Understanding Tox21/ToxCast High-Throughput Screening Data and Application to Modeling: Concentration-response Modeling in High-throughput Transcriptomics

Richard Judson, PhD

US EPA

RTP, NC

Phone: 919-449-7514

Email: judson.richard@epa.gov

Conflict of Interest Statement

The author declares no conflict of interest.

Disclaimer: The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA

Abbreviations

- HTTr – High Throughput Transcriptomics
- BMD – Benchmark Dose
- Tcpl – ToxCast Pipeline

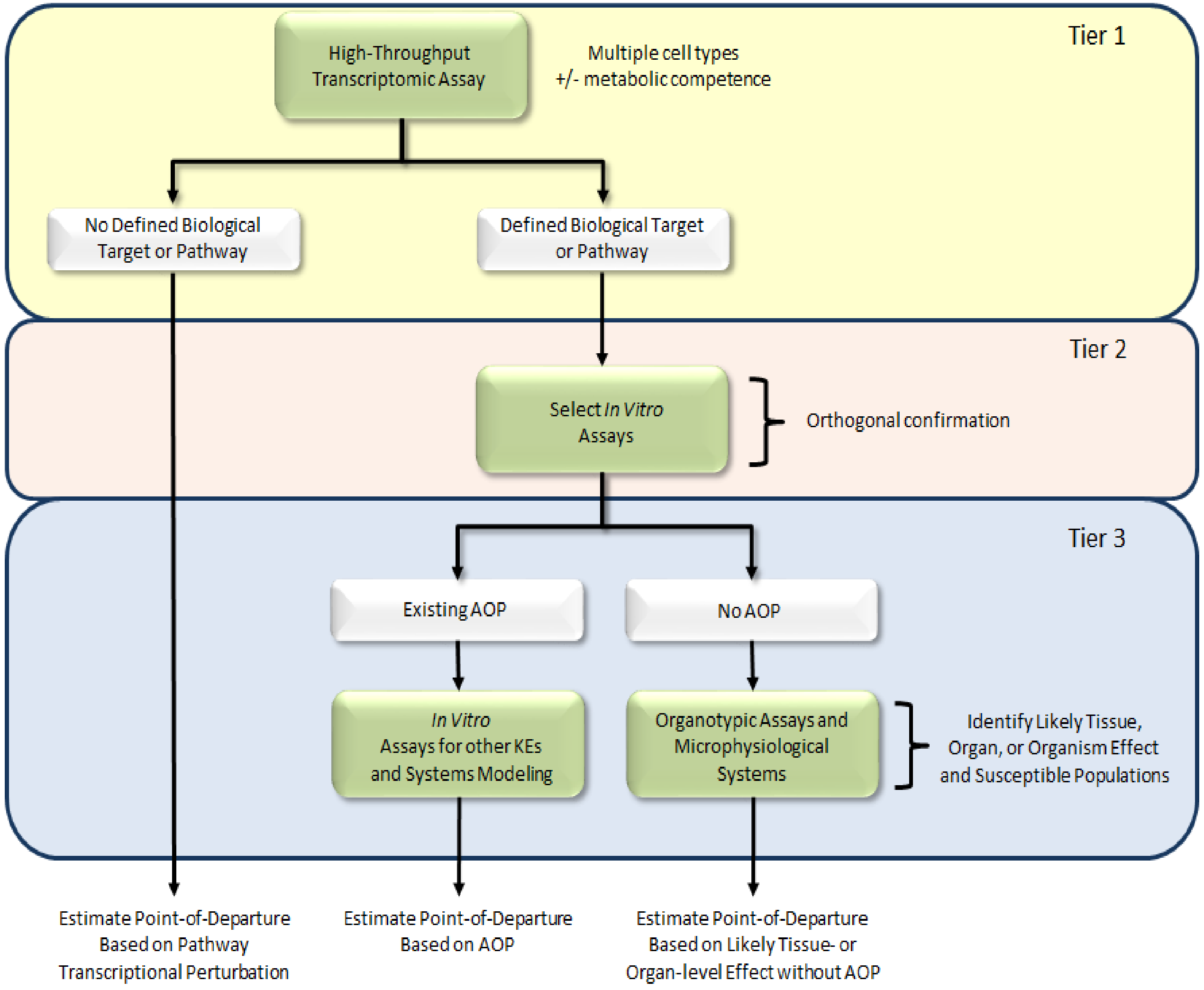
Objectives

- A flexible, portable and cost-efficient platform to comprehensively evaluate the potential biological pathways and processes impacted by chemical exposure

  → High-throughput transcriptomics (HTTr)

- Identify the concentration at which biological pathways / processes begin to be impacted
- Assign putative biological targets for chemicals

A strategic vision and operational road map for computational toxicology at the U.S. Environmental Protection Agency



