United States
Environmental Protection
Agency

Using High-Throughput Transcriptomics to Analyze Chemical Safety

Imran Shah
National Center for
Computational Toxicology
NCCT

Workshop on “TempO-Seq data analysis”
4-5 October 2018, Leiden, the Netherlands

EUTOXRISK

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 NCCT is using high-throughput transcriptomics
- Overall workflow and team
- Experimental analysis
- Computational analysis
    - Overview of different computational workflows / use-cases
    - NCCT HTTr workflow
        - Evaluate Data quality
        - Identify concentration-dependent effects of chemicals
        - Analyze putative molecular / pathway targets of chemicals

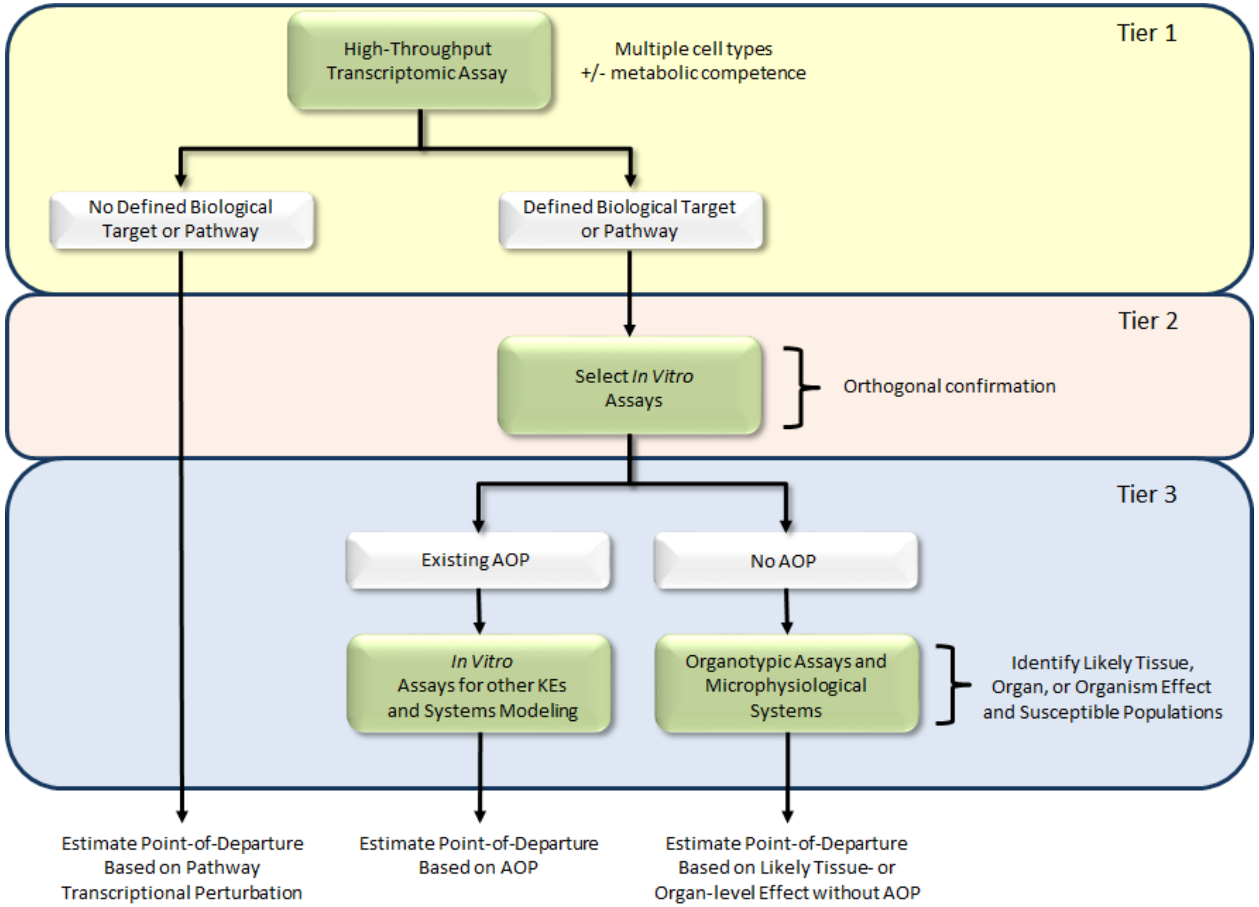
Objectives

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

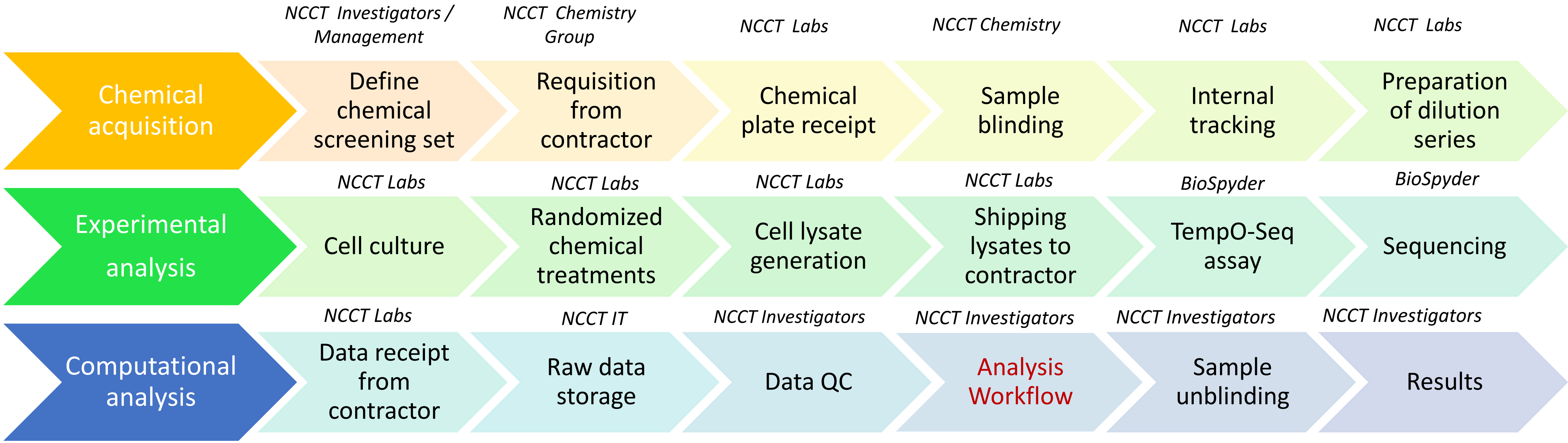
  → High-throughput transcriptomics (HTTr)

- Identify the concentration at which biological pathways/processes begin to be impacted
- Predict biological targets for chemicals with specific modes-of-action

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

HTTr Workflow



	NCCT Investigators / Management	*NCCT Chemistry Group*	*NCCT Labs*	*NCCT Chemistry*	*NCCT Labs*	*NCCT Labs*
Chemical acquisition	Define chemical screening set	Requisition from contractor	Chemical plate receipt	Sample blinding	Internal tracking	Preparation of dilution series
	NCCT Labs	*NCCT Labs*	*NCCT Labs*	*NCCT Labs*	*BioSpyder*	*BioSpyder*
Experimental analysis	Cell culture	Randomized chemical treatments	Cell lysate generation	Shipping lysates to contractor	TempO-Seq assay	Sequencing
	NCCT Labs	*NCCT IT*	*NCCT Investigators*	*NCCT Investigators*	*NCCT Investigators*	*NCCT Investigators*
Computational analysis	Data receipt from contractor	Raw data storage	Data QC	Analysis Workflow	Sample unblinding	Results

NCCT HTTr Project Team

National Center for Computational Toxicology

Joshua
Harrill
Toxicologist

Clinton
Willis
NSSC (JH)

Imran
Shah
Computational Systems Biologist

R. Woodrow
Setzer
Mathematical Statistician

Derik
Haggard
ORISE Fellow

Richard
Judson
Bioinformatician

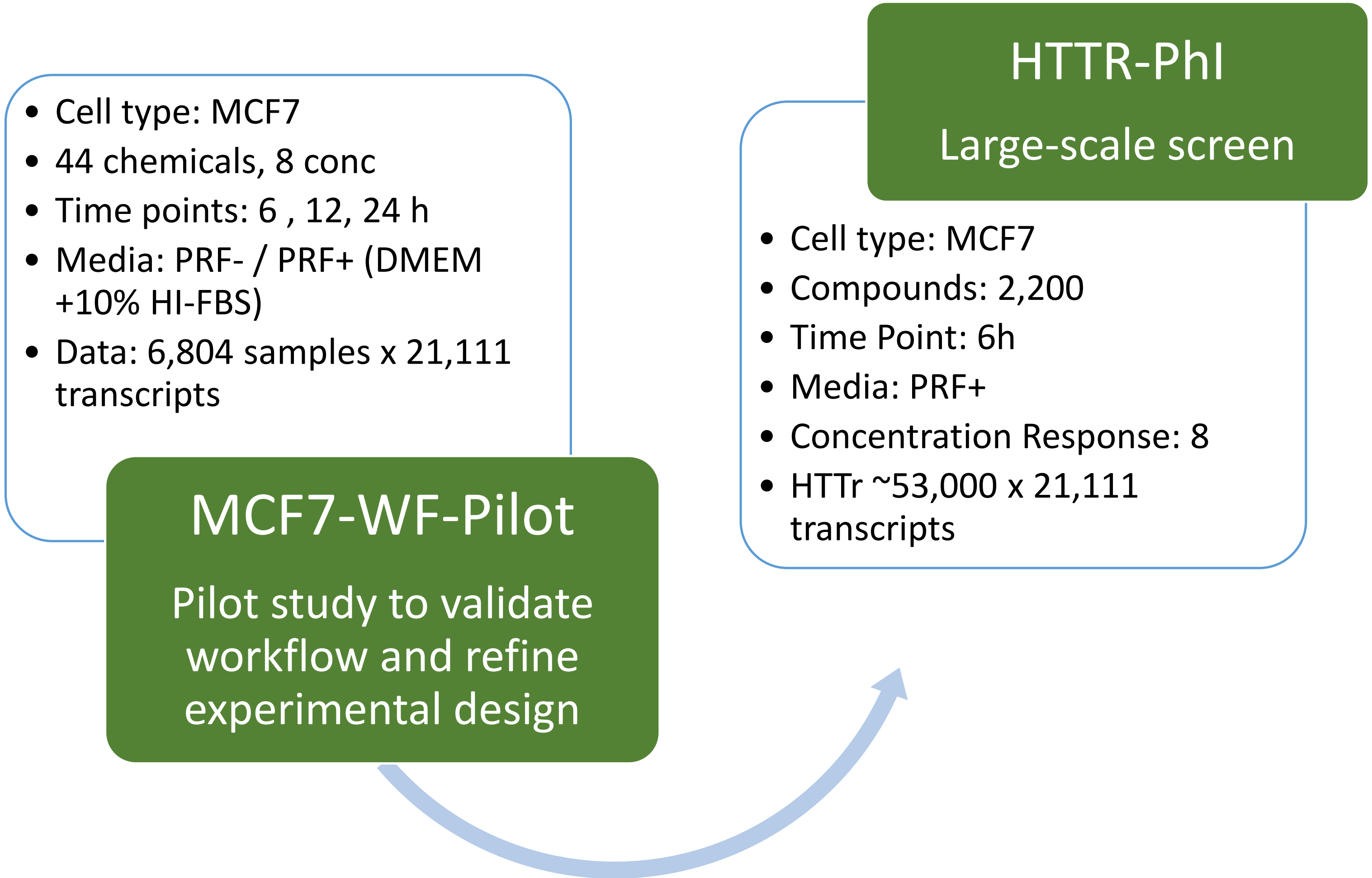
Russell
Thomas
Director

Experimental

Computational

Experimental Analysis

Two Main HTTr Experiments (so far)



- Cell type: MCF7
- 44 chemicals, 8 conc
- Time points: 6 , 12, 24 h
- Media: PRF- / PRF+ (DMEM +10% HI-FBS)
- Data: 6,804 samples x 21,111 transcripts

