EPA United States Environmental Protection Agency

Concentration-response Modeling in High-throughput Transcriptomics

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

SRA DRSG May 5, 2020

Office of Research and Development
National Center for Computational Toxicology

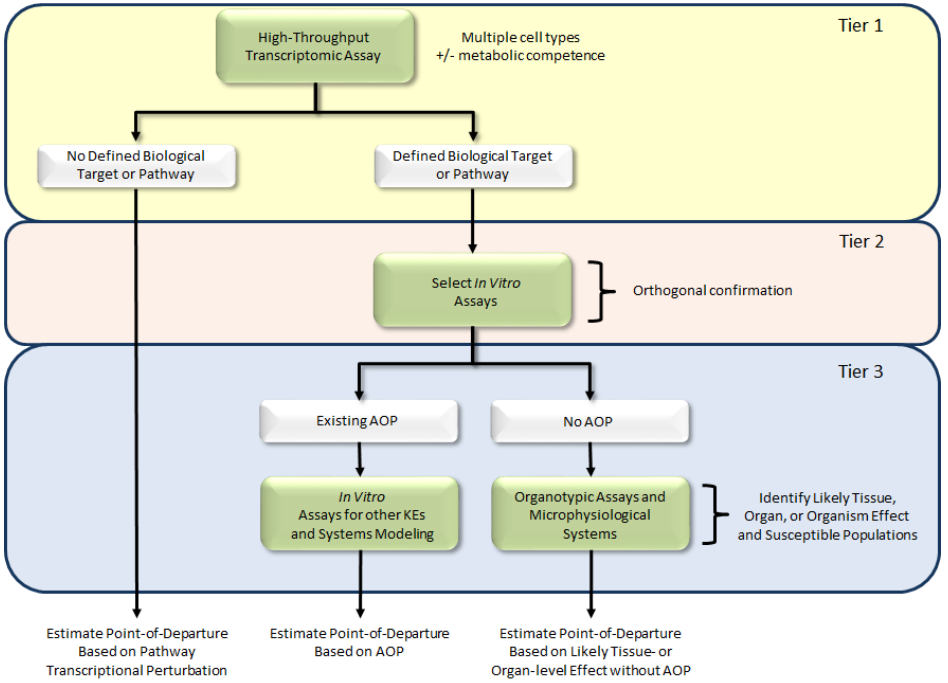
Outline

- 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]

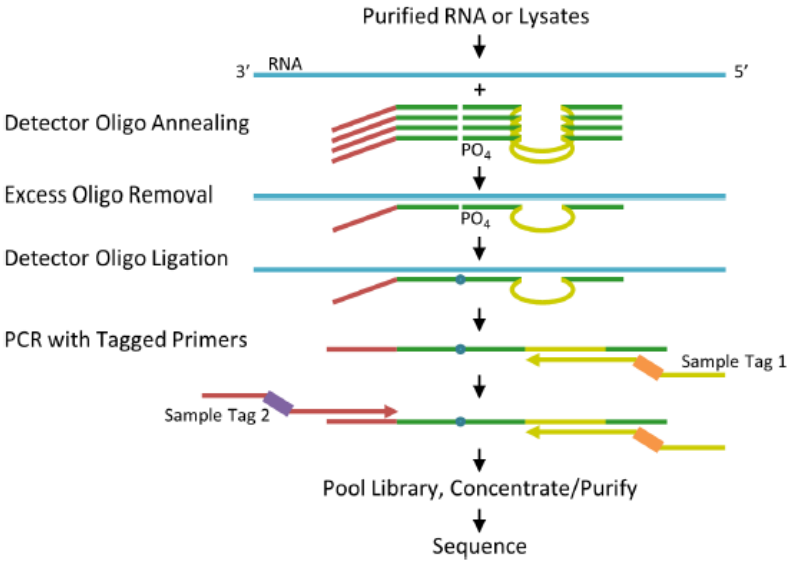


R. Thomas

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: PRF- / PRF+ (DMEM +10% HI-FBS)
- Concentration Response: 8
- Replicates: 3
- Data: 6,804 samples x 21,111 transcripts

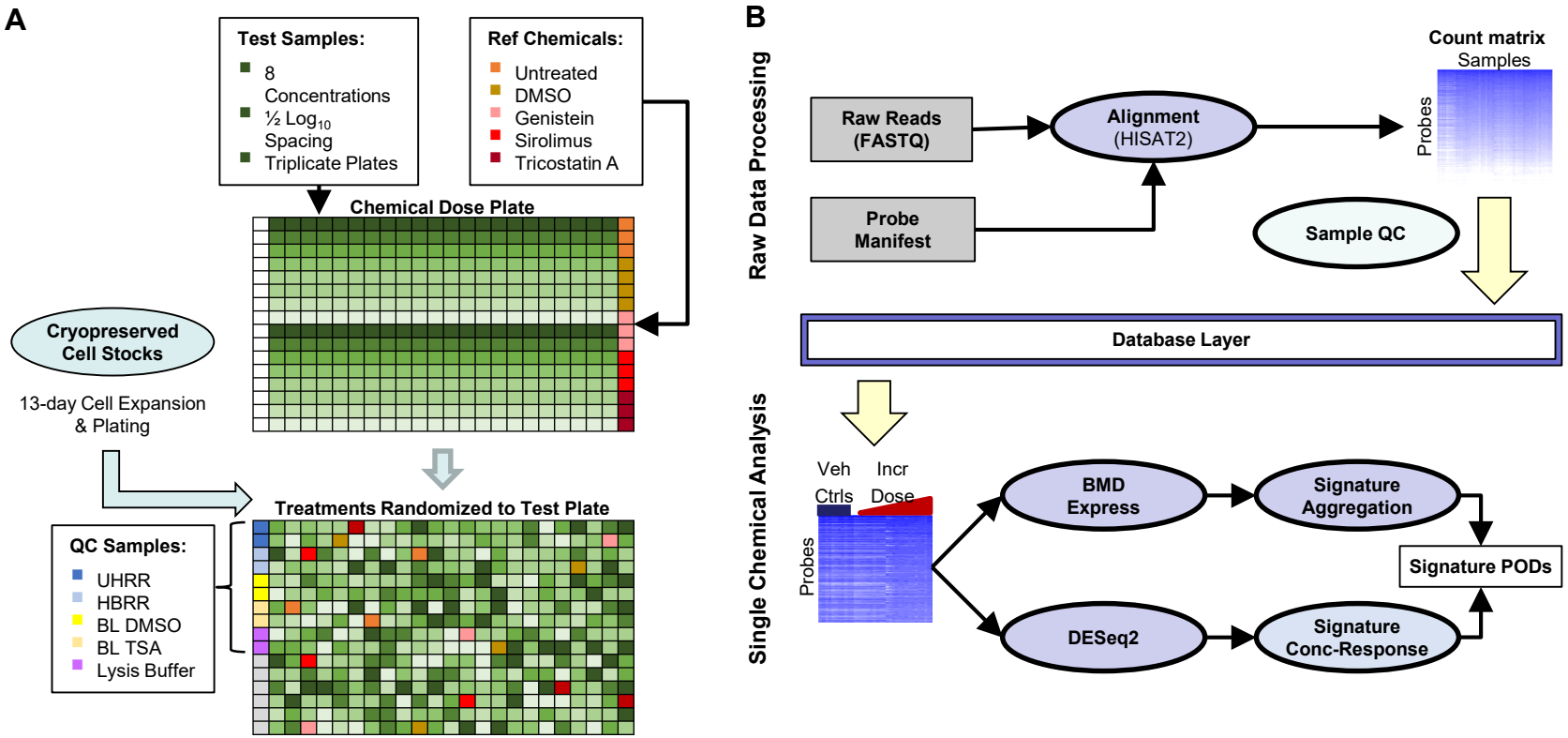
HTTR-PhI

Large-scale screen (Ongoing)