Tier 1

High-Throughput Transcriptomic Assay

Multiple cell types
+/- metabolic competence

No Defined Biological Target or Pathway

Defined Biological Target or Pathway

Tier 2

Select *In Vitro* Assays

Orthogonal confirmation

Tier 3

Existing AOP

No AOP

In Vitro Assays for other KEs and Systems Modeling

Organotypic Assays and Microphysiological Systems

Identify Likely Tissue, Organ, or Organism Effect and Susceptible Populations

Estimate Point-of-Departure Based on Pathway Transcriptional Perturbation

Estimate Point-of-Departure Based on AOP

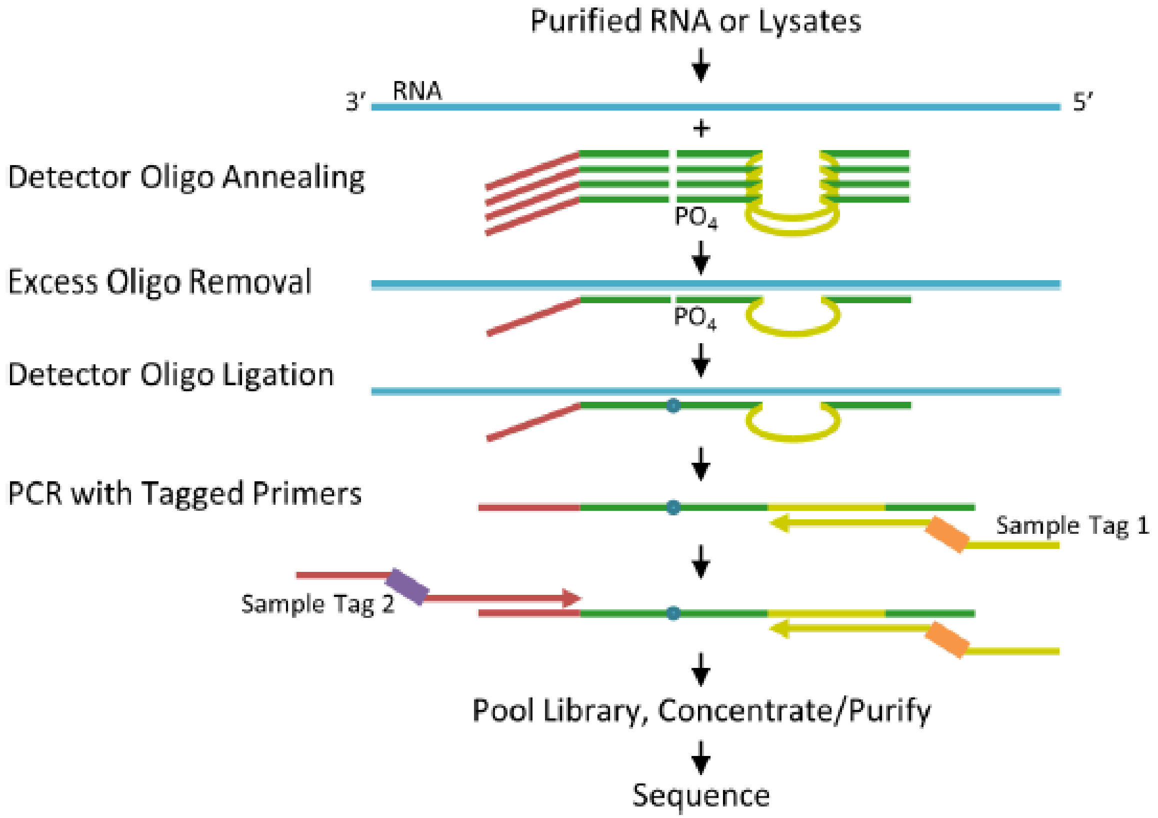
Estimate Point-of-Departure Based on Likely Tissue- or Organ-level Effect without AOP

R. Thomas

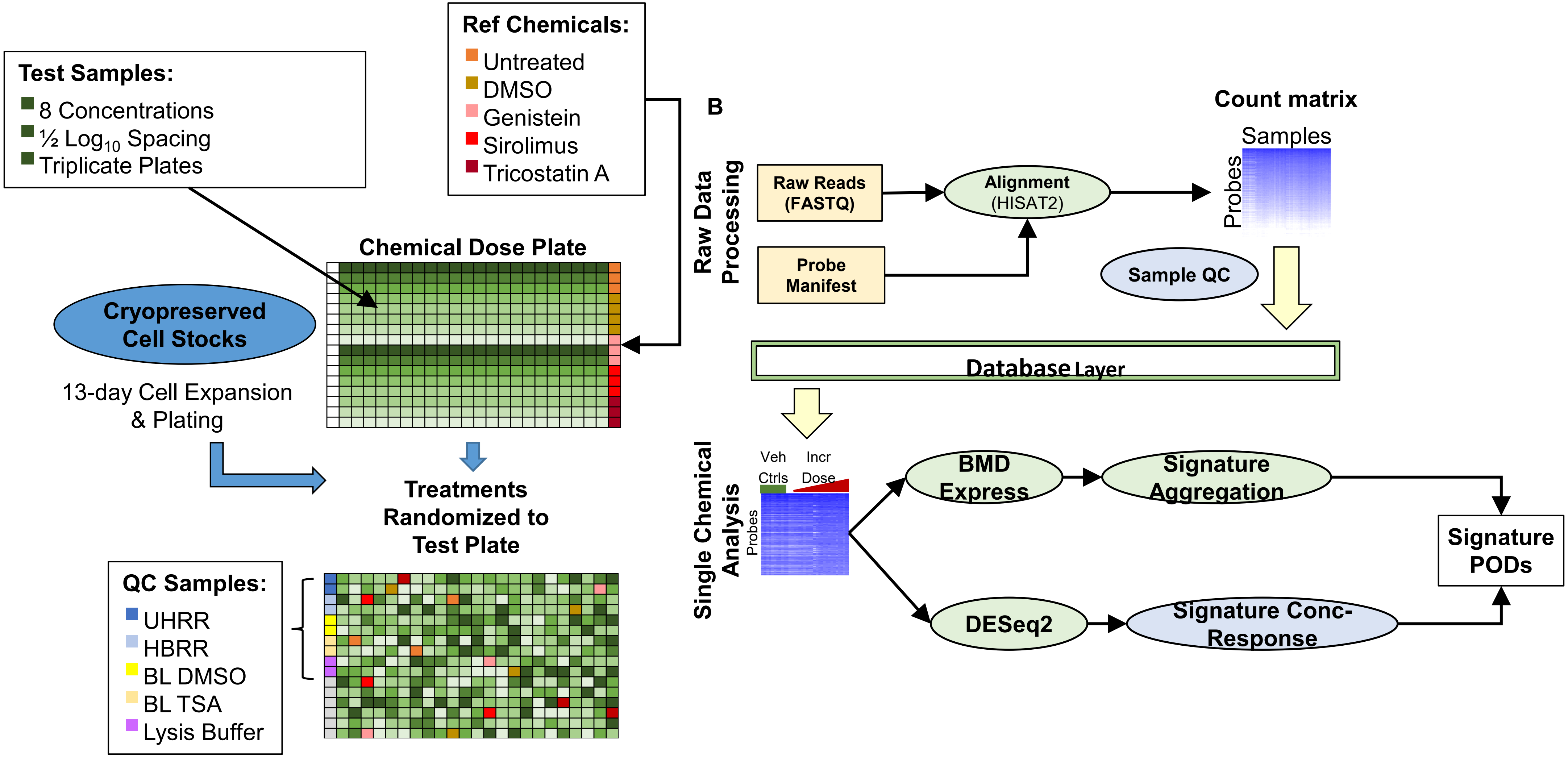
TempO-Seq for HTTr

- The **TempO-Seq** human whole transcriptome assay measures the expression of ~21,100 transcripts.
- Requires only picogram amounts of total RNA per sample.
- Compatible with purified RNA samples or **cell lysates**.
- Transcripts in cell lysates generated in 384-well format barcoded to well position
- Scalable, targeted assay:
  - Measures transcripts of interest
  - Greater throughput and requires lower read depth than RNA-Seq
  - Ability to attenuate highly expressed genes

TempO-Seq Assay Illustration



HTTr Overall Process



Test Samples:
- 🟩 8 Concentrations
- 🟩 ½ Log_{10} Spacing
- 🟩 Triplicate Plates

Ref Chemicals:
- 🟧 Untreated
- 🟨 DMSO
- 🟥 Genistein
- 🟥 Sirolimus
- 🟥 Tricostatin A

Chemical Dose Plate

Cryopreserved Cell Stocks

13-day Cell Expansion & Plating

Treatments Randomized to Test Plate

QC Samples:
- 🟦 UHRR
- 🟦 HBRR
- 🟨 BL DMSO
- 🟨 BL TSA
- 🟪 Lysis Buffer

B

Raw Data Processing

Raw Reads (FASTQ)