MCF7-WF-Pilot

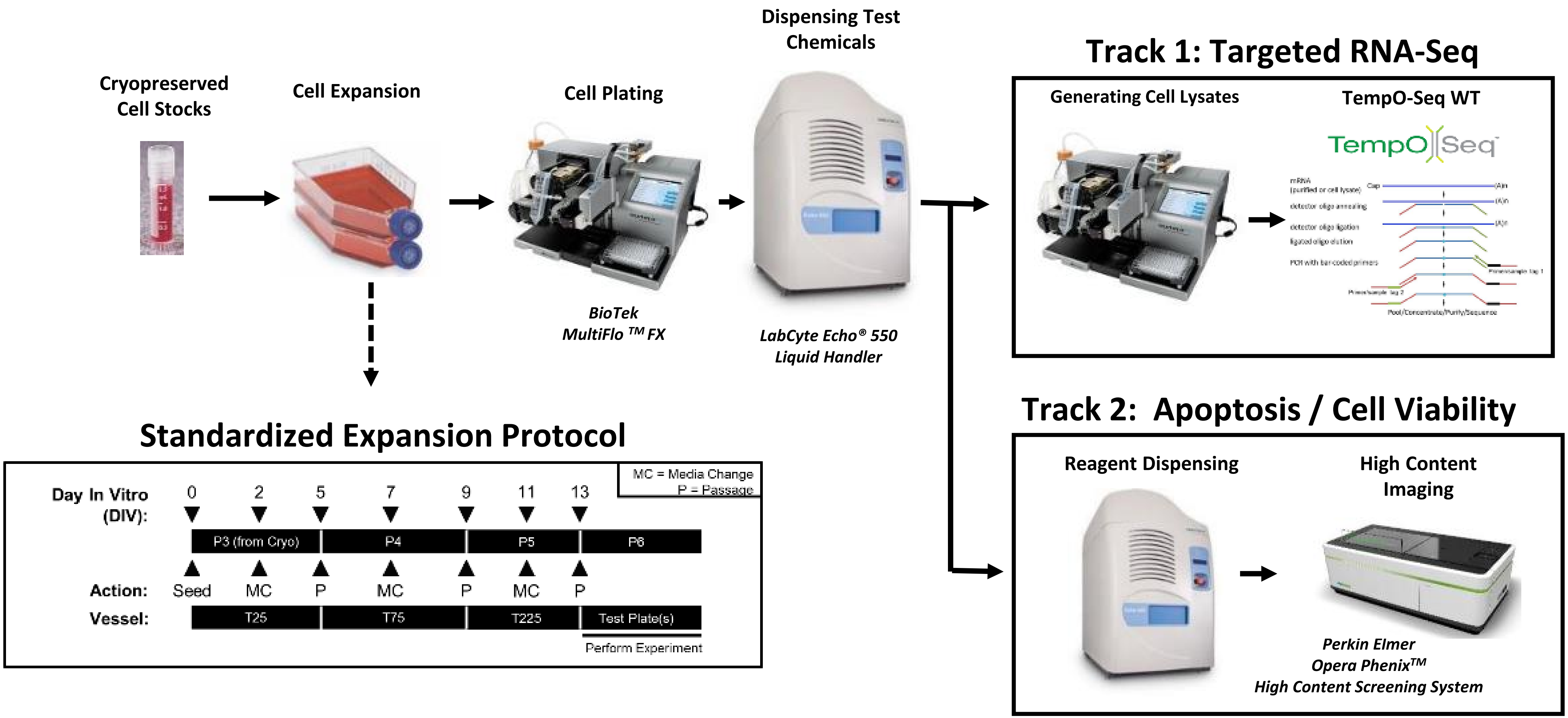
Pilot study to validate workflow and refine experimental design

HTTR-PhI

Large-scale screen

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

Lab Workflow



Cryopreserved Cell Stocks

Cell Expansion

Cell Plating

BioTek MultiFlo TM FX

Dispensing Test Chemicals

LabCyte Echo® 550 Liquid Handler

Track 1: Targeted RNA-Seq

Generating Cell Lysates

TempO-Seq WT

Track 2: Apoptosis / Cell Viability

Reagent Dispensing

High Content Imaging

Perkin Elmer Opera Phenix^{TM} High Content Screening System

Standardized Expansion Protocol

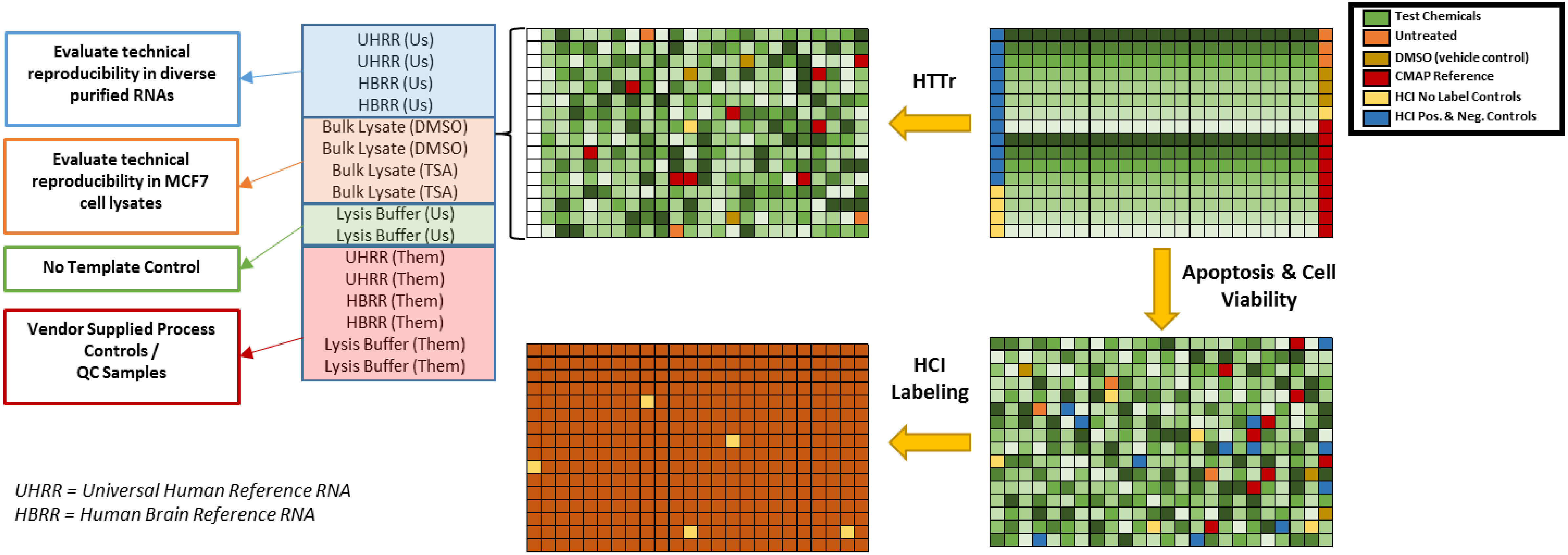
MC = Media Change
P = Passage

Day In Vitro (DIV):	0	2	5	7	9	11	13
	P3 (from Cryo)		P4		P5		P6
Action:	Seed	MC	P	MC	P	MC	P
Vessel:	T25		T75		T225		Test Plate(s)
							Perform Experiment

Quality Control Samples and Reference Standards for Performance-Based Validation

Treatment Randomization: *Each test plate uniquely randomized with respect to treatment.*

QC Samples: *Quality Control samples included on each plate*



Evaluate technical reproducibility in diverse purified RNAs

Evaluate technical reproducibility in MCF7 cell lysates

No Template Control

Vendor Supplied Process Controls / QC Samples

UHRR (Us)
UHRR (Us)
HBRR (Us)
HBRR (Us)
Bulk Lysate (DMSO)
Bulk Lysate (DMSO)
Bulk Lysate (TSA)
Bulk Lysate (TSA)
Lysis Buffer (Us)
Lysis Buffer (Us)
UHRR (Them)
UHRR (Them)
HBRR (Them)
HBRR (Them)
Lysis Buffer (Them)
Lysis Buffer (Them)

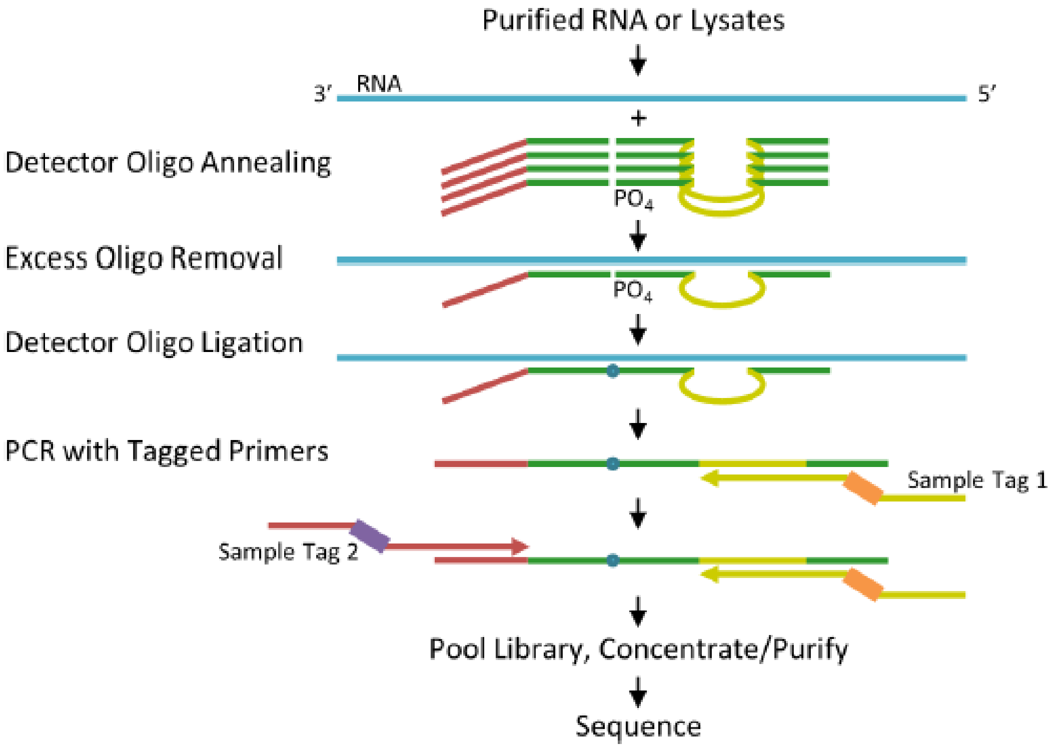
🟩 Test Chemicals
🟧 Untreated
🟫 DMSO (vehicle control)
🟥 CMAP Reference
🟨 HCI No Label Controls
🟦 HCI Pos. & Neg. Controls

UHRR = Universal Human Reference RNA
HBRR = Human Brain Reference RNA

TempO-Seq for HTTr

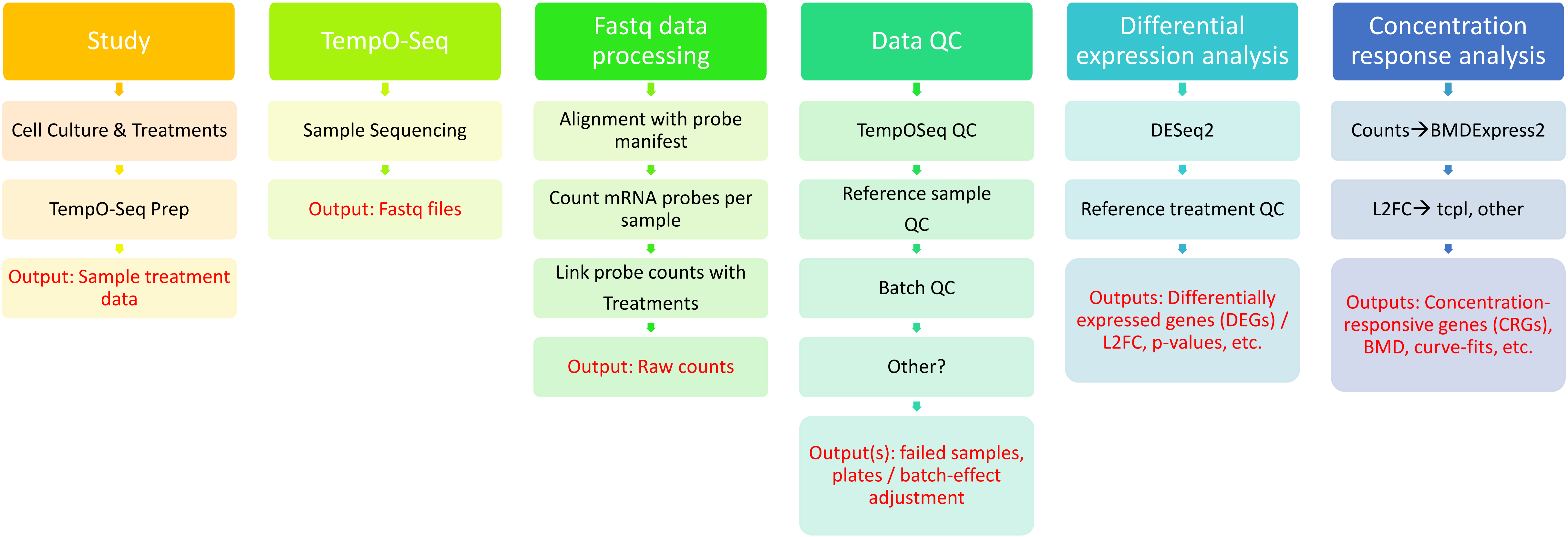
- The **TempO-Seq** human whole transcriptome assay measures the expression of ~21,100 transcripts.
- Requires only picogram amounts of total RNA per sample.
- Compatible with purified RNA samples or **cell lysates**.
- Transcripts in cell lysates generated in 384-well format barcoded to well position
- 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



Computational Analysis Overview

HTTr Computational Analysis Steps



Study
- Cell Culture & Treatments
- TempO-Seq Prep
- Output: Sample treatment data

TempO-Seq
- Sample Sequencing
- Output: Fastq files

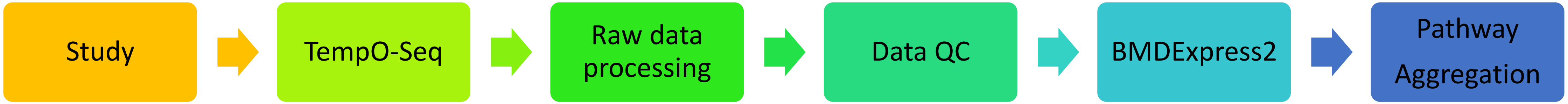
Fastq data processing
- Alignment with probe manifest
- Count mRNA probes per sample
- Link probe counts with Treatments
- Output: Raw counts

Data QC
- TempOSeq QC
- Reference sample QC
- Batch QC
- Other?
- Output(s): failed samples, plates / batch-effect adjustment

Differential expression analysis
- DESeq2
- Reference treatment QC
- Outputs: Differentially expressed genes (DEGs) / L2FC, p-values, etc.

