EPA
United States
Environmental Protection
Agency

High-throughput Transcriptomic Profiling of Chemicals for Risk Assessment Applications

Richard Judson, Imran Shah, Logan Everett, Derik Haggard, Beena Vallanat, Joseph Bundy, Bryant Chambers, Woody Setzer, Joshua Harrill
EPA Center for Computational Toxicology and Exposure

SRA DRSG May 5, 2020

Office of Research and Development
Center for Computational Toxicology and Exposure

EPA United States Environmental Protection Agency

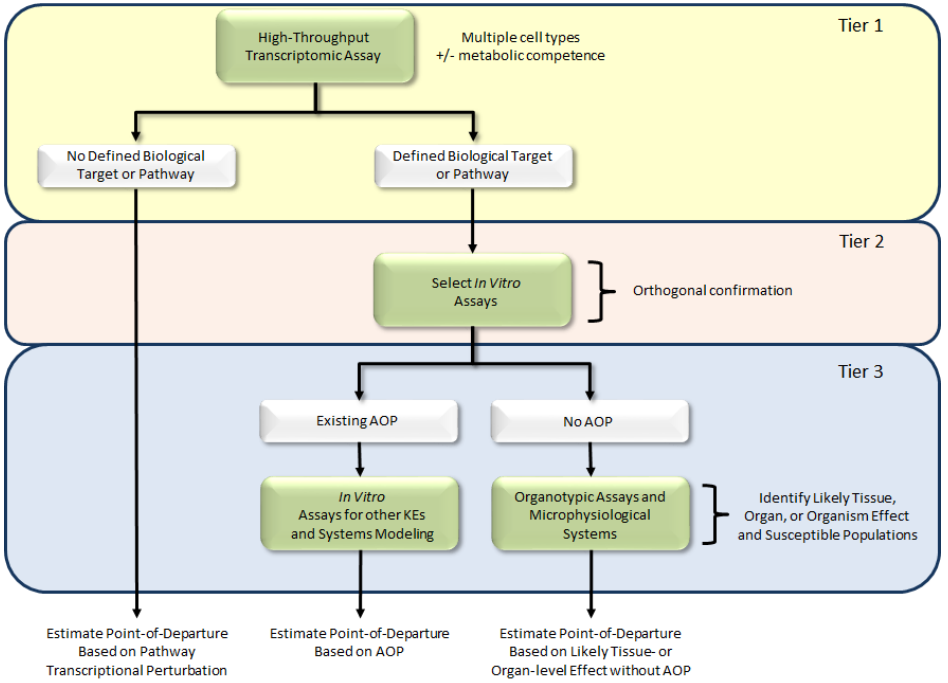
Outline

- Why transcriptomics and TempO-Seq?
- The high-throughput transcriptomics (HTTr) assay
- Processing pipeline and data management
- Platform reproducibility & differential expression
- Concentration-response analysis

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 [DRAFT]

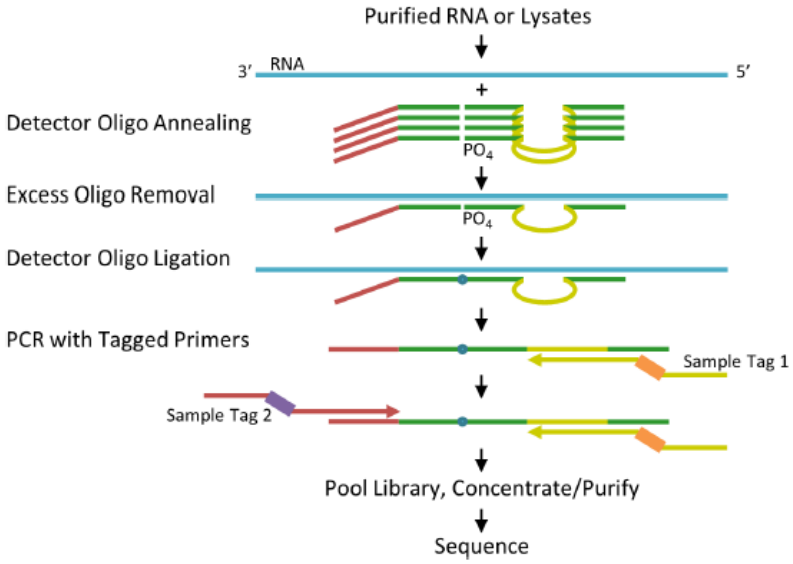


R. Thomas

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 Experiments (more coming in 2020)

MCF7-Pilot

Pilot study to validate workflow, refine experimental design, and develop analysis pipeline

- Cell type: MCF7
- Compounds: 44 chemicals
- Time points: 6 , 12, 24 h
- Media: DMEM +10% HI-FBS or PRF-DMEM + 10% CS-HI-FBS
- Concentration Response: 8
- Replicates: 3
- Data: 6,804 samples x 21,111 transcripts

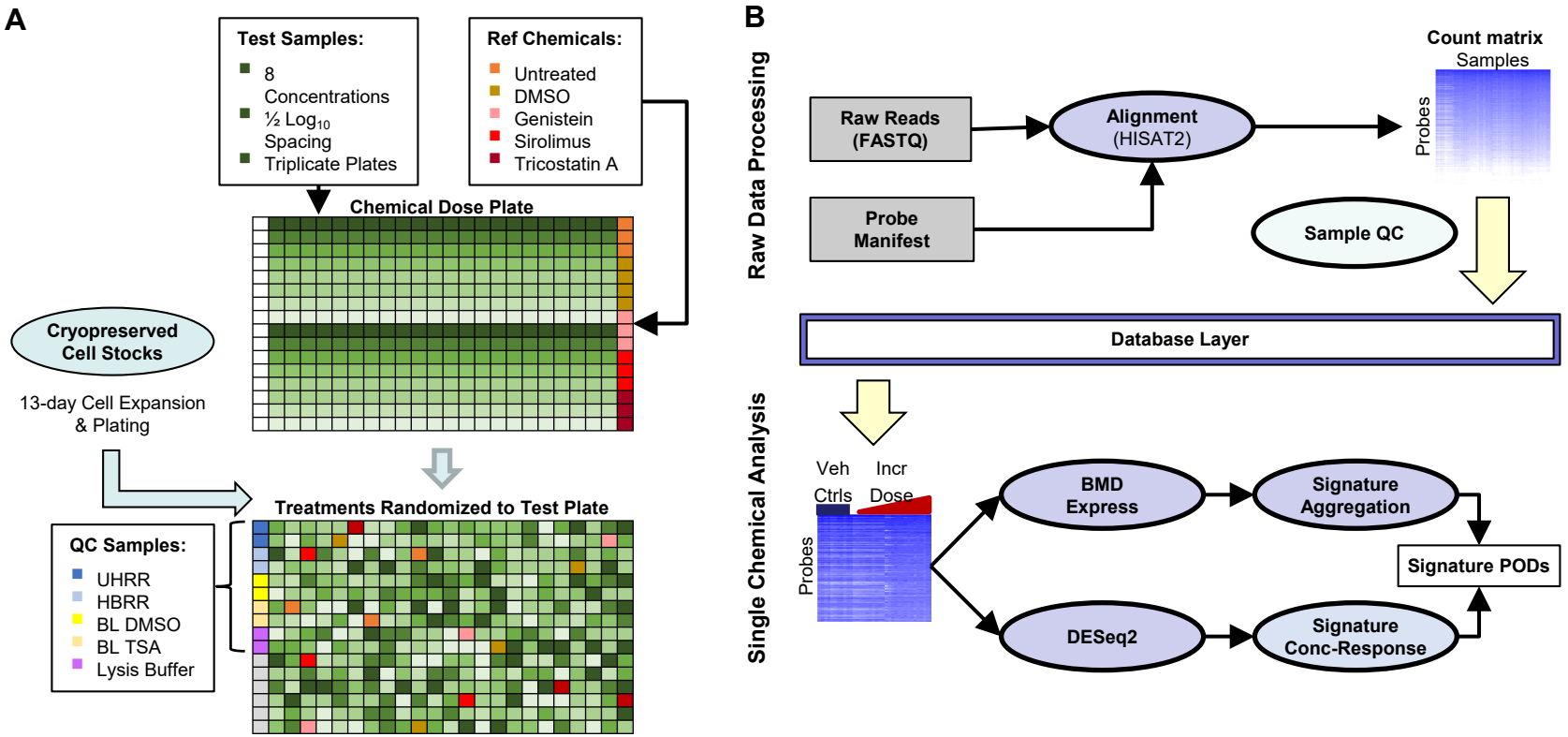
HTTR-PhI

Large-scale screen (Ongoing)