Alignment (HISAT2)

Probe Manifest

Count matrix

Sample QC

Database Layer

Single Chemical Analysis

BMD Express

Signature Aggregation

DESeq2

Signature Conc-Response

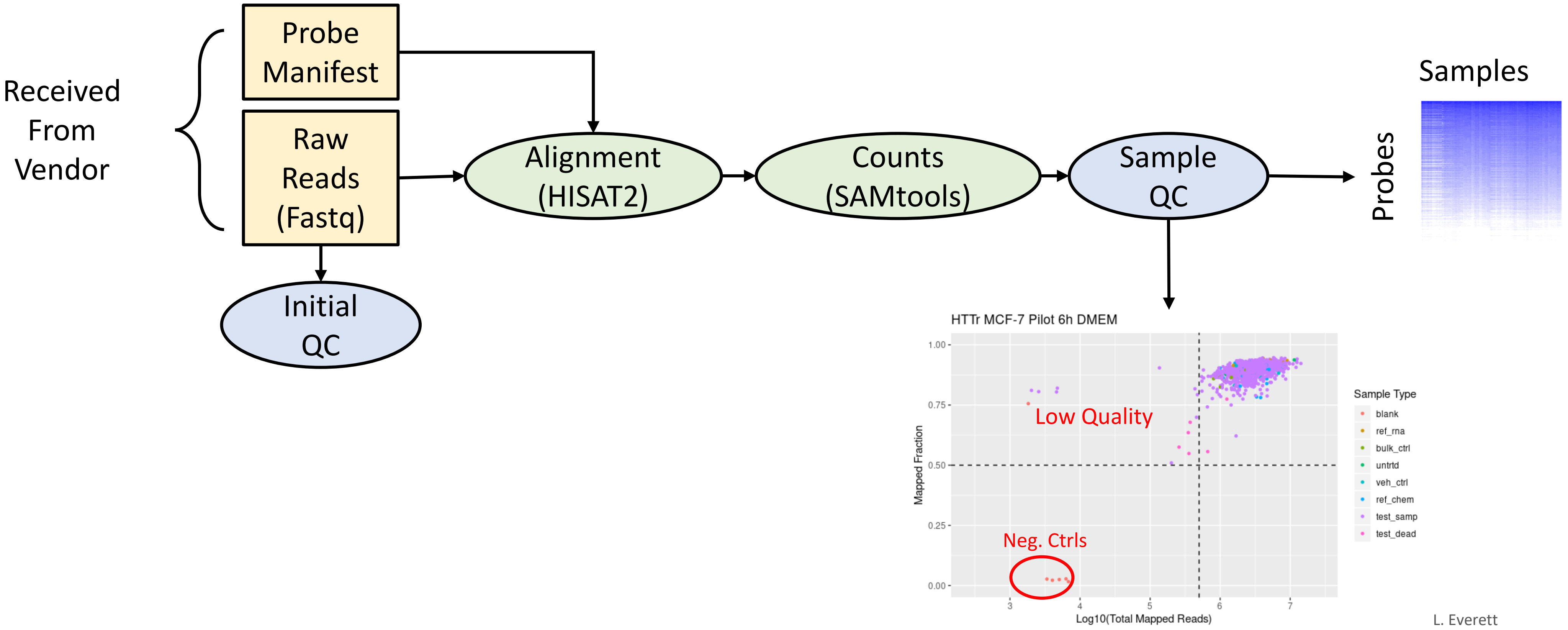
Signature PODs

Raw Processing Options

- Alignment Pipeline – using HISAT2, comparable to STAR
  - Now trims 51bp reads prior to alignment
  - Allowed soft-clipping with per base penalty

- Probe Homology can be an issue
  - Mapped homology within probe manifest (some probes have 49bp overlap)
  - >95% of reads map uniquely to one probe with current parameters
  - HISAT2 was better at resolving unique matches for homologous probes
  - Multi-mapping probes discarded for final counts

Pipeline: Raw Data Processing



Received From Vendor

Probe Manifest

Raw Reads (Fastq)

Initial QC

Alignment (HISAT2)

Counts (SAMtools)

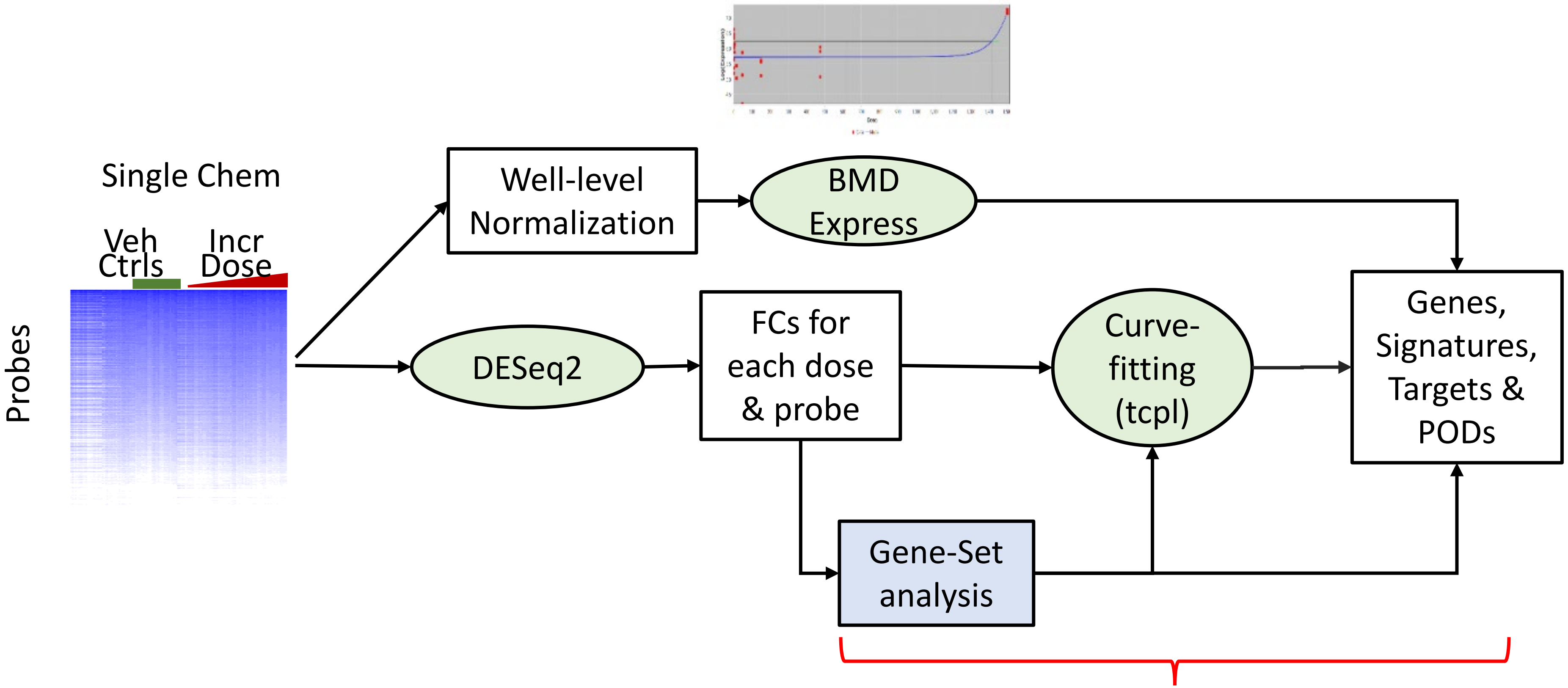
Sample QC

Samples

Probes

L. Everett

Pipeline: Targets & Concentration Response



Single Chem

Veh Ctrls

Incr Dose

Probes

Well-level Normalization

BMD Express

DESeq2

FCs for each dose & probe

Curve-fitting (tcpl)