Concentration response analysis
- Counts → BMDExpress2
- L2FC → tcpl, other
- Outputs: Concentration-responsive genes (CRGs), BMD, curve-fits, etc.

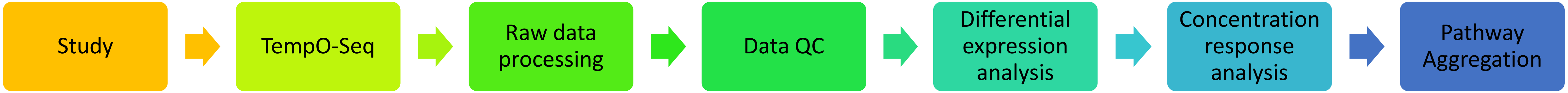
Basic HTTr Analysis Workflow



Study → TempO-Seq → Raw data processing → Data QC → BMDExpress2 → Pathway Aggregation

- Use-case: identify the most sensitive pathway perturbations
- Study design: One cell type, multiple chemicals, multiple conc, single time point
- Approach:
  - TempO-Seq HTTr data generation
  - Process raw data to generate probe level counts
  - Conduct TempO-Seq QC (read depth, mapped fraction, etc.)
  - Filter probes by average/maximum/median count (to exclude very low level counts)
  - Normalise counts for each sample (e.g by read-depth scale to $3x10^6$)
  - Conc-response analysis using BMDExpress2 (choice of filters, fits, and thresholds output conc-responsive probes and BMD values)
  - Pathway level aggregation by genes and BMD values (summarised as accumulation plots)

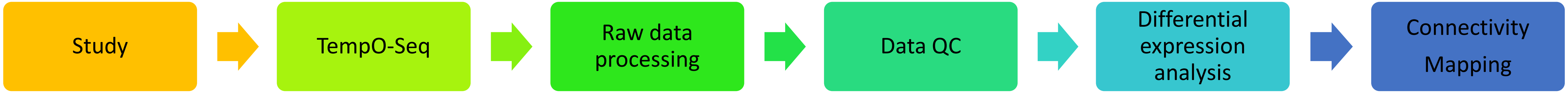
Intermediate HTTr Analysis Workflow



Study → TempO-Seq → Raw data processing → Data QC → Differential expression analysis → Concentration response analysis → Pathway Aggregation

- Use-case: identify the most sensitive pathway perturbations
- Study design: One cell type, multiple chemicals, multiple conc, single time point
- Approach:
  - TempO-Seq HTTr data generation
  - Process raw data to generate probe level counts
  - Conduct TempO-Seq QC (read depth, mapped fraction, etc.)
  - Filter probes by average/maximum/median count (to exclude very low level counts)
  - Differential expression analysis using DESeq2 (produces L2FC, p-values, mean-counts, etc.)
  - Concentration response analysis using L2FC data and tcpl (ToxCast curve-fitting pipeline)
  - Pathway level aggregation of conc-responsive genes using BMD_{10}

Intermediate HTTr Analysis Workflow

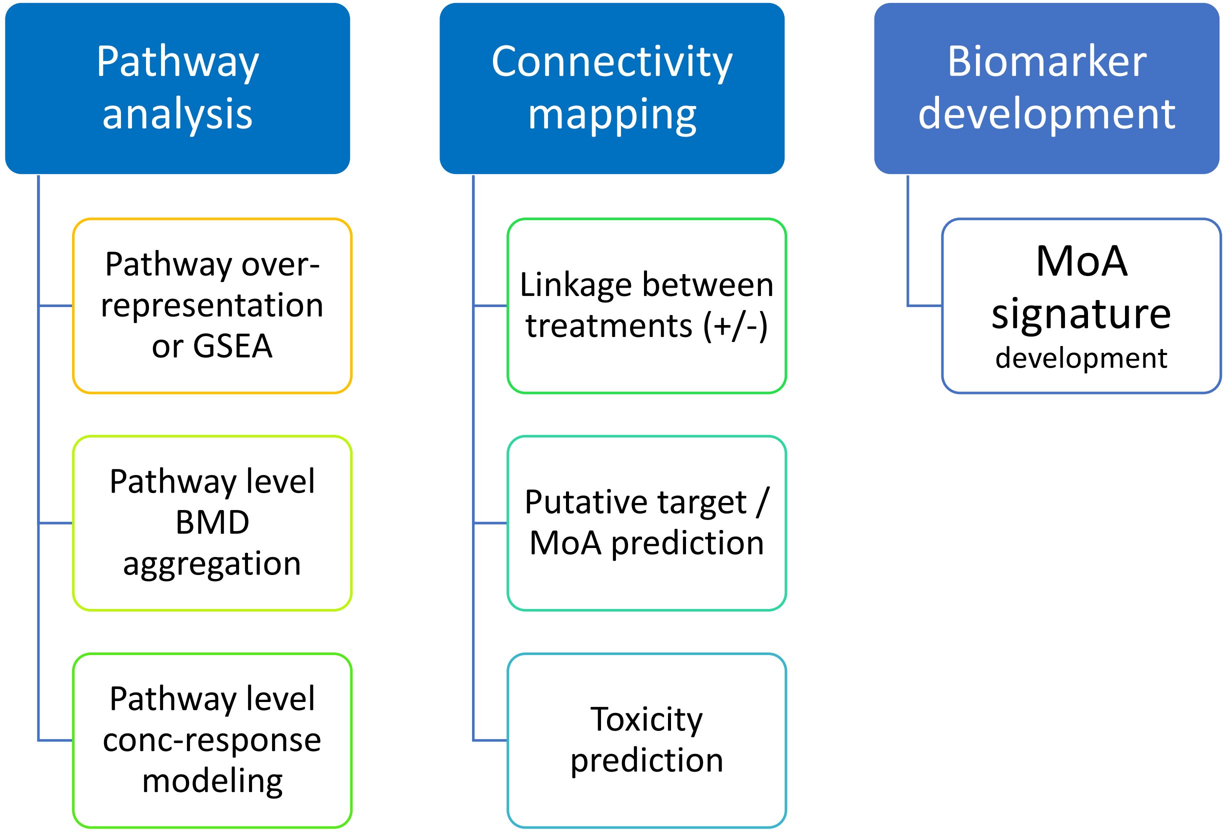


Study → TempO-Seq → Raw data processing → Data QC → Differential expression analysis → Connectivity Mapping

- Use-case: identify the putative molecular targets of chemicals
- Study design: One cell type, multiple chemicals, multiple conc, single time point
- Approach:
  - TempO-Seq HTTr data generation
  - Process raw data to generate probe level counts
  - Conduct TempO-Seq QC (read depth, mapped fraction, etc.)
  - Filter probes by average/maximum/median count (to exclude very low level counts)
  - Differential expression analysis using DESeq2 (produces L2FC, p-values, mean-counts, etc.)
  - Generate DEG signatures for GSEA analysis with CMap reference database
  - Link CMap hits to putative targets

Interpretation – many options

Some interpretation options that can use either CRGs or DEGs



Pathway analysis

- Pathway over-representation or GSEA
- Pathway level BMD aggregation
- Pathway level conc-response modeling

Connectivity mapping

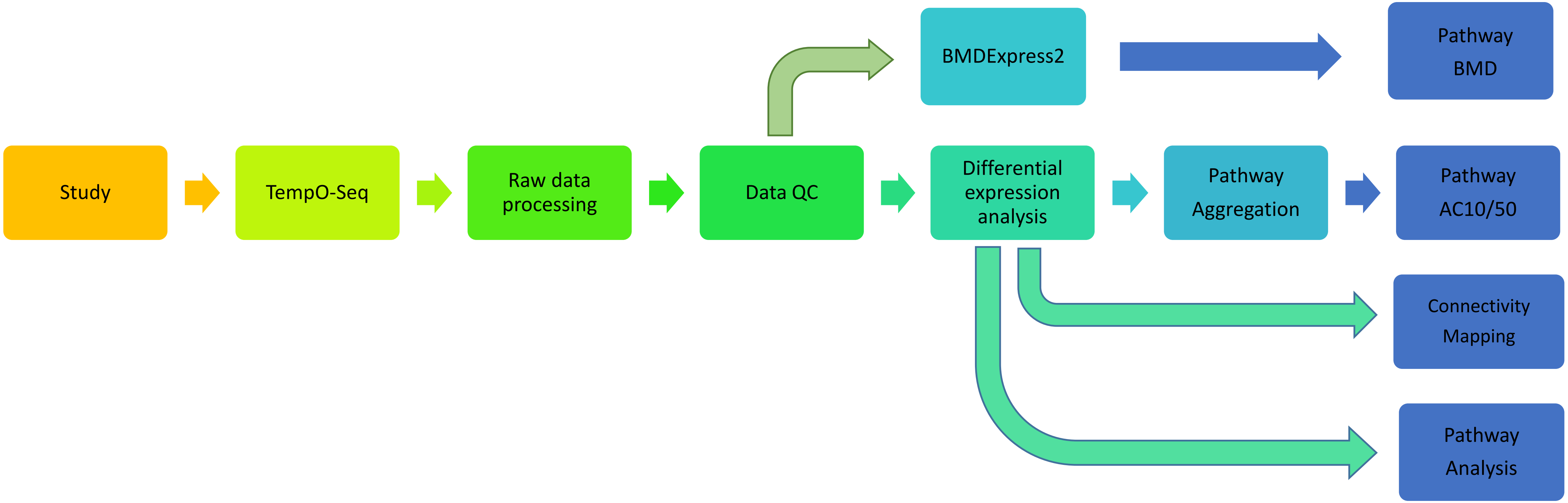
- Linkage between treatments (+/-)
- Putative target / MoA prediction
- Toxicity prediction