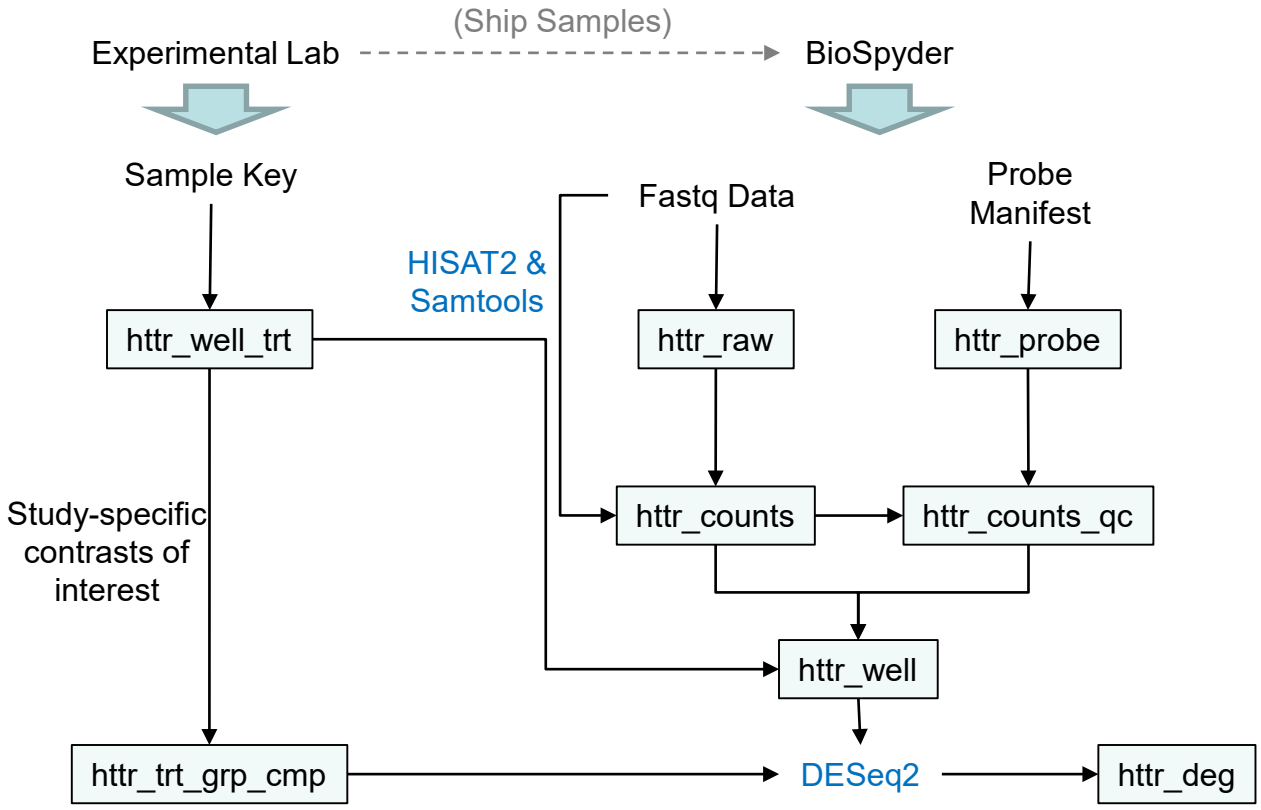
- Cell type: MCF7
- Compounds: 2,200
- Time Point: 6h
- Media: PRF+
- Concentration Response: 8
- Replicates: 3
- Data: ~53,000 samples x 21,111 transcripts