Genes, Signatures, Targets & PODs

Gene-Set analysis

Focus of this presentation

Differential Gene Expression Analysis

- Most recent version of DESeq2 (v1.24.0)
  - Evaluated questions about choice of plate effect and shrinkage using reference chemicals
  - Newer shrinkage methods (Ashr, Apeglm) results less reliable
- Analyze one chemical at a time with matched DMSO controls
- DEG analysis by four DESeq2 options:-
  1. Plate effect - , Shrinkage -
  2. Plate effect - , Shrinkage +
  3. Plate effect + , Shrinkage -
  4. Plate effect + , Shrinkage + (Recommended)

HTTr Datasets

Dataset	MCF7 Pilot	MCF7 Screen	HepaRG Screen	U2OS Screen
Tissue	Breast	Breast	Liver	Bone
Chemicals	44	1593 [3]	1323	1324
Samples [1]	350	12959	10825	10766
Genes [2]	10149	9137	12116	11815

Notes:

[1] Includes 8 concentrations / chemical and replicates, but not reference chemicals

[2] There may be more than one probe per gene. At least 95% of samples must have at least 5 counts for probe to be included

[3] After samples from bad plate groups were removed

Signature Scoring

- Start with matrix of samples x genes with l2fc from DESeq2
- For each concentration of each sample, calculate score for each signature using
  - GSEA (ssGSEA)
  - FC (mean(l2fc|in signature) – mean(l2fc|out of signature))
- Distribution of signature scores are zero centered
- For bidirectional signatures collapse score to that of parent
  - Score(chemical, concentration, parent)=score(up) – score(down)
  - Retains directionality
- For unidirectional signatures, parent score=signature score

Gene Sets: “Signatures”

- Pathways from MSigDB, BioPlanet, DisGeNET
- CMAP:
  - For each chemical treatment, select top 100 genes most up regulated and 100 genes most down regulated
  - Create paired up and down signatures
- Random gene sets
  - Select gene sets with random sets of genes with frequency and gene-gene co-occurrence frequencies matching the rest of the gene signatures
  - Select 1000 of these
- Each signature has a hand-annotated “super target” class to help with annotation
- 11,006 signatures
- ~2000 super targets
  - Hand annotated, subject to revision

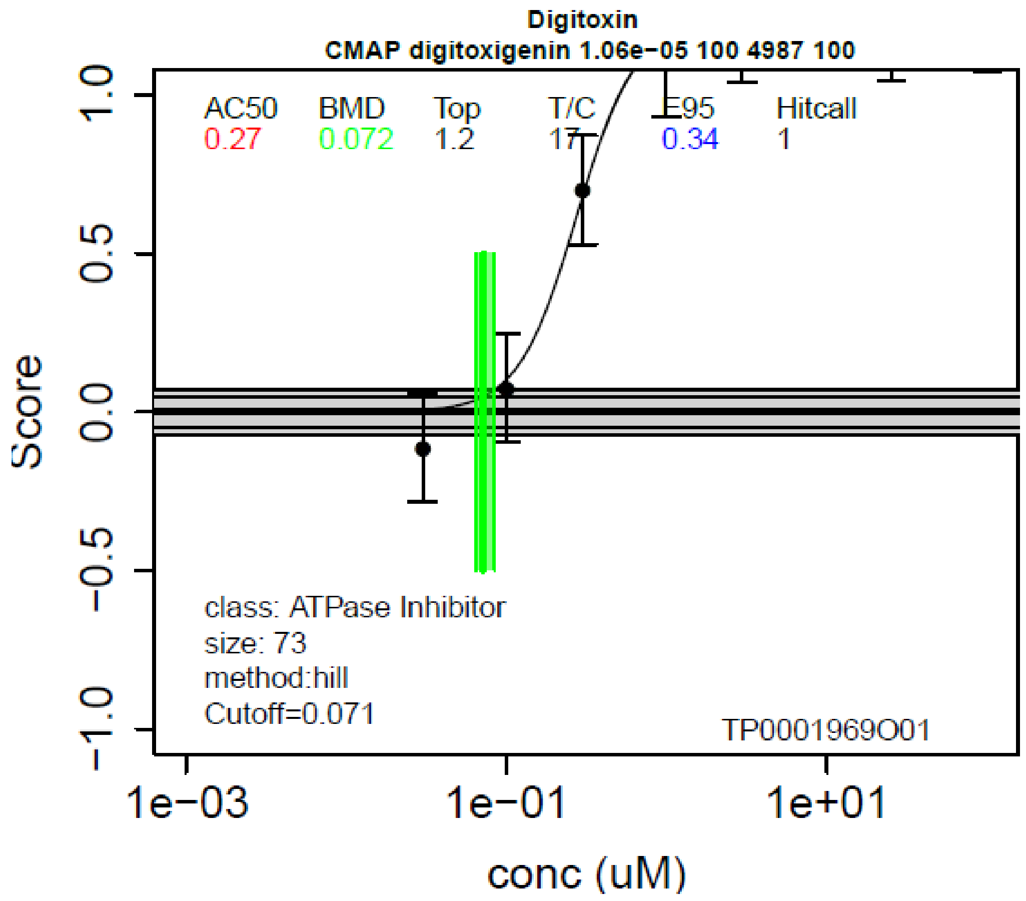
Concentration-response modeling

- Use variant of ToxCast tcpl concentration-response fitting method
- Expanded to include all models used in BMDExpress
  - cnst, hill, gnls, poly1, poly2, pow, exp2, exp3, exp4, exp5
  - Fitting in both up and down directions
  - Model with the lowest AIC is selected
- Produces a continuous hit call value
- Implemented in R package tcplFit2 – public soon
- Create null distribution of 1000 randomly select “chemicals” created by permuting columns of sample x gene matrix
  - “Cutoff” = 95% CI of this distribution of scores
  - Helps determine if a real signature is active