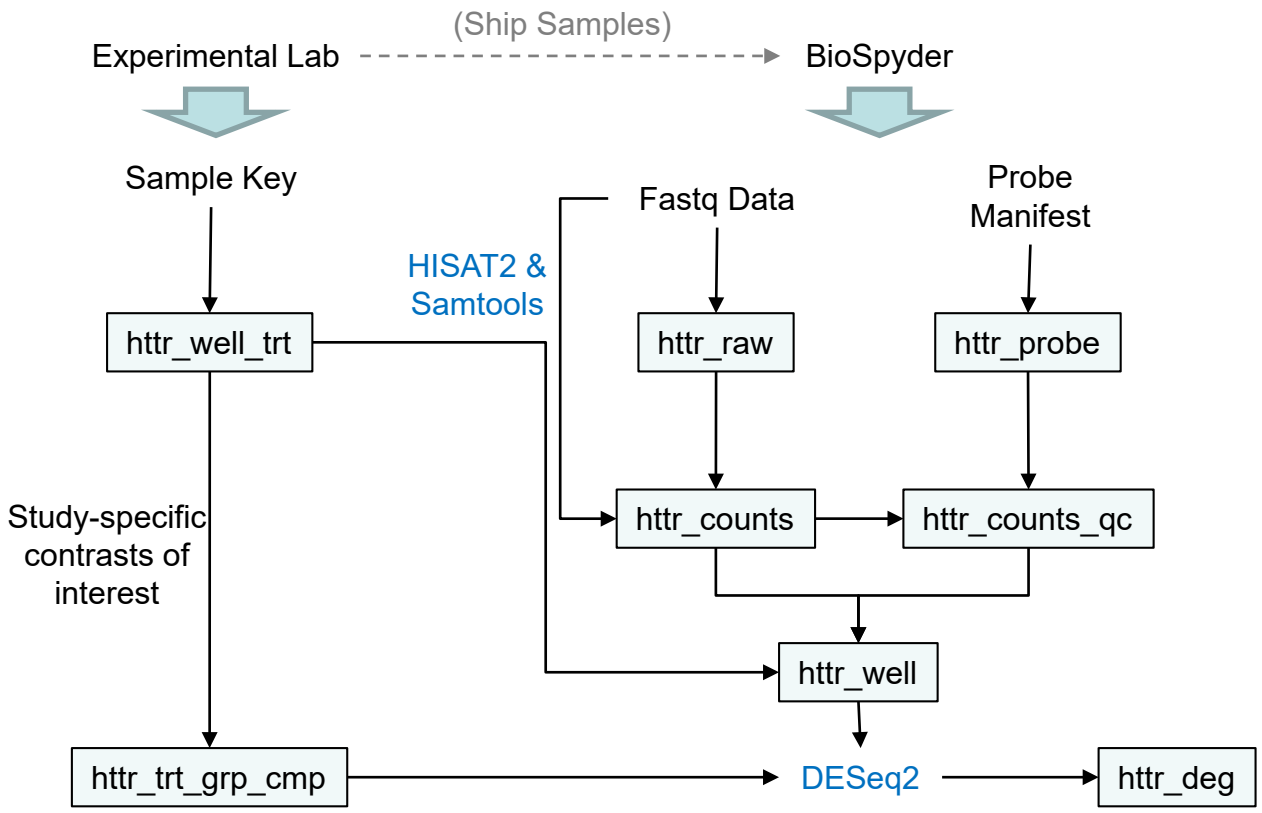
- Cell type: MCF7
- Compounds: 2,200
- Time Point: 6h
- Media: DMEM + 10% HI-FBS
- Concentration Response: 8
- Replicates: 3
- Data: ~53,000 samples x 21,111 transcripts

