTr Studies at EPA

Cell Type	# Chemicals Screened	Conditions	Publication
MCF-7	44 ToxCast chemicals	6h exposures	Harrill, et al. Toxicol Sci 2021
MCF-7	44 ToxCast chemicals	6, 12, 24h exposures x 2 cell media types	*manuscript in preparation*
MCF-7	1,577 ToxCast chemicals	6h exposures	*manuscript in preparation*
U-2 OS	1,201 ToxCast chemicals 137 PFAS chemicals	24h exposures	*manuscript & white paper in preparation*
HepaRG	1,201 ToxCast chemicals 137 PFAS chemicals	24h exposures	*manuscript & white paper in preparation*
BEAS-2B, pHBEC	8 volatile chemicals	Air-Liquid Interface (ALI), 2h exposures	Speen, et al. *in review*

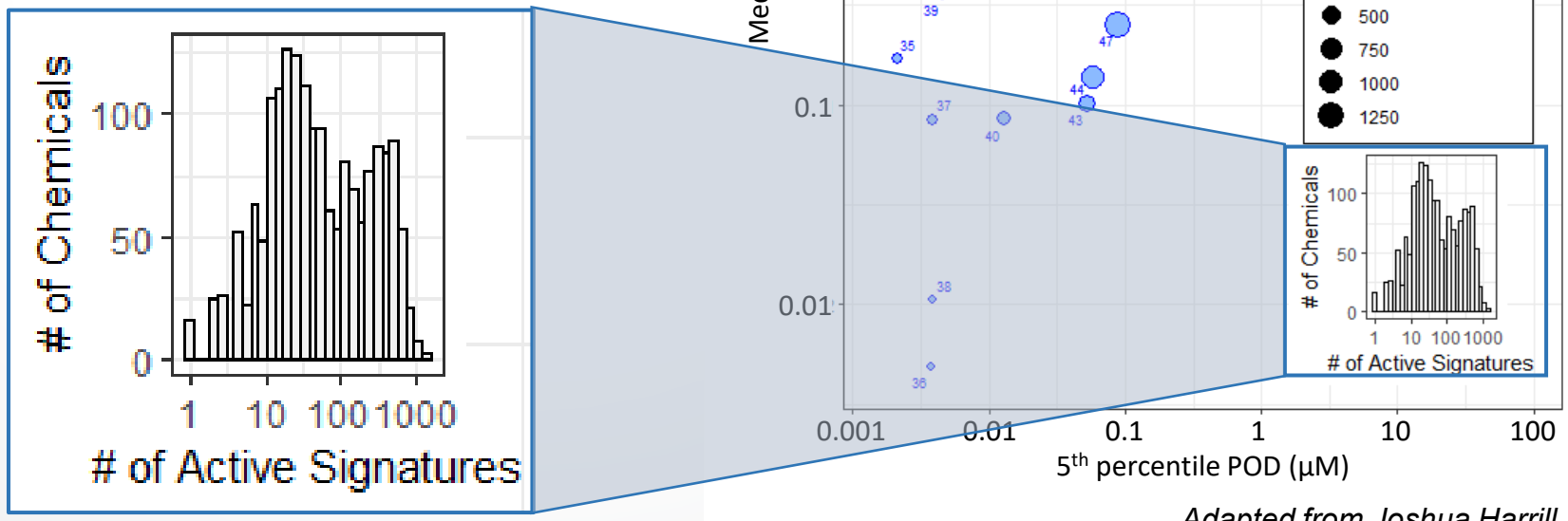
Global View of Bioactivity



- Compute # of Differentially Expressed Genes (DEGs) in response to each concentration of each chemical
  - Based on DESeq2 analysis
  - 10% False Discovery Rate (FDR)
- Each boxplot shows distribution of DEG counts across all tested chemicals
- Majority of chemicals inactive at lowest concentration tested
- Majority of chemicals perturb gene expression at highest concentration tested
  (Tens to thousands of genes)

Full HTTr Screen Results (MCF-7)

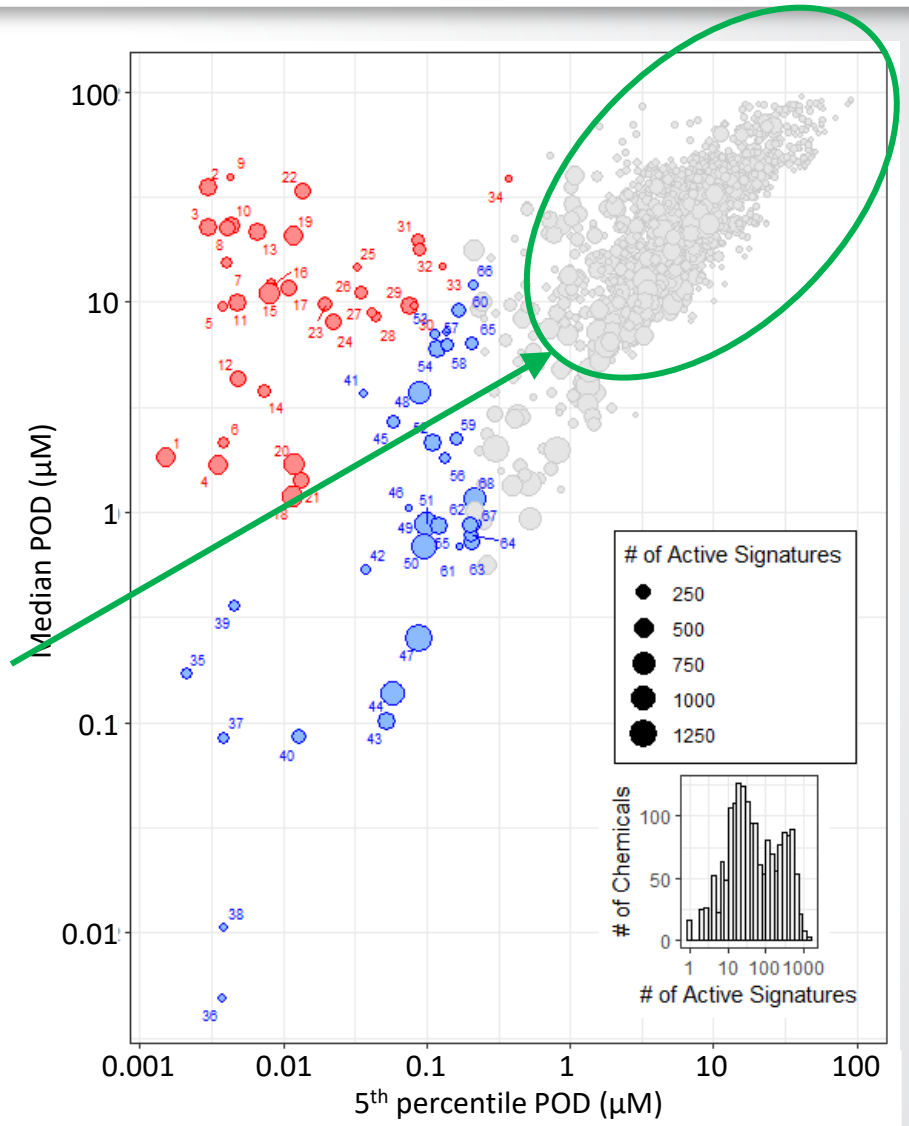
- Performed conc-response analysis with full signature catalog on all ~1,500 chemicals
- Filter to Active Signatures (Hit Call > 0.9)
  - Majority of chemicals have >10 active signatures
  - Many chemicals have 100-1,000 active signatures



Adapted from Joshua Harrill

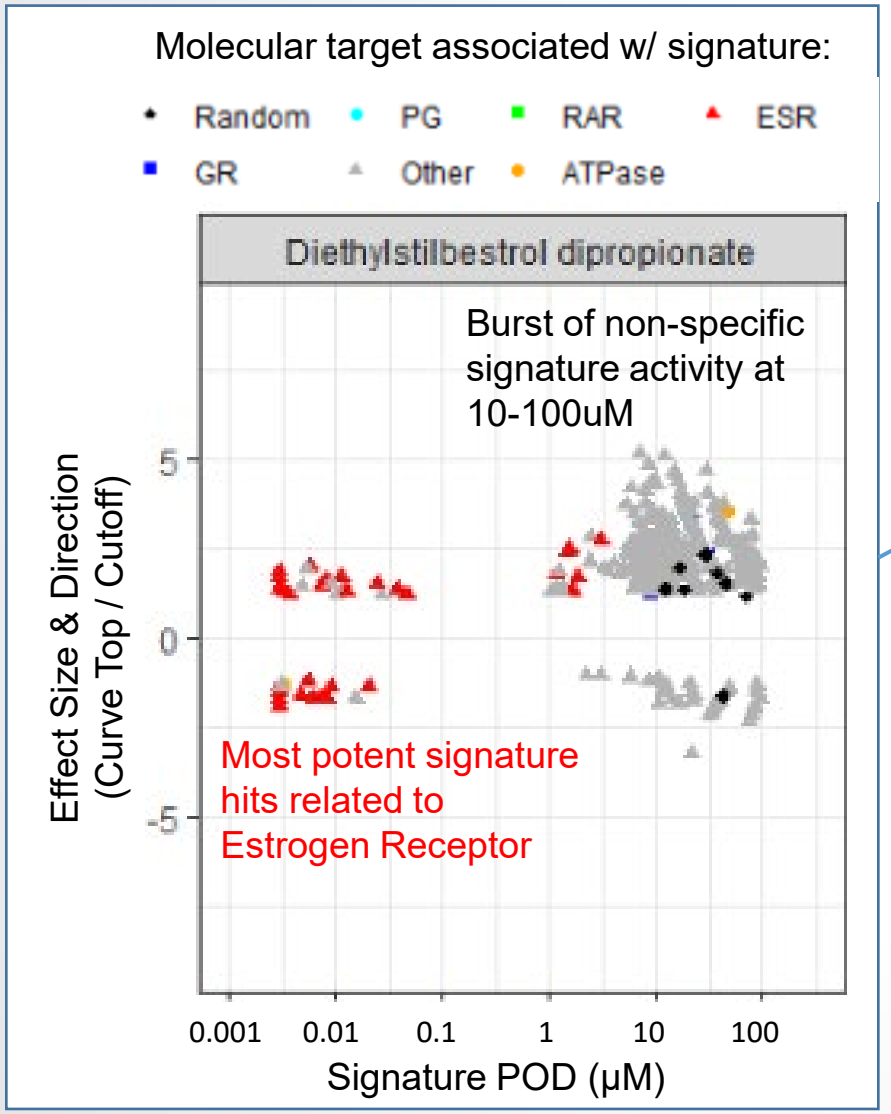
Full HTTr Screen Results (MCF-7)