Concentration-Response Output

- For each chemical sample / signature
  - Hitcall – in range of 0 to 1
    - recommended cutoff = 0.9 for actives
  - BMD – potency estimate in uM
  - Top – maximum efficacy
  - Top / Cutoff maximum efficacy relative to the null distribution 95% CI
    - recommended cutoff = 1.5 for actives
  - Winning fit model, e.g. hill or poly2

Example Concentration-response plot



CI around points from the fitting error term

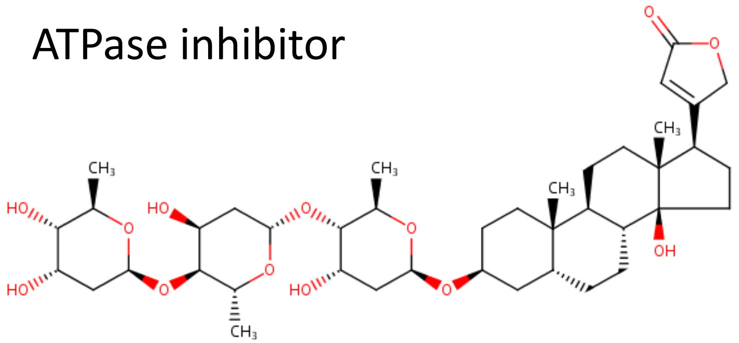
Outer gray band is 95% CI of null dist.
Inner lines are benchmark response

