United States
Environmental Protection
Agency

High-Throughput Transcriptomics (HTTr): Pipeline Updates and Concentration-Response Modeling

EU-ToxRisk Seminar

Dec 18, 2019

Imran Shah

Center for Computational Toxicology & Exposure

The views expressed in this presentation are those of the author[s] and do not necessarily reflect the views or policies of the U.S. 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

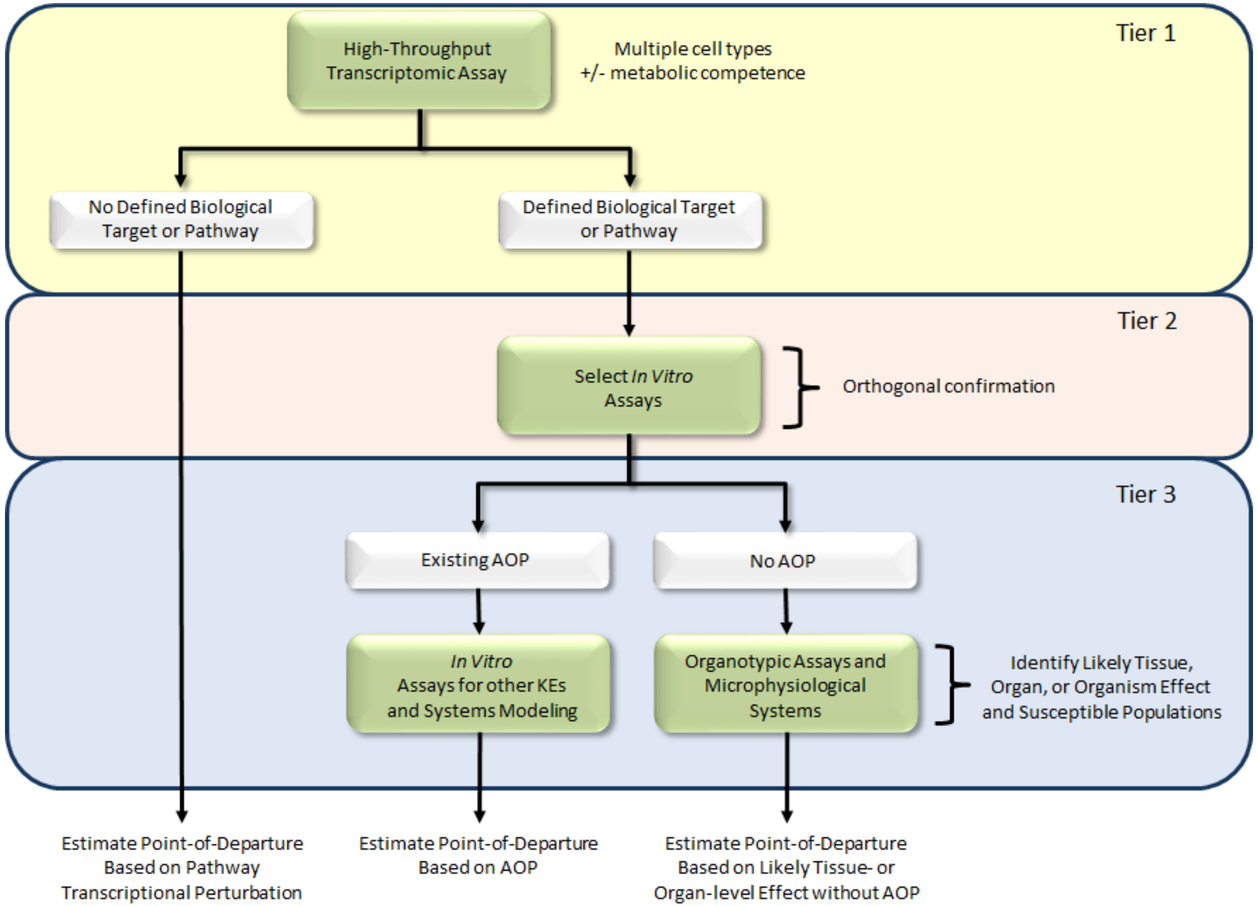
Objectives

- 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

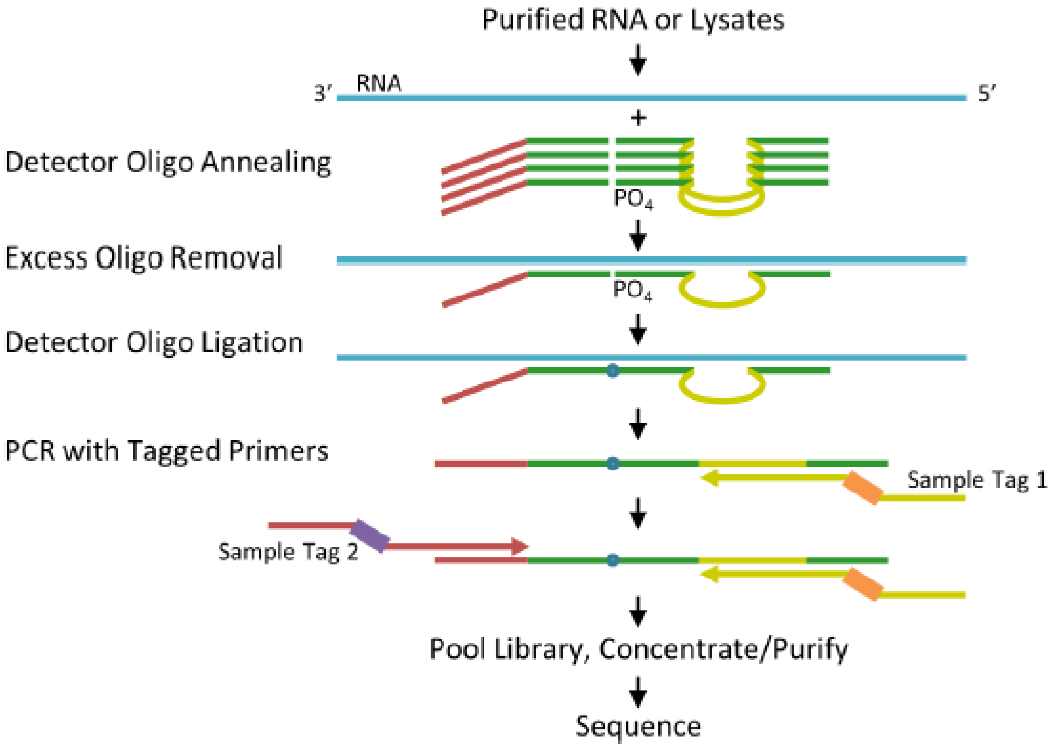
A strategic vision and operational road map for computational toxicology at the U.S. Environmental Protection Agency [DRAFT]


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

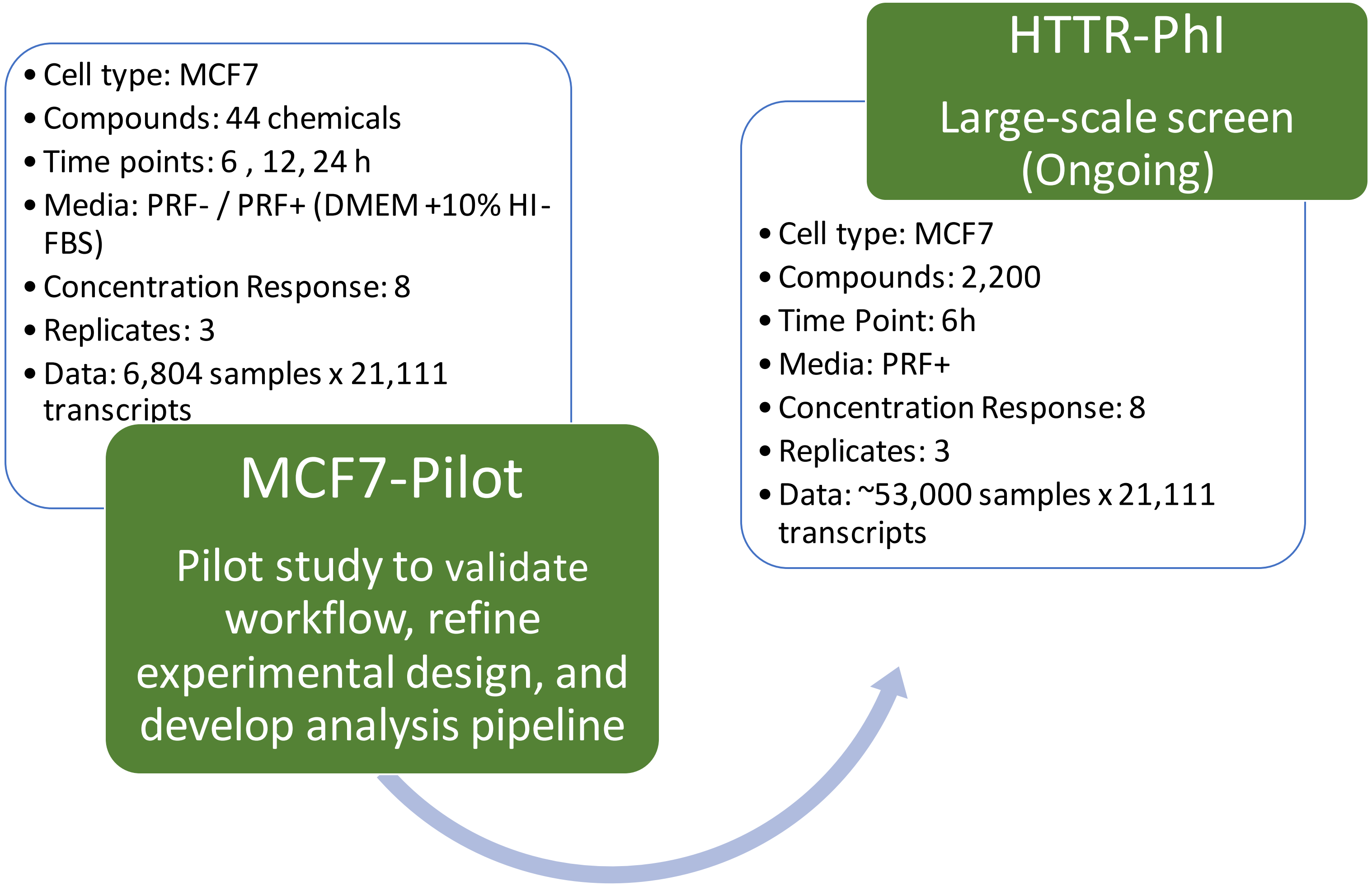
TempO-Seq for HTTr

- 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)