United States Environmental Protection Agency

Figure 1



HTTr Data Management



Scheduled backups
Recovery plan
Rapid export
Open-source tech

L. Everett

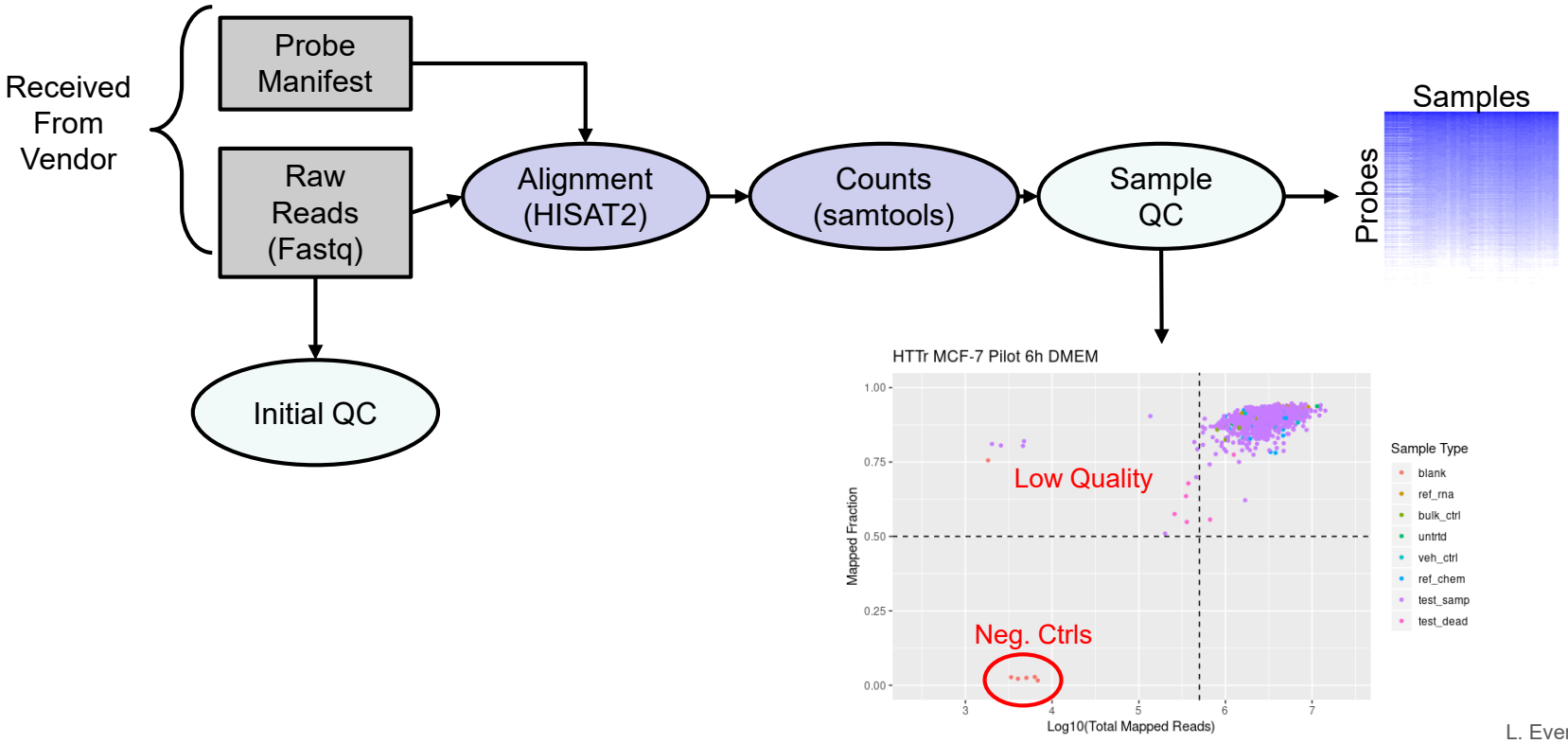
Raw Processing Options

- 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

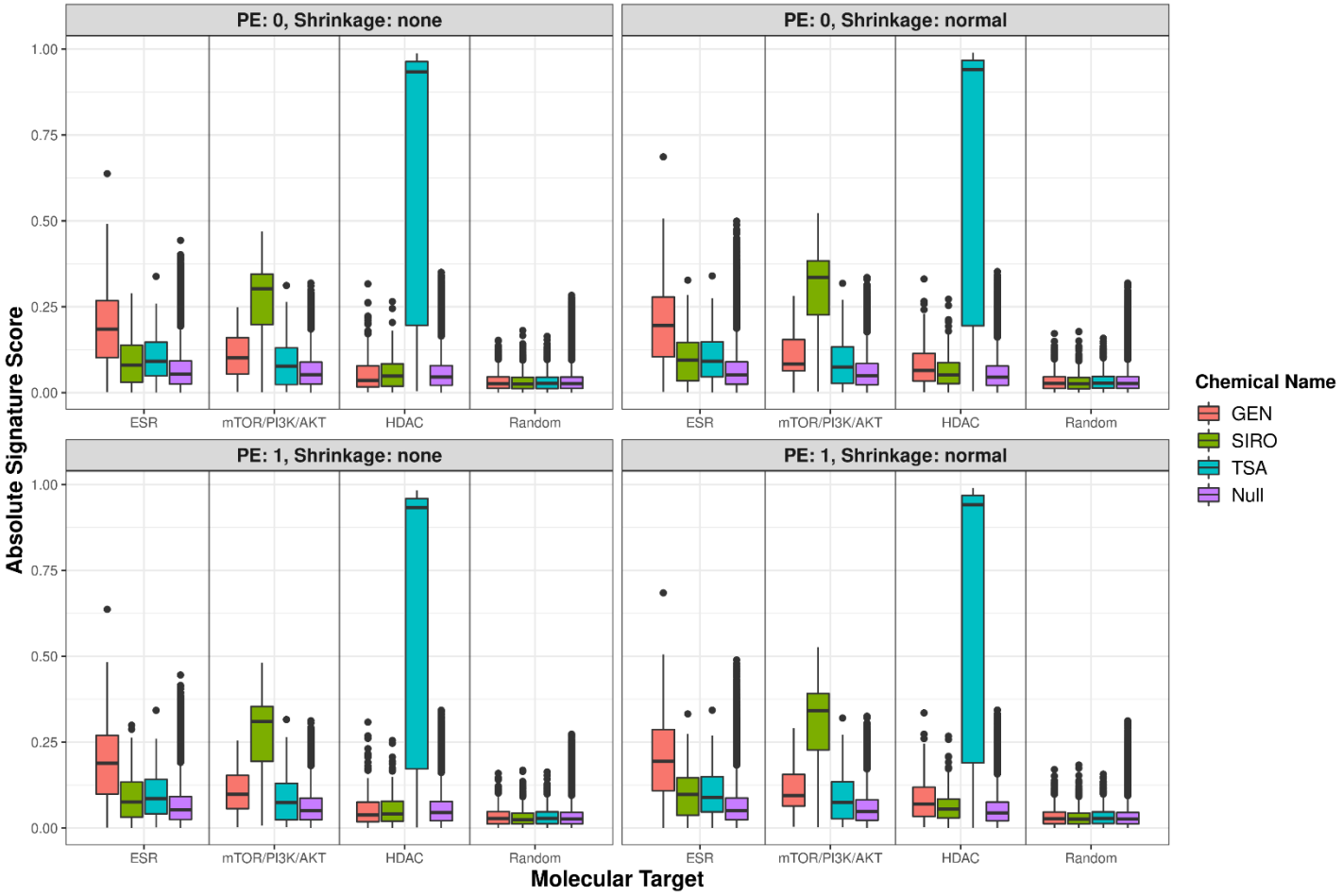
L. Everett

Pipeline: Raw Data Processing

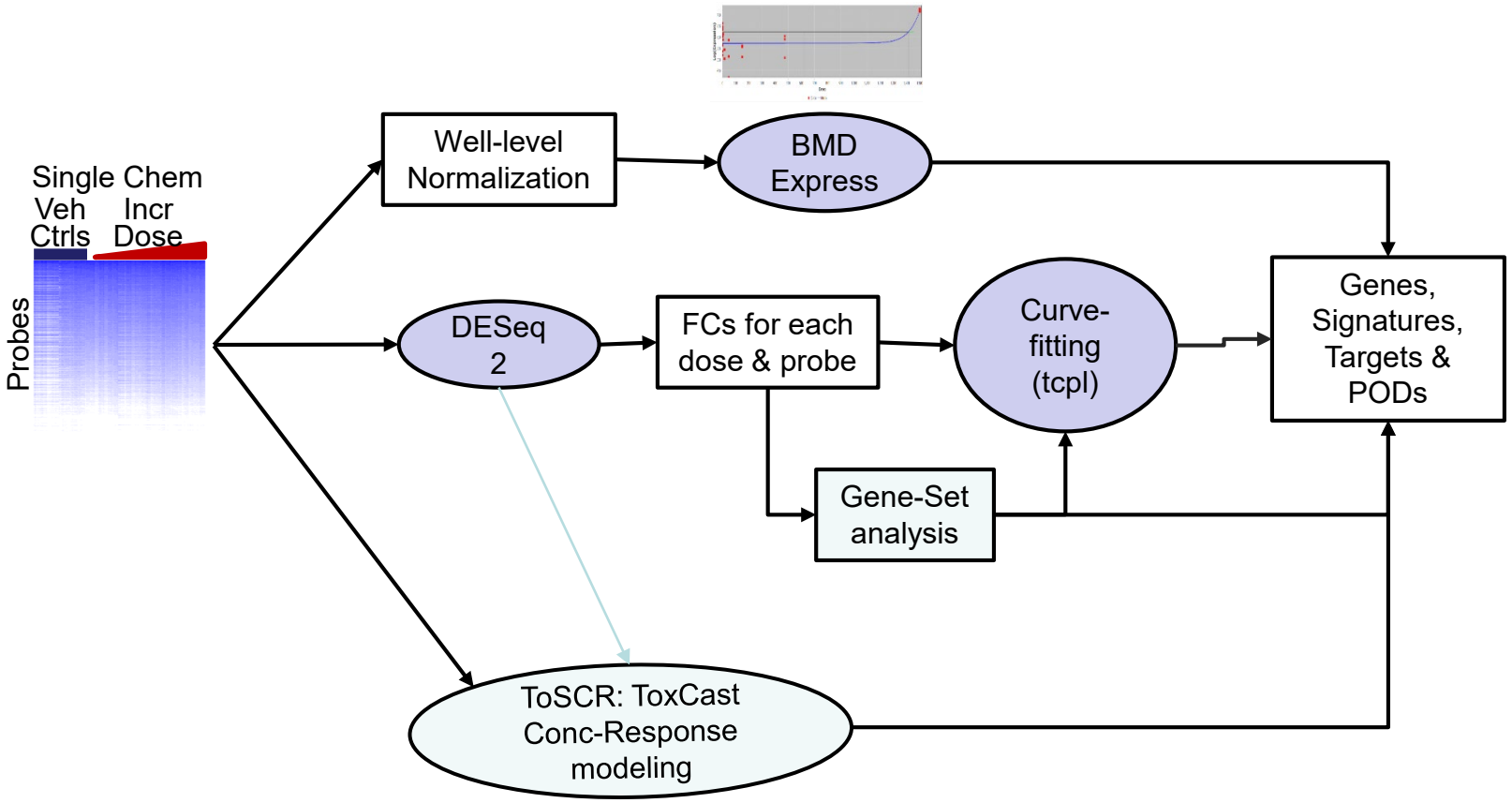


L. Everett