Green vertical band is BMD and 95% CI

Digitoxin: Cardiac Glycoside

For treatment of heart failure and as cancer chemotherapeutic

ATPase inhibitor

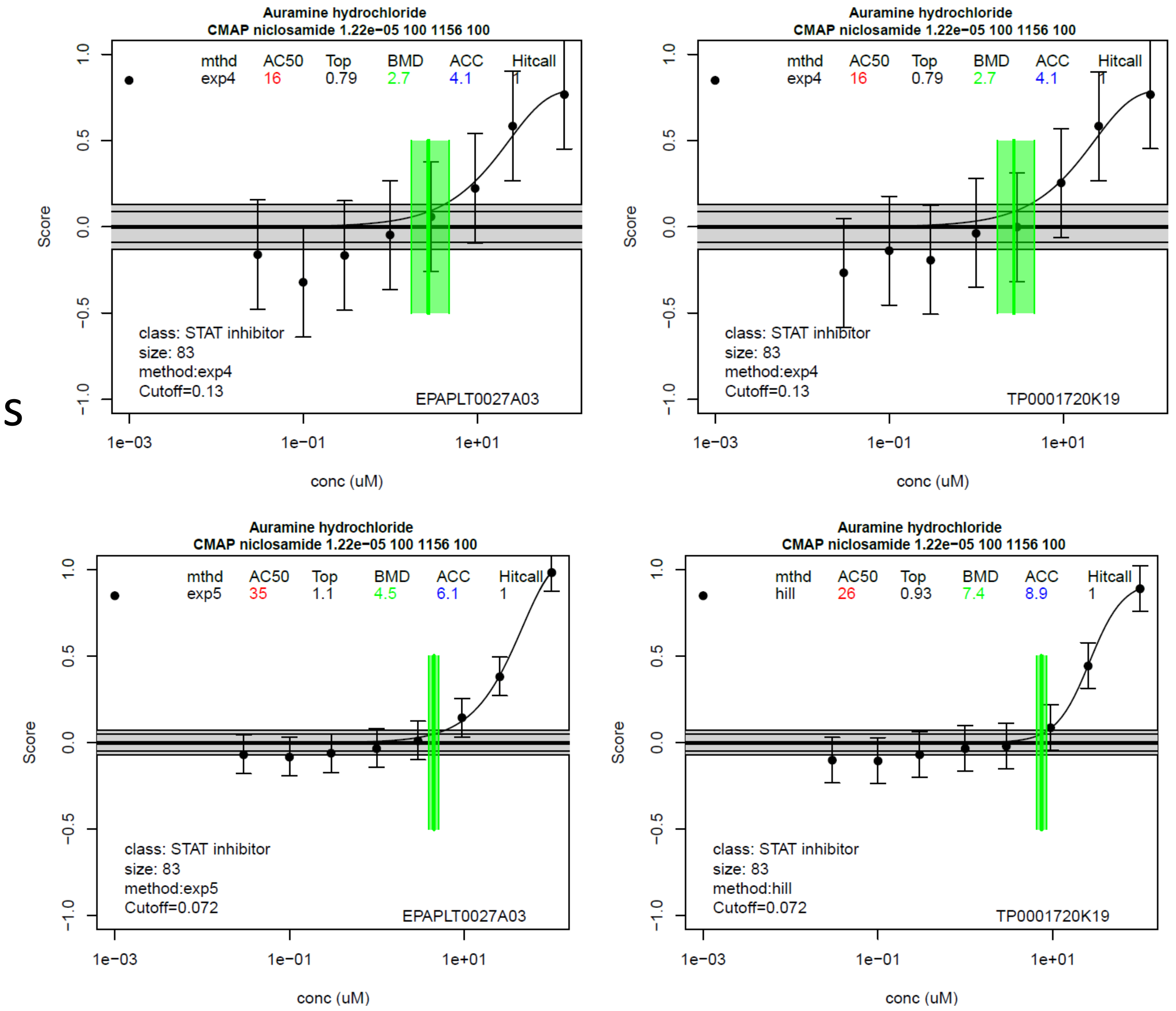


Concentration-response: GSEA vs FC

2 samples

2 scoring methods

Same signature



GSEA – lower S/N

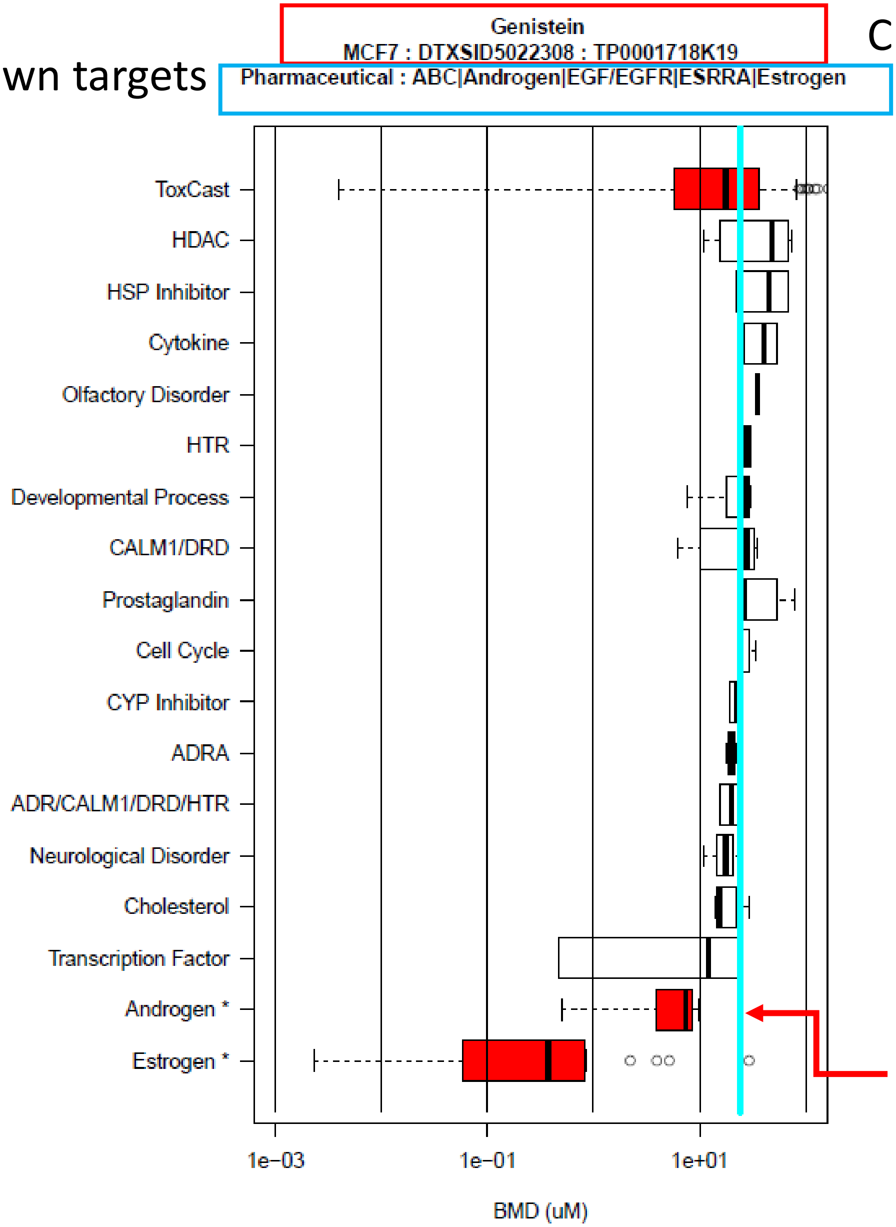
FC

Super Target Summary Plot

Chemical name, DTXSID and sample ID

Use class and known targets

Super targets



Boxplot shows range of BMD values for signatures for the super targets

Red indicates that the super target is a target of the chemical

Median of all super target BMD medians

Running the Code

- R package “httrpathway”
- Input
  - L2fc data with
    - Chemical ID
    - Sample ID
    - Probe ID
    - Concentration
  - Signatures
    - Library of gene sets
    - Catalog with names, super targets and other annotations
- Standard directory structure

Preparing the L2FC Input

```
buildFCMAT1.fromDB <- function(dataset="mcf7_ph1_pe1_normal_block_123",
                               dir="../input/fcdata/new_versions/",
                               infile,
                               pg.filter.file,
                               do.load=T)
```

Would need to be customized if not using EPA processing method

All further files will be labeled with the dataset name

The package assumes the existence of a standard set of directories

```
buildFCMAT2.fromDB <- function(dataset="mcf7_ph1_pe1_normal_block_123",
                               time=6,
                               media="DMEM",
                               dir="../input/fcdata/",
                               method="gene",
                               do.read=T)
```

Produces FCMAT2 and CHEM_DICT files

Running Signature Concentration Response

```
driver <- function(dataset=" mcf7_ph1_pe1_normal_block_123",
                   sigcatalog="signatureDB_master_catalog 2020-10-22",
                   sigset="screen_large",
                   nullset=NULL,
                   nrandom.chems=1000,
                   normfactor=7500,
                   mc.cores=20,
                   bmr_scale=1.349,
                   method="fc",
                   do.build.random=F,
                   do.run.random=F,
                   do.run.all=F,
                   do.scr.plots=F,
                   do.st.plots=F)
```

Signature information

Runs in parallel under Linux

Flags to run each step

Some Current Challenges

- Underlying data has interesting noise properties which we are still exploring
- Many concentration-response profiles have magnitude just outside of the null-distribution band
  - Are these real hits?
- Need to deal with multiple comparison issues
  - Can we determine the likely target of an unknown chemical?
- What is the best way to estimate the chemical-level POD?

Conclusions

- It is now possible to perform concentration-response profiling using high-throughput transcriptomics for thousands of chemicals
- Points of departure are
  - Reproducible
  - Seem to provide accurate relative scaling between chemicals
  - Match results from other technologies
- Chemicals often activate signatures with the correct target before most other classes of targets
- Statistical and data interpretation challenges remain

Acknowledgements

- Josh Harrill
- Logan Everett
- Imran Shah
- Rusty Thomas
- Richard Judson
- Derik Haggard
- Joseph Bundy
- Beena Vallanat
- Bryant Chambers
- Woody Setzer
- Thomas Sheffield
- Laura Taylor
- Clinton Willis
- Richard Brockway
- Johanna Nyffeler
- Megan Culbreth
- Dan Hallinger
- Terri Fairley
- Matt Martin
- Agnes Karmaus