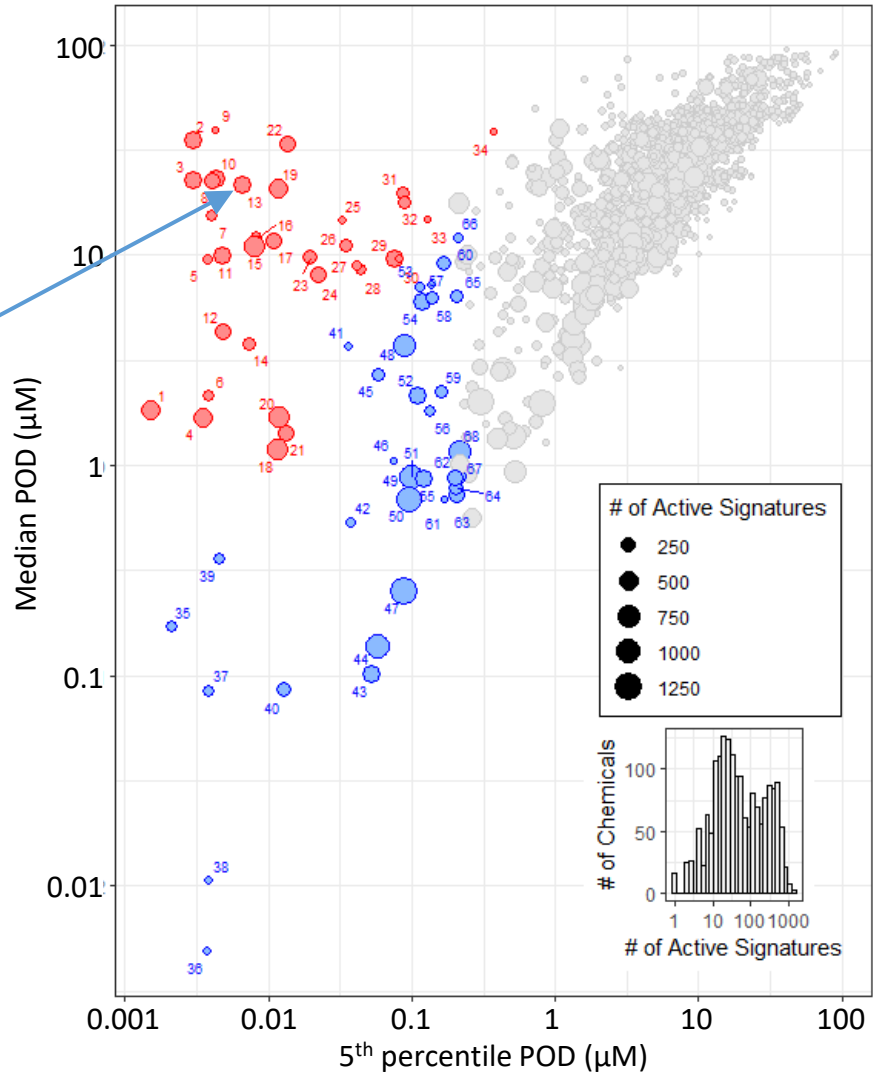
- Performed conc-response analysis with full signature catalog on all ~1,500 chemicals
- Filter to Active Signatures (Hit Call > 0.9)
  - Majority of chemicals have >10 active signatures
  - Many chemicals have 100-1,000 active signatures
- Majority of active signatures are low potency (POD > 10µM)



Adapted from Joshua Harrill

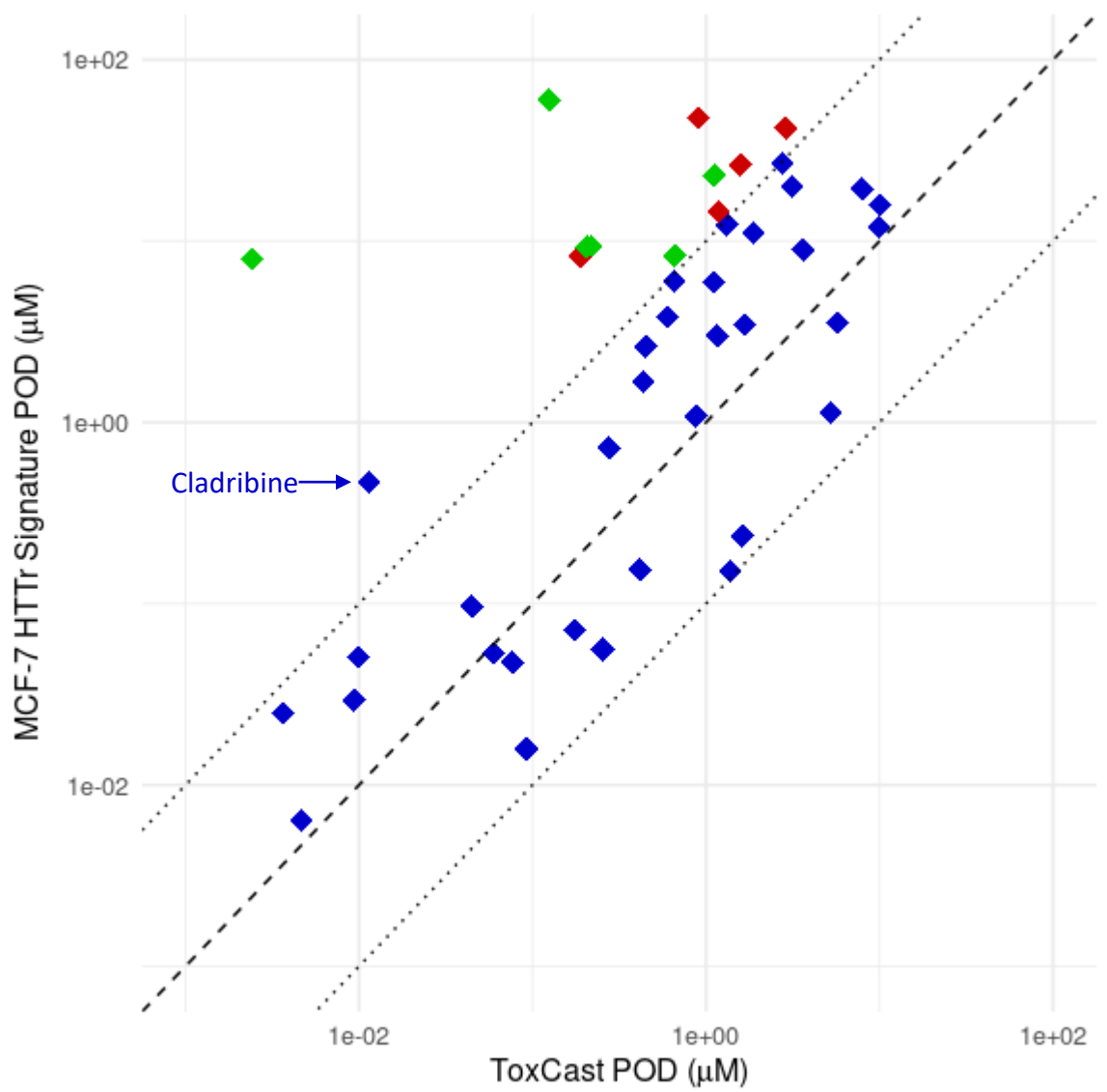
Full HTTr Screen Results (MCF-7)





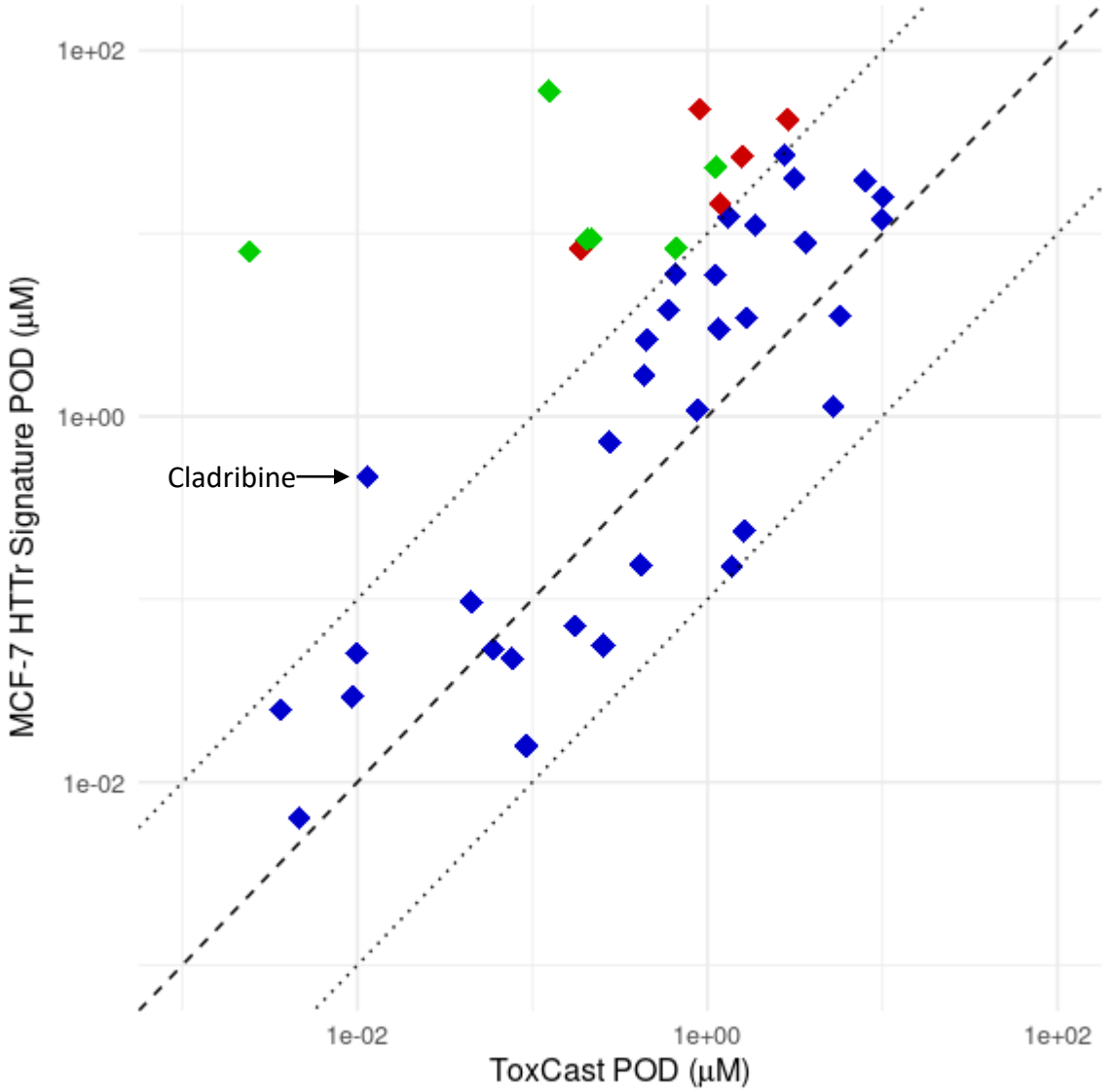
Adapted from Joshua Harrill

HTTr vs ToxCast Targeted Assays



- Pilot study of 44 well-characterized chemicals in MCF-7 cells, 6h exposure ***(Harrill, et al. Toxicol Sci, 2021)***
- Compared HTTr-derived PODs to previous ToxCast targeted assay results *(multiple cell types, assays, and exposure lengths)* (***Paul-Friedman, et al. Toxicol Sci 2020)***
- Signature-based PODs are highly concordant with ToxCast results for the majority of test chemicals in pilot study