- 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

MCF7-Pilot

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

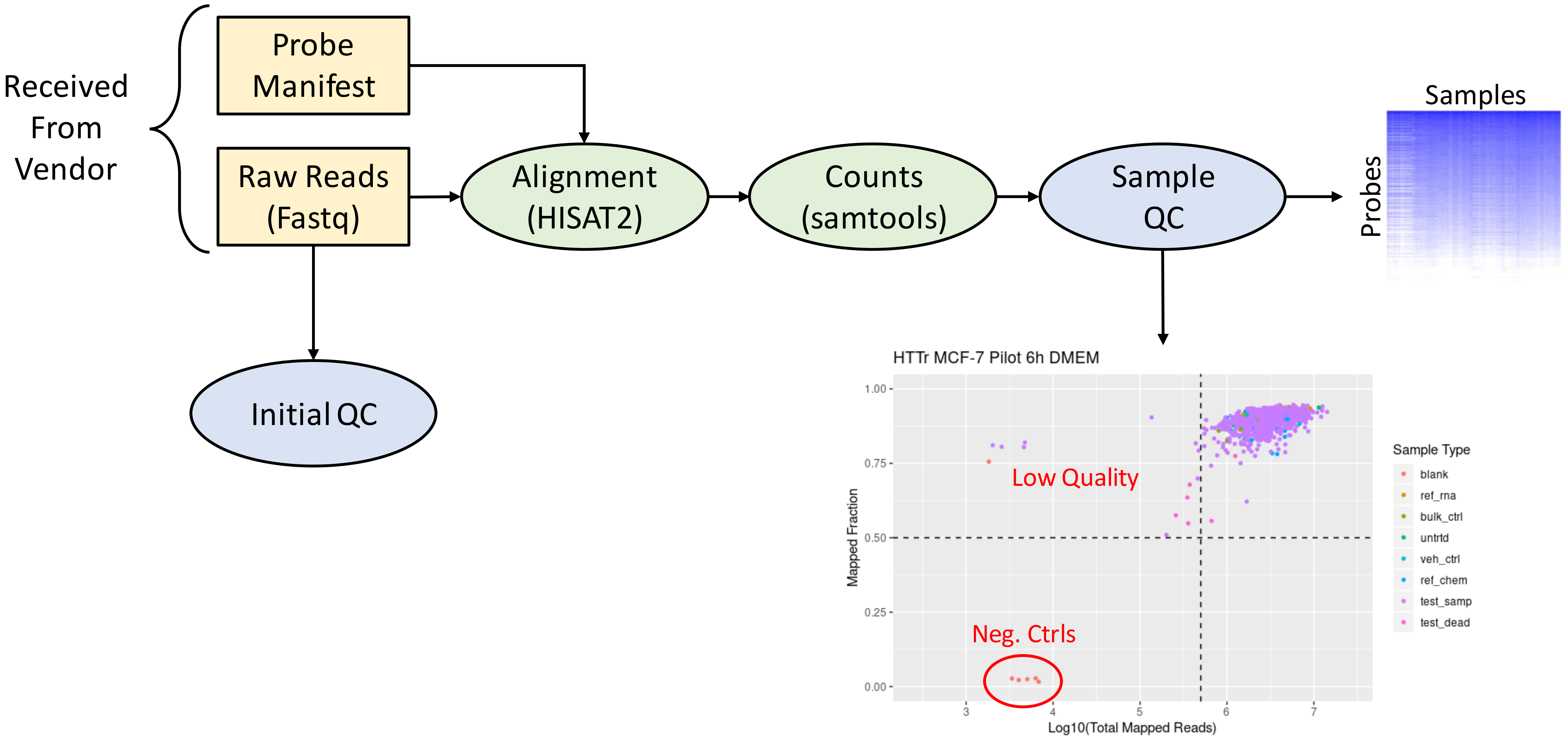
HTTR-PhI

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

HTTr Processing Pipeline

Pipeline: Raw Data Processing



Received From Vendor

Probe Manifest

Raw Reads (Fastq)

Alignment (HISAT2)

Counts (samtools)

Sample QC

Initial QC

Samples

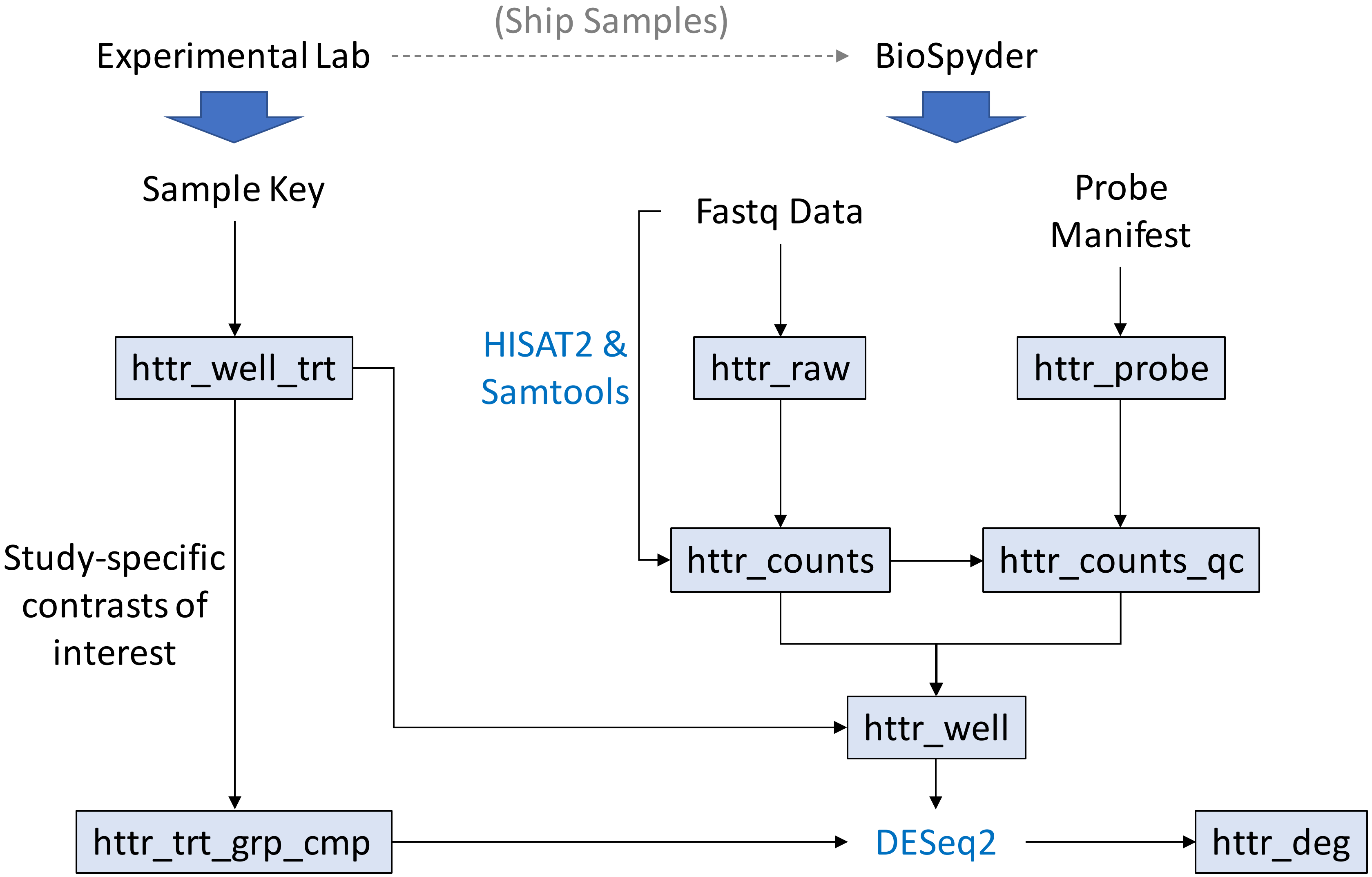
Probes

HTTr MCF-7 Pilot 6h DMEM

Low Quality

Neg. Ctrls

HTTr Data Management



(Ship Samples)

Experimental Lab

BioSpyder

Sample Key

Fastq Data

Probe Manifest

httr_well_trt

HISAT2 & Samtools

httr_raw

httr_probe

Study-specific contrasts of interest

httr_counts

httr_counts_qc

httr_well

httr_trt_grp_cmp

DESeq2

httr_deg

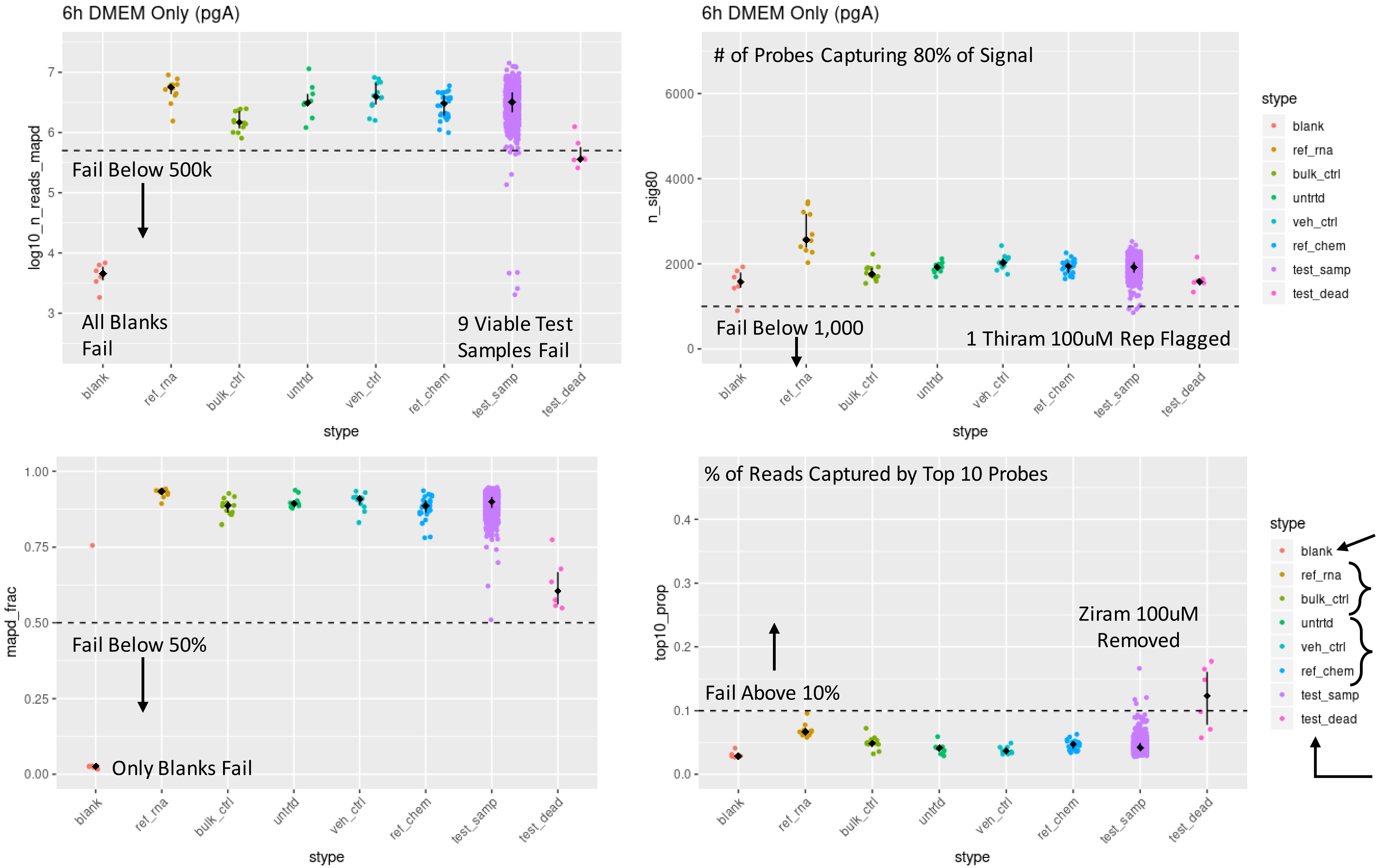
Scheduled backups
Recovery plan
Rapid export
Open-source tech