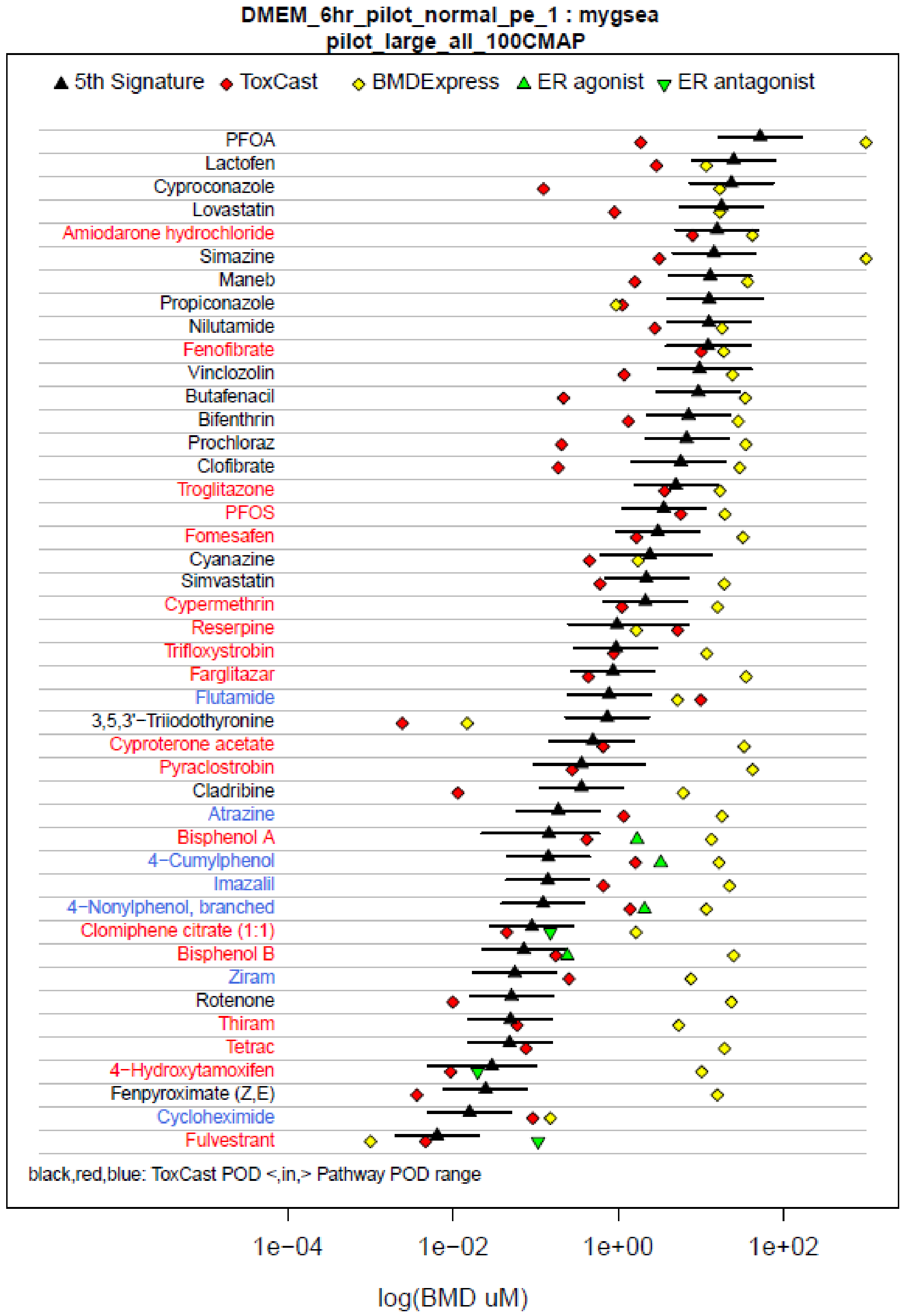
References

- Harrill, JA et al. “High-Throughput Transcriptomics Platform for Screening Environmental Chemicals”, Tox Sci, in press (2021)

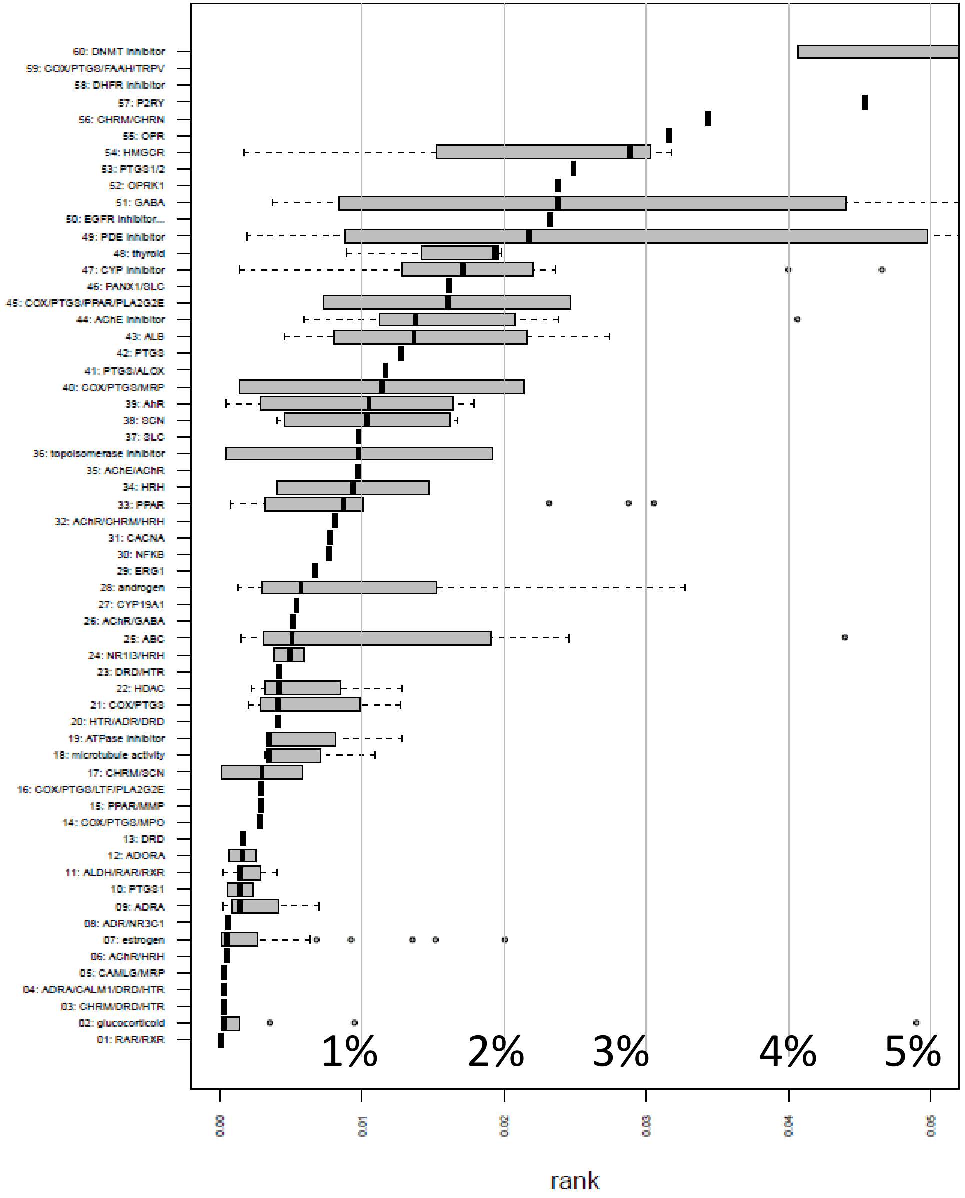
Extra Slides

Chemical-wise PODs

Black: lowest 5%-ile signature
Red: ToxCast 5% POD
Yellow: BMD Express
Green: ToxCast ER Model