HTTr vs ToxCast Targeted Assays



- 6 chemicals with targets that have low/absent expression in MCF-7 cells
  - 3,5,3'-triiodothyronine (Thyroid Receptor)
  - Cyproconazole (pan-CYP inhibitor)
  - Butafenacil (pan-CYP inhibitor)
  - Prochloraz (pan-CYP inhibitor)
  - Imazalil (pan-CYP inhibitor)
  - Propiconazole (pan-CYP inhibitor)
- 5 chemicals where most potent assays in ToxCast do not match known target(s)
  - Lovastatin
  - Clofibrate
  - Maneb
  - Lactofen
  - Vinclozolin
- Cladribine (2-chloro-2'-deoxyadenosine) is a DNA synthesis inhibitor

(Harrill, et al. Toxicol Sci, 2021)

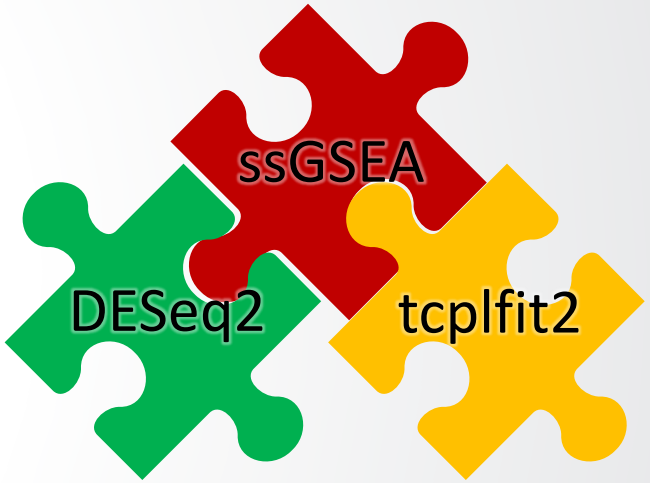
Comparing POD Analysis Methods

Gene-Centric Approaches

- BMDExpress (NTP)
- tcplFit2 (CCTE)
- BIFROST (Unilever)

Compare approaches across multiple types of studies, "best" method may be context-dependent

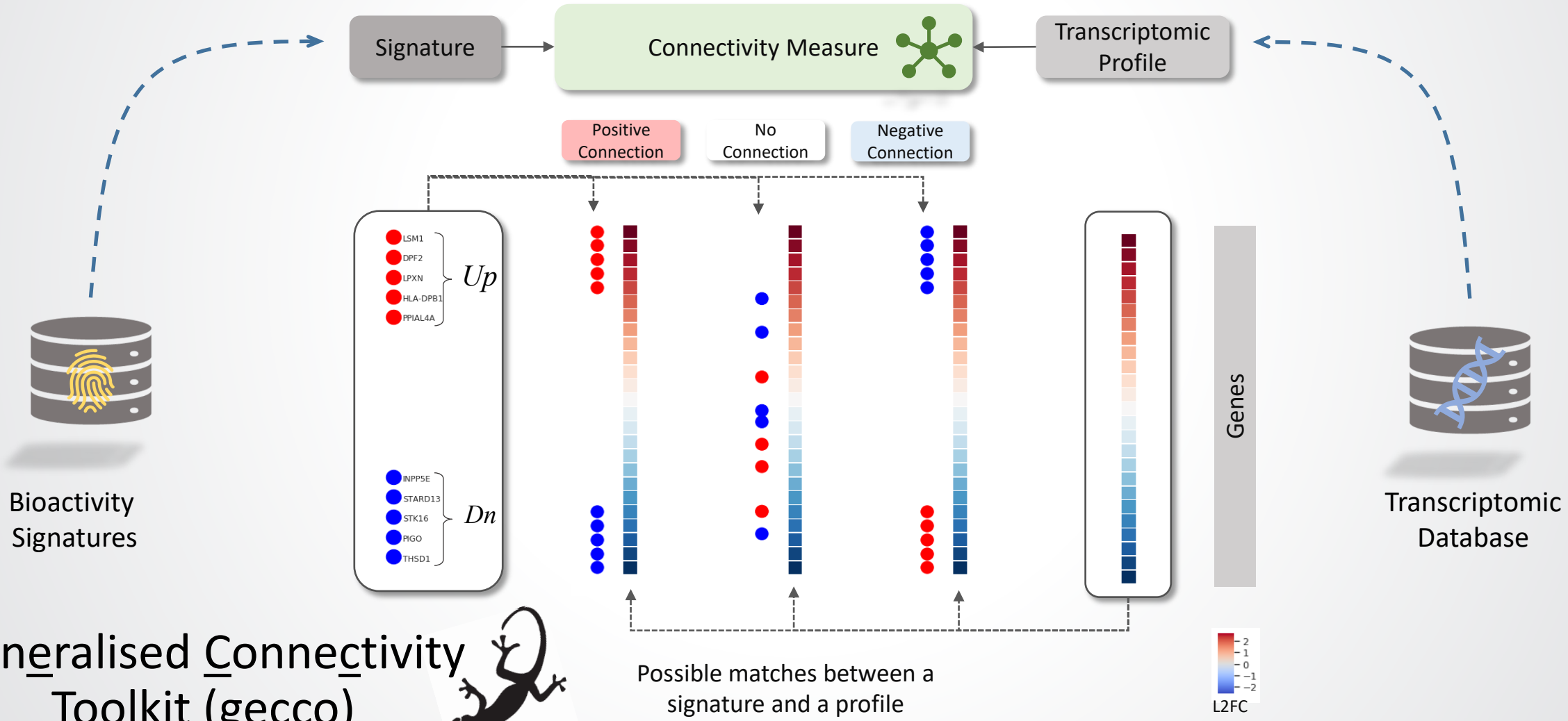
Signature Conc-Response



github.com/USEPA/CompTox-httrpathway
(Richard Judson)

Improved integration through HTTr pipeline & database development

Connectivity-mapping with gene signatures



Generalised Connectivity Toolkit (gecco)

Imran Shah

Summary

- EPA/ORD has developed reliable and cost-efficient workflow for generating HTTr data from thousands of chemicals across multiple cell lines

- Preliminary/pilot analysis demonstrates that overall results are concordant with previous assays (ToxCast/HTS) and known chemical targets
  Harrill, et al. Toxicol Sci 2021

- Ongoing research efforts focused on:
  - Data generation in complementary cell models
  - Methods to summarize signature-level/overall PODs from high-dimensional data
  - Predictive models of MIEs/pathways relevant to toxicity

Acknowledgements

Questions? everett.logan@epa.gov

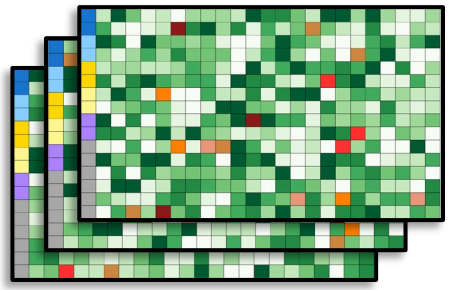
Joshua Harrill
Richard Judson
Imran Shah
Woody Setzer
Katie Paul Friedman
Derik Haggard
Beena Vallanat
Joseph Bundy
Bryant Chambers
Laura Taylor
Clinton Willis
Thomas Sheffield
Jesse Rogers
Johanna Nyffeler
Mark Higuchi
Adam Speen

CCTE Leadership

Rusty Thomas
Sid Hunter
Drew Watkins
John Cowden
Kimberly Slentz-Kesler

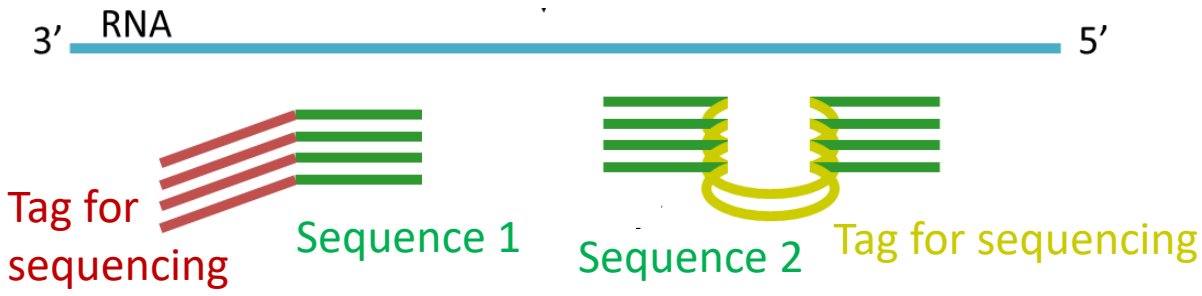
TempO-seq Assay

Triplicate Test Plates

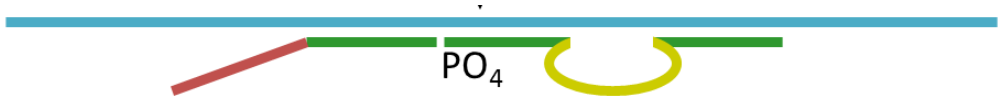


1. Cells lysed, RNA available for assay

2. Paired sequences that match nearby areas of each human gene RNA sequence added

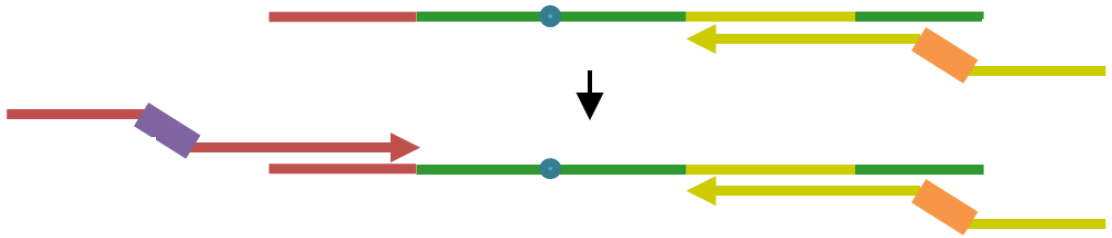


3. Paired sequences hybridize to target RNA when present in sample



4. If paired sequences bind, they are connected into full probe sequence

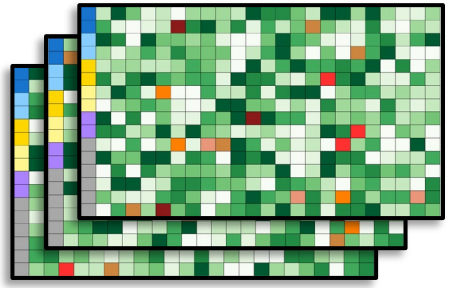
5. RNA removed, PCR using tags to amplify sequences, library of sequences pooled and read