Biomarker development

- MoA signature development

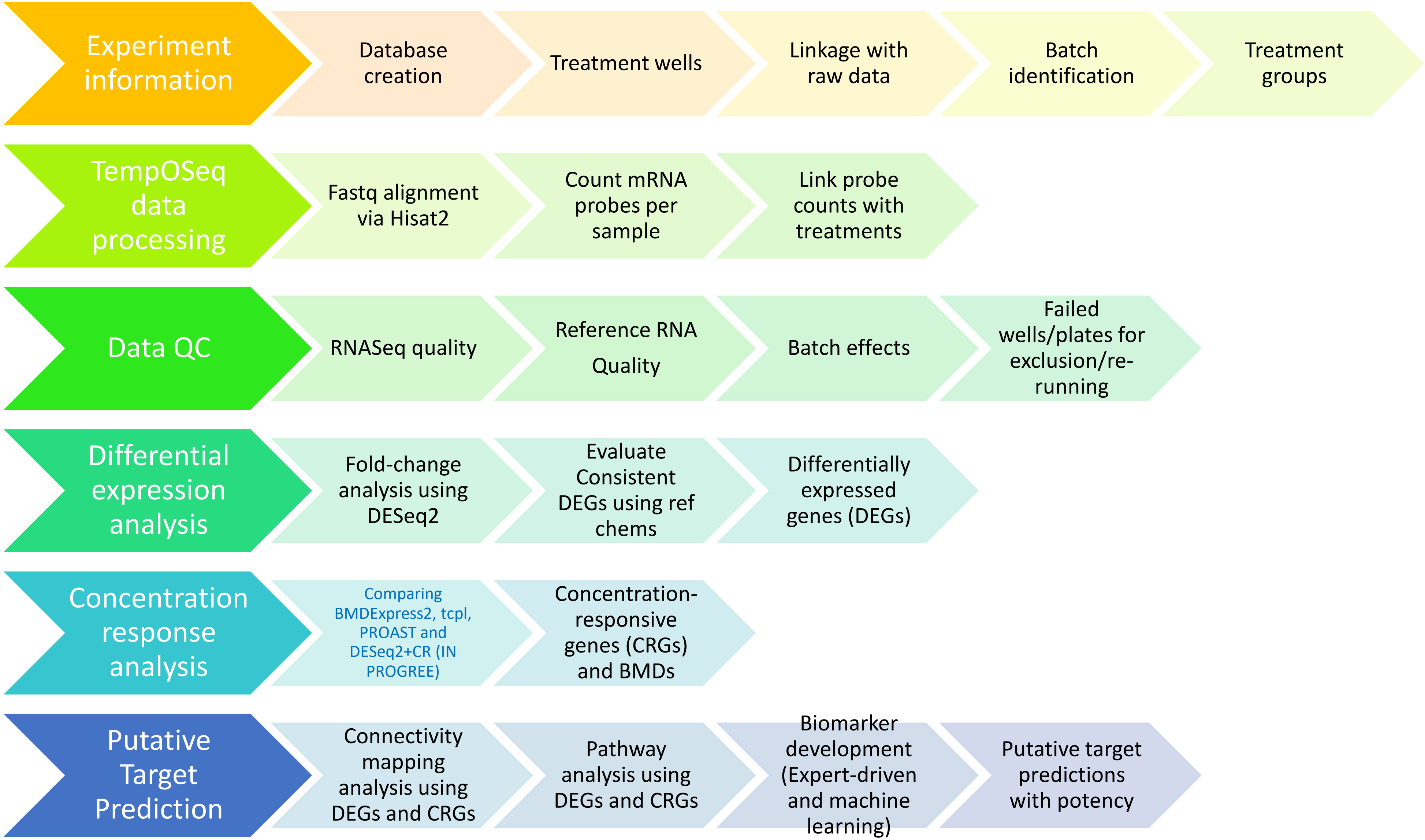
NCCT HTTr Analysis Workflow

NCCT HTTr Analysis Workflow



- Use-case: **Evaluate chemical potency, putative targets and pathways using HTTr**
- Study design: MCF7 cells, 2100 chemicals, 8 conc, 6 h time point
- Approach: “Exploratory”

Analysis Pipeline (June 2018)



Experiment information → Database creation → Treatment wells → Linkage with raw data → Batch identification → Treatment groups

TempOSeq data processing → Fastq alignment via Hisat2 → Count mRNA probes per sample → Link probe counts with treatments

Data QC → RNASeq quality → Reference RNA Quality → Batch effects → Failed wells/plates for exclusion/re-running

Differential expression analysis → Fold-change analysis using DESeq2 → Evaluate Consistent DEGs using ref chems → Differentially expressed genes (DEGs)

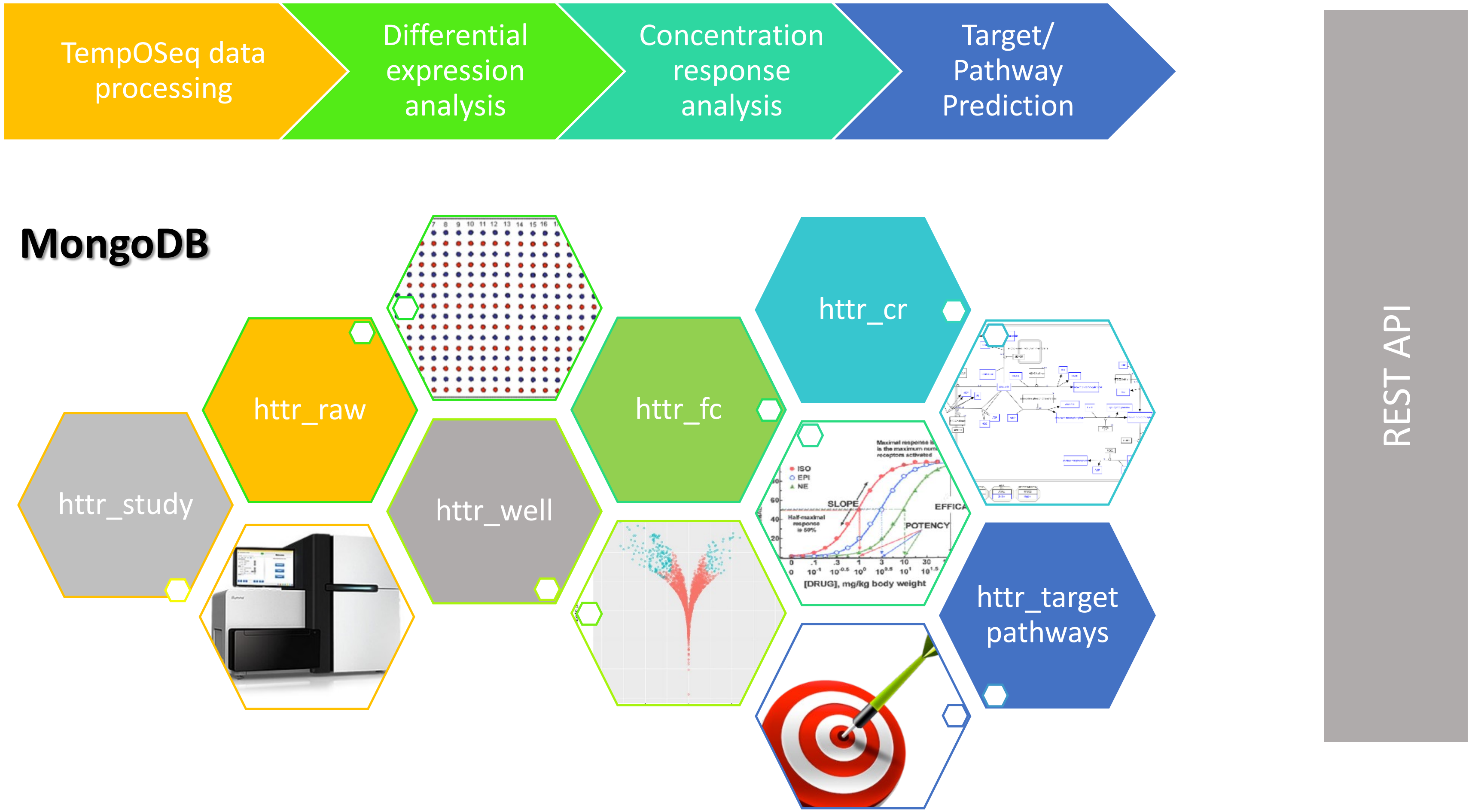
Concentration response analysis → Comparing BMDExpress2, tcpl, PROAST and DESeq2+CR (IN PROGREE) → Concentration-responsive genes (CRGs) and BMDs

Putative Target Prediction → Connectivity mapping analysis using DEGs and CRGs → Pathway analysis using DEGs and CRGs → Biomarker development (Expert-driven and machine learning) → Putative target predictions with potency

HTTr Analysis, Storage and Dissemination

(Internal EPA)

Python & R analysis pipeline



TempOSeq data processing

Differential expression analysis

Concentration response analysis

Target/ Pathway Prediction

MongoDB

httr_study

httr_raw

httr_well

httr_fc

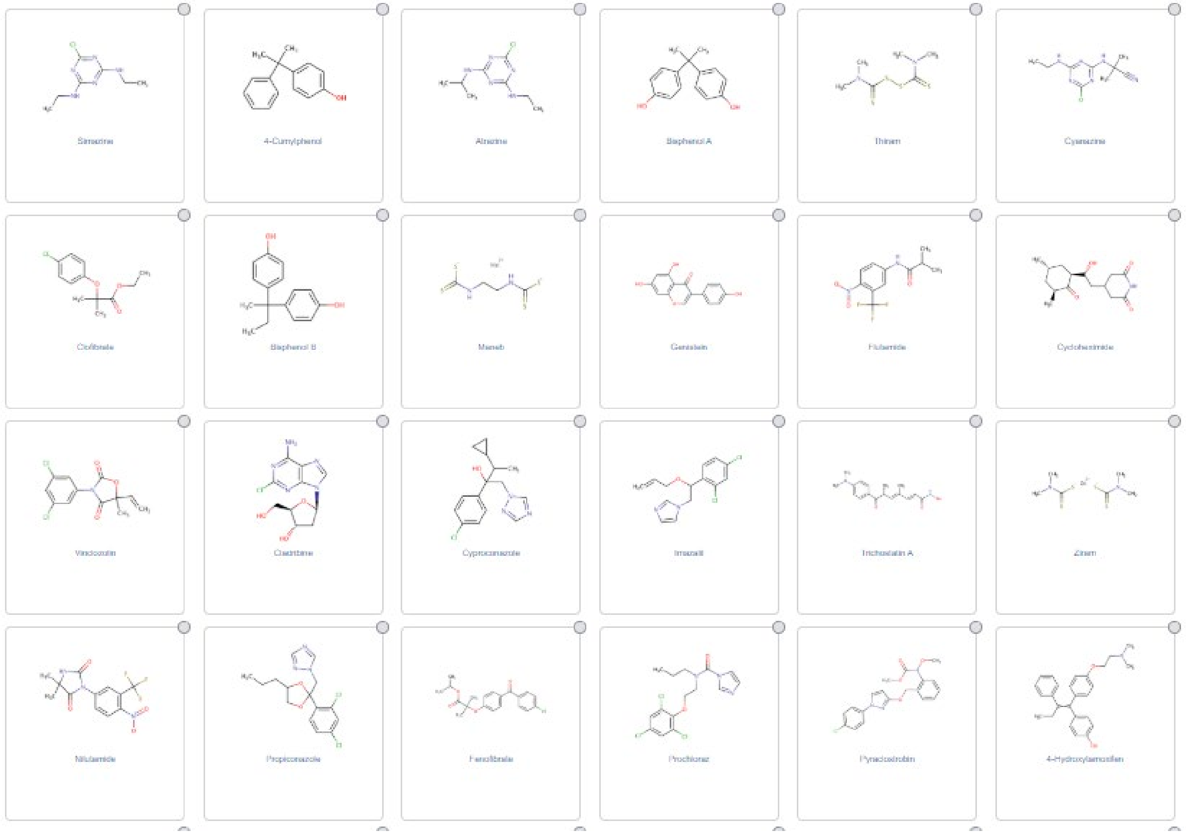
httr_cr

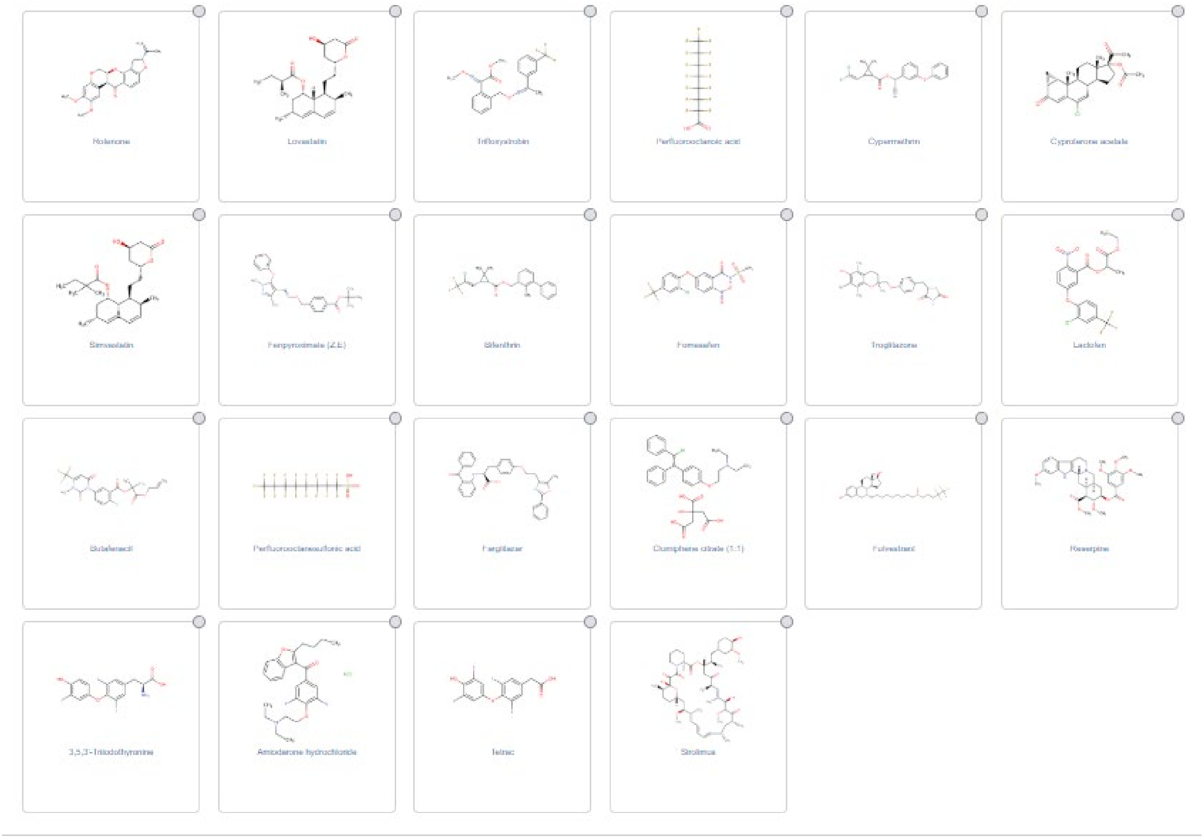
httr_target pathways

REST API

ISO
EPI
NE
Maximal response is the maximum number of receptors activated
SLOPE
Half-maximal response is 50%
EFFICA
POTENCY
[DRUG], mg/kg body weight