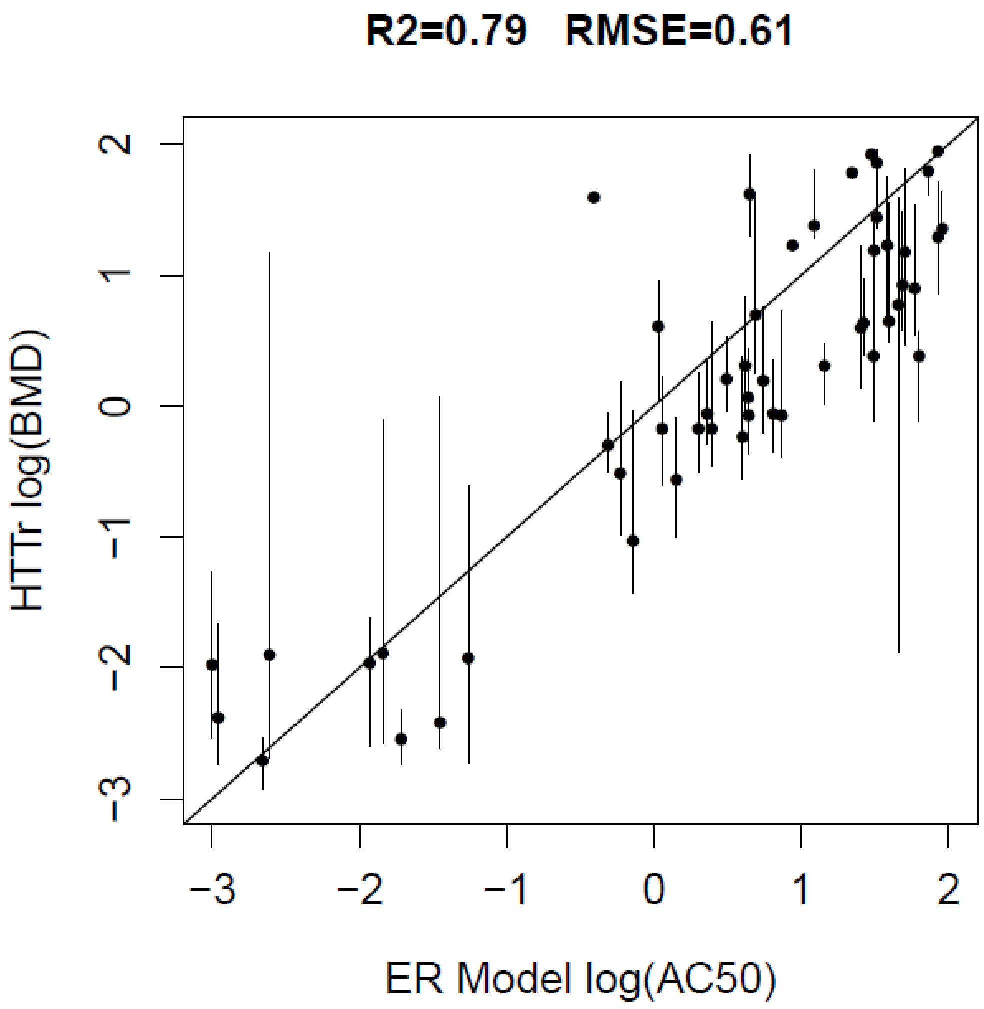
Measuring how well the signatures ID the chemical target



Fraction of signatures more active than the first on-target signature

Lowest set are all GPCR or nuclear receptor target families

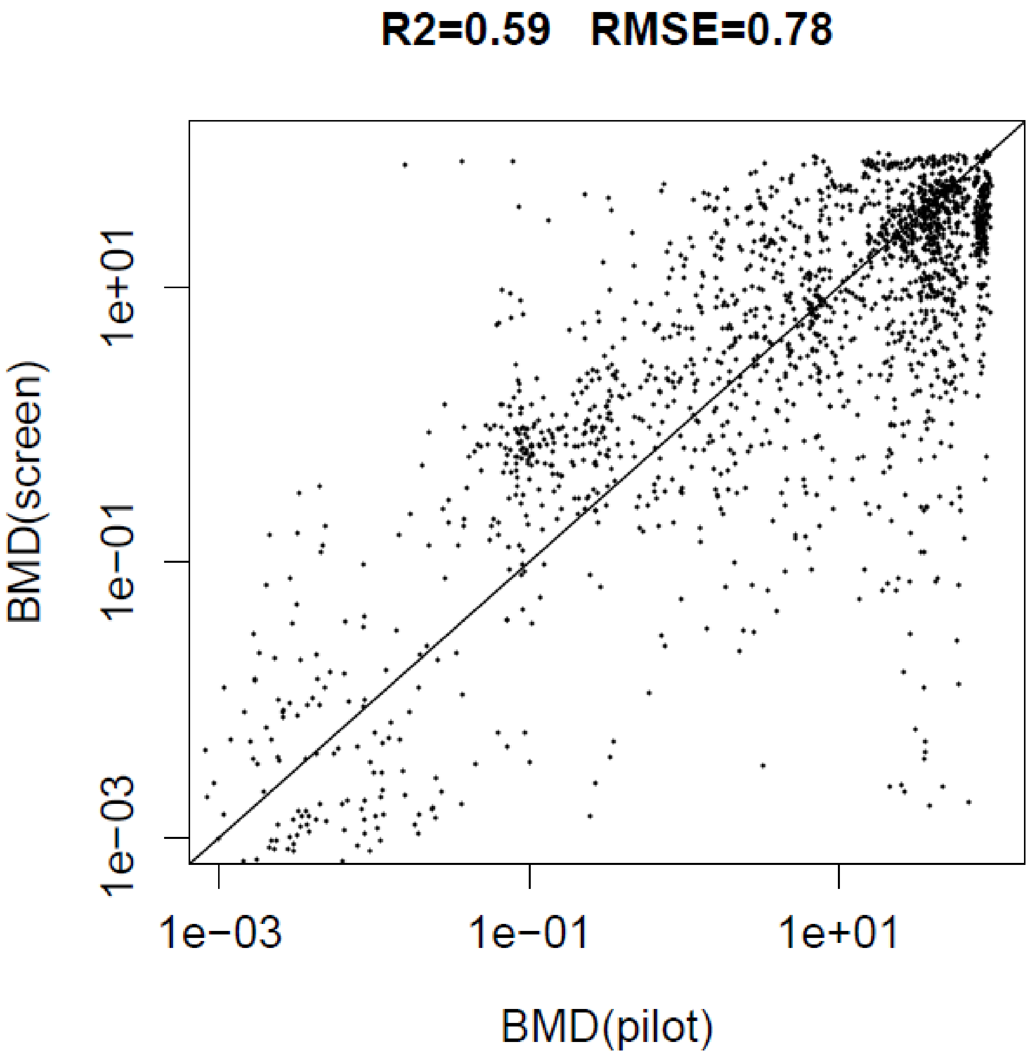
How do potencies compare with other in vitro assays?



Compare potency with estimates from ToxCast ER model using 18 in vitro agonist and antagonist assays.

HTTr values are BMDs from 10 ER signatures active in the 10 most potent ER reference compounds

How Replicable are Potencies?



43 chemicals were run in both the MCF7 pilot and screen studies, > 1 year apart, slightly different protocols

Compare potencies for all signatures that were active in both pilot and screen

A point is one chemical-signature pair