Yeakley, et al. PLoS ONE 2017

HTTr Data Generation

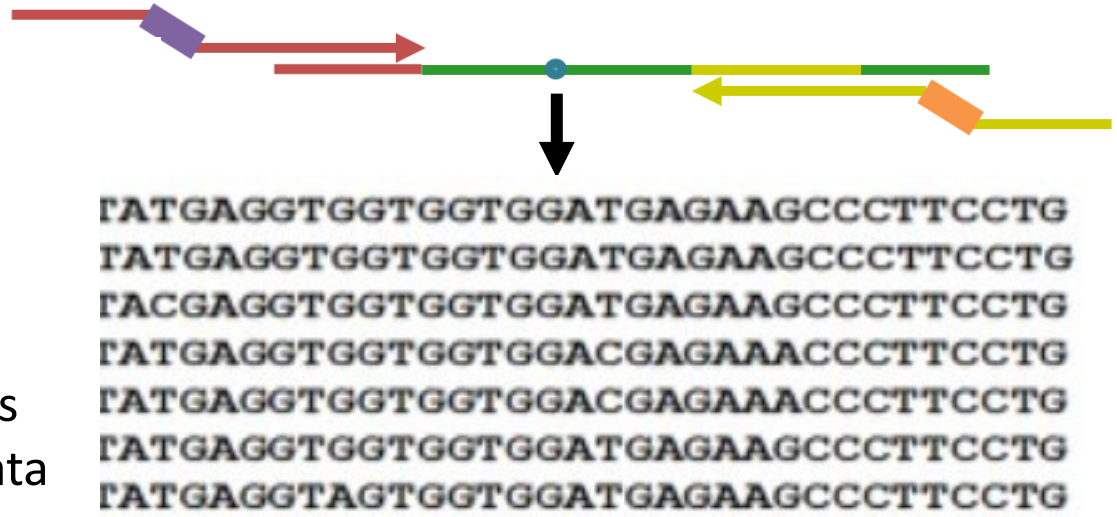
Triplicate Test Plates



5. Library of sequences pooled and read

Resulting Data:

- Millions of 50 nucleotide reads per sample
- 1,000 chemical screen generates ~27,000 samples = ~4 TB raw data

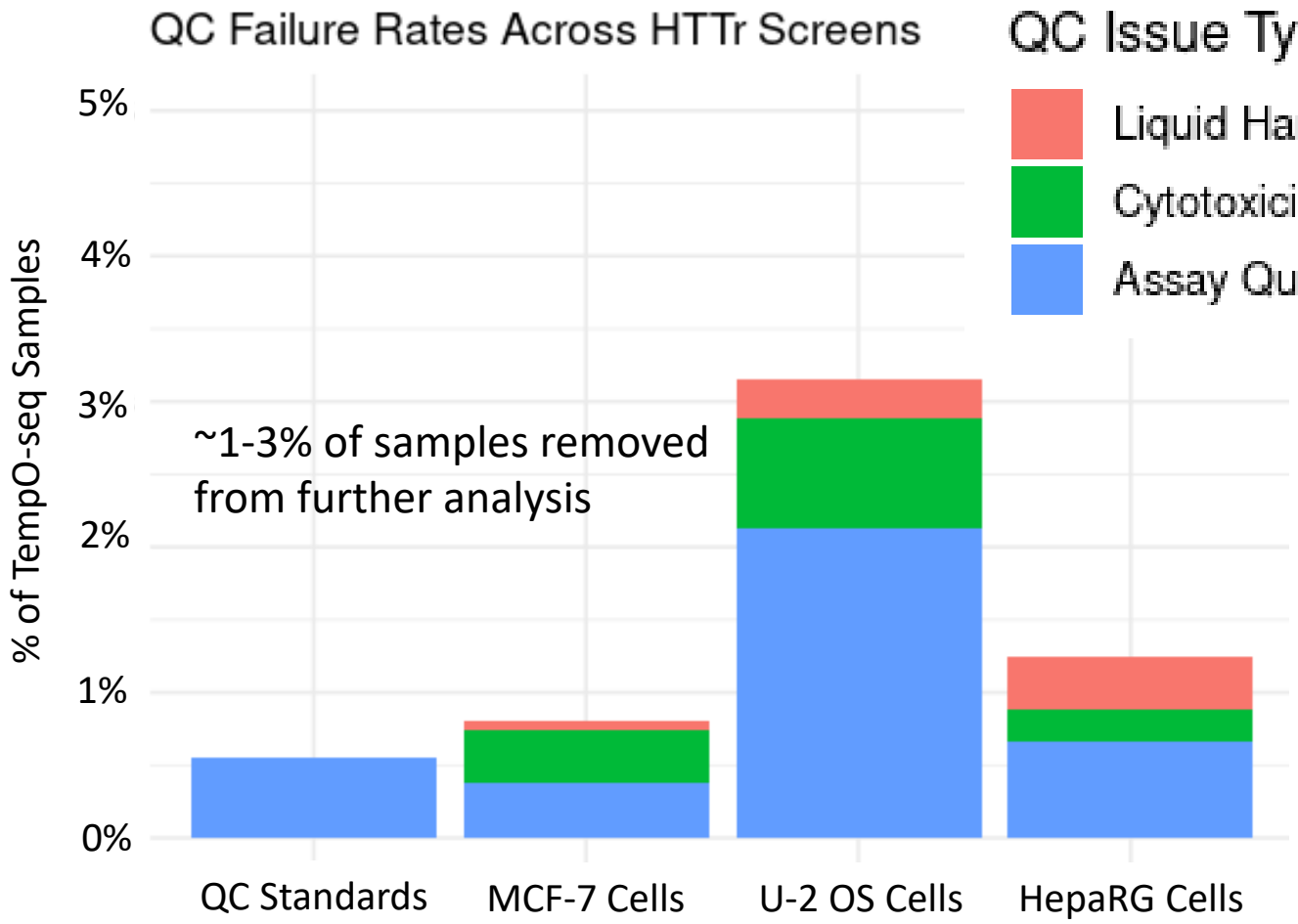


End goal: Determine which chemicals, at what concentrations, show relevant biological responses

Bioinformatics Pipeline needed to rapidly & reproducibly:

Align sequence reads to probe set sequences	Perform sample-level quality control	Manage large-scale study data in database	Estimate changes in gene expression	Identify concentration-responsive genes & pathways

HTTr Quality Control (QC)



Acoustic dispenser logs identify problems with chemical handling

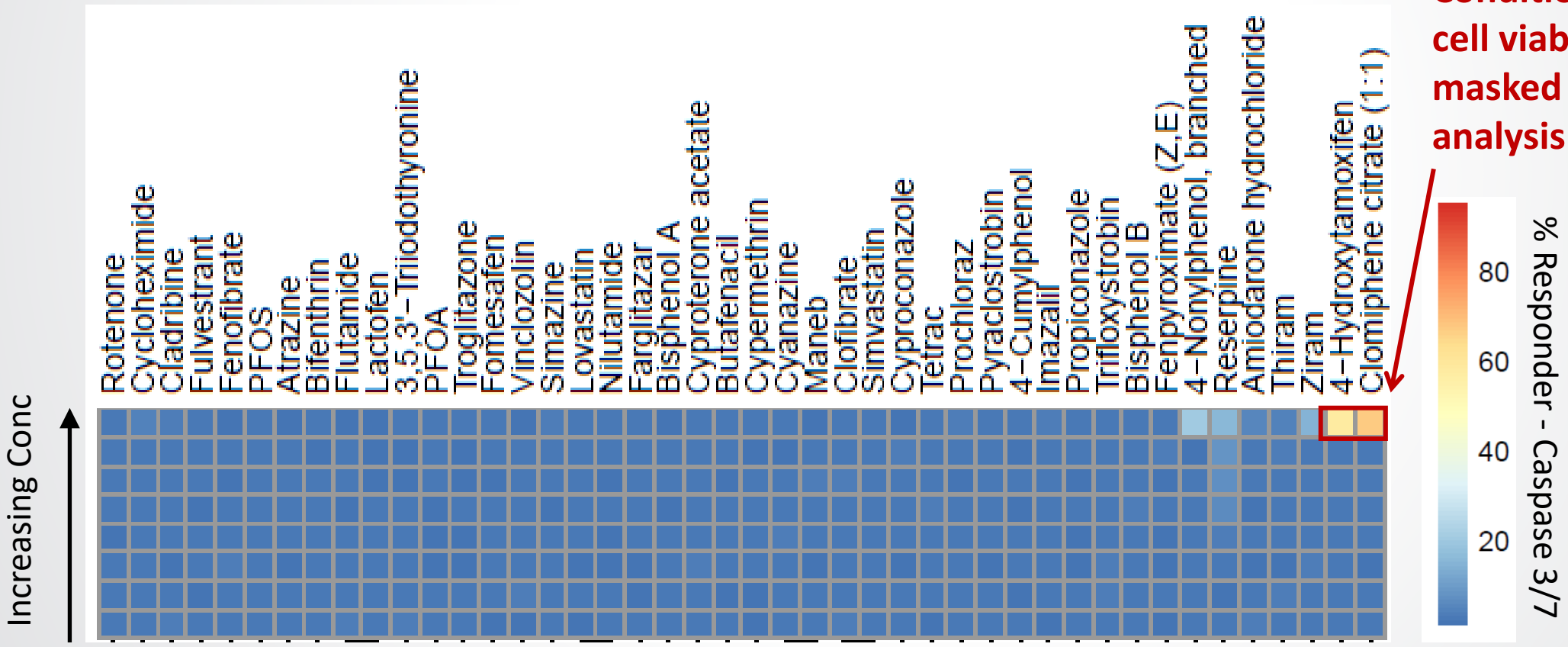
Cell imagining assays identify concentrations causing >50% cell death or cytotoxicity