United States Environmental Protection Agency

Figure 1



HTTr Data Management



Scheduled backups
Recovery plan
Rapid export
Open-source tech

L. Everett

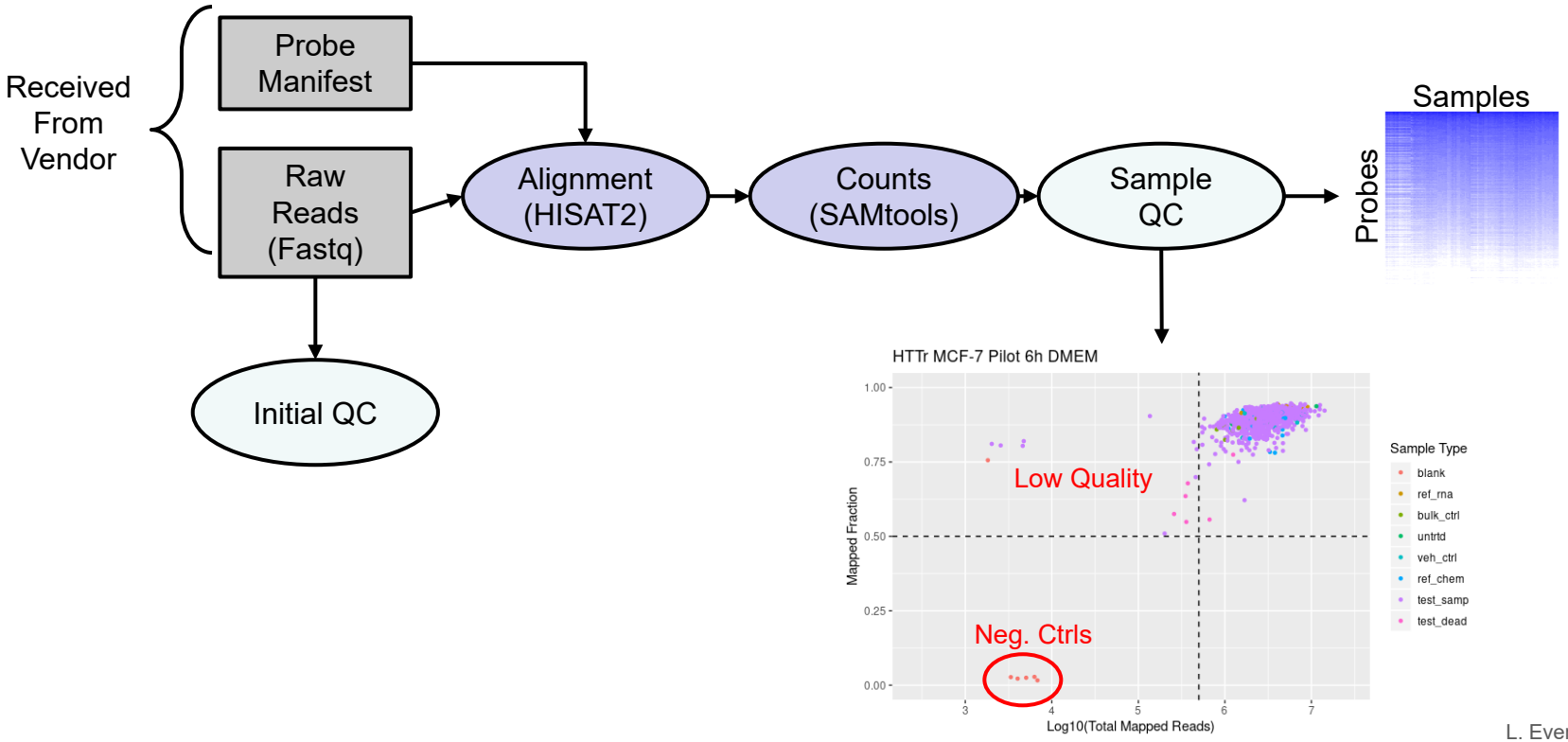
Raw Processing Options

- 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

L. Everett

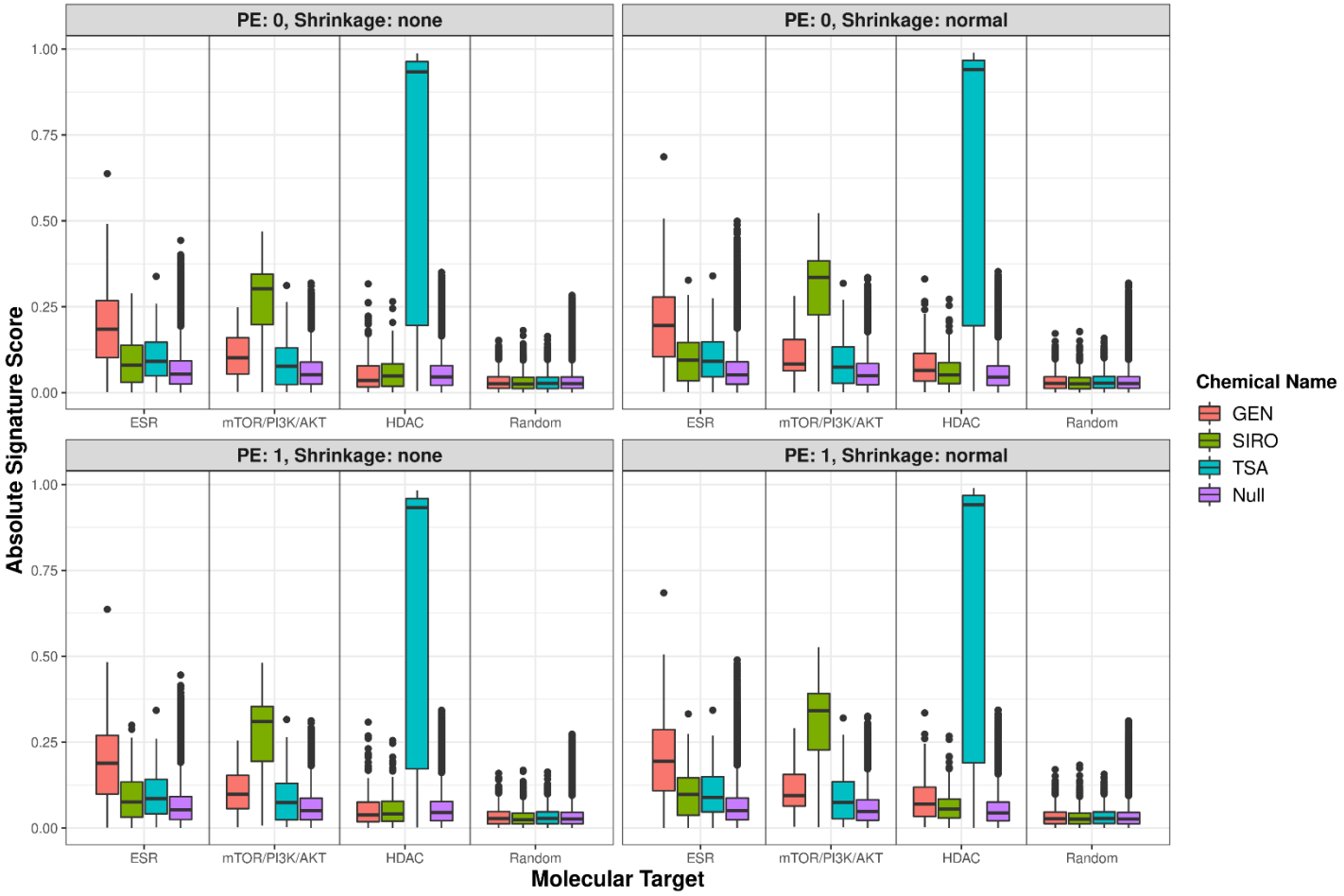
Pipeline: Raw Data Processing



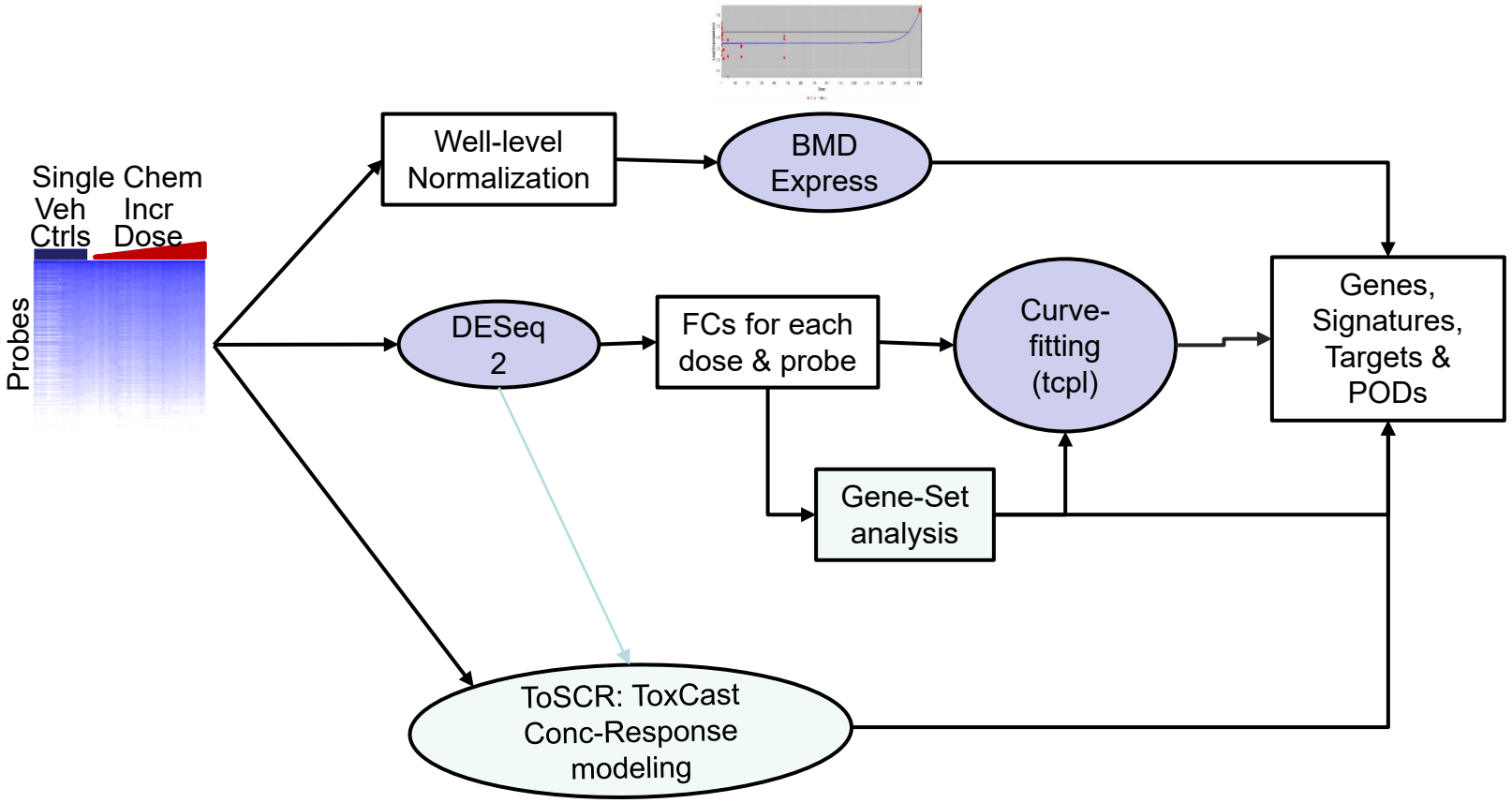
L. Everett

United States Environmental Protection Agency

Plate-wise reference samples



Pipeline: Targets & Concentration Response



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)

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 MyGSEA (ssGSEA)
- 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 Set Selection: “Signatures”

- Select 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