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 issues
  - 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

QC Metrics to Filter Samples



1,039 (98%) test samples pass all QC checks

Other QC Metrics:

- Ncov5 = Number of probes with at least 5 reads
- Gini Coefficient = Measure of inequality

Track with metrics shown

Lysis Buffer Only

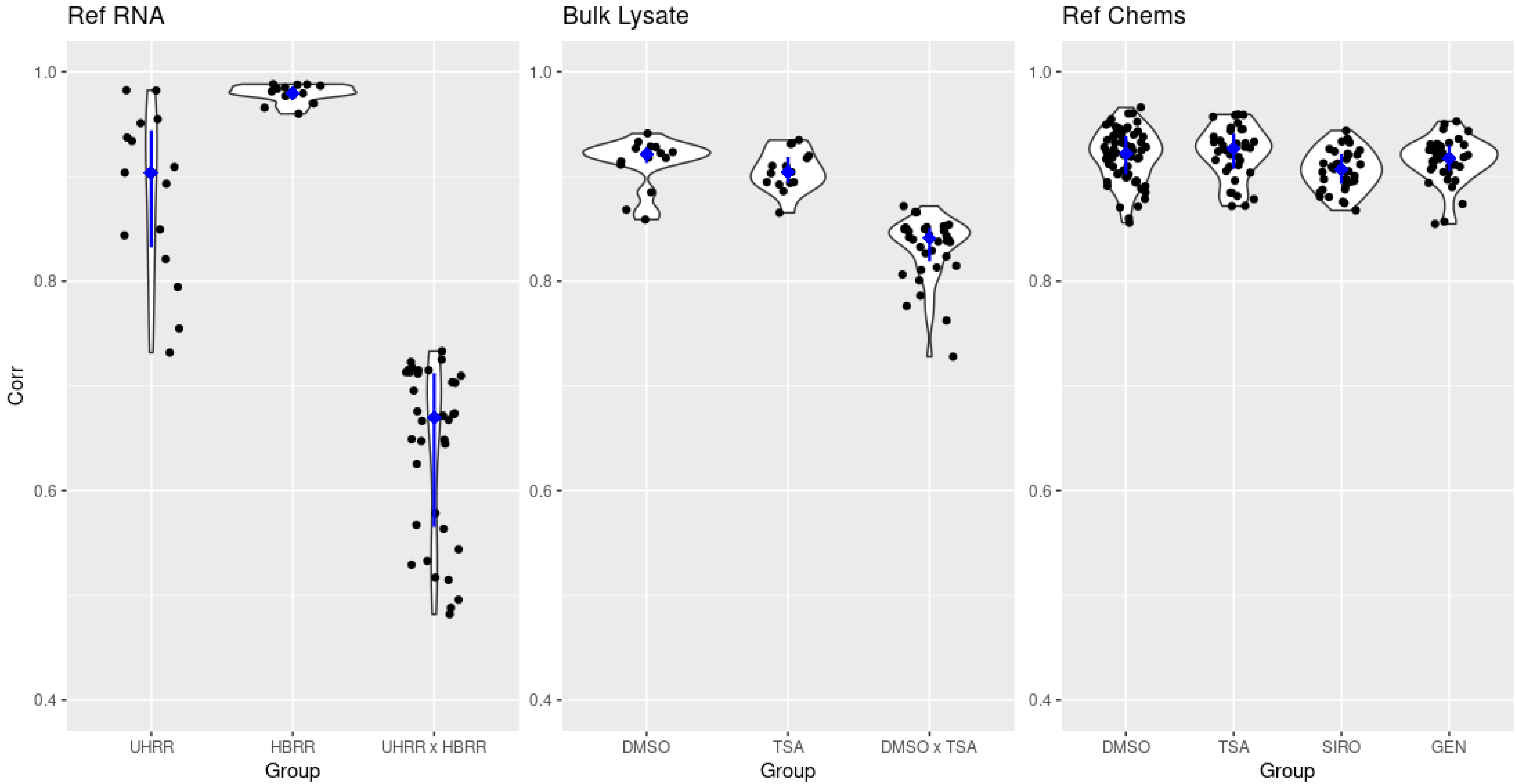
Standards

Plate Controls

Test Samples

>50% Cell Death on HCI Plates

Reproducibility: MCF7 Pilot DMEM 6h



Differential Gene Expression Analysis

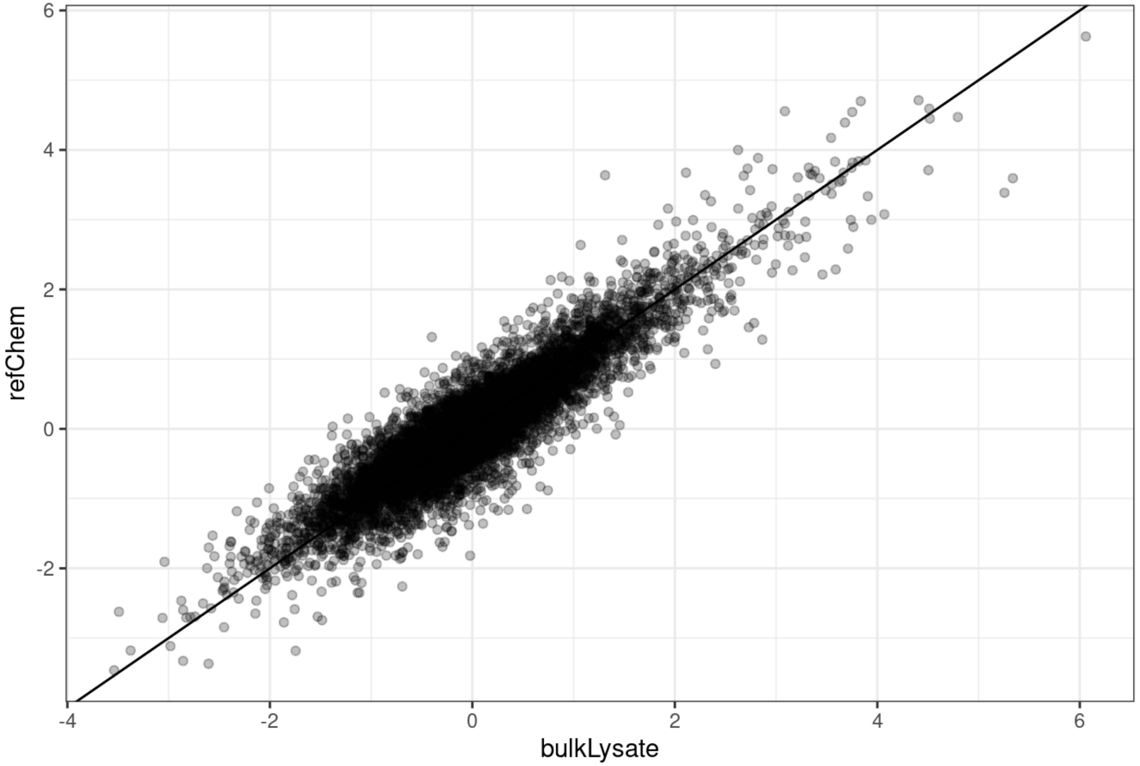
- Most recent version of DESeq2 (v??)
  - Evaluated questions about choice of plate effect and shrinkage using reference chemicals
  - Newer shrinkage methods (Ashr, Apeglm) results less reliable
- DEG analysis by four DESeq2 options:-
  1. Plate effect - , Shrinkage -
  2. Plate effect - , Shrinkage +
  3. Plate effect + , Shrinkage -
  4. Plate effect + , Shrinkage + (Recommended)

Reproducibility: MCF7 Pilot DMEM 6h

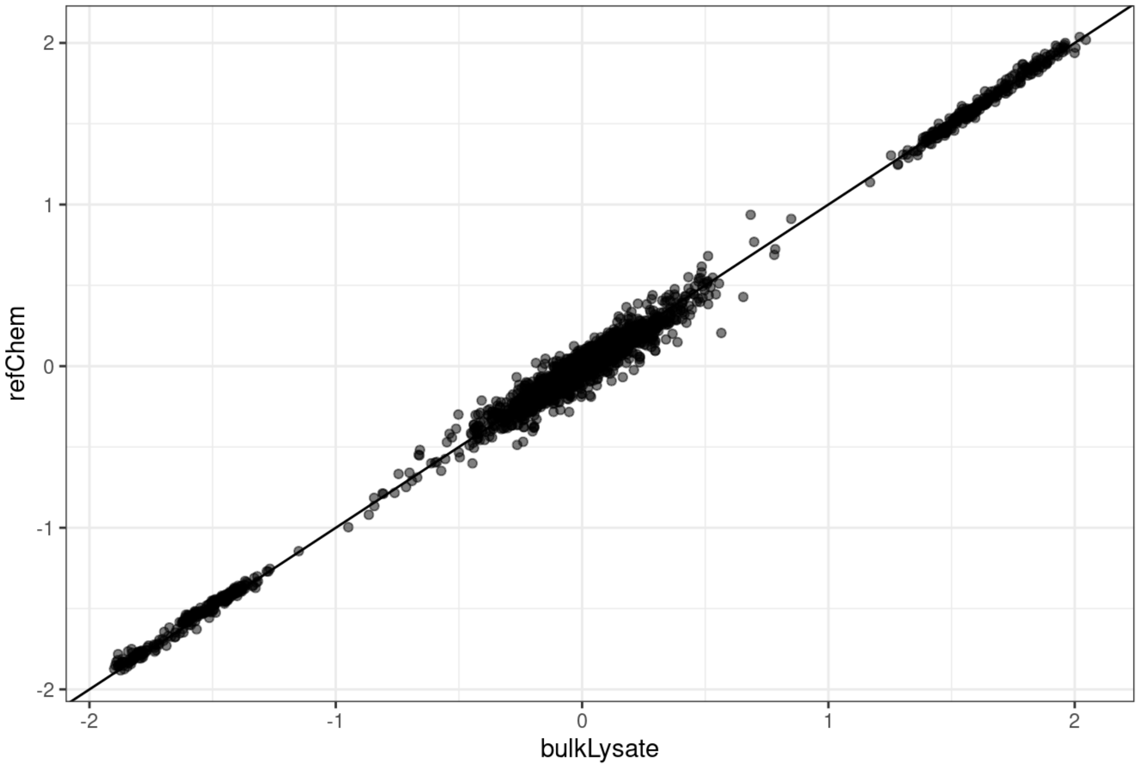
- TSA Treatment Effect: Bulk Lysate Control vs Plated Reference