Bioinformatic QC checks remove:

- Low read depth samples
- High rate of alignment failure
- Samples with low gene coverage
- Samples with irregular count distributions

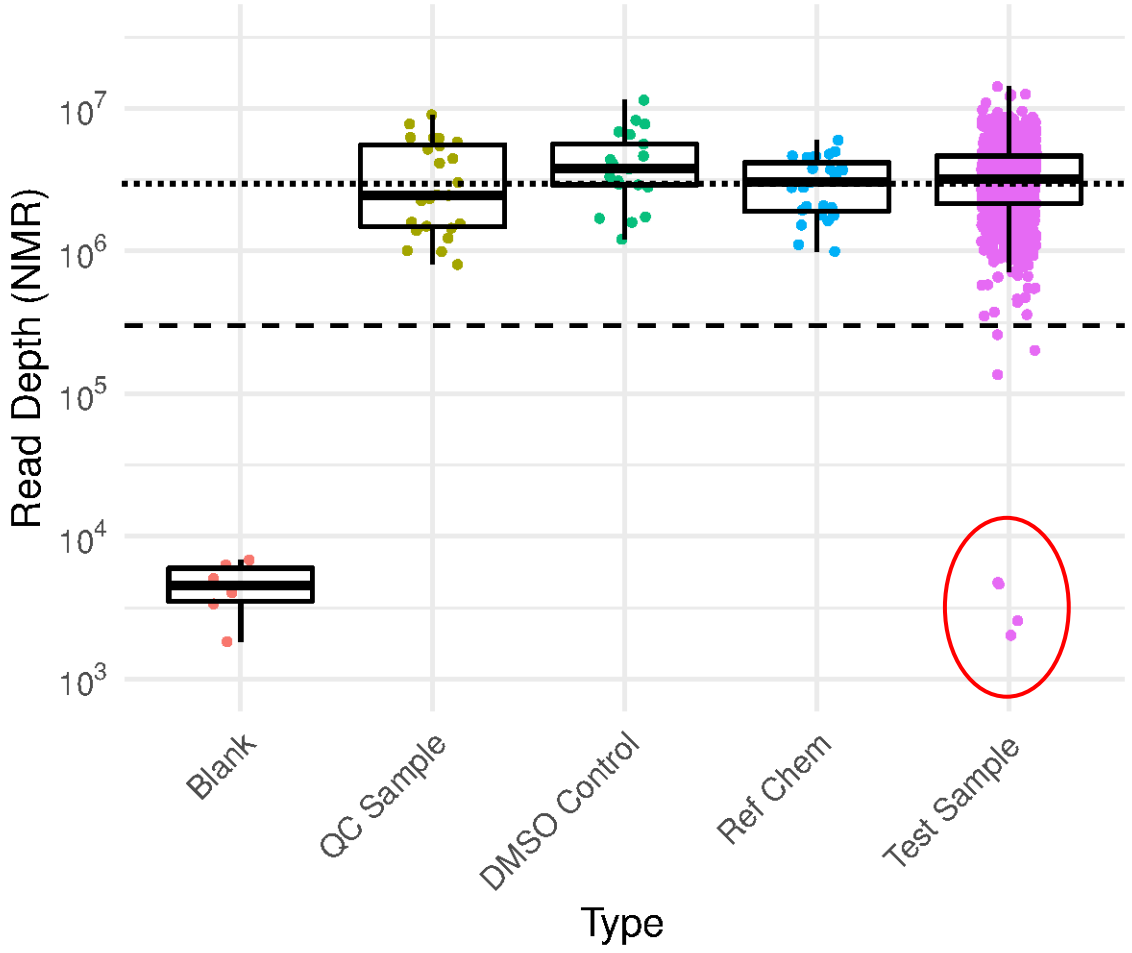
Summary of data from screening >1,000 chemicals in 3 cell lines

Cell Viability



Conditions causing cell viability loss >50% masked from further analysis.

QC Metrics: Read Depth



Depends on probe set, cell type, attenuation

Target Depth = 3M Reads

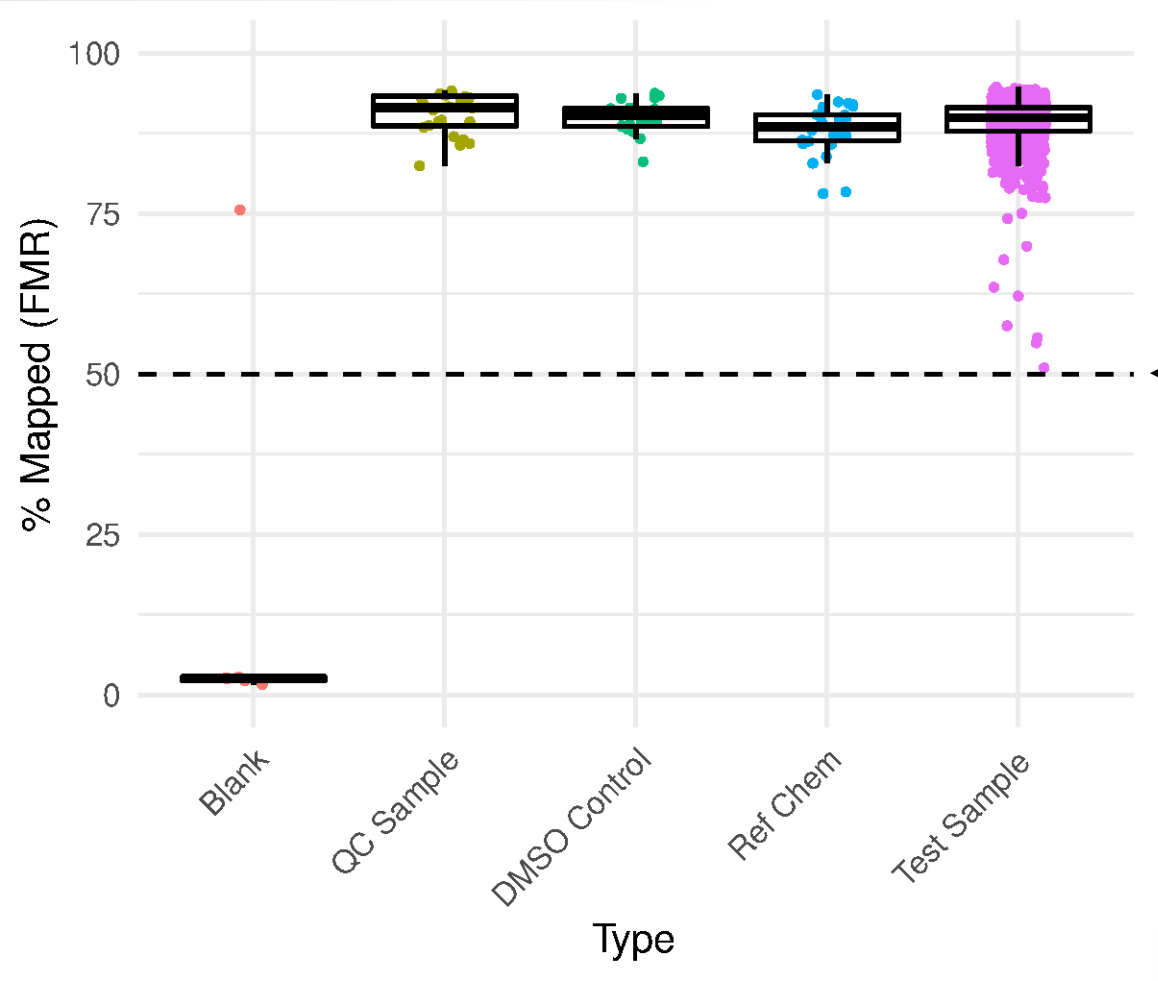
Threshold = 10% of Target Depth

More variability for cell lysates than purified RNA

Reasons for low read depth:

- **Cytotoxicity**
- **Sample degradation**
- **Low input**
- **Assay failure**

QC Metrics: Mapping Rate



- **Each read mapped to known probe sequences**
- **Only uniquely mapped reads used for analysis**

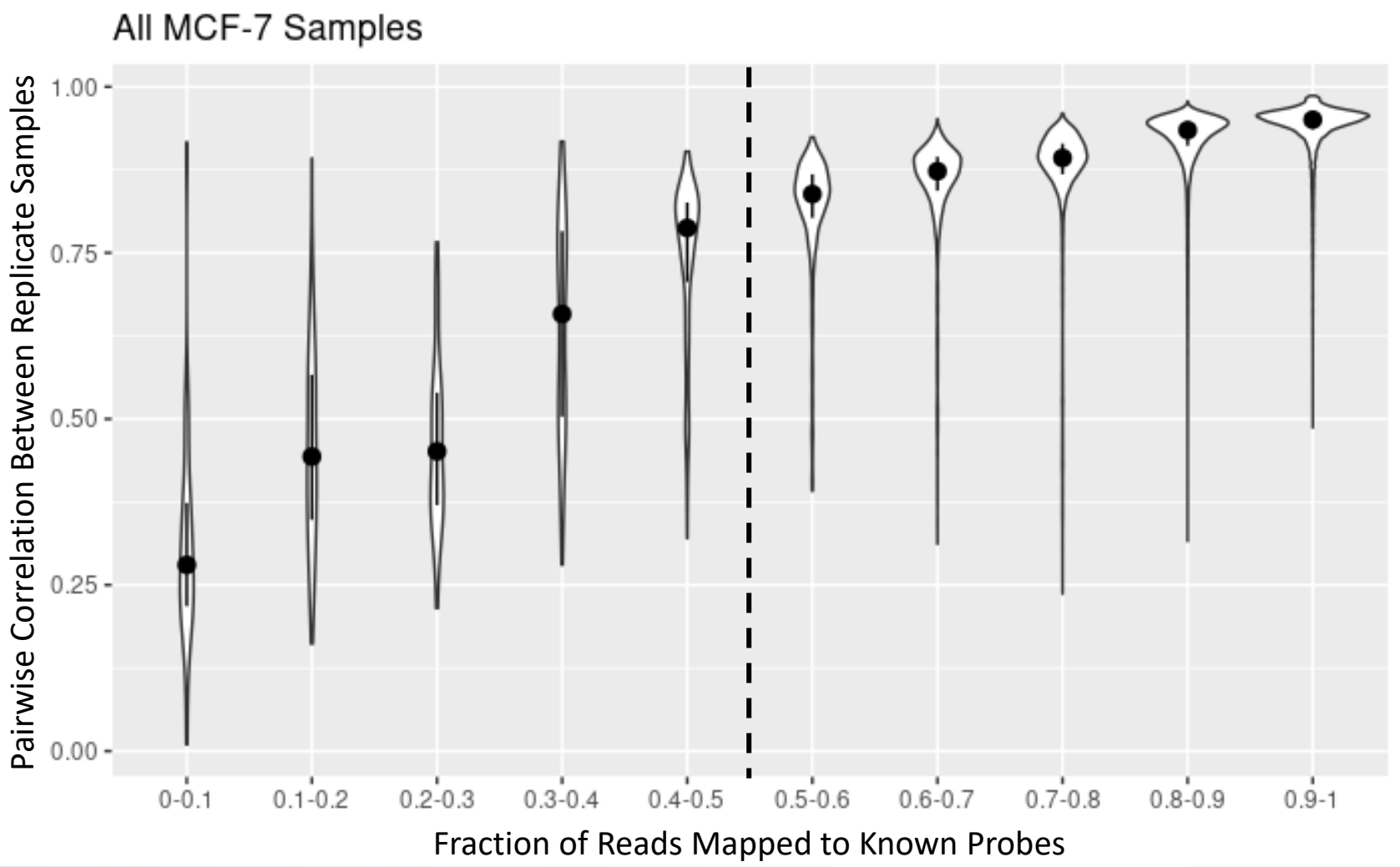
Threshold = 50% Mapping Rate

May depend on media/lysate condition, cell type

Reasons for low mapping rate:

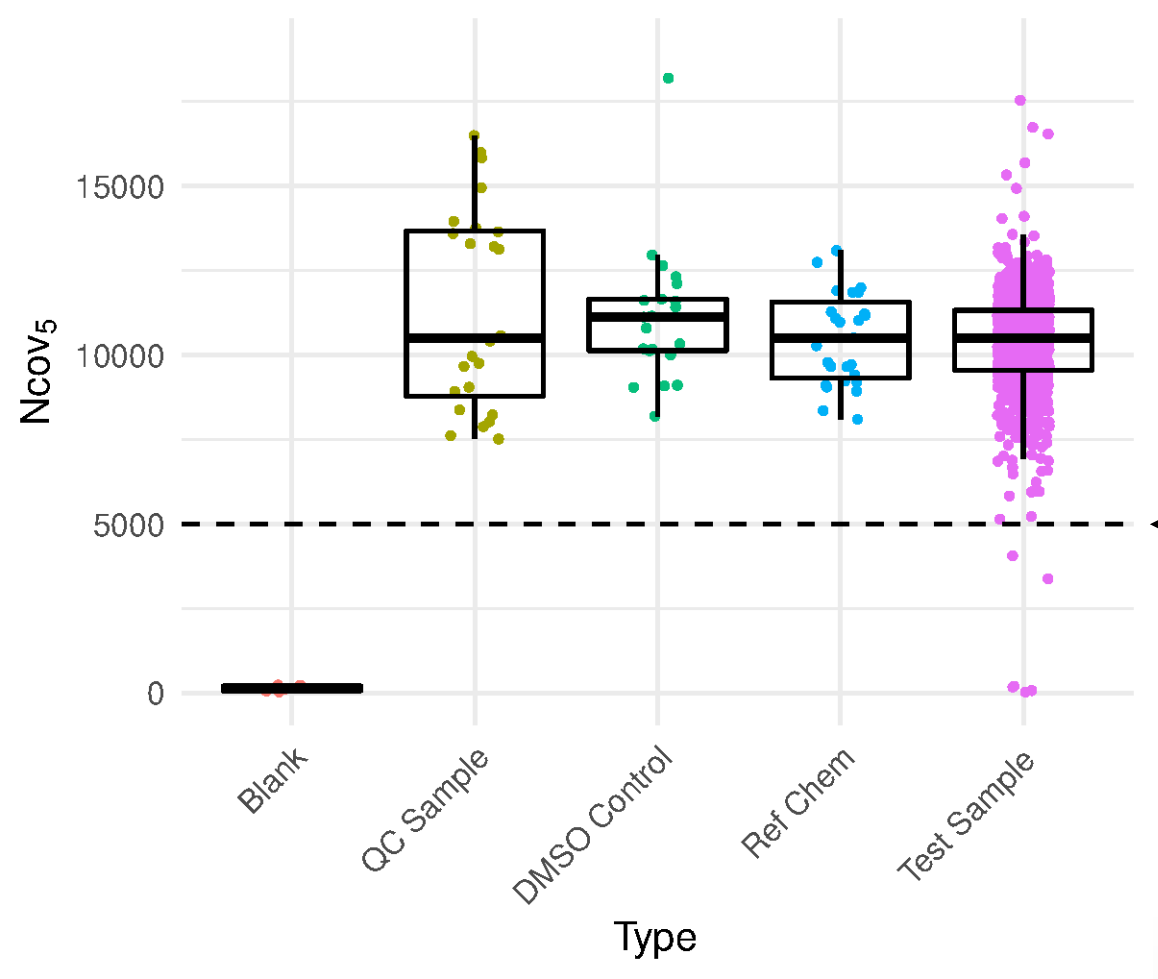
- **Cytotoxicity**
- **Sample degradation**
- **Low input**
- **Assay failure**

QC Metrics: Mapping Rate



- Replicate correlation drops off when <50% of reads mapped uniquely to probe sequences
- Lower mapping rate leads to lower depth
- May also indicate sample quality issues (e.g. RNA degradation or incomplete cell lysis)

QC Metrics: Transcriptome Coverage



$Ncov_5$ = # of probes with at least 5 reads

Threshold = 5,000 Probes (MCF-7)

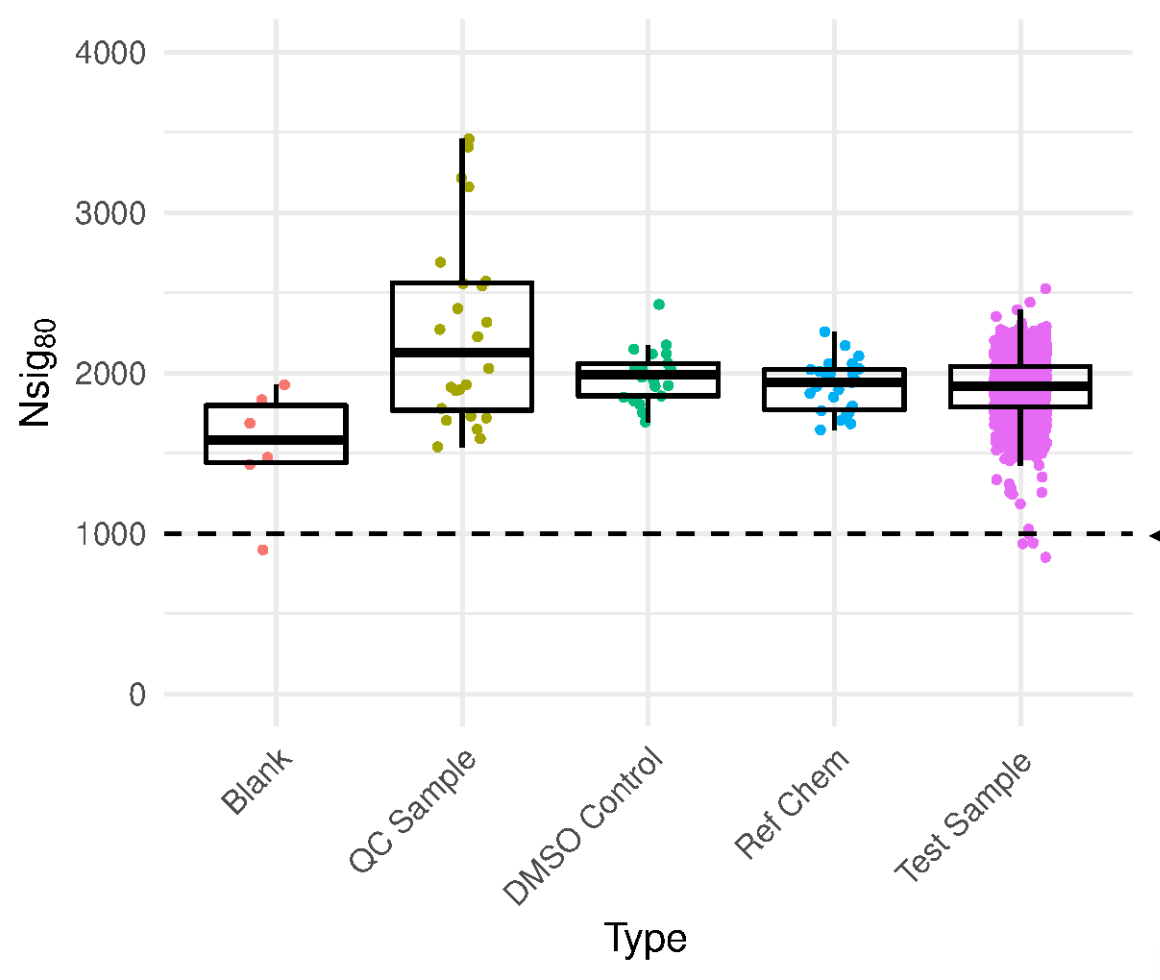
Based on "outer fence" principle (Tukey, 1976)

Re-evaluated on new cell types, probe sets, and attenuation strategies

Reasons for low coverage samples:

- **Low read depth**
- **Sample degradation**
- **Low input**
- **Assay failure**

QC Metrics: Signal Distribution



- **$Nsig_{80}$ = # of probes capturing top 80% of signal**
- **Low values = reads highly concentrated among small number of probes**

Threshold = 1,000 Probes (MCF-7)