- Pilot: select 7,586 signatures related to targets of chemicals
- Screen: select 22,343 signatures
- Each signature has a hand-annotated “super target” class to help with annotation

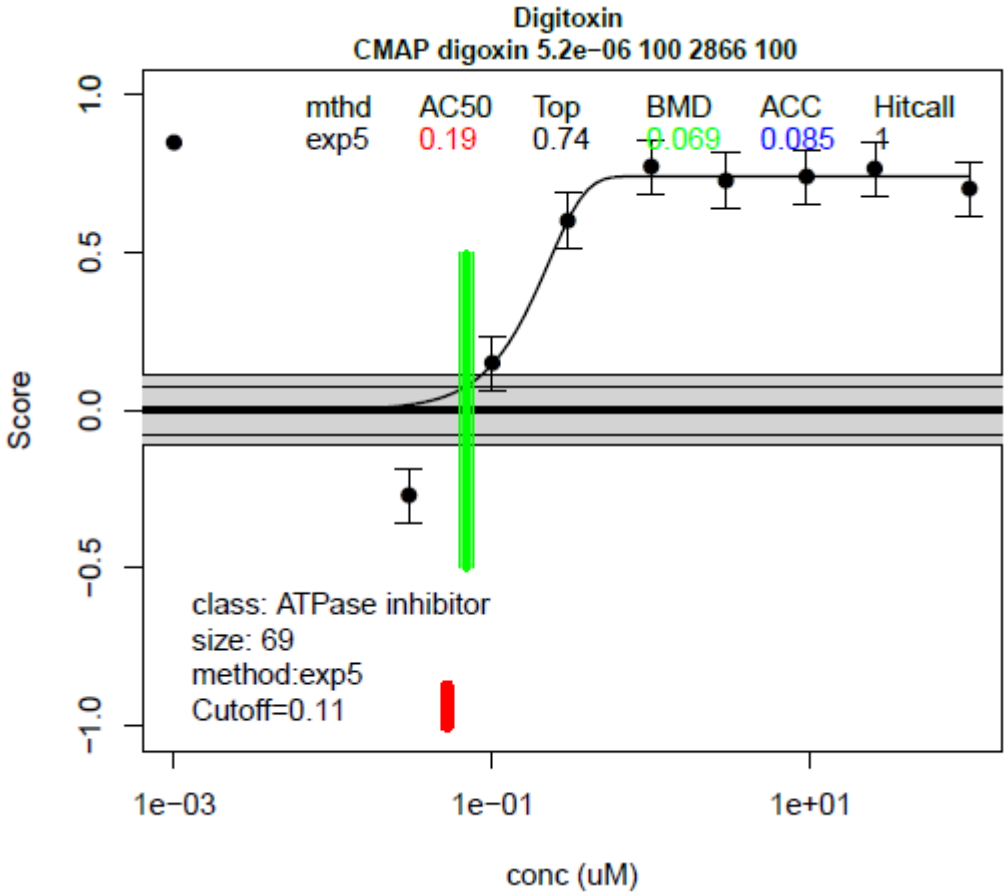
Signature files

- signatureDB_genelists.Rdata
  - List of lists
  - Top level is signature name
  - Second level is a vector of genes
- signatureDB_master_catalog.xlsx
  - Contains all signature annotation
  - Lots of hand-editing is required and this will continue to be updated
  - Contains columns for named signature sets
  - To add a new set of signatures for some analysis, just add a new column and set desired signatures to 1

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 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
- Real chemical response has to exceed 95% CI of the null distribution