MCF7 Pilot Study Chemicals



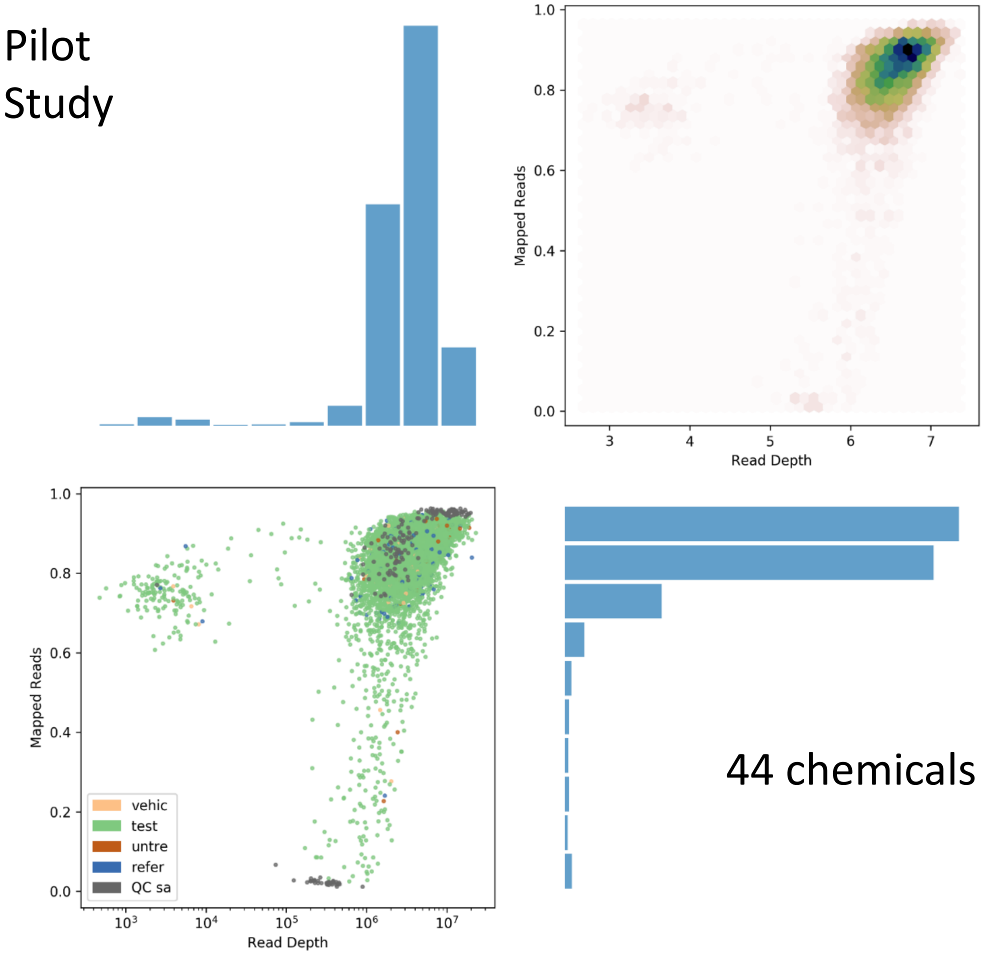


Data Quality

TempO-Seq Quality

Quality metrics:

- Read depth: number of mRNAs sequenced
  - Ideal value = $3x10^6$
- Mapped reads: fraction of sequenced mRNAs that map to a specific probe/gene
  - Ideal value = 100%

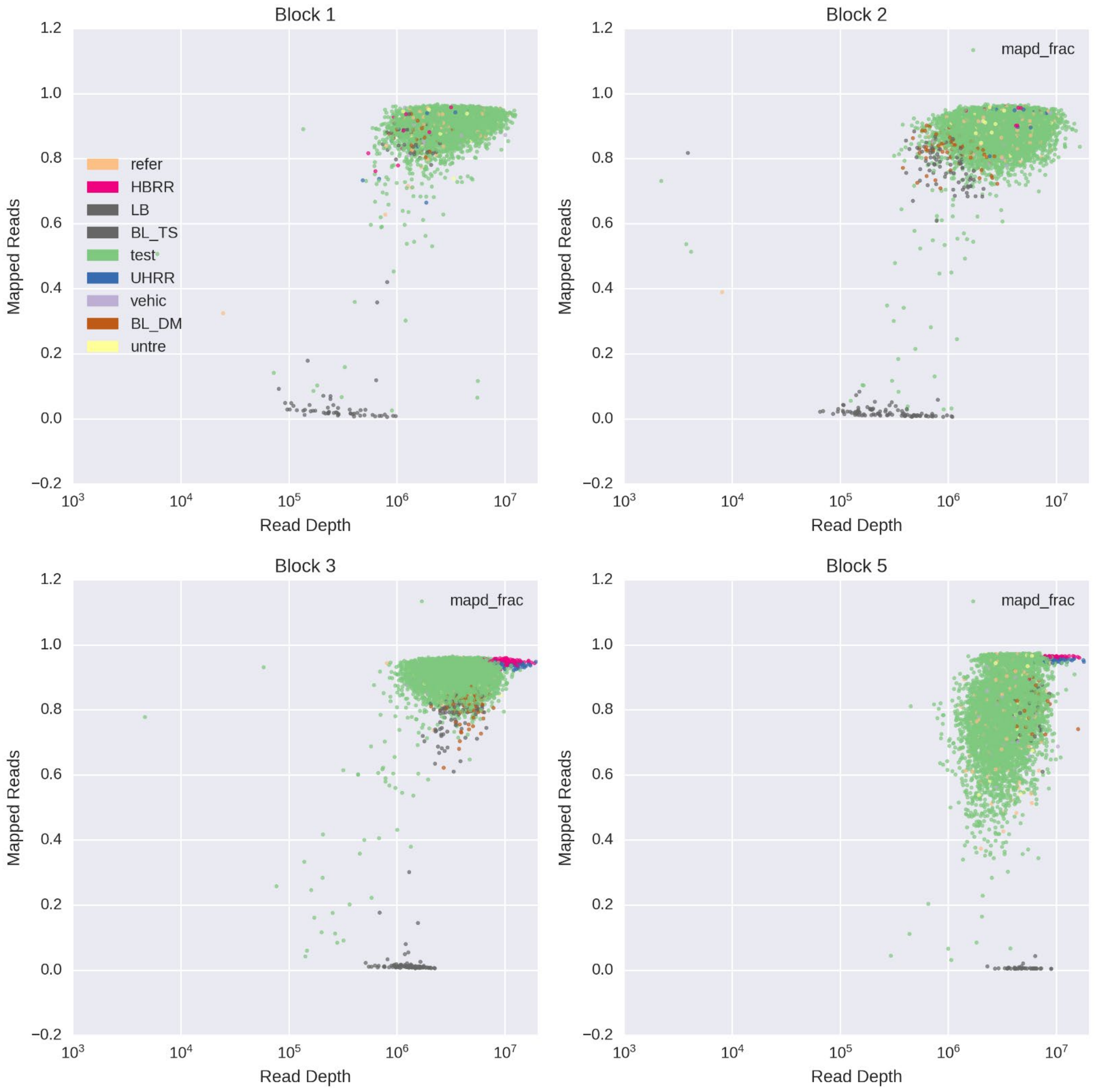


Pilot Study

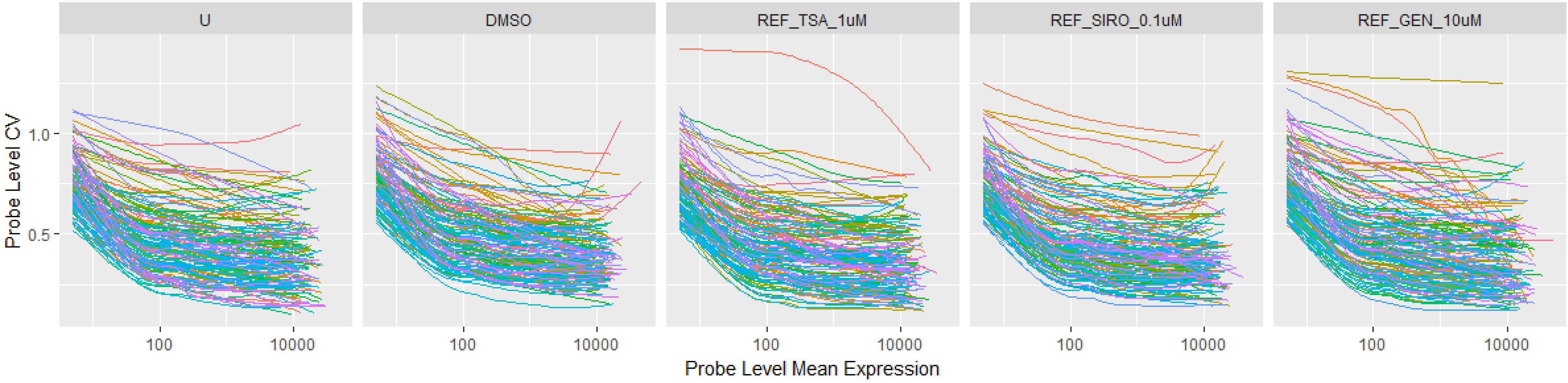
44 chemicals

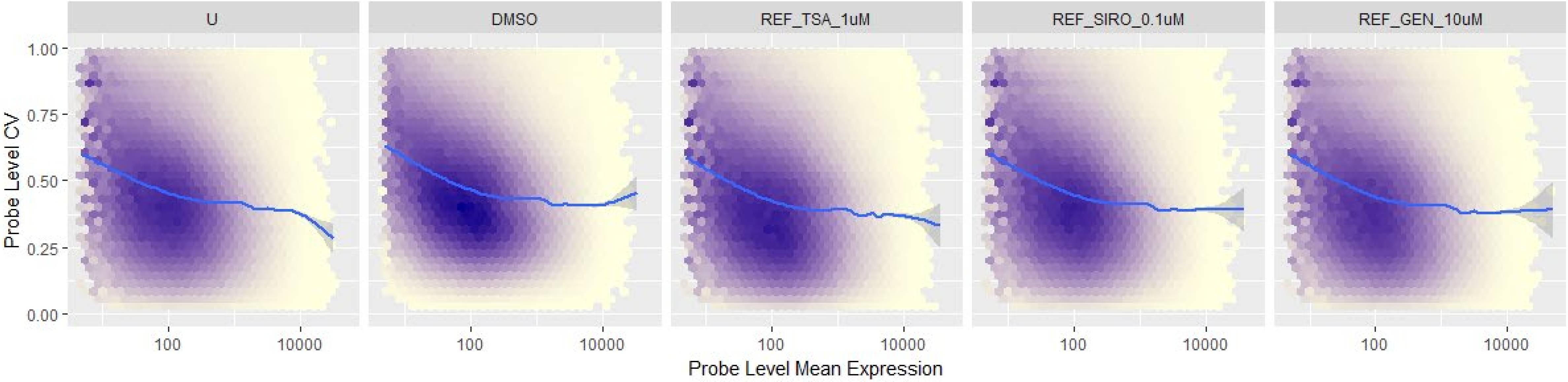
TempO-Seq quality

	Mapped %		Read depth	
block_id	**mean**	**std**	**mean**	**std**
1	0.908	0.077	3.33E+06	1.60E+06
2	0.892	0.078	3.53E+06	1.64E+06
3	0.909	0.076	3.72E+06	1.56E+06
5	**0.797**	**0.124**	3.77E+06	1.64E+06