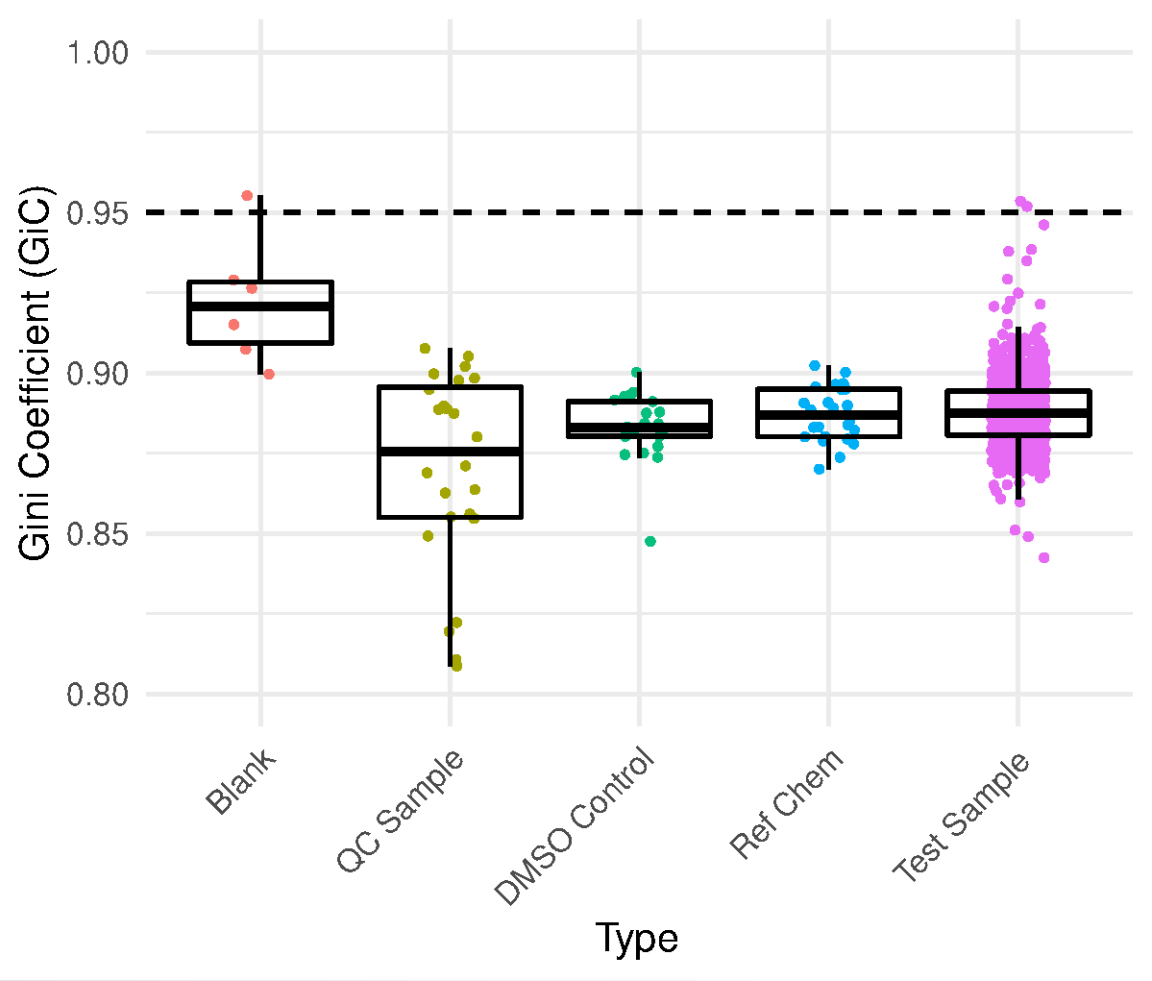
Based on "outer fence" principle (Tukey, 1976)

Should be re-evaluated on new cell types, probe sets, and attenuation strategies

Reasons for low values:

- **Sample degradation**
- **Low input**
- **Assay failure**

QC Metrics: Signal Distribution



Reasons for high values:

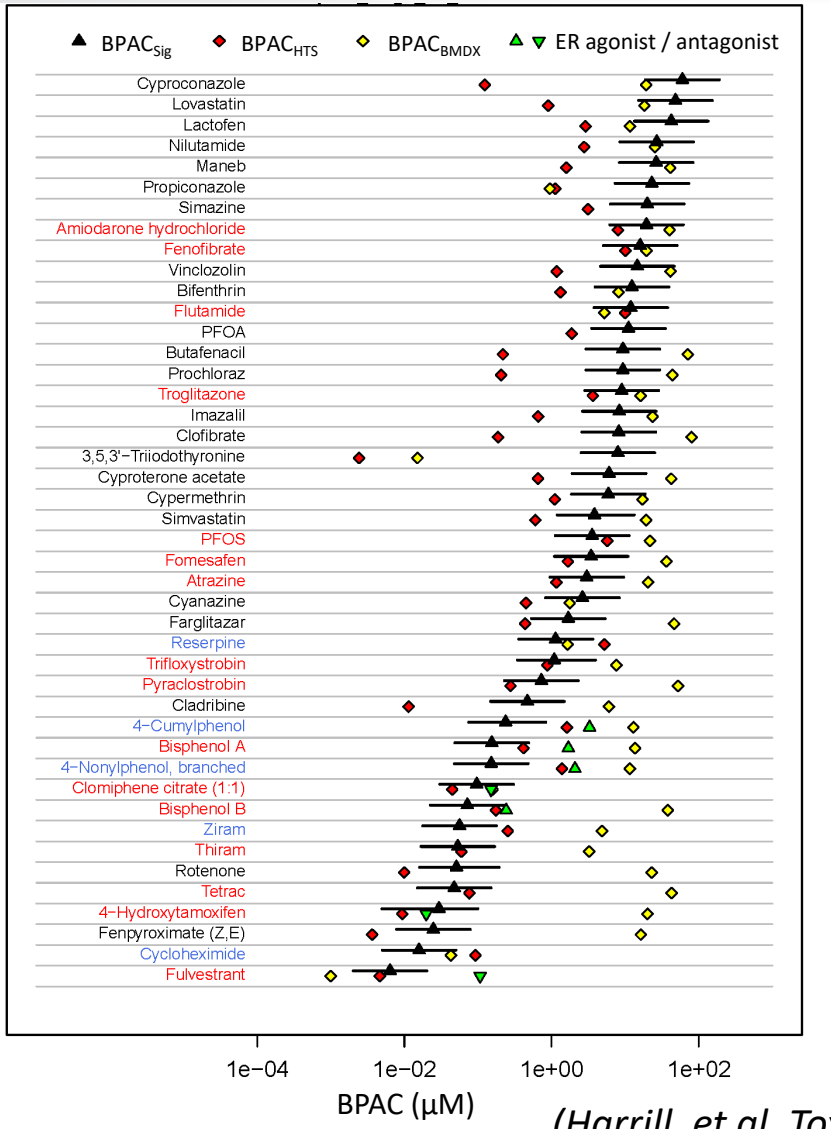
- **Sample degradation**
- **Low input**

Threshold = 0.95

Based on "outer fence" principle (Tukey, 1976)
Should be re-evaluated on new cell types, probe sets, and attenuation strategies

- **Gini coefficient = measure of inequality or skewness in a distribution**
- **High values = most reads coming from few probes (Max 1: All reads from 1 probe)**
- **Lower values = closer to uniform distribution of reads across all probes (Min 0, not expected for expression data)**
- **Expect samples from same cell type to be similar**

HTTr MCF-7 Pilot Analysis



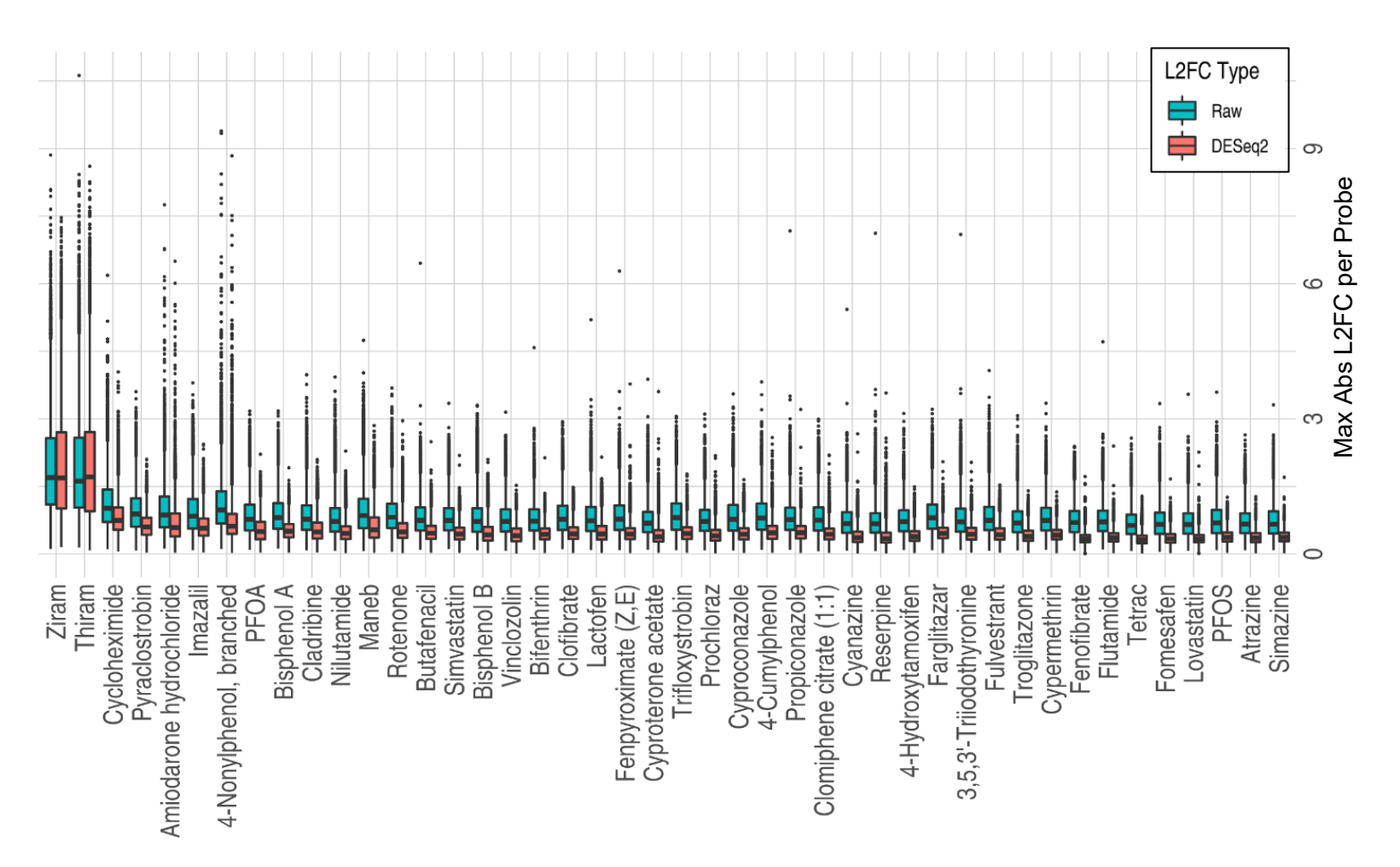
- Also calculated BPAC/PODs using NTP approach with BMDExpress2
  (NTP Research Report 5, 2018; Phillips, et al. 2019)