Example Concentration-response plot

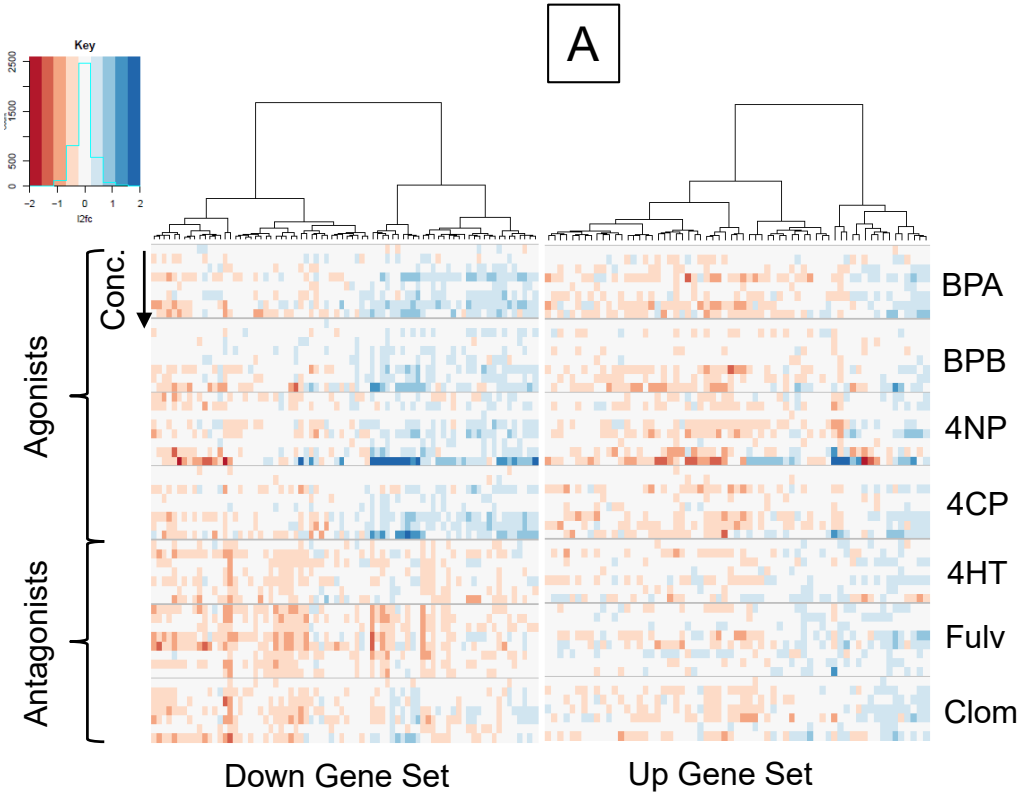


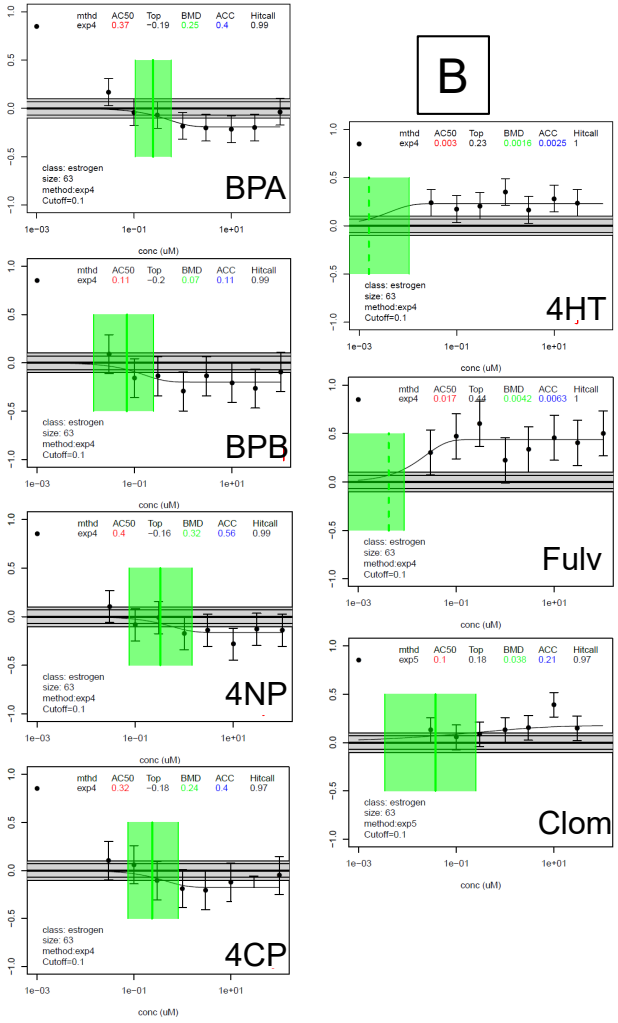
CI around points from the fitting error term