Related Complete Slide Deck

DESeq2 log2 FC (plate effect+, shrinkage+)



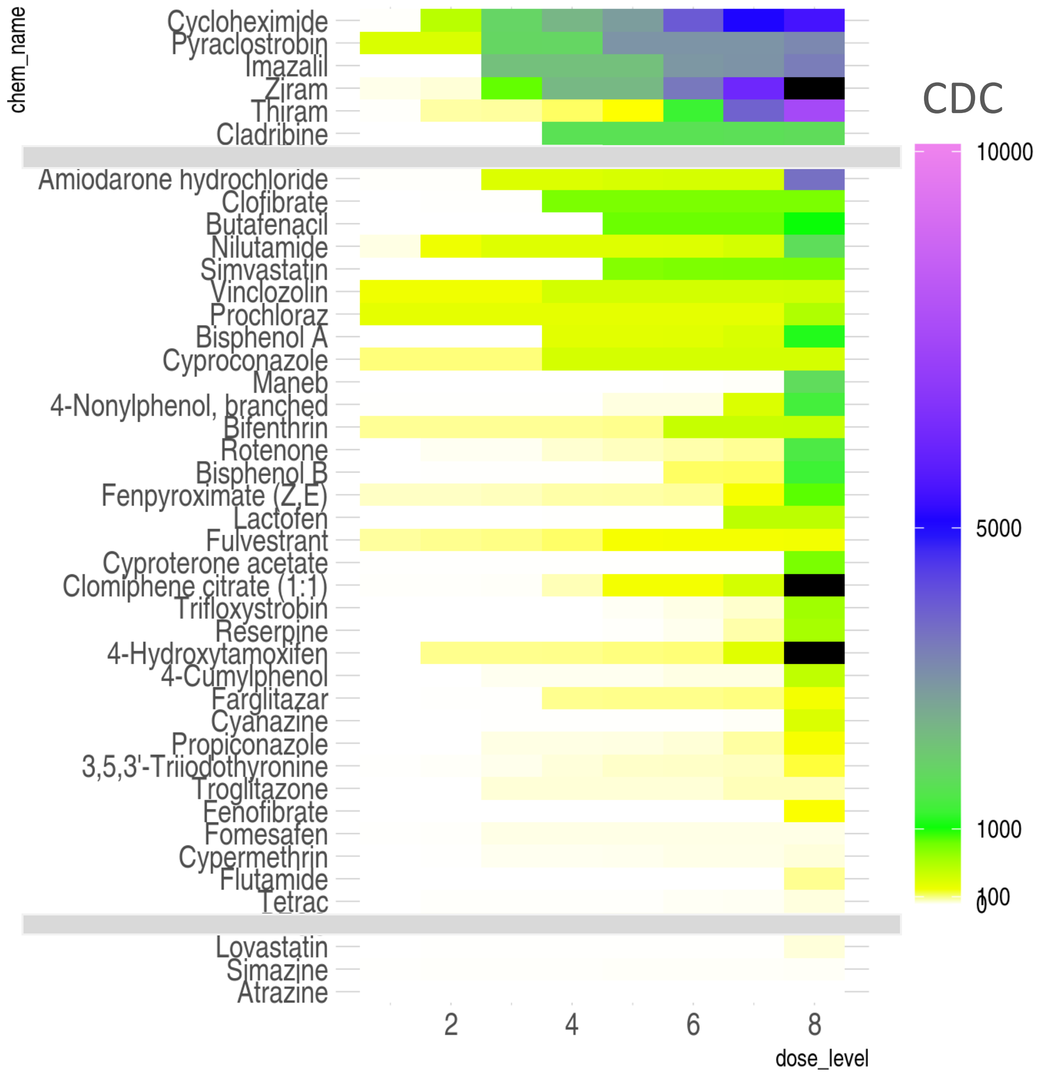
Pathway Scores (plate effect+, shrinkage+)



MCF7 Pilot DMEM 6h DEGs

- Summarize DEGs for all chemicals & concentrations
- Propose DEG Metric = sum(probes w/ DESeq2 q value < 5% FDR)
- Cumulative DEG Count (CDC)

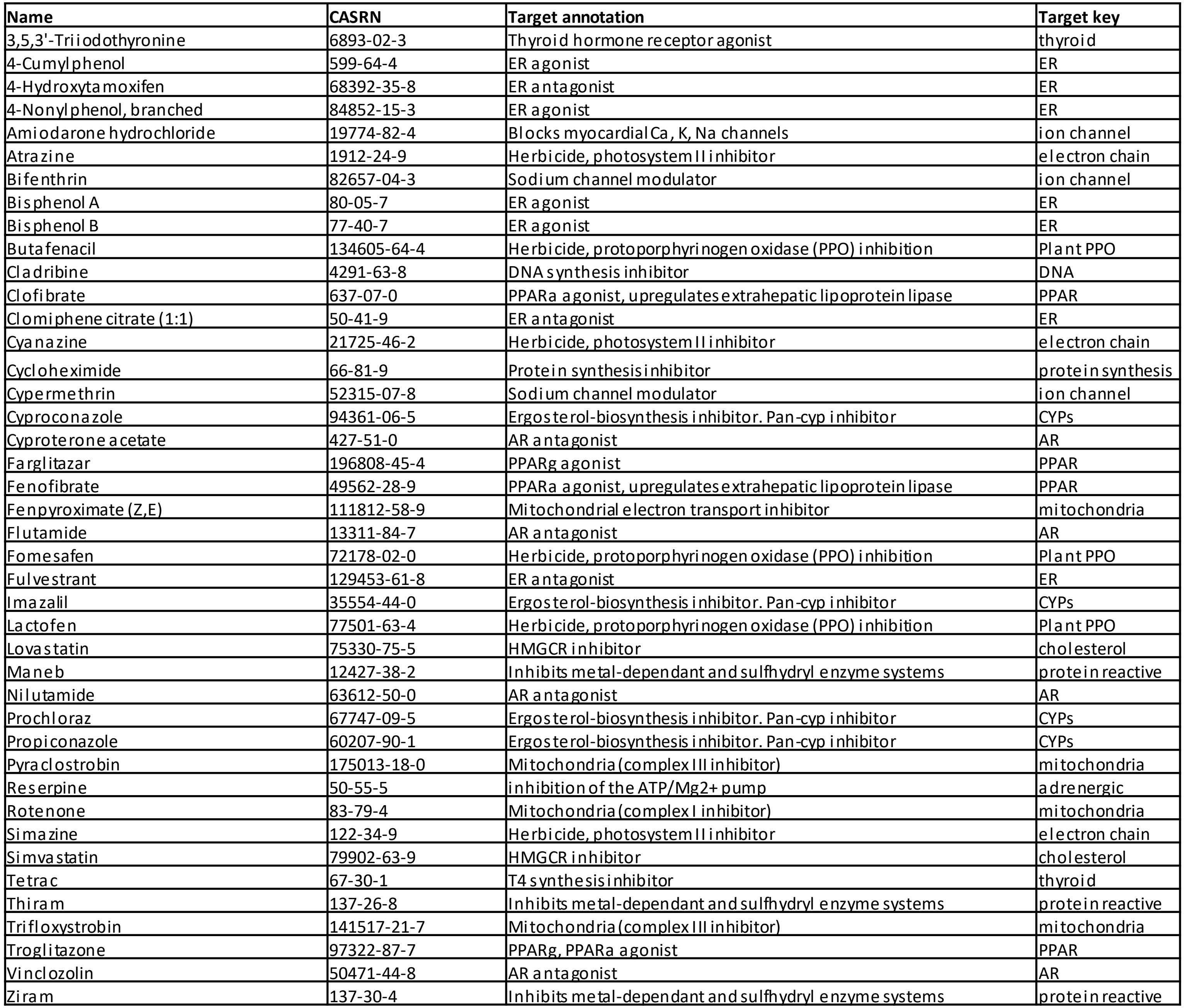
```
sum(unique(
          probes w/ DESeq2 q < 0.05
          in current dose,
          probes w/ DESeq2 q < 0.05
          in any lower dose
          ))
```