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

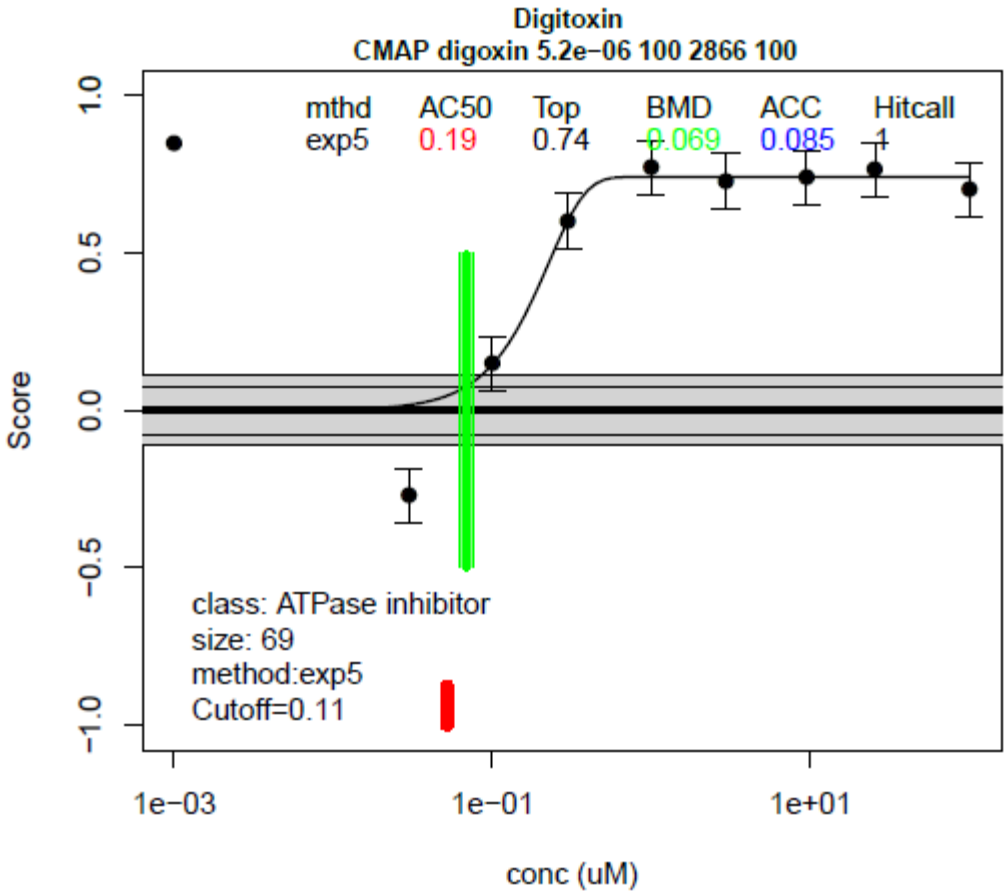
Signature files

- 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

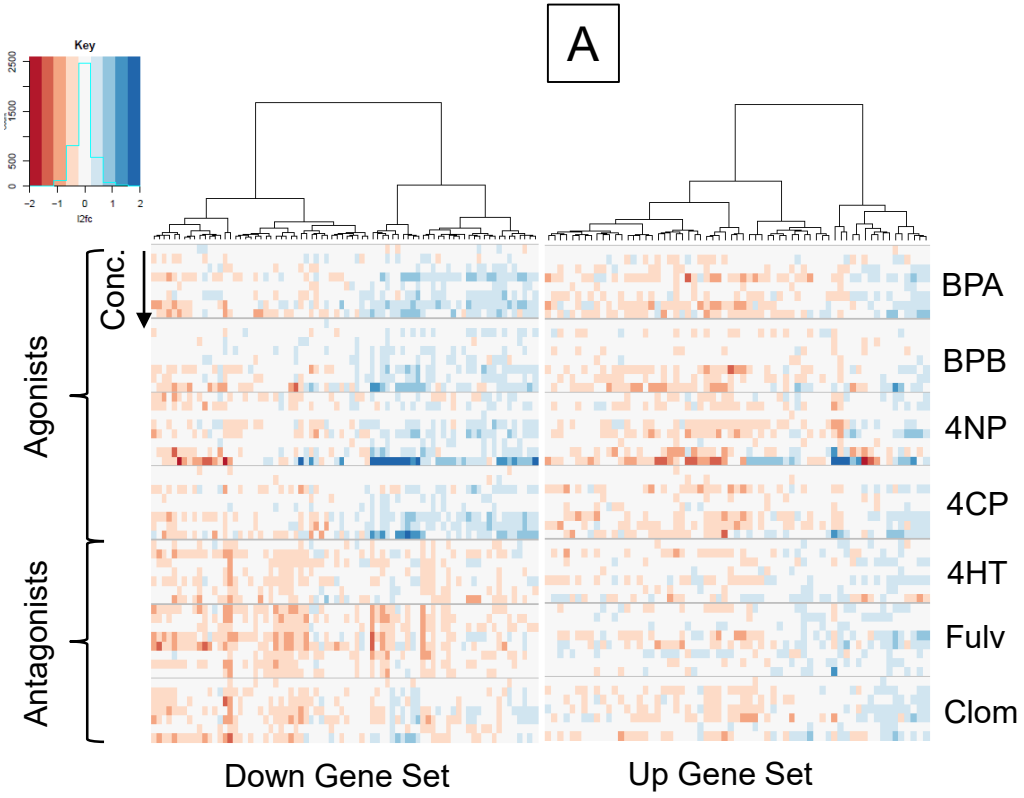
United States Environmental Protection Agency

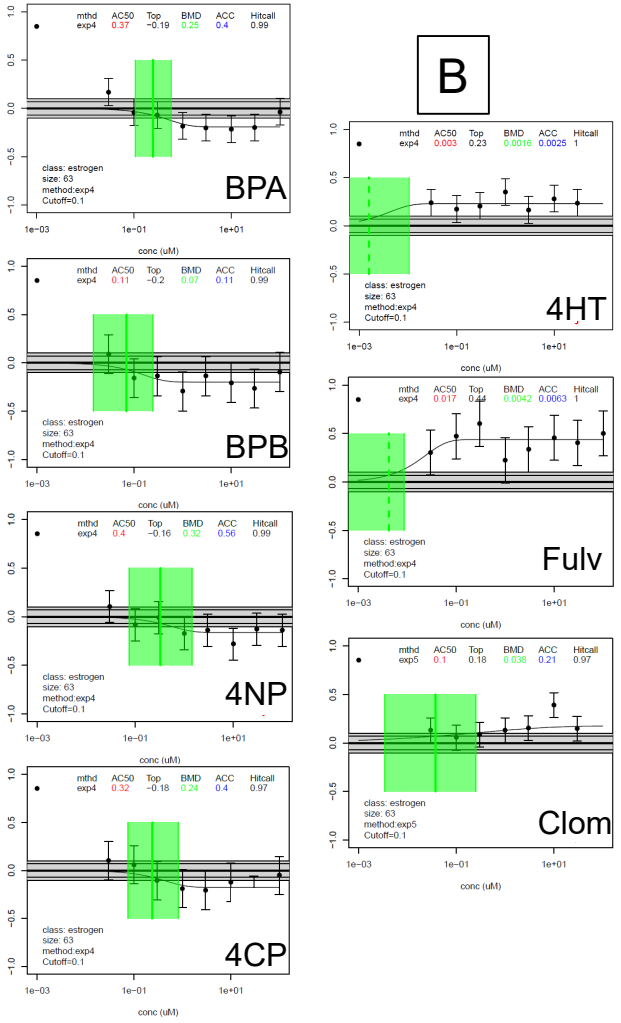
MCF7 Pilot:

Cell type: MCF7
Compounds: 44 chemicals
Time points: 6 h
Media: DMEM
Concentrations: 8
Replicates: 3
Data: 6,804 samples x 21,111 transcripts