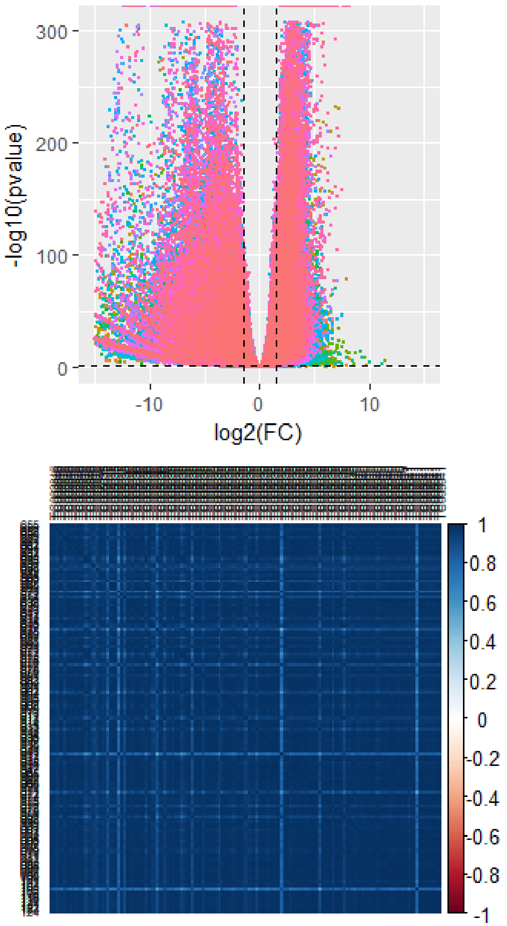
Coefficient of Variation Vs. Transcript Abundance



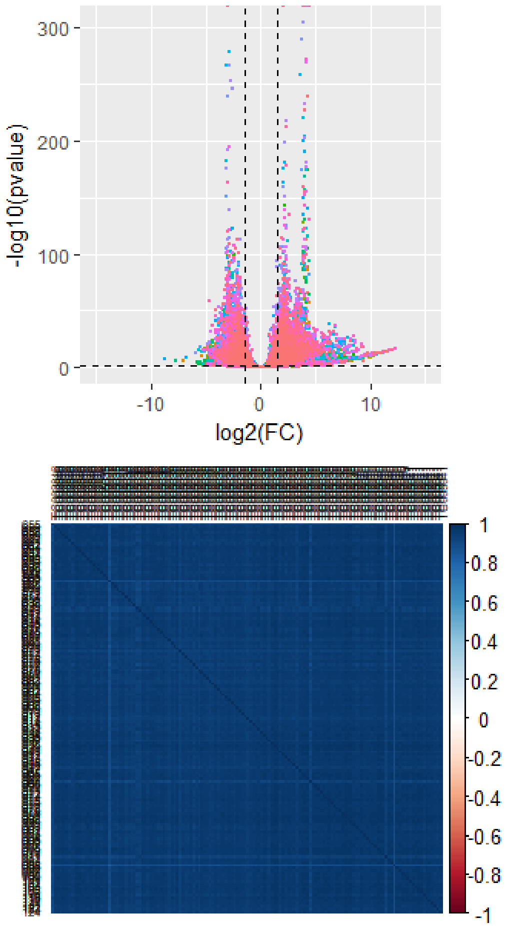


Reproducibility of Log$_2$(FC) Estimates

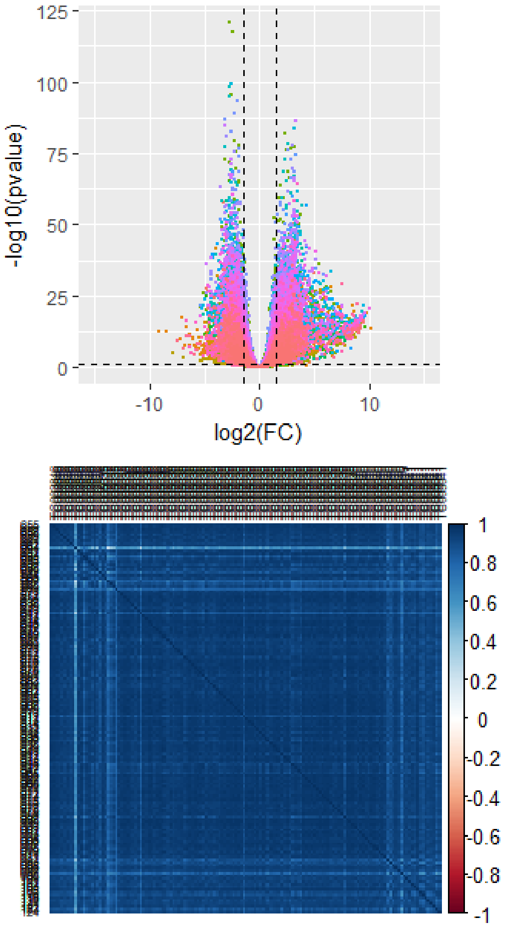
UHRR vs. HBRR



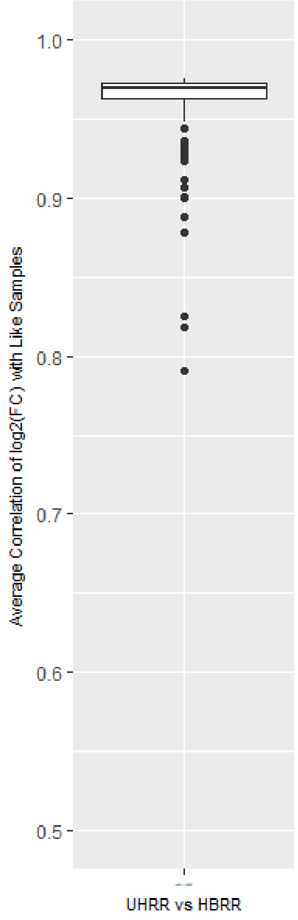
Bulk Lysate



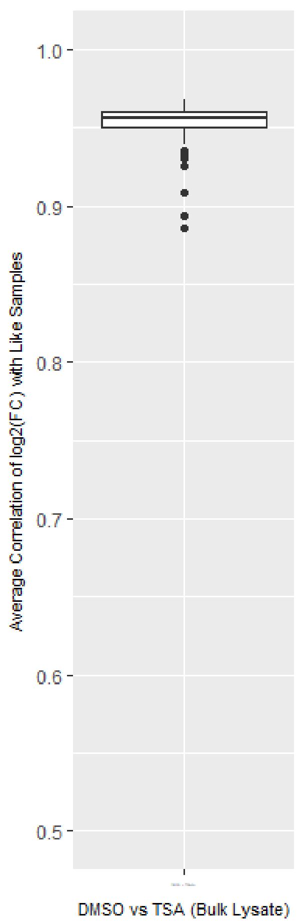
Reference (DMSO vs. TSA)



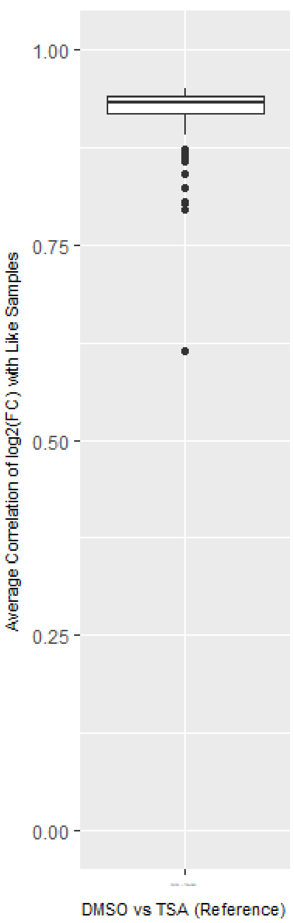
UHRR vs. HBRR



Bulk Lysate



Ref. (DMSO vs TSA)



- High correlation of log2 FC estimates across plates and screening blocks.

Concentration-Response Analysis

BMDExpress2

Benchmark Dose Modeling

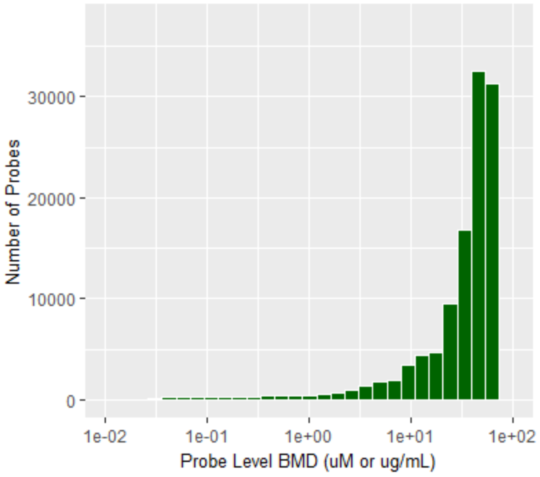
Parameter	Criteria [a]
Pre-filter:	Williams trend test
Models	Hill, Exponential 2, *poly2, power, linear*
BMR Factor:	1.349 (10 %)
Best Model Selection:	Lowest AIC
Hill Model Flagging [b]:	'k' < 1/3 Lowest Positive Dose Retain Flagged Models
Pathway Analysis:	Genes with BMD <= Highest Dose ≥ 3 ≥ 1% Gene Set Coverage
Gene Set Collections [c]:	MSigDB_C2 MSigDB_H Reactome BioPlanet KEGG

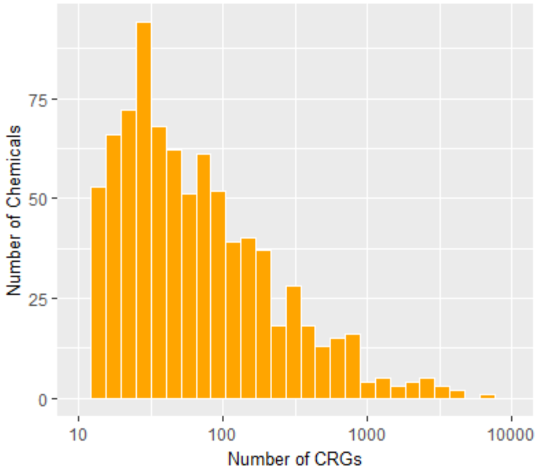
[a] Exploratory analysis – modeling criteria not finalized

[c] Gene Set Collections:

- **MSigDB_C2:** Curated gene sets from online pathway databases, publications and knowledge of domain experts (n = 4738).
- **MSigDB_H:** Coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes (n = 50).
- **Reactome:** Open-source, curated and peer reviewed pathway database with hierarchical pathway relationships in specific domains of biology. (n = 1764). Some pathways included in MSigDB_C2.
- **BioPlanet** (n = 1700): Curated pathway set developed by National Toxicology Program.