Putative Targets, Pathways & Potencies

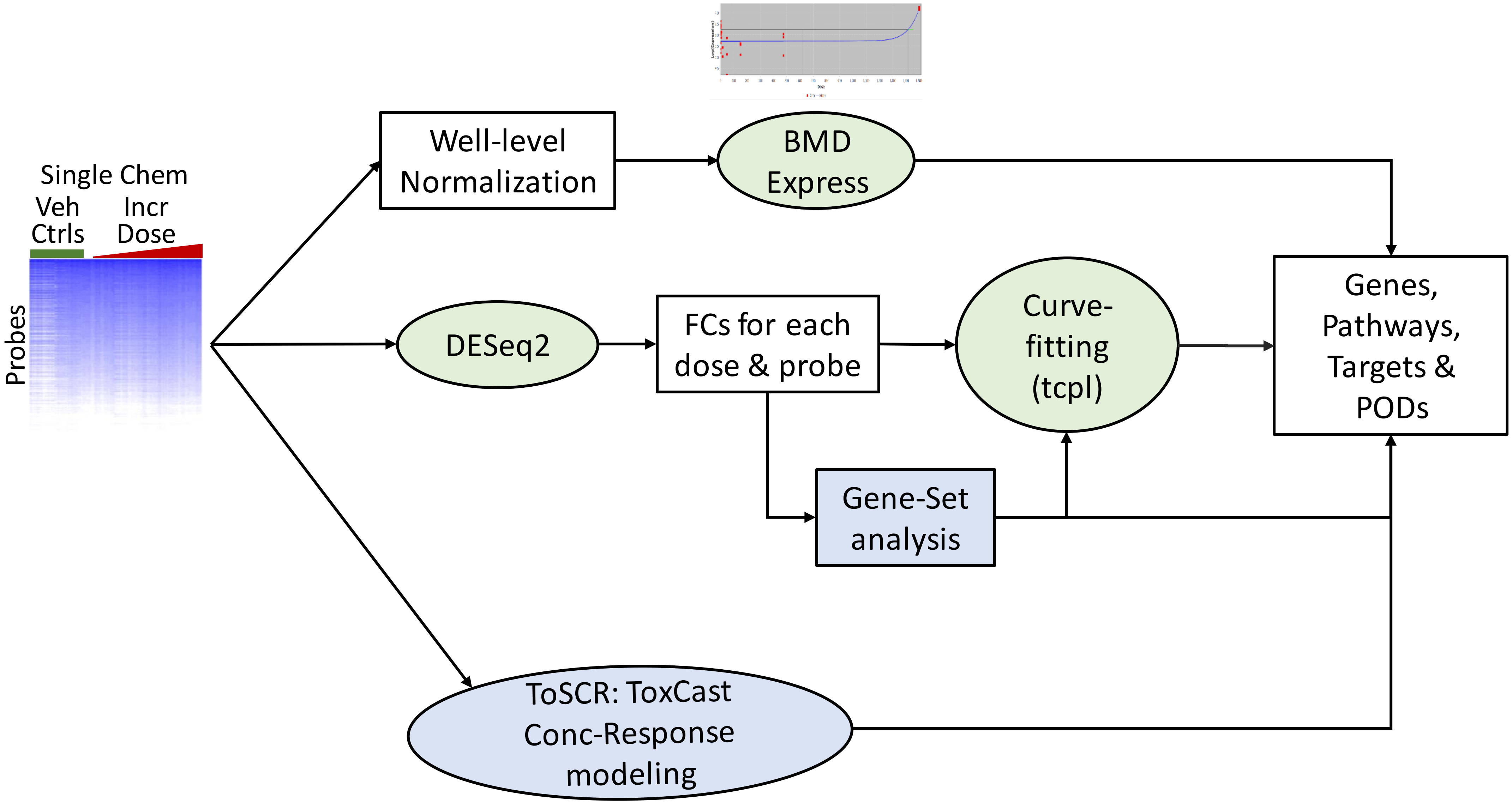
MCF7 Pilot:

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

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)

Gene-Set analysis

Genes, Pathways, Targets & PODs

ToSCR: ToxCast Conc-Response modeling

Gene Set Selection: Pathways and Treatments

Canonical Pathway gene sets

- Select 500 pathways from MSigDB and BioPlanet related to chemical targets
- **Randomly select** another 500 gene sets/pathways from MSigDB, BioPlanet
- Create CMap gene sets with chemicals in class of the 44 chemicals
- Add the ER-specific pathways
- Total canonical pathways = 2277

Random Gene sets

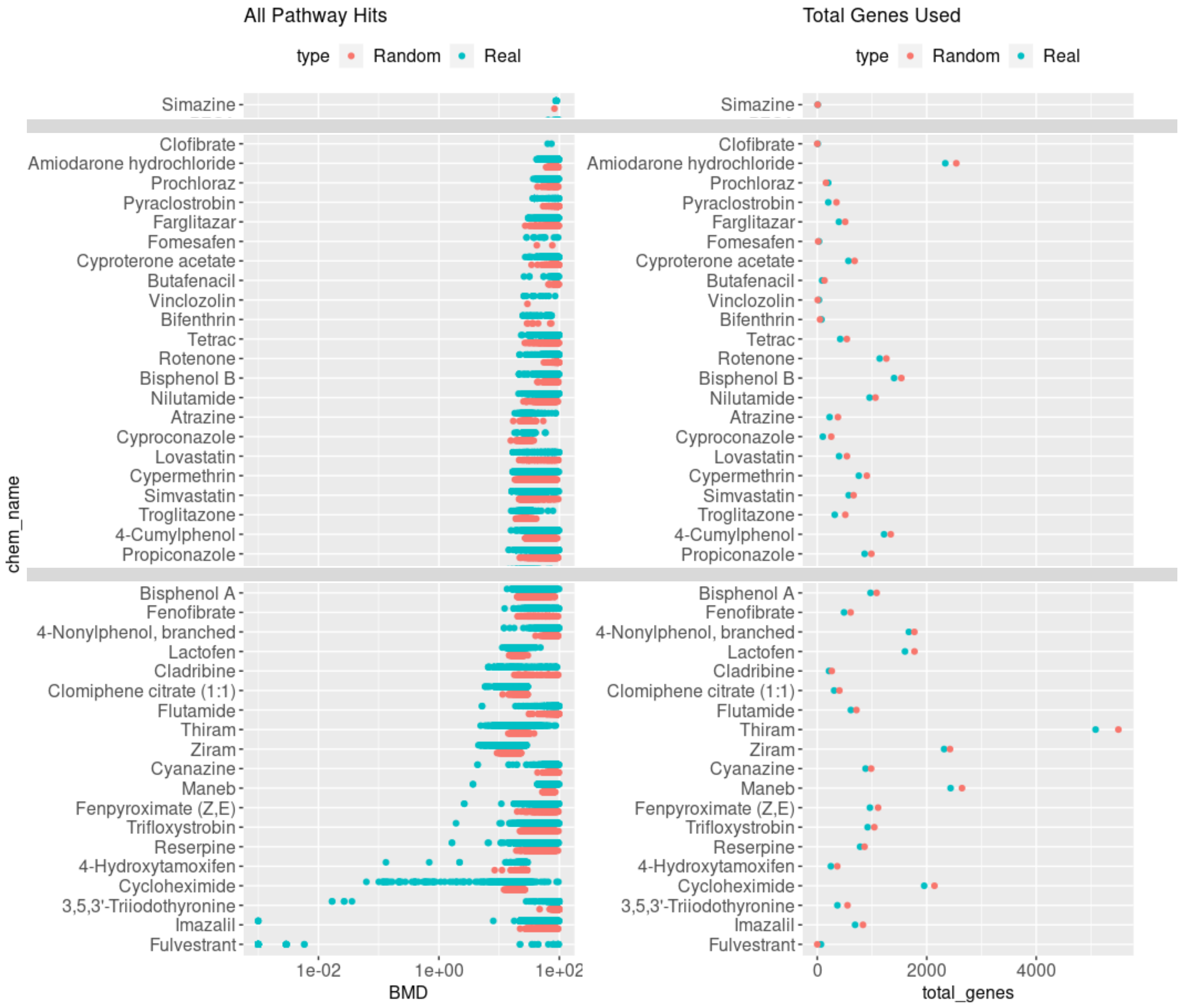
- For null distribution
- **Create** 500 random gene sets with mean 100, SD=40
- Total random pathways = 500

Total pathways = 2777

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



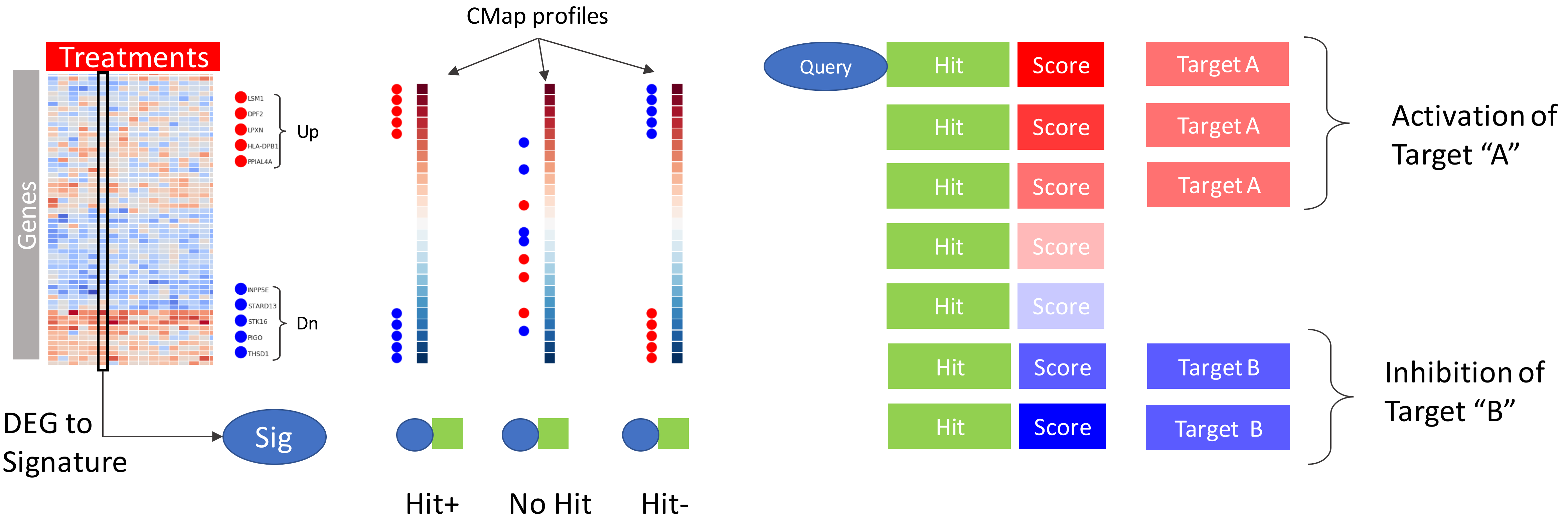
Putative Targets by Gene Set Connectivity

MCF7 Pilot DEG
(TempO-Seq)

Connectivity
(CMap v2, Affymetrix)