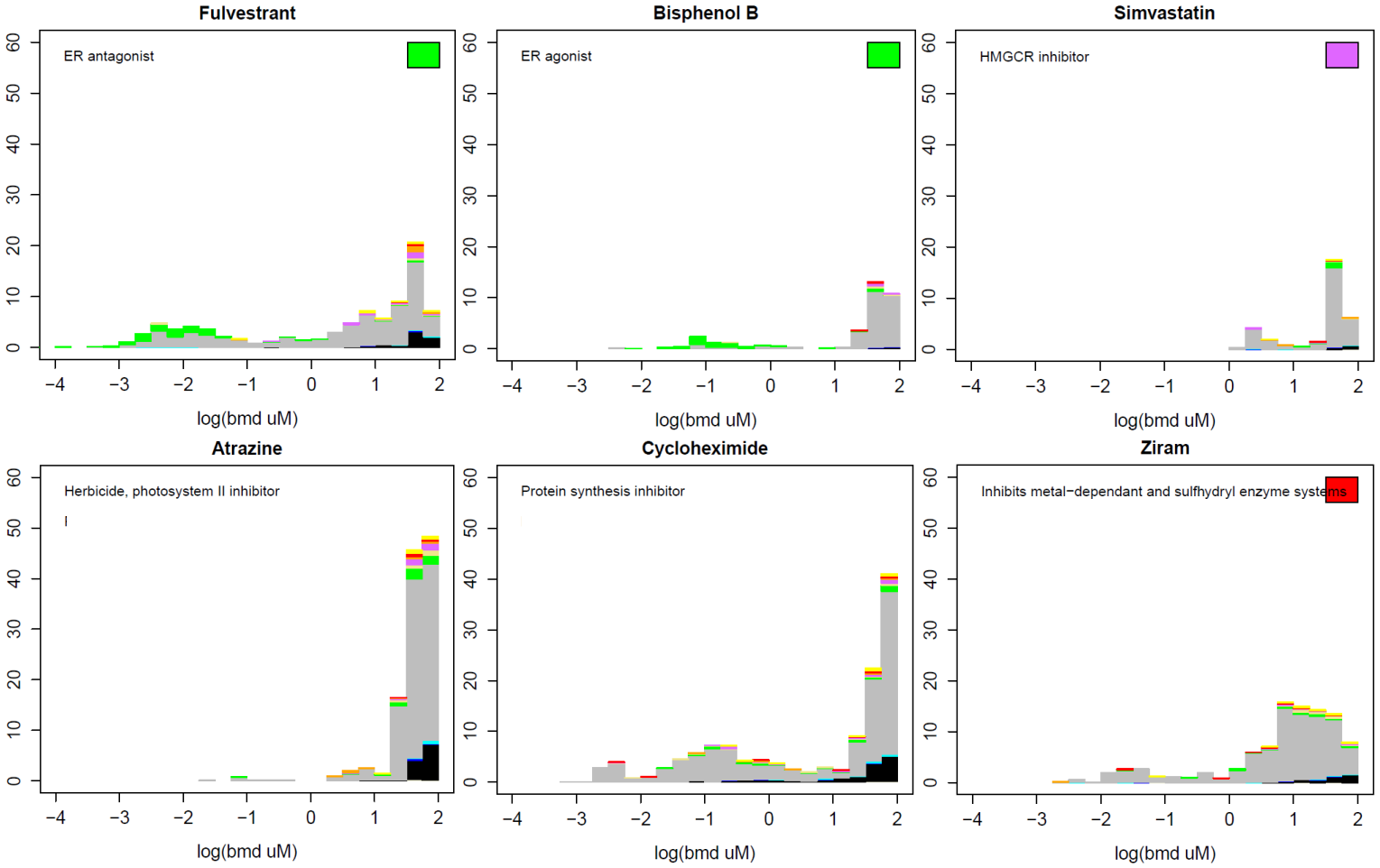
Name	CASRN	Target annotation	Target key
3,5,3'-Triiodothyronine	6893-02-3	Thyroid hormone receptor agonist	thyroid
4-Cumylphenol	599-64-4	ER agonist	ER
4-Hydroxytamoxifen	68392-35-8	ER antagonist	ER
4-Nonylphenol, branched	84852-15-3	ER agonist	ER
Amiodarone hydrochloride	19774-82-4	Blocks myocardial Ca, K, Na channels	ion channel
Atrazine	1912-24-9	Herbicide, photosystem II inhibitor	electron chain
Bifenthrin	82657-04-3	Sodium channel modulator	ion channel
Bisphenol A	80-05-7	ER agonist	ER
Bisphenol B	77-40-7	ER agonist	ER
Butafenacil	134605-64-4	Herbicide, protoporphyrinogen oxidase (PPO) inhibition	Plant PPO
Cladribine	4291-63-8	DNA synthesis inhibitor	DNA
Clofibrate	637-07-0	PPARa agonist, upregulates extrahepatic lipoprotein lipase	PPAR
Clomiphene citrate (1:1)	50-41-9	ER antagonist	ER
Cyanazine	21725-46-2	Herbicide, photosystem II inhibitor	electron chain
Cycloheximide	66-81-9	Protein synthesis inhibitor	protein synthesis
Cypermethrin	52315-07-8	Sodium channel modulator	ion channel
Cyproconazole	94361-06-5	Ergosterol-biosynthesis inhibitor. Pan-cyp inhibitor	CYPs
Cyproterone acetate	427-51-0	AR antagonist	AR
Farglitazar	196808-45-4	PPARg agonist	PPAR
Fenofibrate	49562-28-9	PPARa agonist, upregulates extrahepatic lipoprotein lipase	PPAR
Fenpyroximate (Z,E)	111812-58-9	Mitochondrial electron transport inhibitor	mitochondria
Flutamide	13311-84-7	AR antagonist	AR
Fomesafen	72178-02-0	Herbicide, protoporphyrinogen oxidase (PPO) inhibition	Plant PPO
Fulvestrant	129453-61-8	ER antagonist	ER
Imazalil	35554-44-0	Ergosterol-biosynthesis inhibitor. Pan-cyp inhibitor	CYPs
Lactofen	77501-63-4	Herbicide, protoporphyrinogen oxidase (PPO) inhibition	Plant PPO
Lovastatin	75330-75-5	HMGCR inhibitor	cholesterol
Maneb	12427-38-2	Inhibits metal-dependant and sulfhydryl enzyme systems	protein reactive
Nilutamide	63612-50-0	AR antagonist	AR
Prochloraz	67747-09-5	Ergosterol-biosynthesis inhibitor. Pan-cyp inhibitor	CYPs
Propiconazole	60207-90-1	Ergosterol-biosynthesis inhibitor. Pan-cyp inhibitor	CYPs
Pyraclostrobin	175013-18-0	Mitochondria (complex III inhibitor)	mitochondria
Reserpine	50-55-5	inhibition of the ATP/Mg2+ pump	adrenergic
Rotenone	83-79-4	Mitochondria (complex I inhibitor)	mitochondria
Simazine	122-34-9	Herbicide, photosystem II inhibitor	electron chain
Simvastatin	79902-63-9	HMGCR inhibitor	cholesterol
Tetrac	67-30-1	T4 synthesis inhibitor	thyroid
Thiram	137-26-8	Inhibits metal-dependant and sulfhydryl enzyme systems	protein reactive
Trifloxystrobin	141517-21-7	Mitochondria (complex III inhibitor)	mitochondria
Troglitazone	97322-87-7	PPARg, PPARa agonist	PPAR
Vinclozolin	50471-44-8	AR antagonist	AR
Ziram	137-30-4	Inhibits metal-dependant and sulfhydryl enzyme systems	protein reactive

Gene-level to signature score





Chemical Level Signature Summary Plots

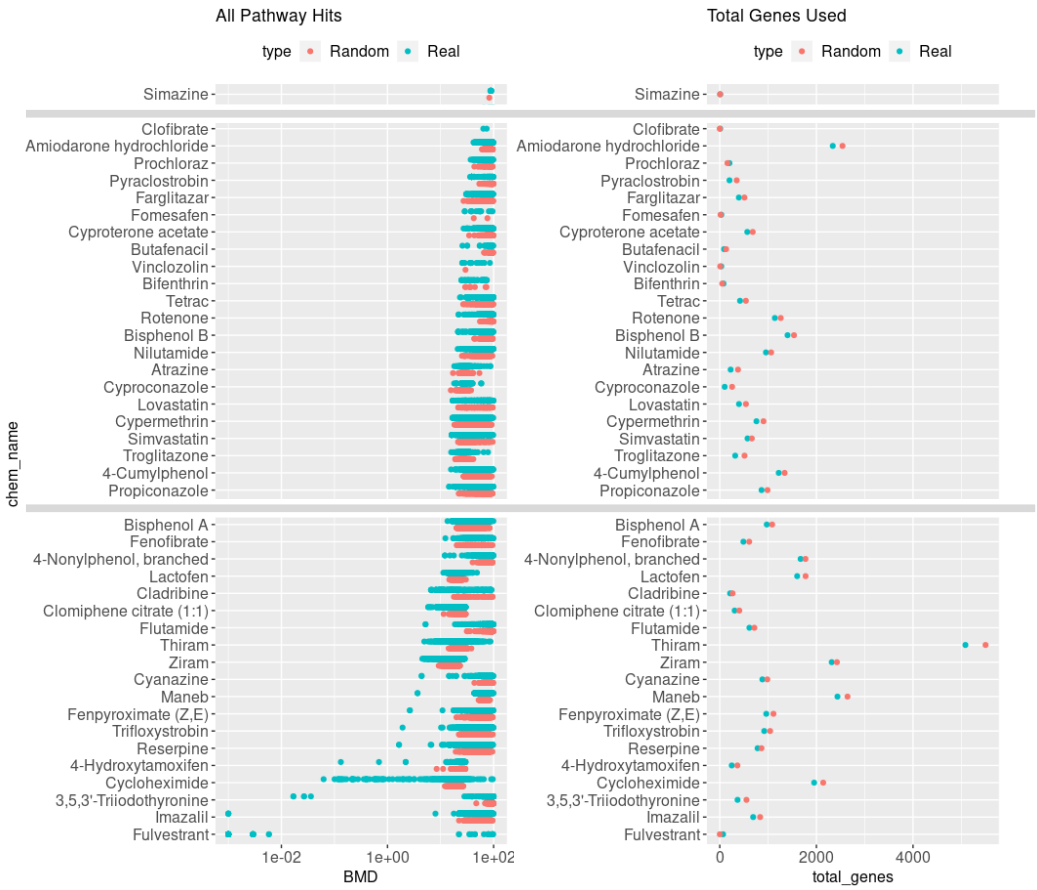


Green=ER
Black=random
Red=cell stress

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 Richard's 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