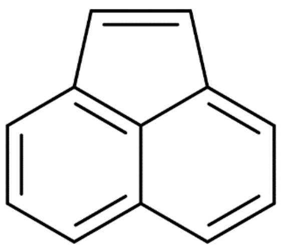
Benchmark Dose Modeling Summary & Inducible Genes

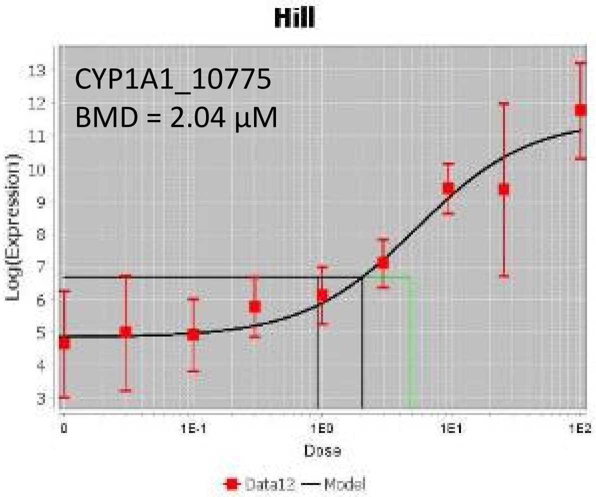




CYP1A1_10775
(n = 473)

Acenaphthylene
208-96-8 | DTXSID3023845

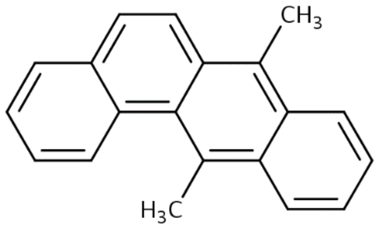


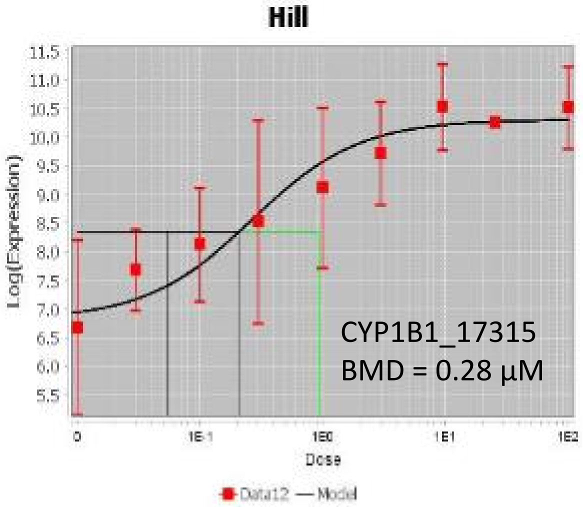


CYP1A1_10775
BMD = 2.04 µM

CYP1B1_17315
(n = 279)

7,12-Dimethylbenz(a)anthracene
57-97-6 | DTXSID1020510

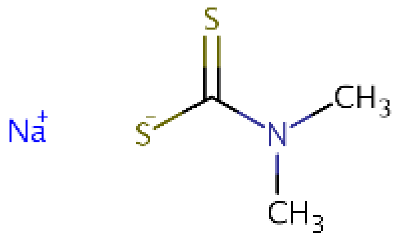


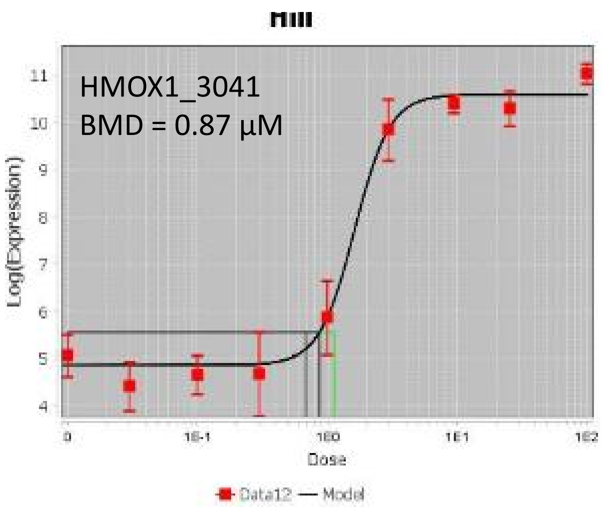


CYP1B1_17315
BMD = 0.28 µM

HMOX1_3041
(n = 174)

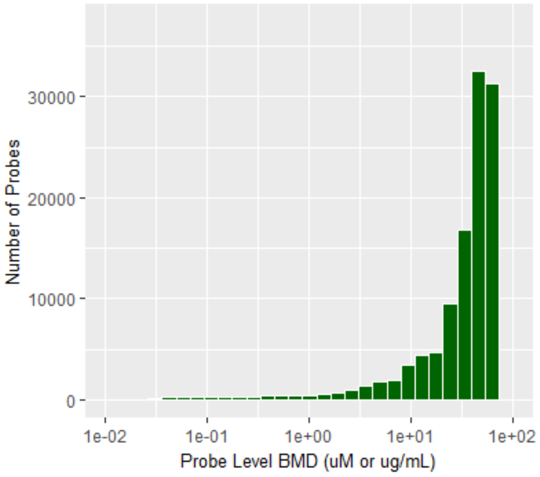
Sodium dimethyldithiocarbamate
128-04-1 | DTXSID6027050

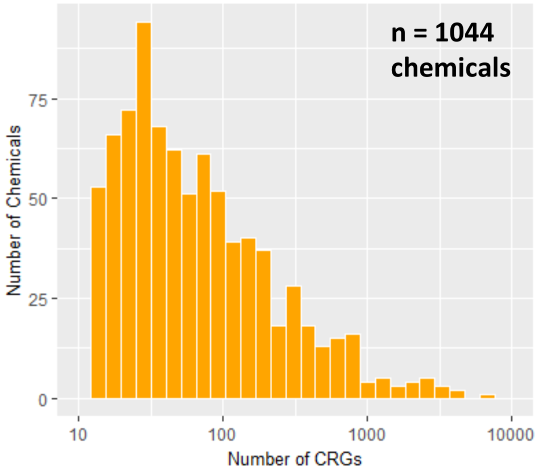




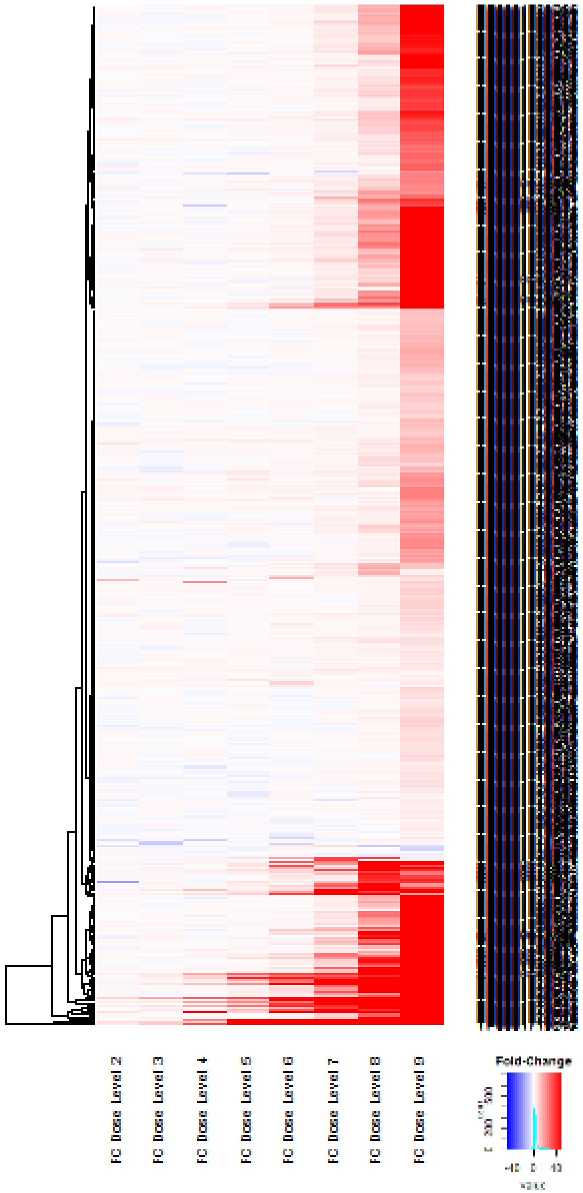
HMOX1_3041
BMD = 0.87 µM

Benchmark Dose Modeling Summary & Inducible Genes

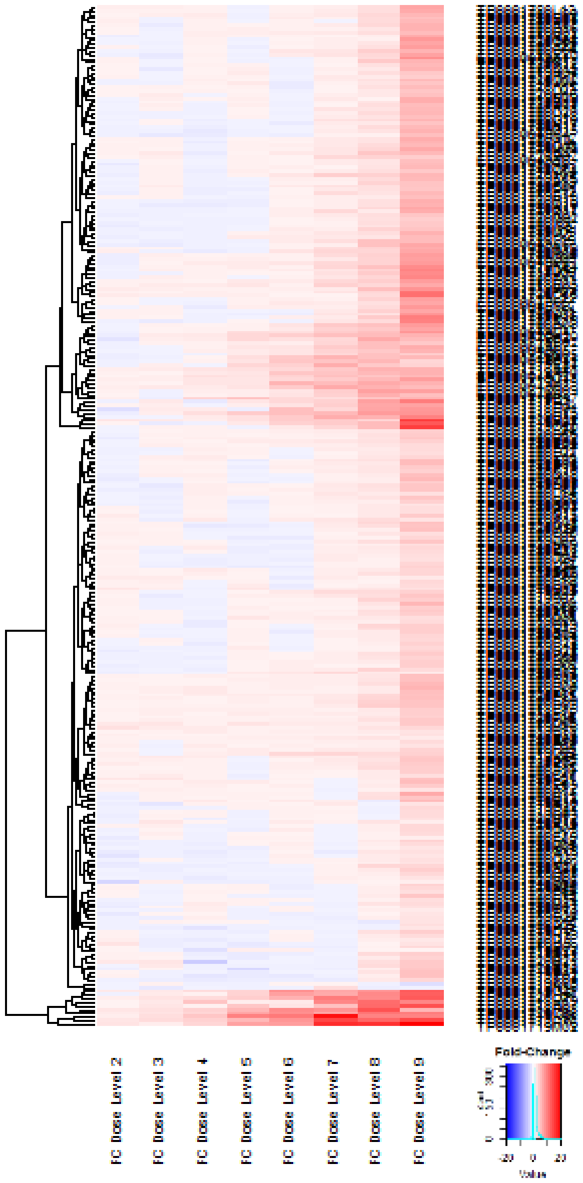



n = 1044 chemicals

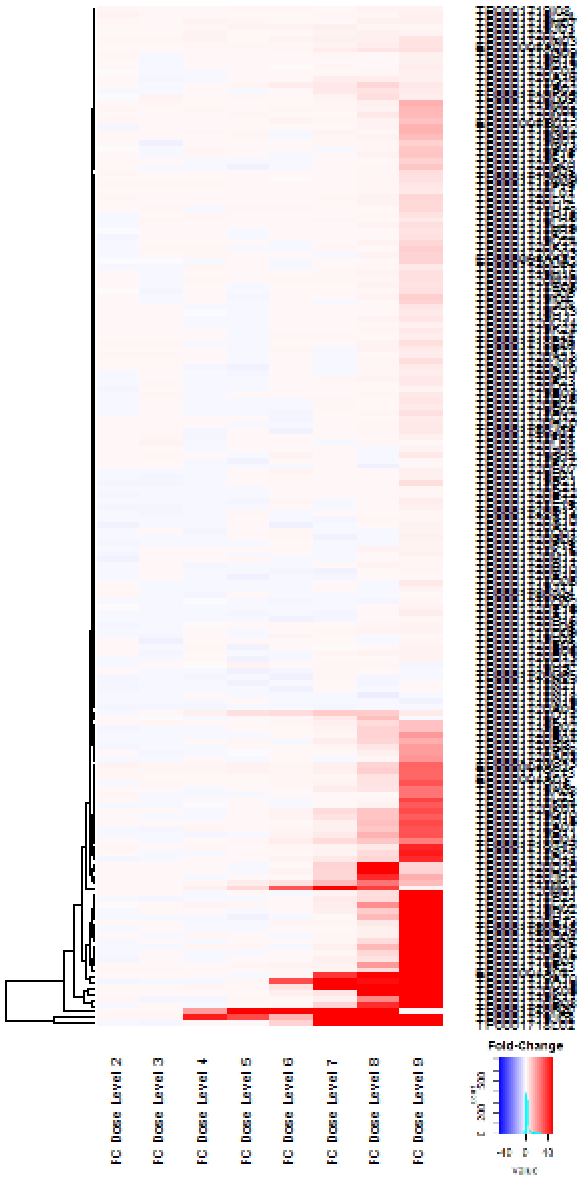
CYP1A1_10775 (n = 473)



CYP1B1_17315 (n = 279)