CMap Hits & Targets

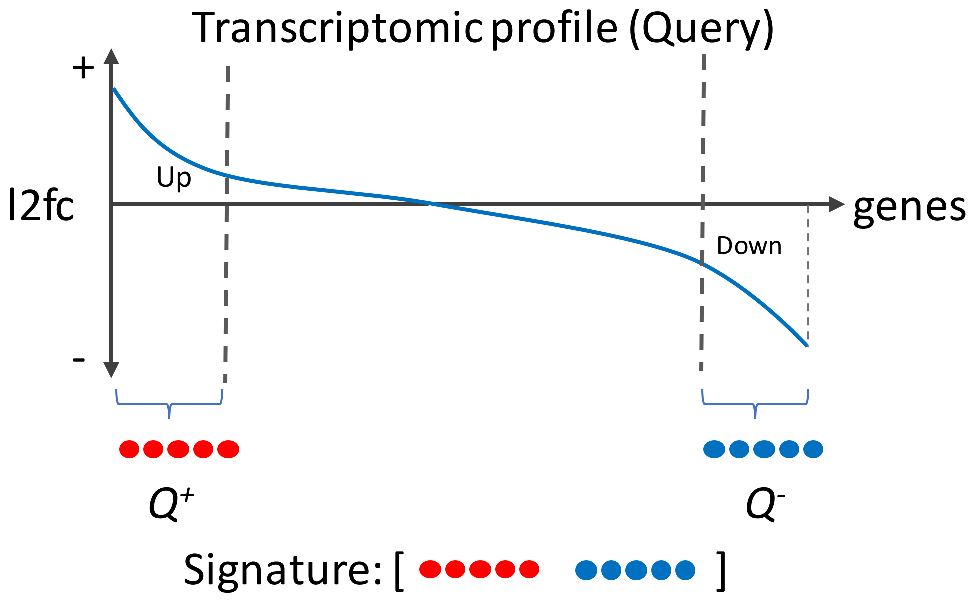
Putative Targets

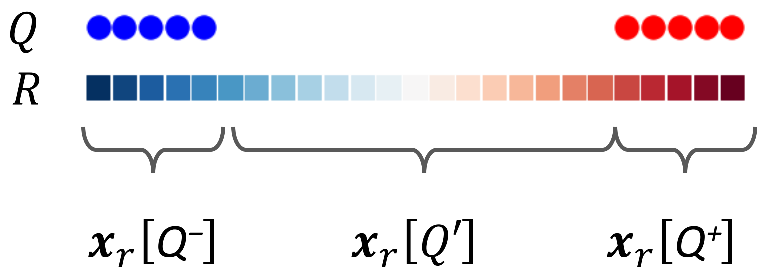


Cmap v2: MCF7: 1294 chems
Profiles for all cell types 6100

Connectivity Analysis

- A query signature *Q* containing *q* genes
  - $Q = \{g_1, g_2, .. g_j, .., g_q\}$
  - A directional signature (i.e. defined by Q^+ and Q^-)
- A query vector $\boldsymbol{x}_\mathrm{q}$ containing l2fc or Z-scores
- A reference transcriptomic profile $\boldsymbol{x}_r$ containing *m* genes (where *m*>*q*)
- A reference transcriptomic signature
  - $R = \{g_1, g_2, .. g_j, .., g_m\} = \{R^+, R^-\}$
- Genes not in the signature, $Q' = R - Q$
- The subset of the reference transcriptomic profile containing query genes $\boldsymbol{x}_r[Q]$ or not containing query genes $\boldsymbol{x}_r[Q']$



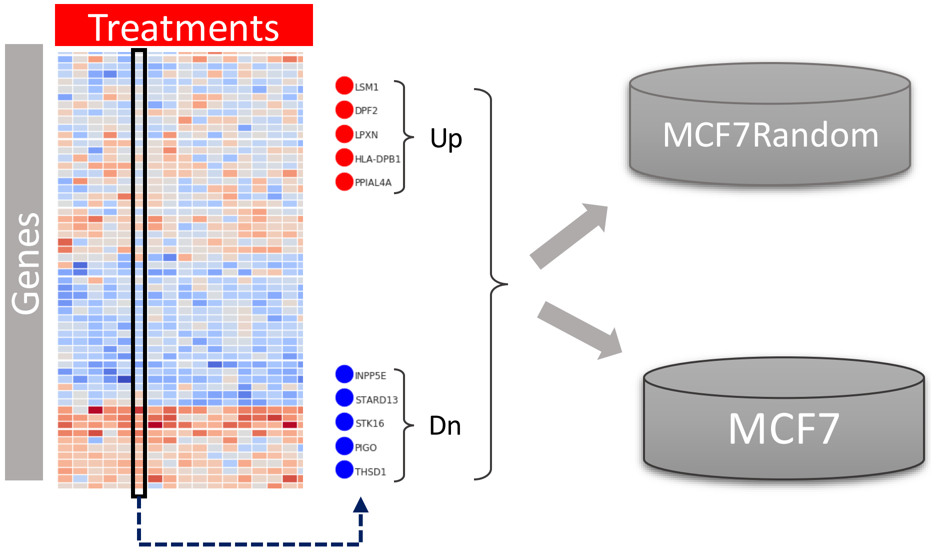


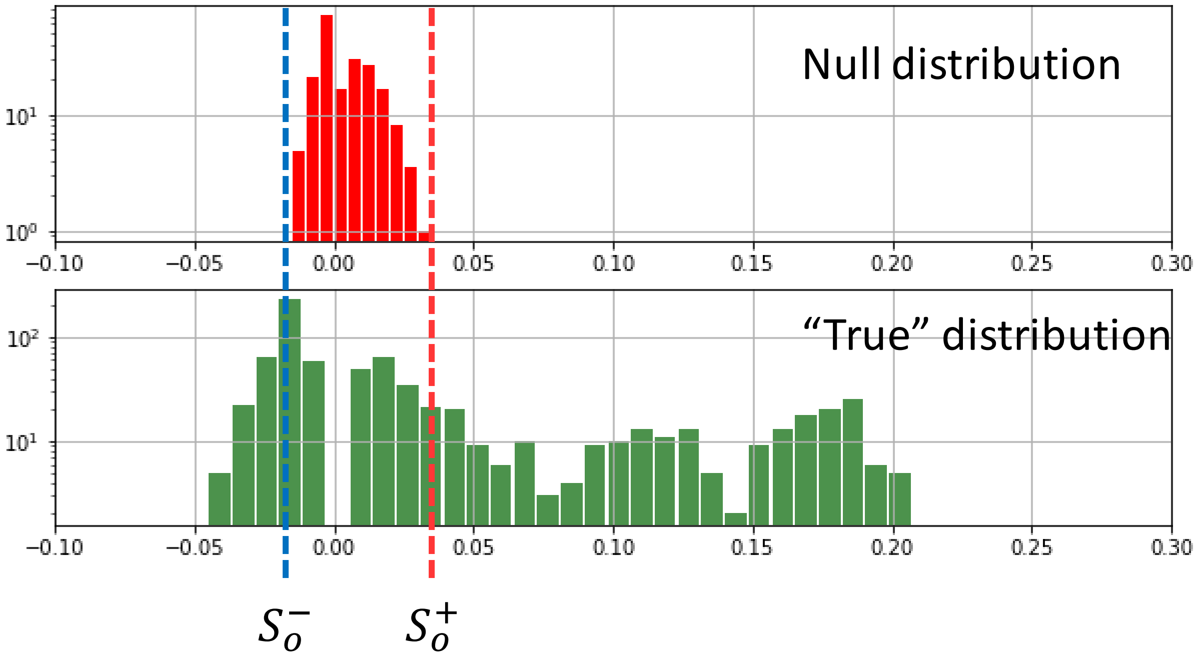
Shah *et al.* in prep

Evaluating Hit Significance Empirically

- Permute DEG matrix for MCF7 Pilot to create random gene expression profiles
- Column shuffle and generate *N* random profiles

- Search signatures against MCF7 Pilot and randomized MCF7 Pilot (to obtain null dist)
- Estimate significance for Up and Down hits separately