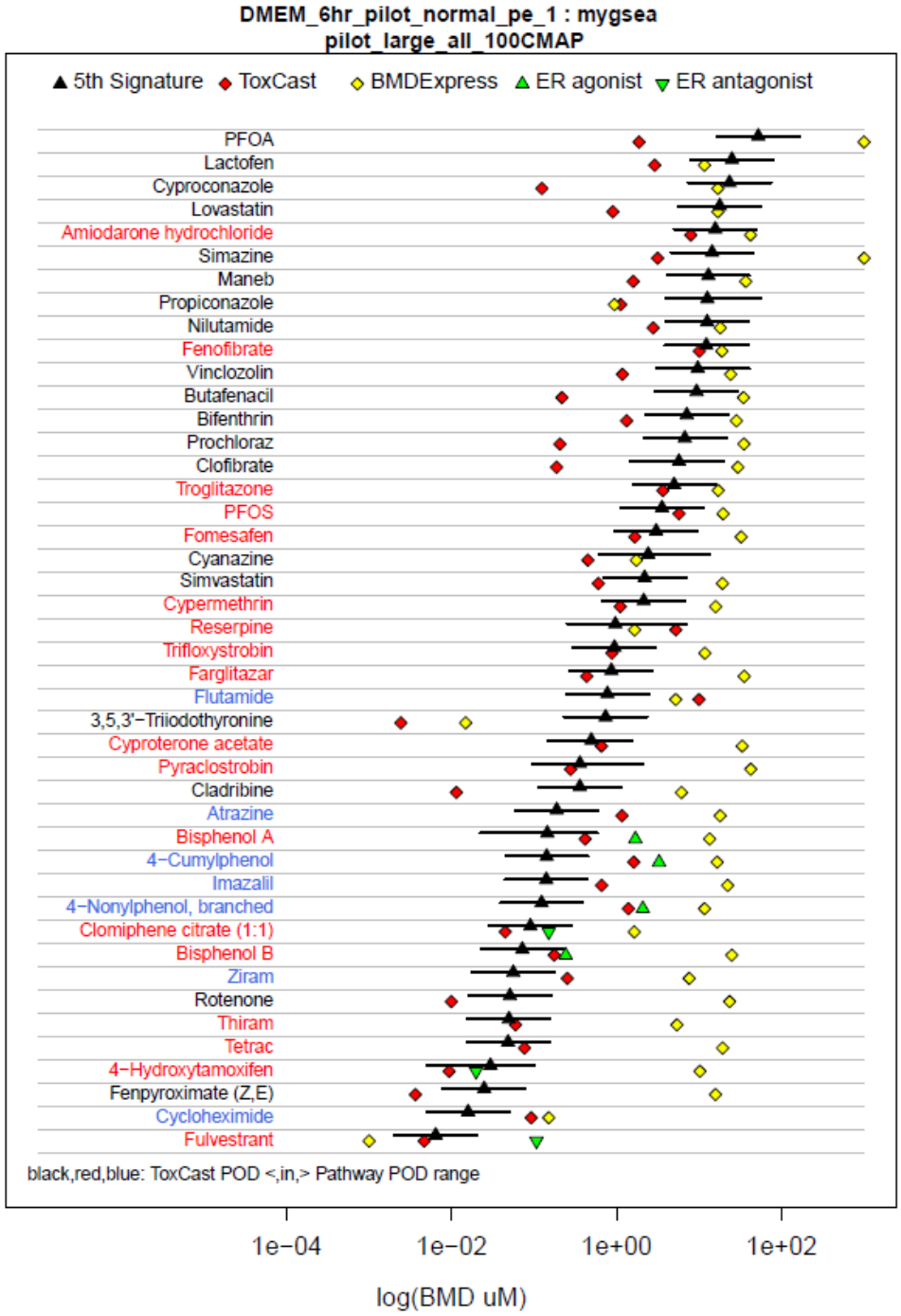
Target Pathway / Gene Set Ranks

Chemical	Target	DESeq2 first on-target pathway	Rank	BMDExpress first on-target pathway	Rank
4-Hydroxytamoxifen	ER	BHAT_ESR1_TARGETS_NOT_VIA_AKT1_UP	1	WILLIAMS_ESR1_TARGETS_UP	1
Clomiphene citrate (1:1)	ER	HALLMARK_ESTROGEN_RESPONSE_EARLY	1	CMAP_up_MCF7-estradiol-6h-1.46e-05M_2017	2
Fulvestrant	ER	CMAP_up_MCF7-fulvestrant-6h-1e-06M_6366	1	CMAP_up_MCF7-estradiol-6h-1e-07M_259	2
Propiconazole	CYPs	CMAP_dn_MCF7-miconazole-6h-9.6e-06M_1095	2	HALLMARK_XENOBIOTIC_METABOLISM	141
Bisphenol B	ER	CMAP_dn_MCF7-tamoxifen-6h-7e-06M_7805	3	CMAP_up_MCF7-fulvestrant-6h-1e-06M_265	23
4-Nonylphenol, branched	ER	DUTERTRE_ESTRADIOL_RESPONSE_24HR_UP	9	YANG_BREAST_CANCER_ESR1_BULK_DN	15
3,5,3'-Triiodothyronine	thyroid	RODRIGUES_THYROID_CARCINOMA_ANAPLASTIC_DN	13	RODRIGUES_THYROID_CARCINOMA_DN	564
Cyproterone acetate	AR	CMAP_up_MCF7-flutamide-6h-1.44e-05M_3991	16	CMAP_up_MCF7-testosterone-6h-1.16e-05M_5596	79
Nilutamide	AR	CMAP_dn_MCF7-flutamide-6h-1.44e-05M_4466	18	CMAP_dn_MCF7-flutamide-6h-1.44e-05M_3991	73
Cypermethrin	ion channel	CMAP_dn_MCF7-amiodarone-6h-5.8e-06M_2902	23	CMAP_dn_MCF7-amiodarone-6h-5.8e-06M_6526	57
Cyproconazole	CYPs	CMAP_up_MCF7-sertaconazole-6h-8e-06M_7935	24	CMAP_dn_MCF7-terconazole-6h-7.6e-06M_2336	25
4-Cumylphenol	ER	CMAP_up_MCF7-estradiol-6h-1e-08M_8238	25	CMAP_up_MCF7-estradiol-6h-1e-08M_874	17
Flutamide	AR	CMAP_up_MCF7-testosterone-6h-1.16e-05M_6582	28	CMAP_up_MCF7-testosterone-6h-1.16e-05M_5597	47
Bifenthrin	ion channel	CMAP_up_MCF7-amiodarone-6h-5.8e-06M_5547	40		
Prochloraz	CYPs	CMAP_up_MCF7-isoconazole-6h-9.6e-06M_8693	44	CMAP_up_MCF7-sertaconazole-6h-8e-06M_3613	161
Bisphenol A	ER	CMAP_dn_MCF7-fulvestrant-6h-1e-06M_8532	45	MASSARWEH_RESPONSE_TO_ESTRADIOL	30
Imazalil	CYPs	CMAP_up_MCF7-terconazole-6h-7.6e-06M_4598	48	CMAP_dn_MCF7-miconazole-6h-9.6e-06M_5175	127
Lovastatin	cholesterol	CMAP_dn_MCF7-bezafibrate-6h-1.1e-05M_1948	51	CMAP_up_MCF7-lovastatin-6h-9.8e-06M_5233	5
Vinclozolin	AR	CMAP_dn_MCF7-testosterone-6h-1.16e-05M_5597	57	CMAP_up_MCF7-testosterone-6h-1.16e-05M_6580	11
Amiodarone hydrochloride	ion channel	CMAP_up_MCF7-amiodarone-6h-5.8e-06M_2902	63	CMAP_up_MCF7-amiodarone-6h-5.8e-06M_6526	49
Troglitazone	PPAR	Peroxisomal lipid metabolism	90		
Clofibrate	PPAR	KEGG_PEROXISOME	99		
Simvastatin	cholesterol	CMAP_dn_MCF7-rosiglitazone-6h-1e-05M_6451	116	HALLMARK_CHOLESTEROL_HOMEOSTASIS	37
Tetrac	thyroid	KEGG_THYROID_CANCER	194	LUI_THYROID_CANCER_CLUSTER_1	651
Trifloxystrobin	mitochondria	MOOTHA_MITOCHONDRIA	199	MOOTHA_MITOCHONDRIA	987
Fenpyroximate (Z,E)	mitochondria	MOOTHA_MITOCHONDRIA	267	Mitochondrial protein import	1890
Rotenone	mitochondria	MOOTHA_MITOCHONDRIA	372	Mitochondrial fatty acid beta-oxidation	24
Fenofibrate	PPAR	REACTOME_PPARA_ACTIVATES_GENE_EXPRESSION	440	REACTOME_PPARA_ACTIVATES_GENE_EXPRESSION	373
PFOS	PPAR	KEGG_PEROXISOME	513	Peroxisome	27
PFOA	PPAR	Peroxisomal lipid metabolism	583		
Farglitazar	PPAR	REACTOME_PPARA_ACTIVATES_GENE_EXPRESSION	1006	Peroxisome	883
Pyraclostrobin	mitochondria	MOOTHA_MITOCHONDRIA	1241	Mitochondrial fatty acid beta-oxidation	250