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

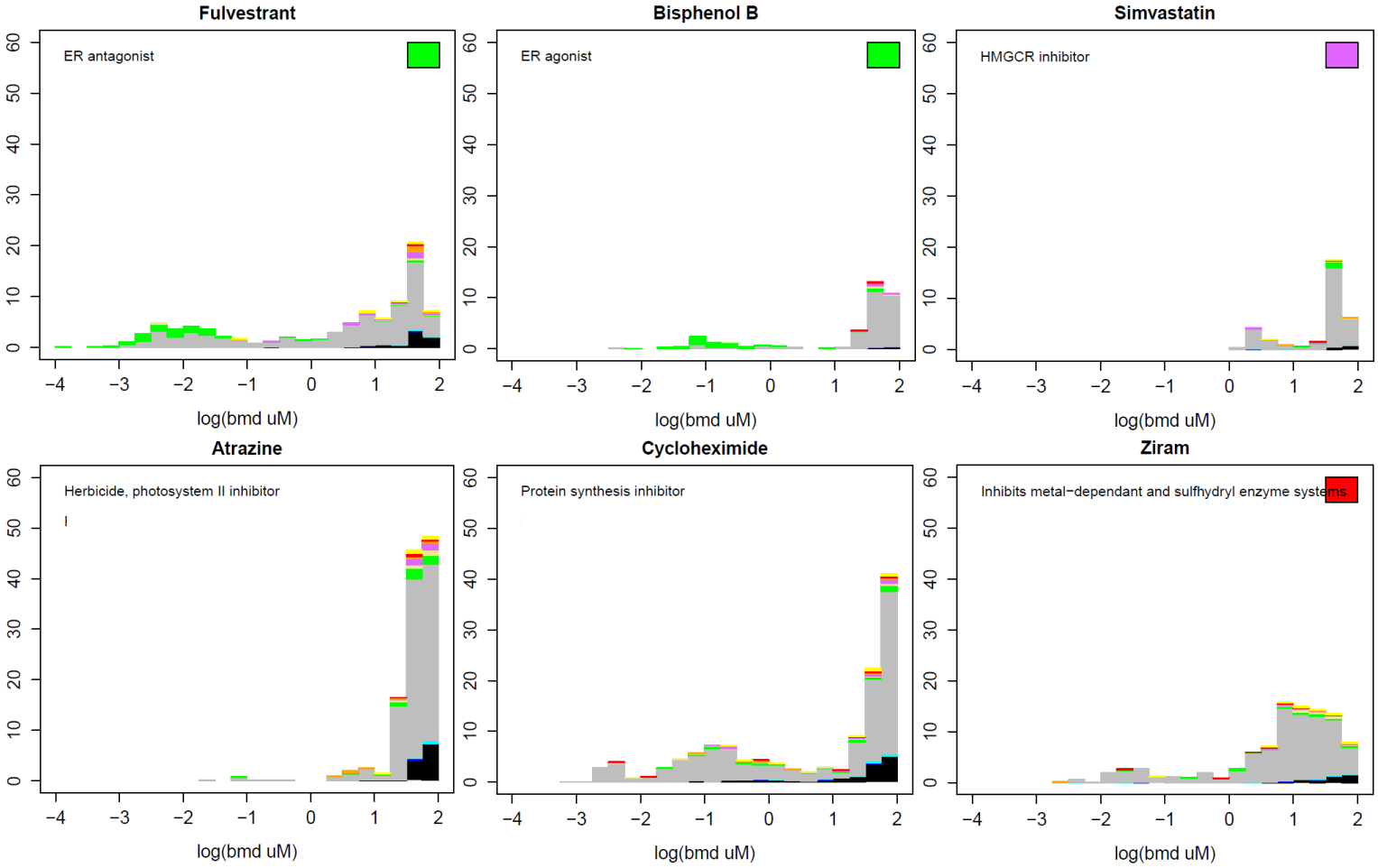
Green vertical band is BMD and 95% CI

Gene-level to signature score





Chemical Level Signature Summary Plots

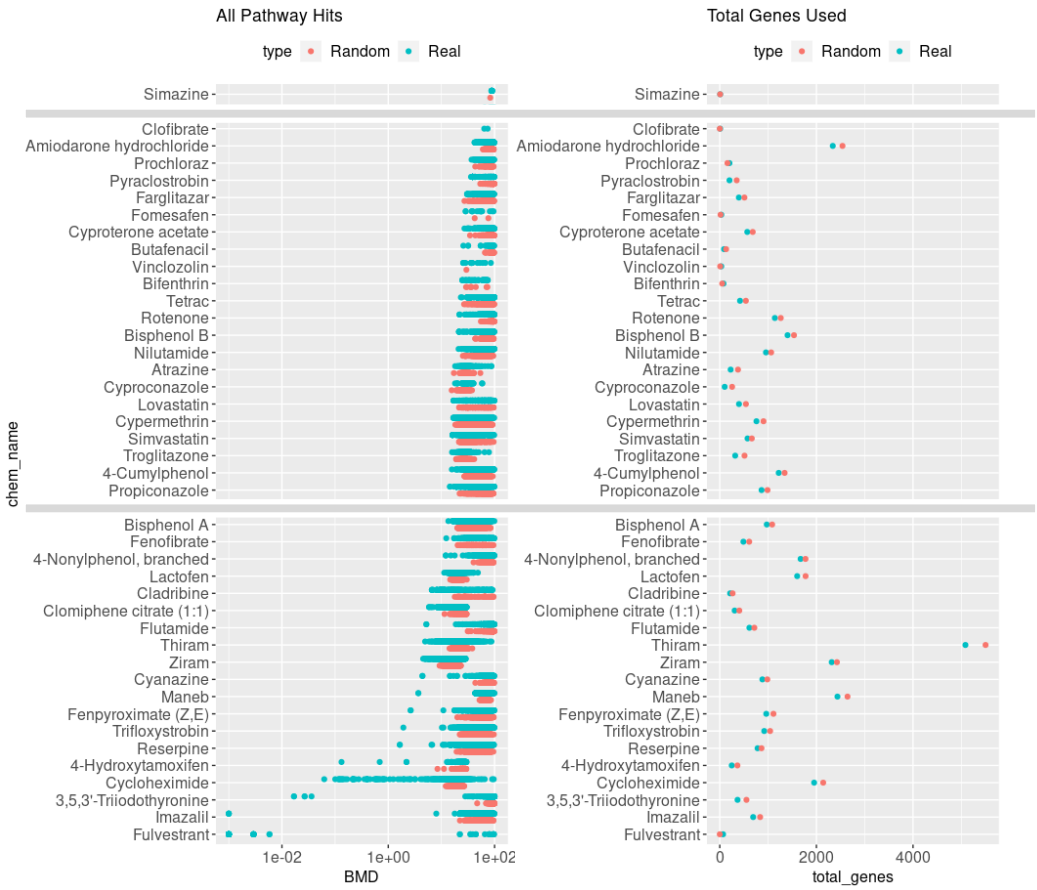


Green=ER
Black=random
Red=cell stress

BMD Express

- Ran BMDExpress using models and parameters specified in NTP RR 5
  - https://ntp.niehs.nih.gov/ntp/results/pubs/rr/reports/rr05_508.pdf
  - Using BMR Factor = 1.349 instead of 1
  - Using fold-change cutoff of 2x, no other pre-filter
- Summarized probe-level BMD values at pathway level following the guidelines in NTP RR 5
  - Consider only BMDs < top dose, BMDU/L < 40, p-value > 0.1
  - Take median of these BMDs for pathways with at least 3 passing genes, 5% coverage
  - Used same pathway collection as for tcpl analysis
  - Included random gene sets but computed min BMD for each chemical separately for random and real gene sets
  - 0.001 uM was used as a minimum limit for pathway level BMDs (Fulvestrant and Imazalil)

MCF7 Pilot DMEM 6h

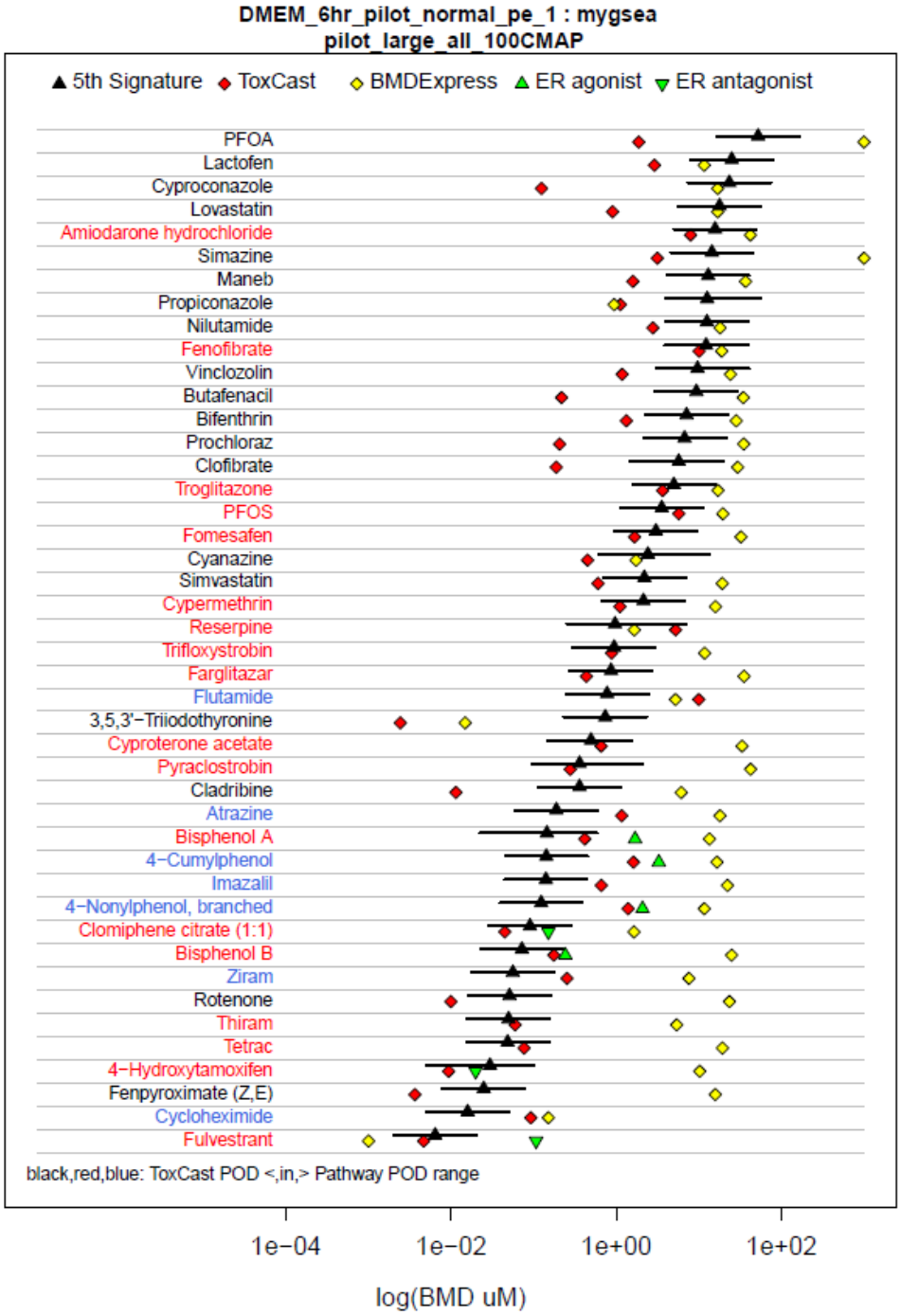


L. Everett

United States Environmental Protection Agency

Chemical-wise PODs

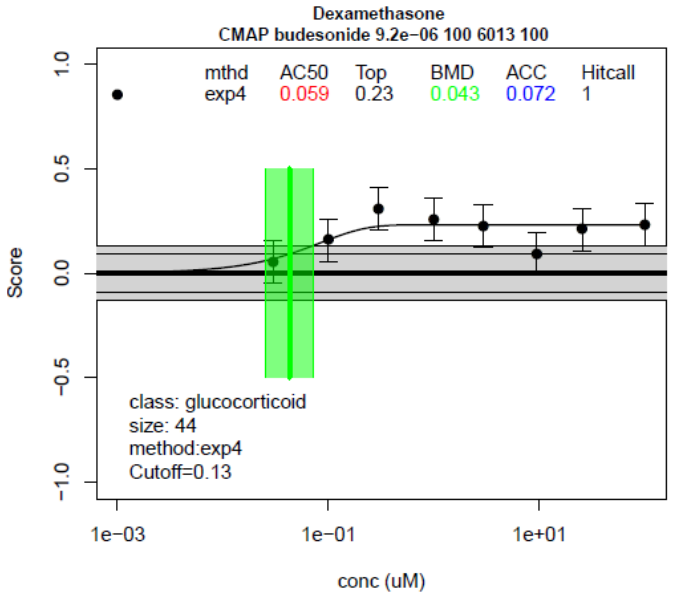
Black: lowest 5%-ile signature
Red: ToxCast 5% POD
Yellow: BMD Express
Green: ToxCast ER Model