Shah *et al.* in prep

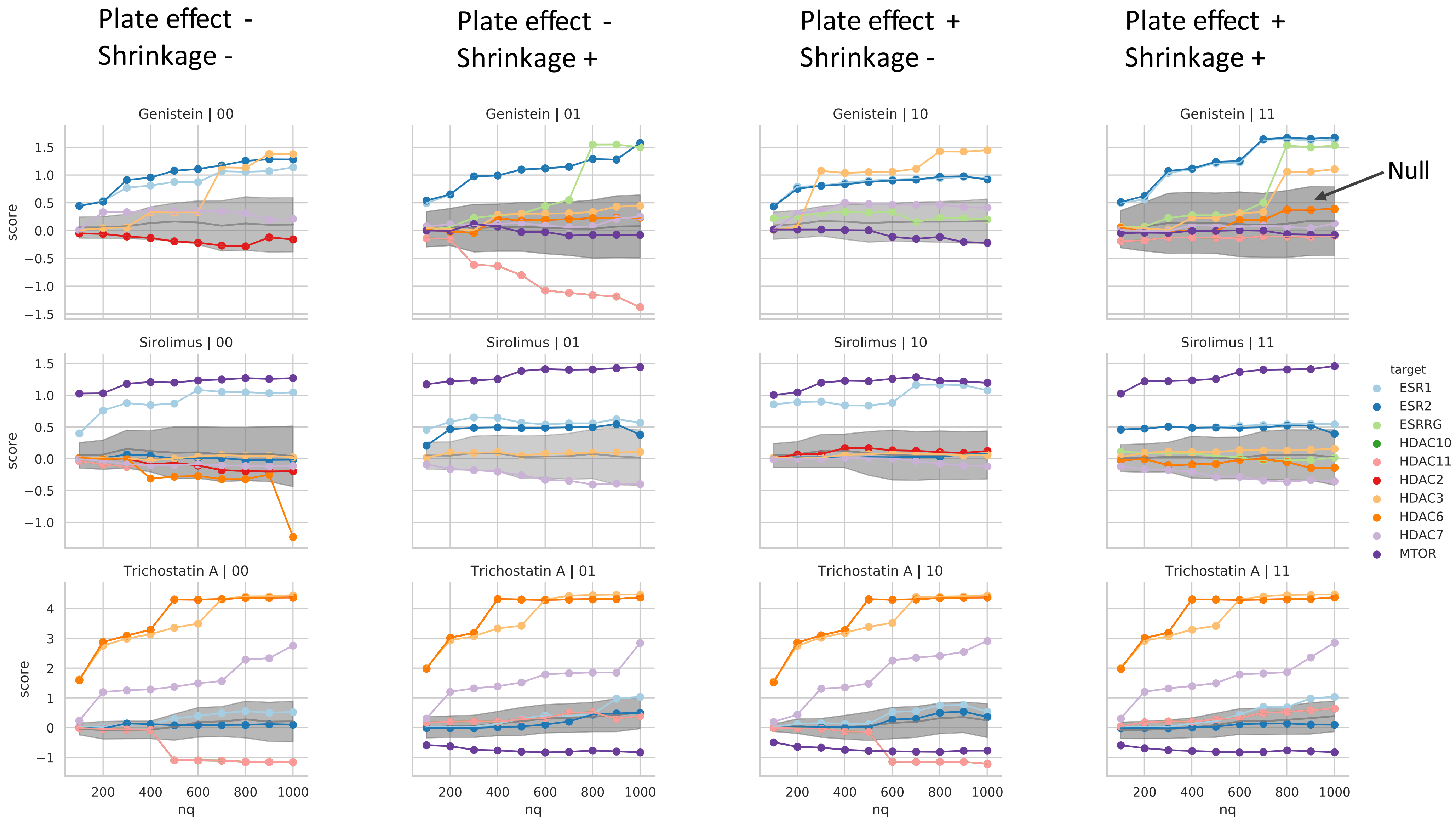
Gene Set Connectivity Scoring Methods

	Score	Method	Reference
➡	T-statistic	$ts = \frac{\overline{x_r[Q]} - \overline{x_r[Q']}}{\sqrt{\frac{\sigma_q^2}{q} + \frac{\sigma_{q'}^2}{q'}}}$; $\sigma_q^2 = \frac{1}{N}\sum_{i \epsilon Q}(x_{ri} - \overline{\boldsymbol{x}_r[Q]})^2$, $\sigma_{q'}^2 = \frac{1}{N}\sum_{i \epsilon Q'}(x_{ri} - \overline{\boldsymbol{x}_r[Q']})^2$	Tian et al. 2005; Goeman et al. 2004, 2005
	Ranksum statistic	$rs = \min\left(qq' + \frac{q(q+1)}{2} - \sum y_r, qq' + \frac{q'(q'+1)}{2} - \sum y'_r\right); y = rank(x)$	Barry, Nobel, and Wright 2005; Gower, Spira, and Lenburg 2011
	Gene Set Enrichment analysis (GSEA)	$\mathrm{ES} = max_{1 \le j \le m}(S_i - S_i'); \ S_i = \sum_{\substack{i \in Q \\ j \le i}} \frac{\lvert x_j \rvert^b}{\sum_{i \in Q} \lvert x_i \rvert^{b'}}, S'_i = \sum_{\substack{i \in Q' \\ j \le i}} \frac{\lvert x_j \rvert^b}{\sum_{i \in Q} \lvert x_i \rvert^b}$	Mootha et al. 2003; Subramanian et al. 2005
➡	Total enrichment score (TES)	$TES = 1 - \frac{ES^+ - ES^-}{2}$	Iorio, Tagliaferri, and Bernardo 2009
➡	eXtreme Pearson correlation (xpc)	$\frac{cov(\boldsymbol{x}_q, \boldsymbol{x}_r)}{\sigma_q \sigma_r}$	Tenenbaum et al. 2008
➡	eXtreme Spearman Correlation (xsc)	$\frac{cov(y_q, y_r)}{\sigma_{y_q}\sigma_{y_r}}, y = rank(x)$	Tanner and Agarwal 2008
	eXtreme Sum (XSum, xs)	$\sum_{i \epsilon Q^+} \boldsymbol{x_{ri}} - \sum_{i \epsilon Q^-} \boldsymbol{x_{ri}}$	Cheng et al. 2014
➡	eXtreme Cosine (XCos, xc)	$\frac{\boldsymbol{x_q} \cdot \boldsymbol{x_r}}{\lvert \boldsymbol{x_q} \rvert \lvert \boldsymbol{x_r} \rvert}$	Cheng et al. 2012
	Jaccard index (ji)	$J(Q, R) = \frac{Q \cap R}{Q \cup R}$	
➡	Signed Jaccard (sji)	$\frac{J(Q^+, R^+) + J(Q^-, R^-) - J(Q^+, R^-) - J(Q^-, R^+)}{2}$	Zichen Wang et al. 2016

Shah *et al.* in prep

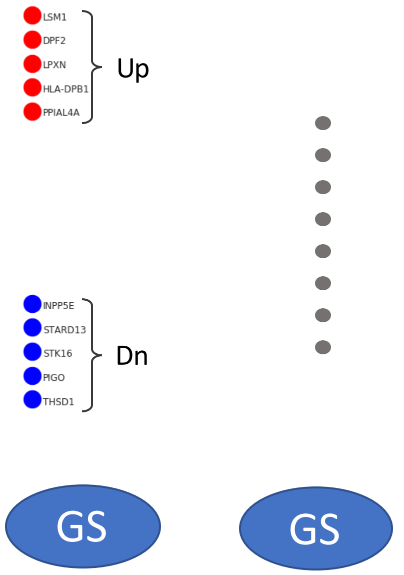
Connectivity Mapping: Reference Chemicals

Gene set-based connectivity mapping correctly identifies targets of reference chemicals