United States Environmental Protection Agency

Chemical-wise PODs

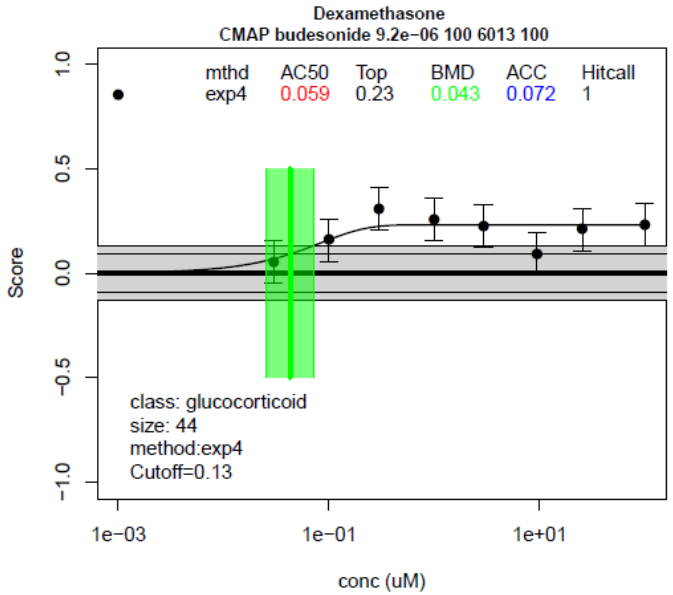
Black: lowest 5%-ile signature
Red: ToxCast 5% POD
Yellow: BMD Express
Green: ToxCast ER Model

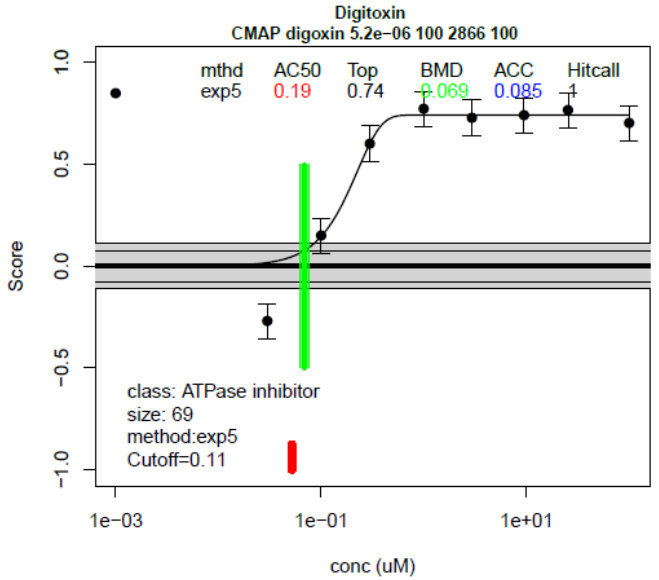


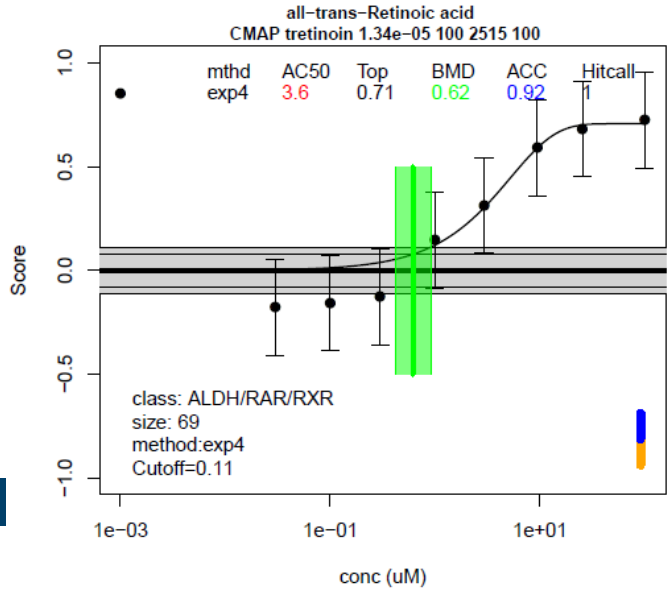
MCF7 Screen, Preliminary Observations

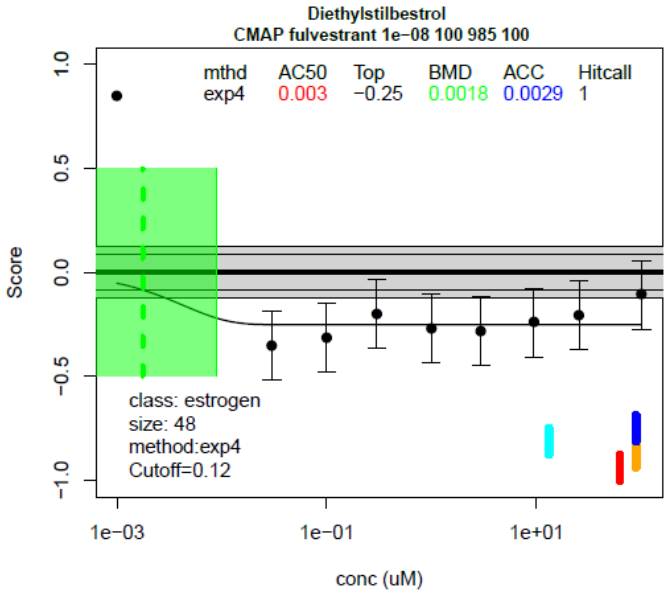
- 2112 Chemicals
- Drugs, food chemicals, pesticides, industrial chemicals
- 8-point concentration-response
- 6 hour exposure
- 22,343 signatures
- 355 chemicals have gene target annotations
  - Used to assess how well the active signatures match the chemical target