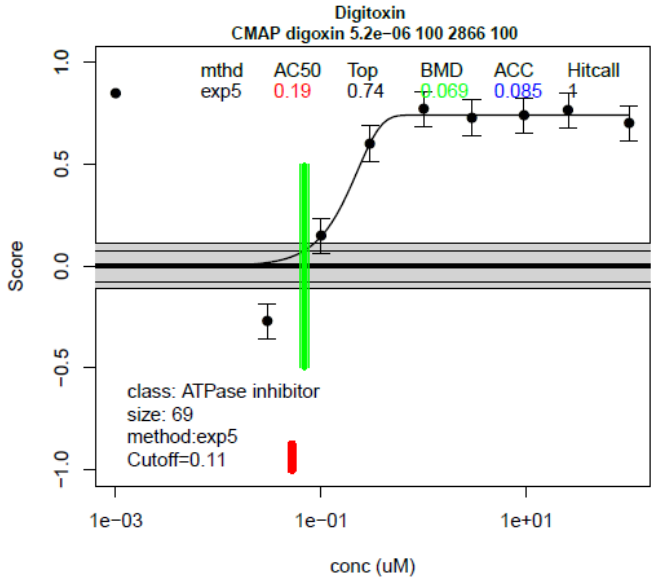


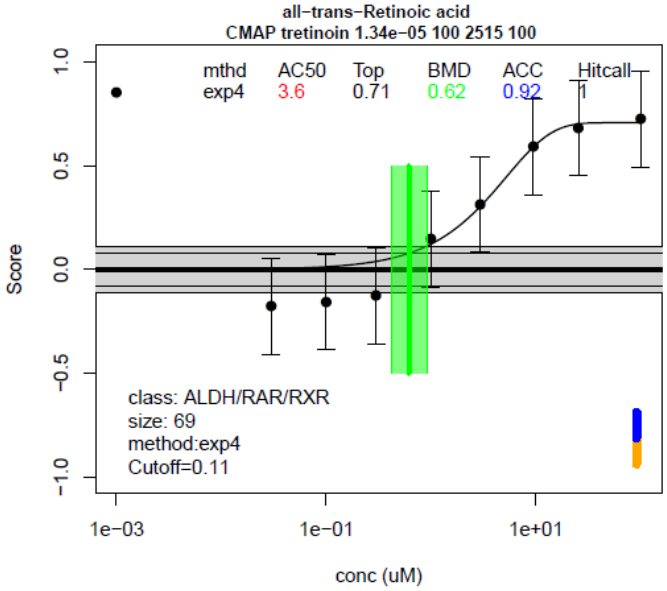
MCF7 Screen, Preliminary Observations

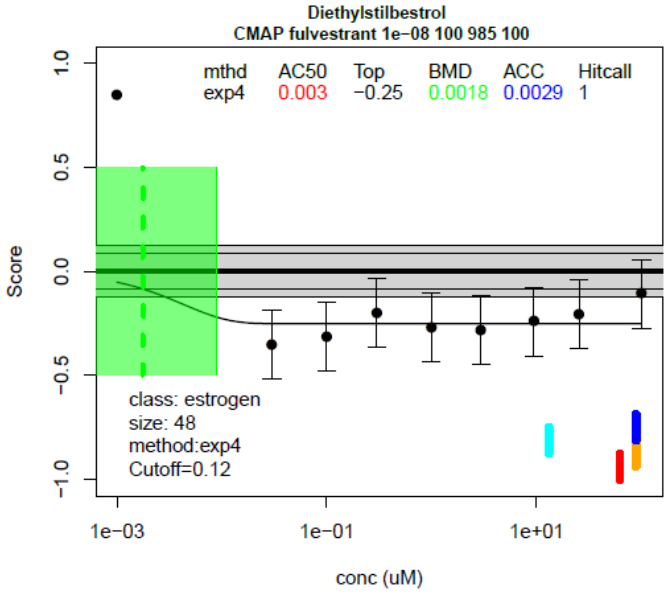
- 2112 Chemicals
- Drugs, food chemicals, pesticides, industrial chemicals
- 8-point concentration-response
- 6 hour exposure
- 22,343 signatures
- 355 chemicals have gene target annotations
  - Used to assess how well the active signatures match the chemical target