- BPAC$_{BMDX}$ (◆) tended to be higher and less concordant with ToxCast PODs
  - Poor signal:noise at gene-level is likely cause

- ***We continue to use BMDExpress for other transcriptomics applications and continue to explore this issue***

(Harrill, et al. Toxicol Sci, 2021)

HTTr MCF-7 Pilot Analysis



- Majority of differential expression is weak (2-4x) for most chemical treatments
  - DESeq2 dampens these further in most cases
- Consistent with previous studies using MCF-7 cells
- Lower effect size results in lower signal:noise
- Signature-level scores (e.g. GSEA) may perform better than probe-level when this is the case

(Harrill, et al. Toxicol Sci, 2021)

Stress Response Gene Signatures

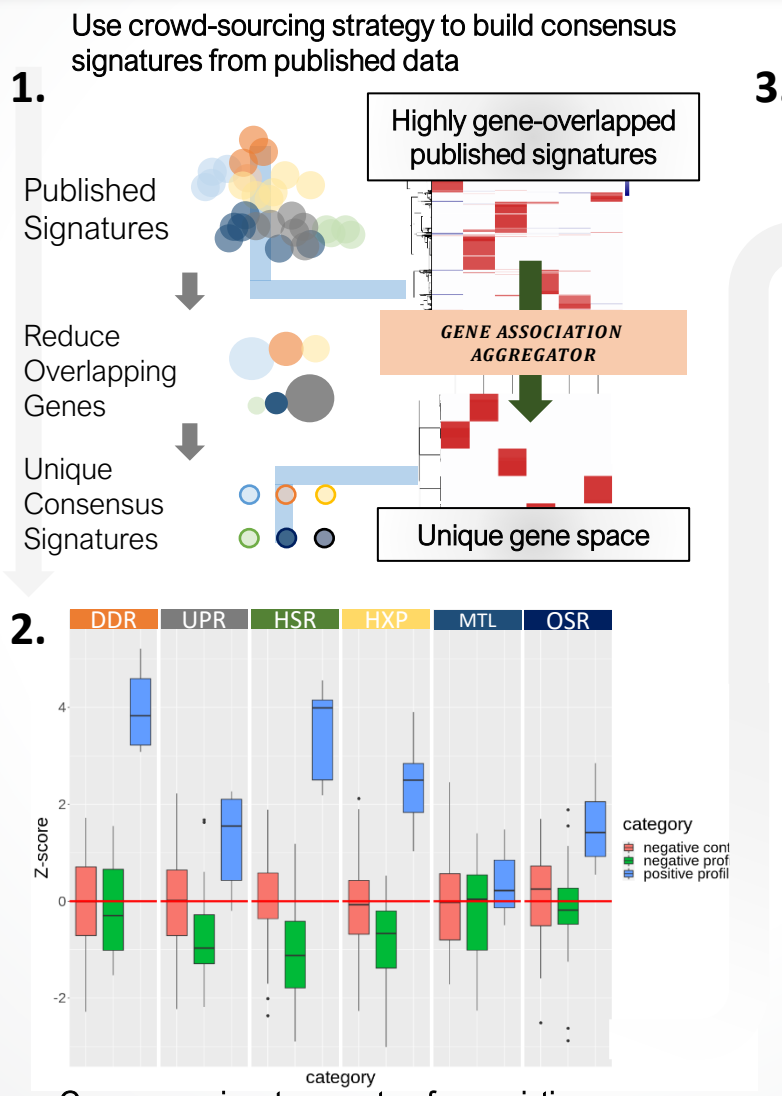
<u>Goal</u>: Develop NAMs to characterize non-specific environmental chemicals that activate stress response pathways (SRPs)

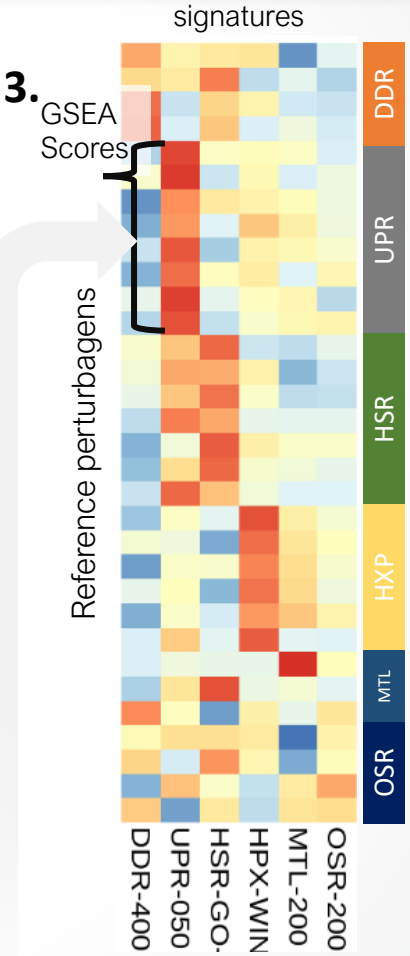
<u>Approach:</u> Characterize chemical hazards using HTTr data to assess SRP gene signature activity

<u>Challenges:</u> Cross-talk in signaling networks makes it difficult to find gene signatures of SRPs

<u>Results:</u> We have developed *consensus* SRP signatures for accurately classifying known stressors

<u>Future:</u> Use signatures to identify cellular states involved in adaptive stress responses and "tipping points" that lead to adversity





DNA Damage Response

Unfolded Protein Response

Heat Shock Response

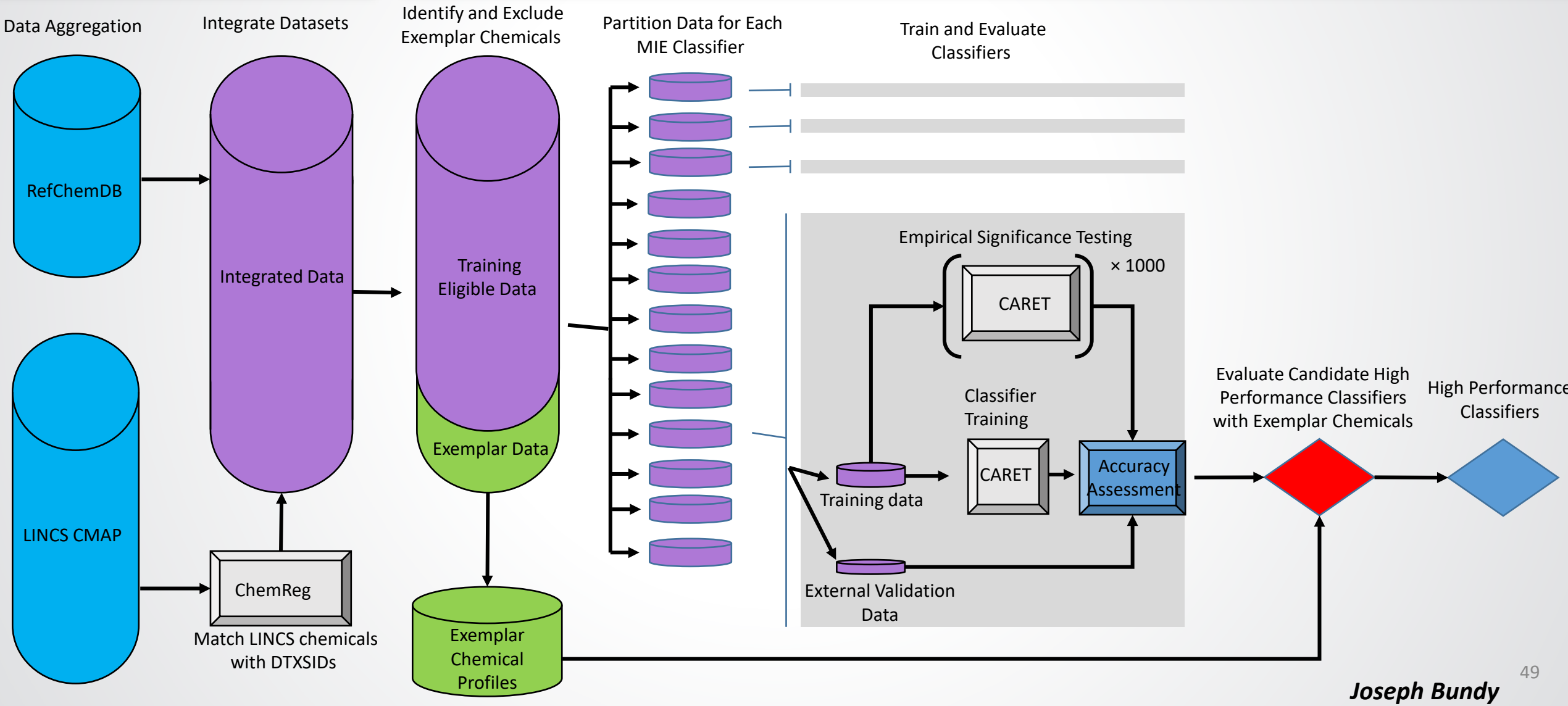
Hypoxic Response

Response to Metals

Oxidative Stress Response

Bryant Chambers & Imran Shah

ML Models for MIE Classification



Joseph Bundy

References

- Thomas RS, Bahadori T, Buckley TJ, et al. *"The Next Generation Blueprint of Computational Toxicology at the U.S. Environmental Protection Agency"*, Toxicol Sci 2019
- Yeakley JM, Shepard PJ, Goyena DE, et al. *"A trichostatin A expression signature identified by TempO-Seq targeted whole transcriptome profiling"*, PLoS ONE 2017
- Harrill J, Everett LJ, Haggard D, et al. *"High-Throughput Transcriptomics Platform for Screening Environmental Chemicals"*, Toxicol Sci 2021 *in press*
- Love MI, Huber W, and Anders S. *"Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2"*, Genome Biol 2014
- Barbie DA, Tamayo P, Boehm JS, et al. *"Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1"*, Nature 2009
- Huang R, Grishagin I, Wang Y, et al. *"The NCATS BioPlanet - An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics"*, Front Pharmacol 2019
- Subramanian A, Narayan R, Corsello SM, et al. *"A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles"*, Cell 2017
- Pinero J, Queralt-Rosinach N, Bravo A, et al. *"DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes"*, Database 2015
- Liberzon A, Birger C, Thorvaldsdottir H, et al. *"The Molecular Signatures Database (MSigDB) hallmark gene set collection"*, Cell Syst 2015
- Paul-Friedman K, Gagne M, Loo LH, et al. *"Utility of In Vitro Bioactivity as a Lower Bound Estimate of In Vivo Adverse Effect Levels and in Risk-Based Prioritization"*, Toxicol Sci 2020