More activity that just Estrogen Receptor

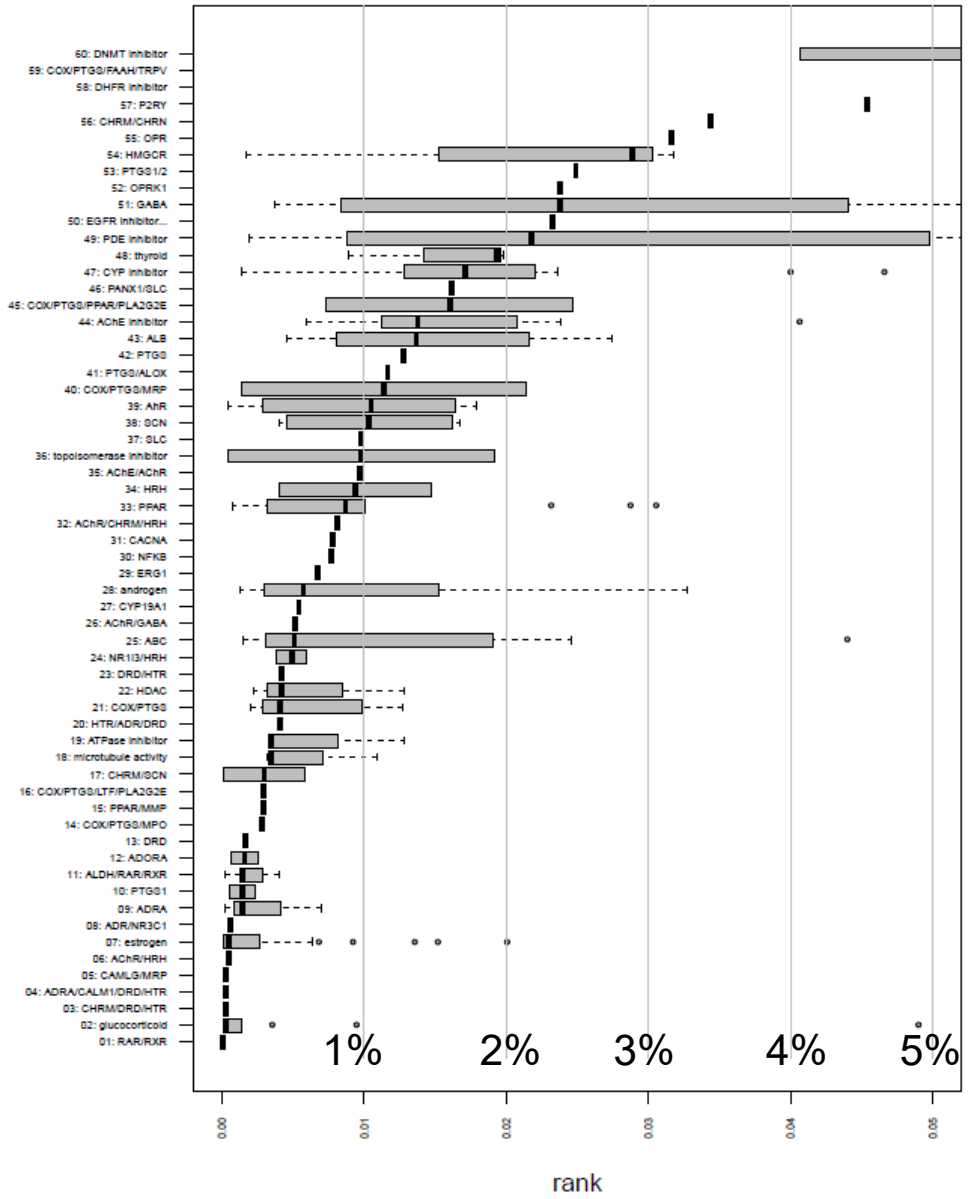








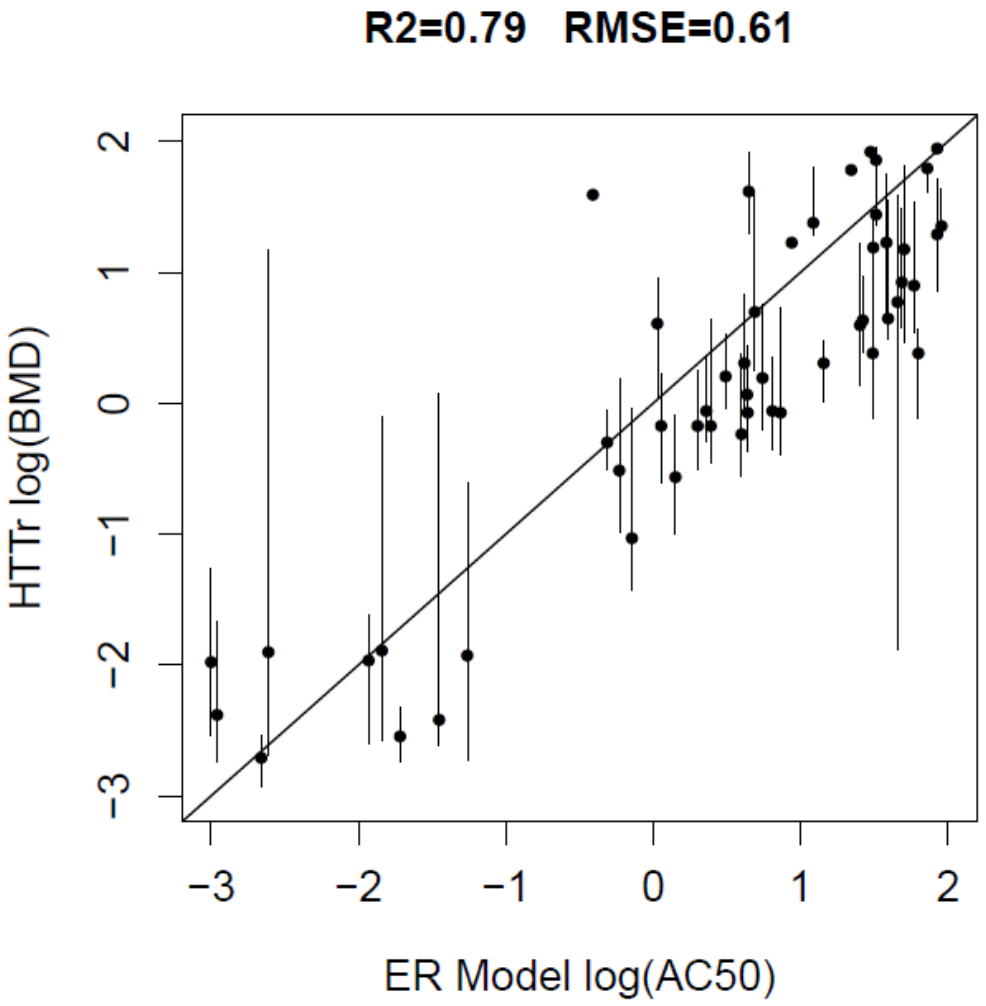
Measuring how well the signatures ID the chemical target



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

Lowest set are all GPCR or nuclear receptor target families

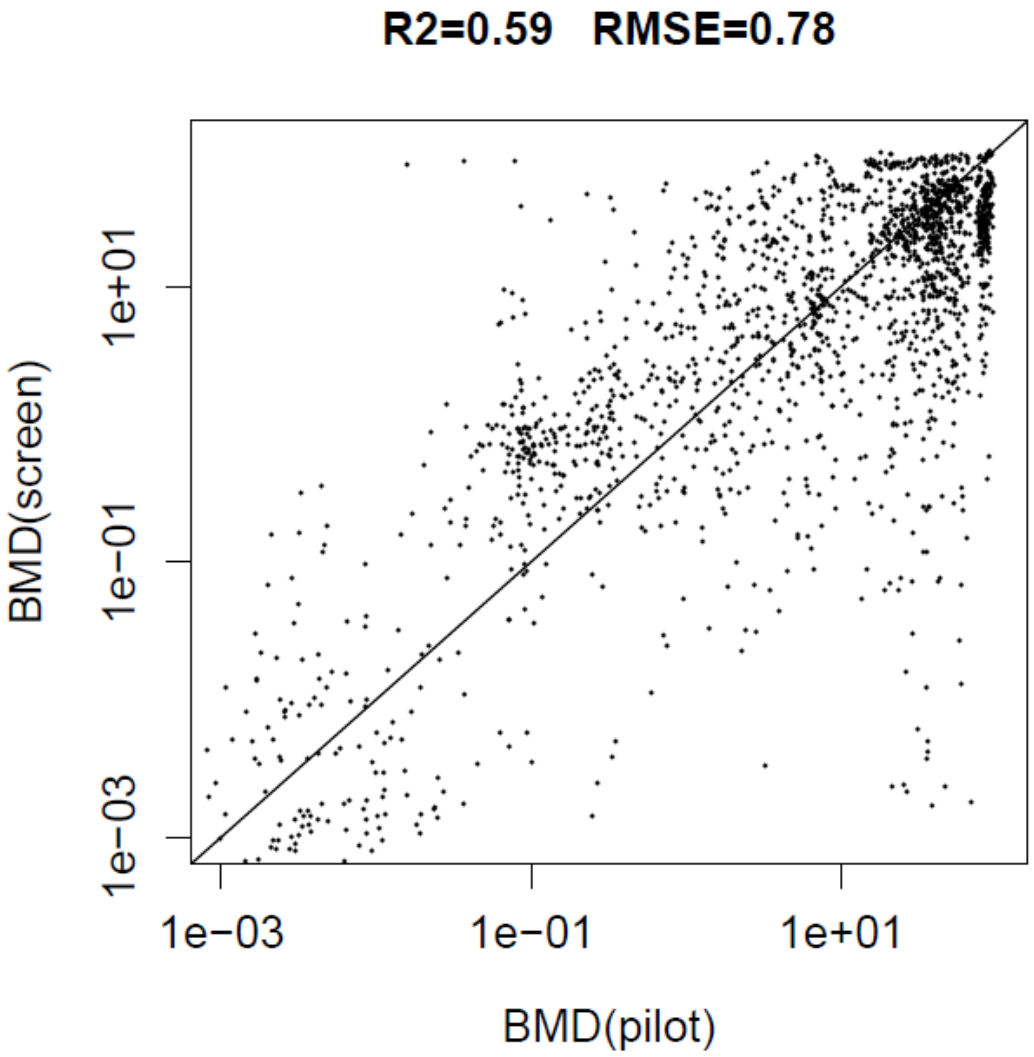
How do potencies compare with other in vitro assays?



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
- Clinton Willis
- Richard Brockway
- Johanna Nyffeler
- Megan Culbreth
- Dan Hallinger
- Terri Fairley
- Matt Martin
- Agnes Karmaus