More activity that just Estrogen Receptor

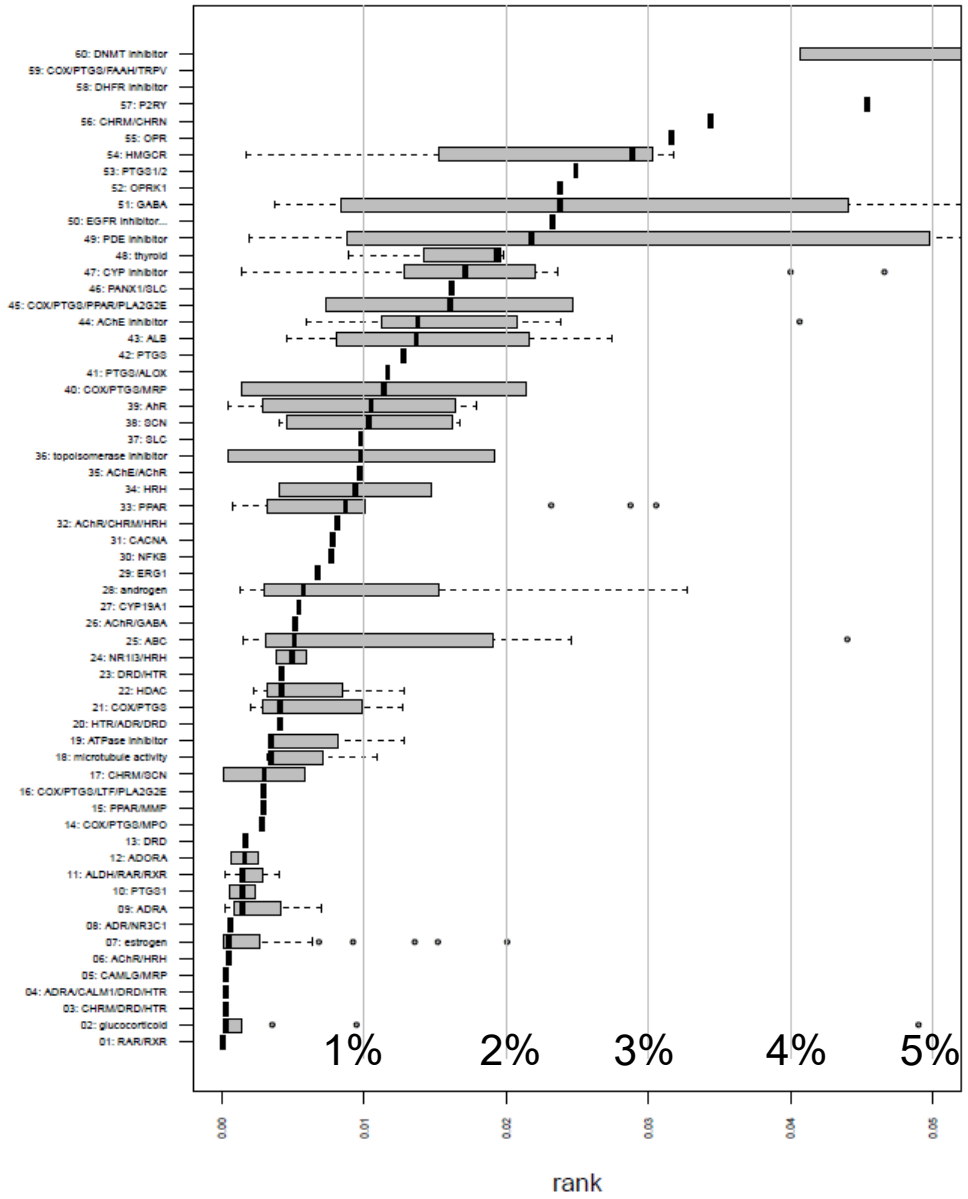








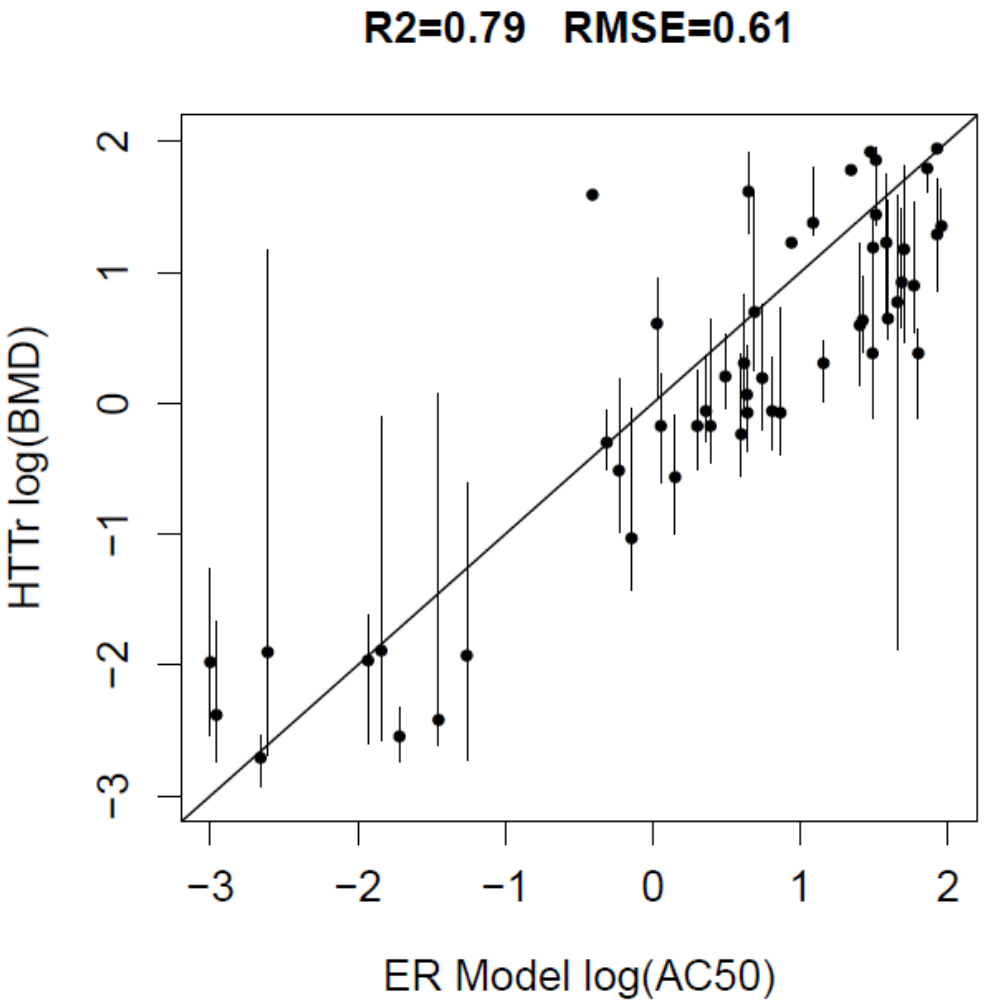
Measuring how well the signatures ID the chemical target



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

Lowest set are all GPCR or nuclear receptor target families

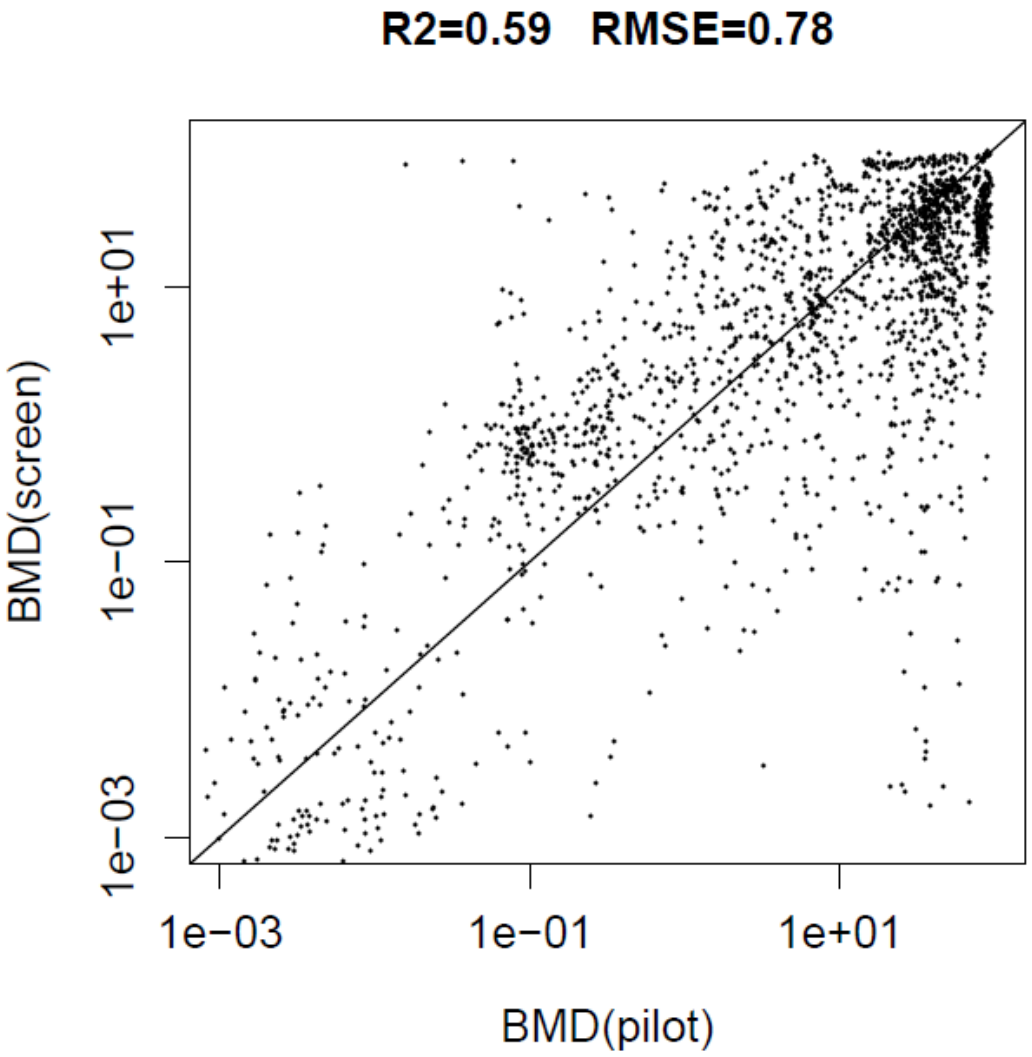
How do potencies compare with other in vitro assays?



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

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?
- How do we summarize the data per chemical or chemical set?
- 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
- Clinton Willis
- Richard Brockway
- Johanna Nyffeler
- Megan Culbreth
- Dan Hallinger
- Terri Fairley
- Matt Martin
- Agnes Karmaus