Gene Set Concentration-Response

Gene Sets

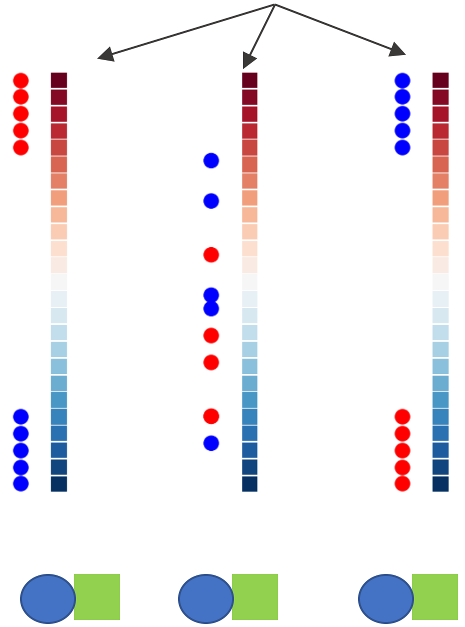


“Treatment” Gene Set

“Pathway” Gene Set

Connectivity

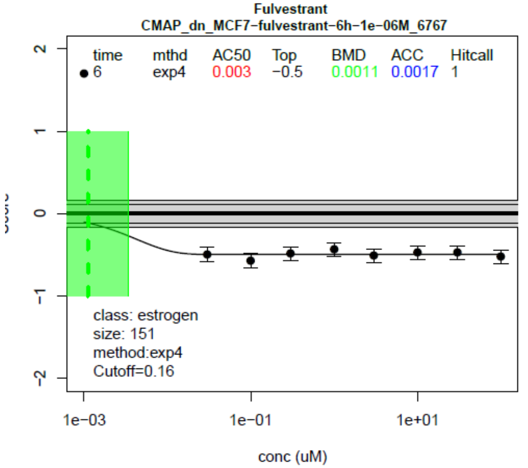
TempO-Seq profiles



Positive Score

Negative Score

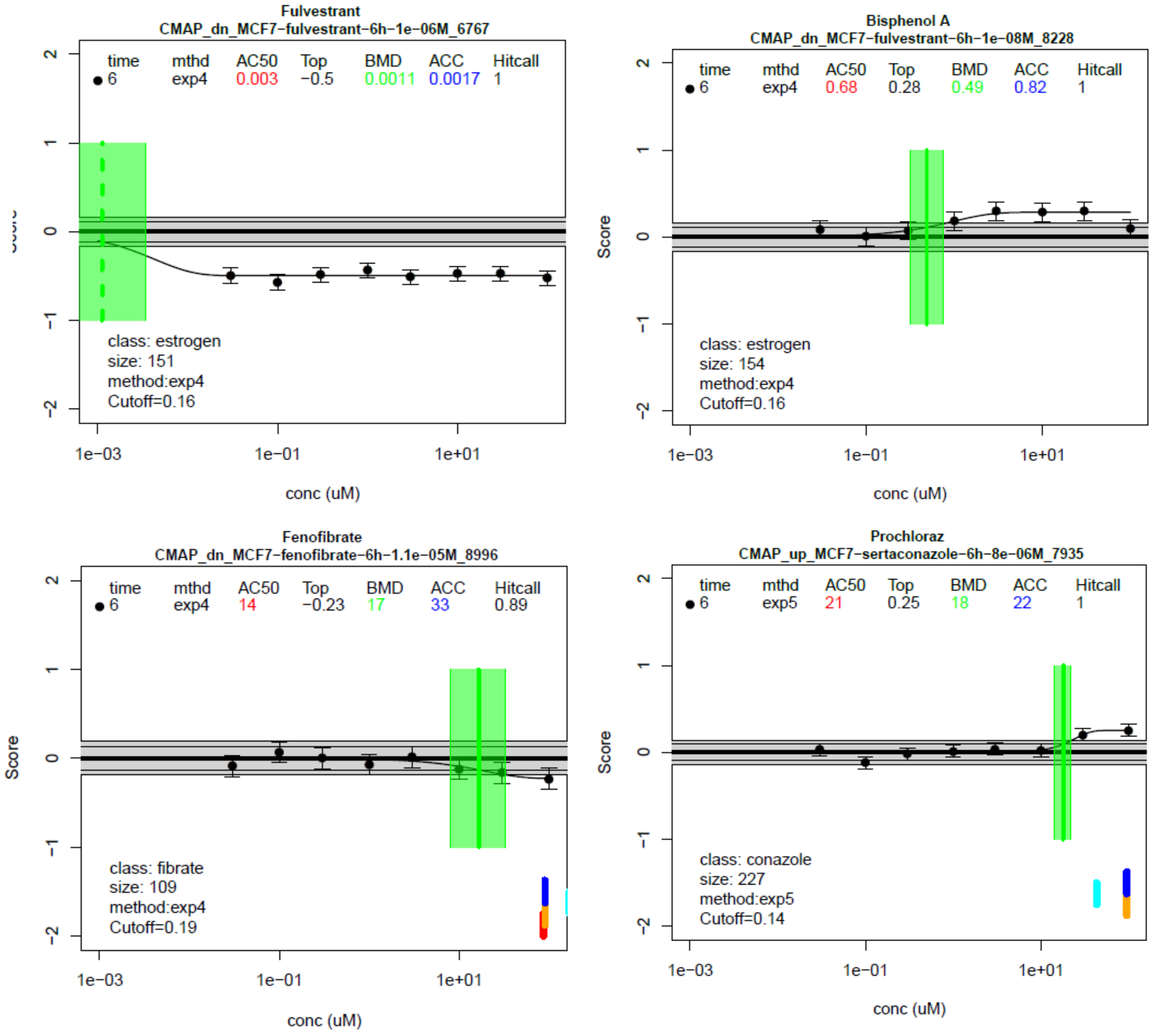
Curve-fitting Gene-Set Scores



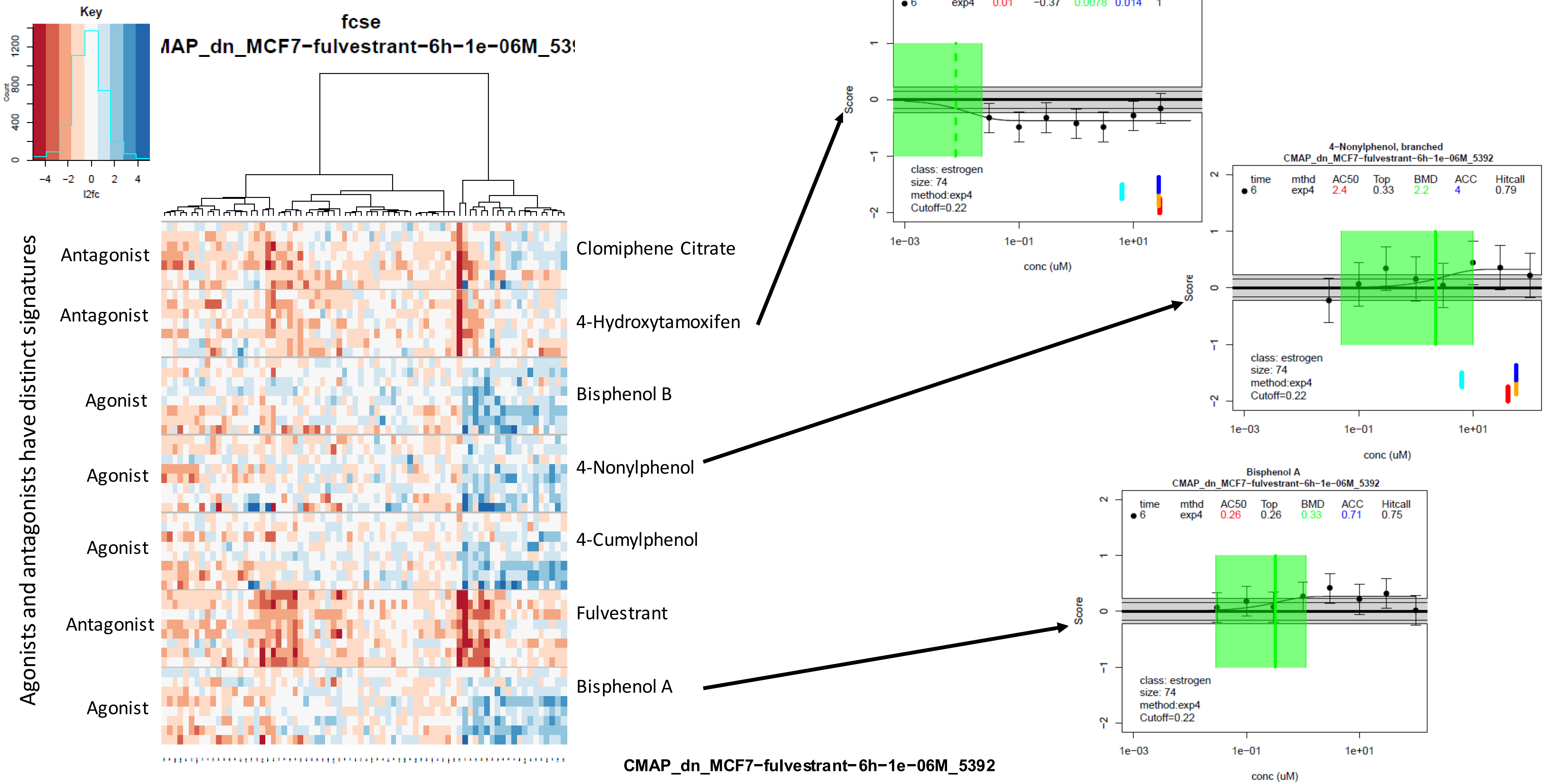
Summarize BMD scores for each
Chemical across gene sets to obtain
A potency distribution

Gene Set Concentration Response

- Calculate the pathway score for each pathway for the 44 real and 1000 random chemicals for each condition and concentration
- Random set forms null distribution for concentration-response modeling
- Do concentration-response modeling for 44+1000 chemicals
- Do post-processing analyses



Example: Estrogen



Agonists and antagonists have distinct signatures

Antagonist

Antagonist

Agonist

Agonist

Agonist

Antagonist

Agonist

Clomiphene Citrate

4-Hydroxytamoxifen

Bisphenol B

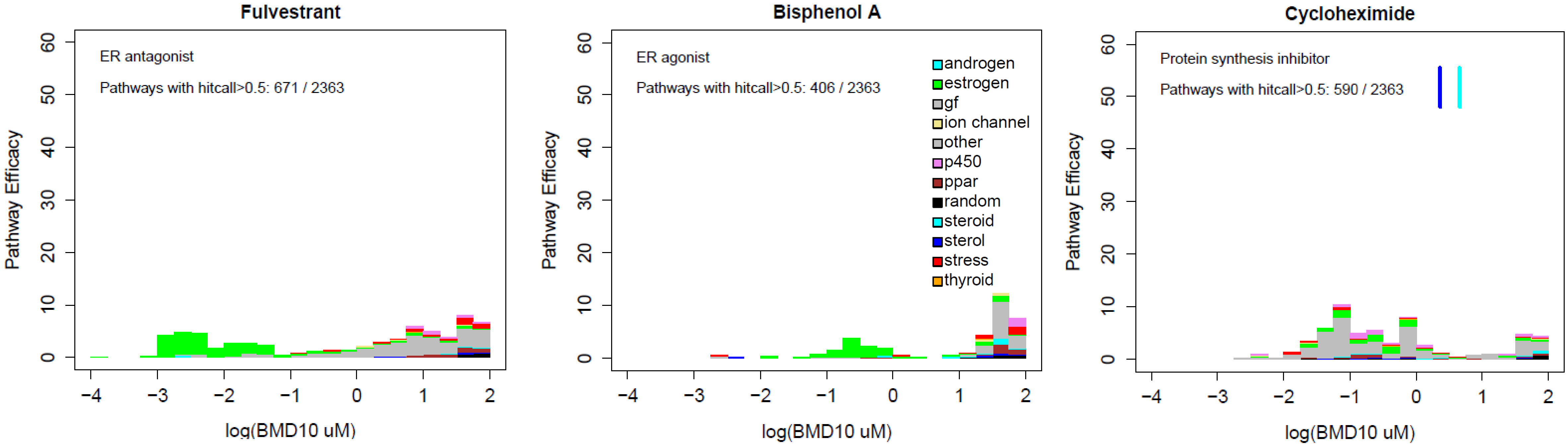
4-Nonylphenol

4-Cumylphenol

Fulvestrant

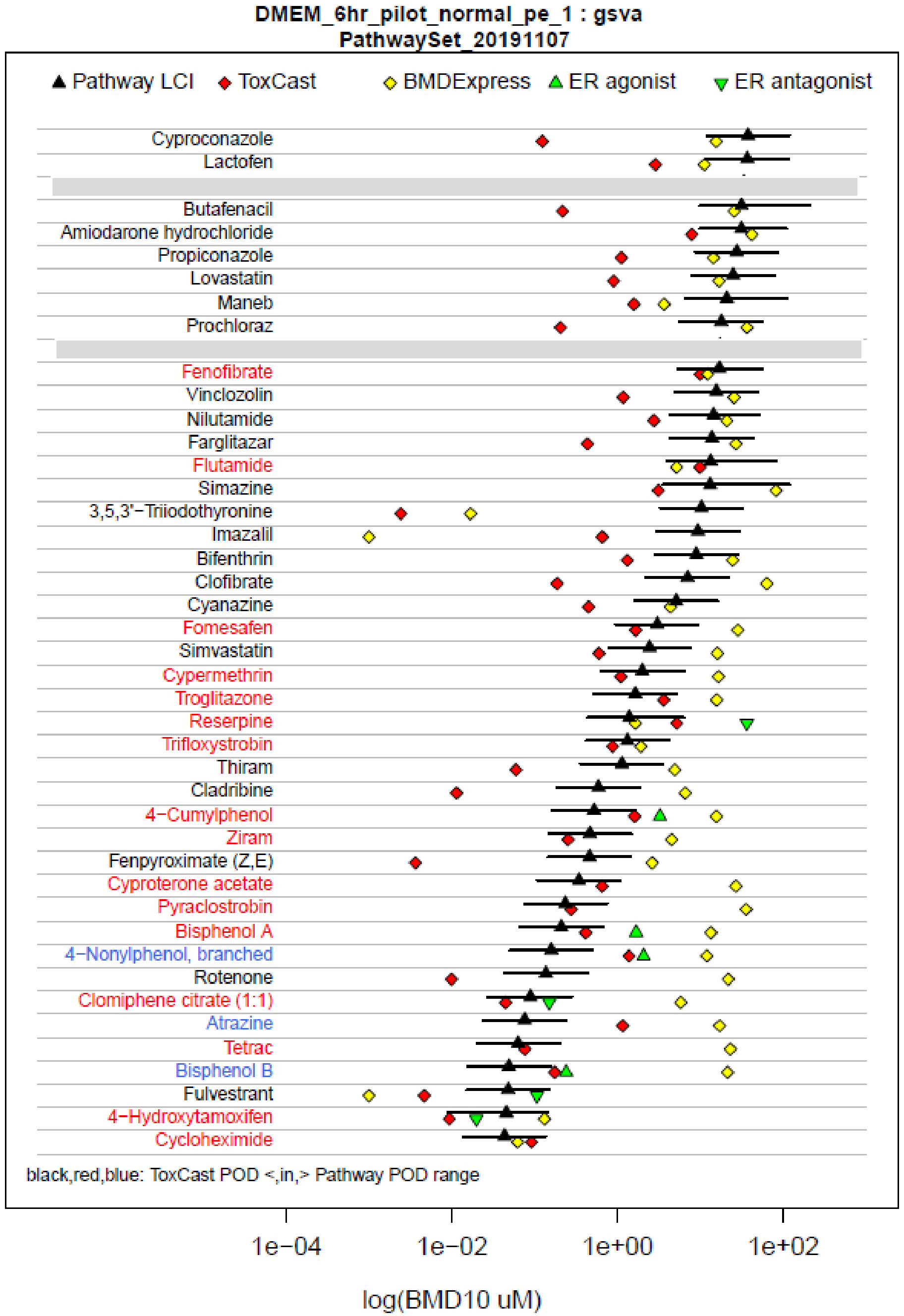
Bisphenol A

Gene Set Classes



Key points:

1. Estrogens have estrogen pathways at low concentrations
2. Most chemicals show stress at high concentrations
3. Random pathways usually only show up at high concentrations



Comparing PODs

Key Points

1. Potent chemicals in DESeq2 HTTr tend to have PODs ~ ToxCast
2. PODs from BMDExpress are mostly at high dose (>DESeq2)
3. Chemicals with significant efficacy (l2fc) tend to have better agreement between DESeq2 and BMDExpress PODs
4. ER pathway PODs from DESeq2 are on average more potent than those from ER Pathway Model / ToxCast ER assays

Understanding where BMDExpress has very potent predictions

<u>Pathways with BMD<1 uM</u>

- Cycloheximide – high efficacy, cell cycle, stress-related pathways
- Fulvestrant - high efficacy, ER pathways (e.g. CMAP Fulvestrant …)
- 4-Hydroxytamoxifen – 2 “real” ER pathways
- 3,5,3'-Triiodothyronine – 5 small gene sets with CYP1A1, CYP1B1
- Imazalil – 4 x 3-gene pathways ~ TNF signaling (TRAD, FADD, JUN)

Summary

- Robust HTTr processing pipeline and data management
- HTTr TempO-Seq platform reproducible
- Results for targets, pathways and potencies as expected
- Gene-set approaches produced more biologically-relevant results *for this data set*
- Ongoing research:
  - Choice of curve-fitting approaches
  - Gene set connectivity scoring methods
  - General approaches for putative target prediction

Acknowledgements

- HTTr Wetlab
  - Joshua Harrill
  - Clinton Willis
- TempO-Seq assay
  - BioSpyder
- HTTr Processing and Analysis
  - Logan Everett
  - Derik Haggard
  - Woody Setzer
  - Richard Judson
  - Thomas Sheffield
  - Bryant Chambers
  - Beena Vallanat
- Other
  - Russell Thomas