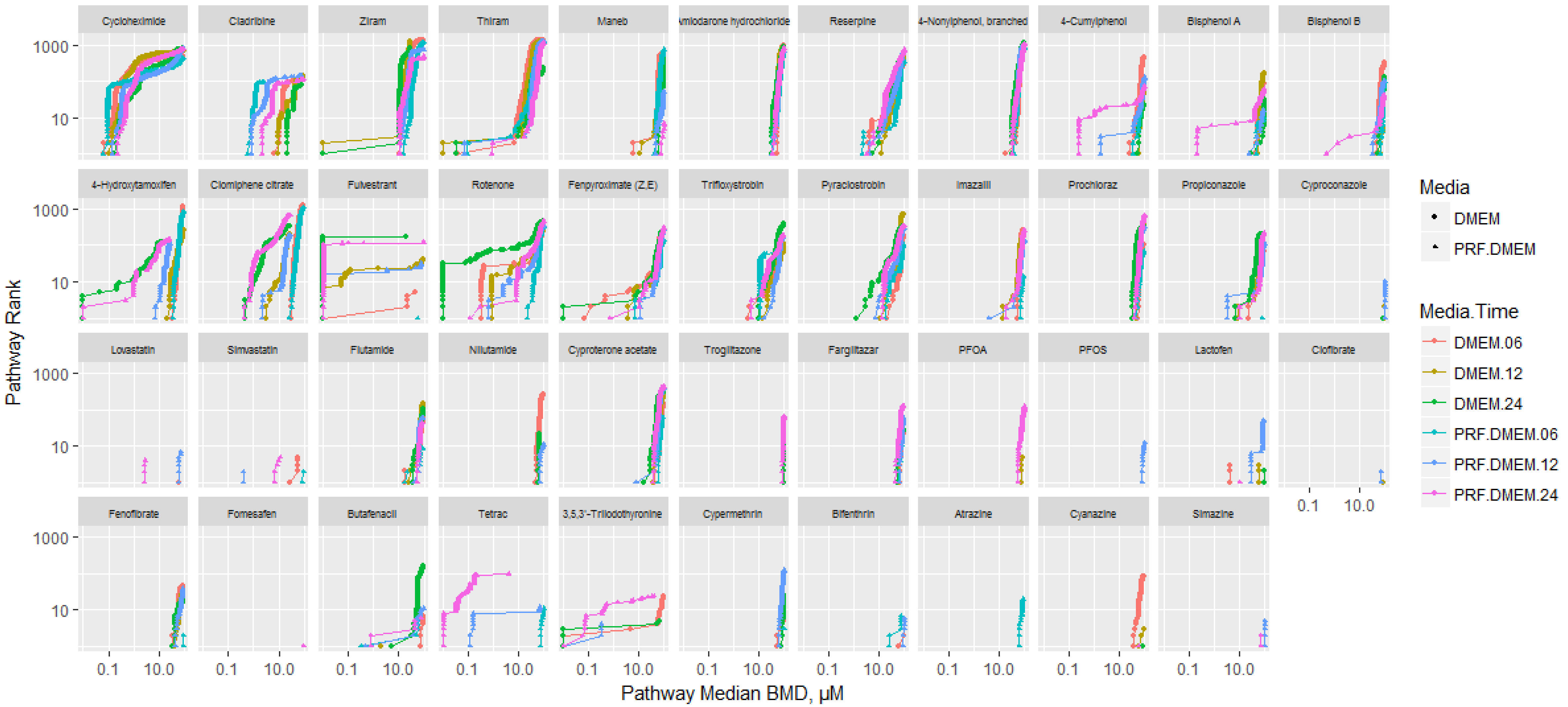


HMOX1_3041 (n = 174)



Gene Set Accumulation Plots (1) *log_{10} x-axis scaling*

Reactome Pathway Collection



- Identification of the most sensitive gene set / pathway (or lower %ile of affected pathways) is a common way to identify bioactivity thresholds in transcriptomics data.
- Some chemicals affect many pathways across a broad concentration range (i.e. cycloheximide, ziram).
- Other affect a comparatively smaller number of pathways within a narrow concentration range (i.e. flutamide, prochloraz).

No affected pathways identified for Vinclozolin

Concentration-Response Analysis

# of Pathways

Concentration Response Modeling

ToxCast Pipeline (tcpl):

- Originally developed for CR modeling of high-throughput targeted screening assays.
- Fits 3 Models:
  - Constant
  - Hill
  - Gain-Loss
- Winning model = Lowest AIC
- “Hits” are defined as curves where:
  - The Hill or Gain-Loss wins
  - Response surpasses an efficacy threshold
- Modified to handle both upwards and downwards trending concentration-response curves.

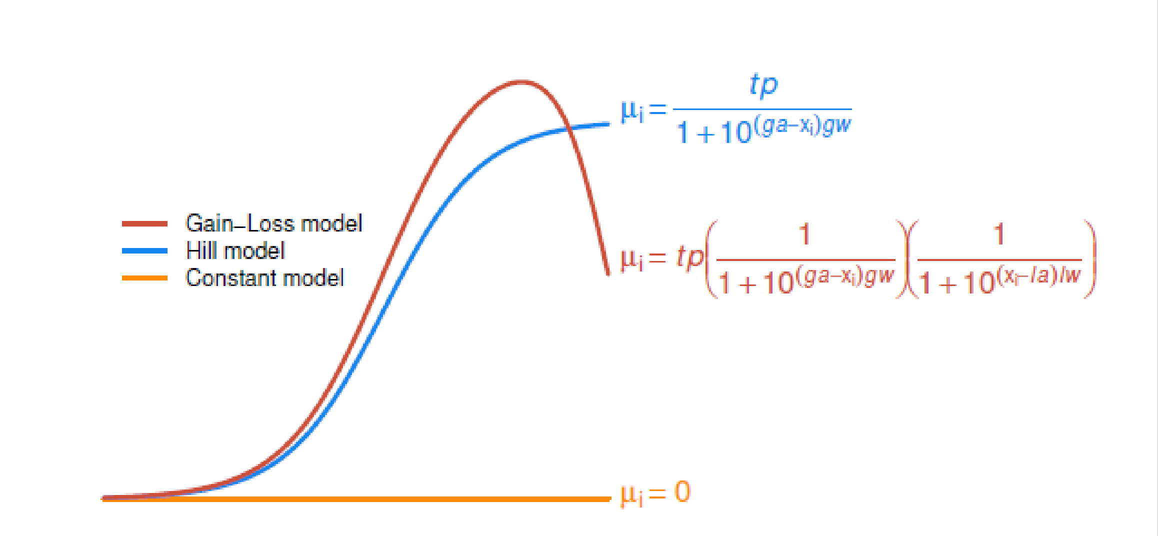


Fig. 1. The three models utilized by the tcpl package. The constant model and its associated formula for μ_i is shown in orange, the Hill model and its associated formula for μ_i is shown in blue, and the gain-loss model and its associated formula for μ_i is shown in blue.

Filer et al. (2017)

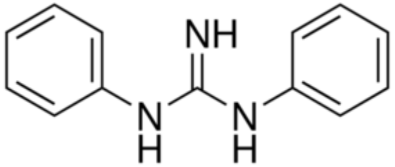
Applications in HTTr:

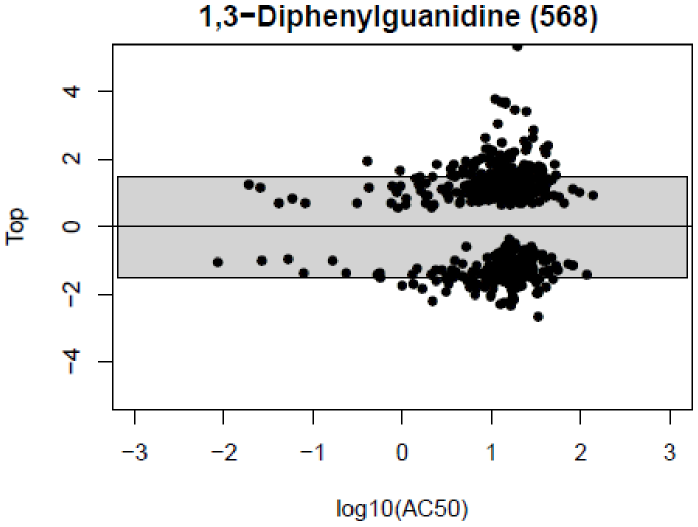
- Gene level concentration-response modeling of DESeq2 FC estimates.
- Pathway level concentration-response modeling

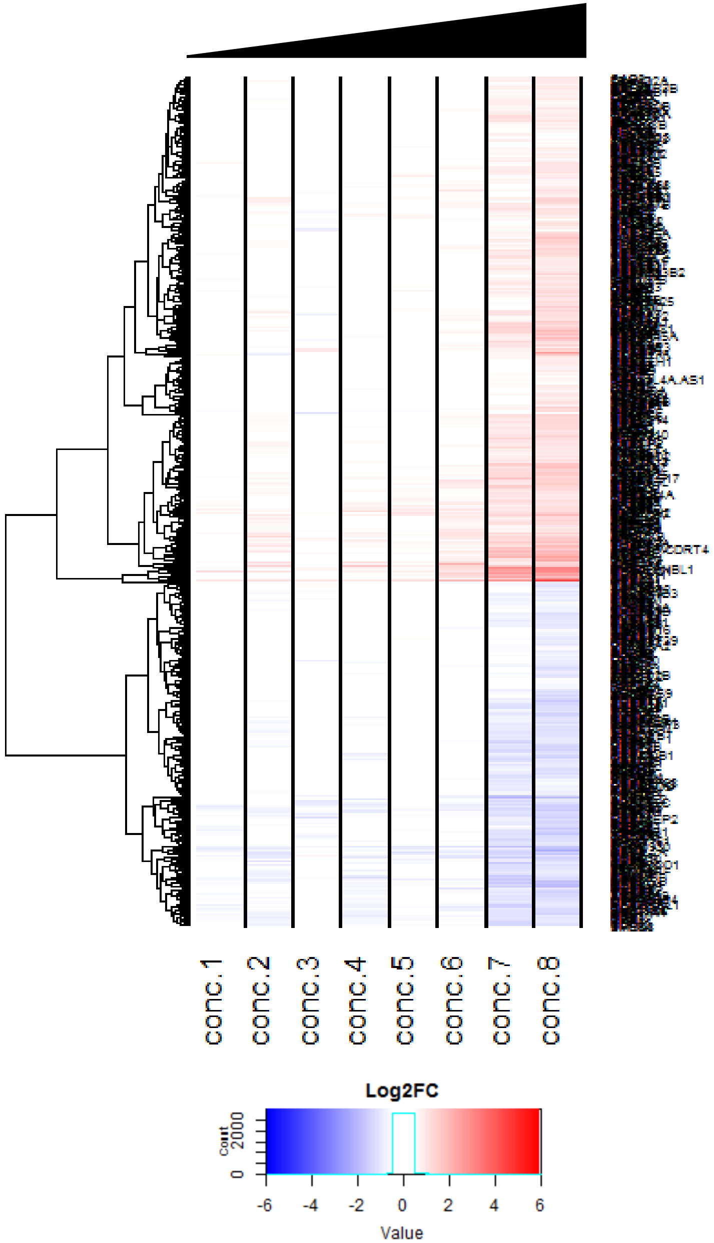
Gene Level CR Modeling Example

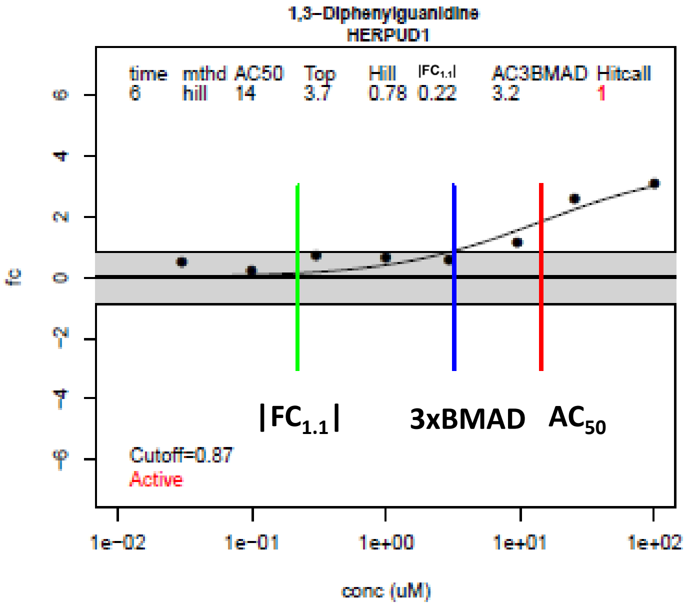
1,3-Diphenylguanidine

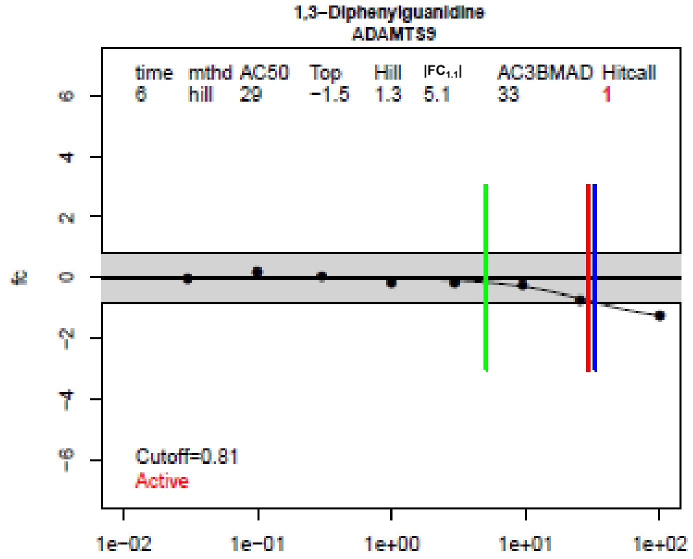
102-06-7 | DTXSID3025178











Gene Set Level Concentration Response Modeling

Step 1: Calculate Response

- A gene set is a list / bag of genes
- Under one condition (chemical x dose) calculate “gene set response” separately for genes in the set and out of the set:

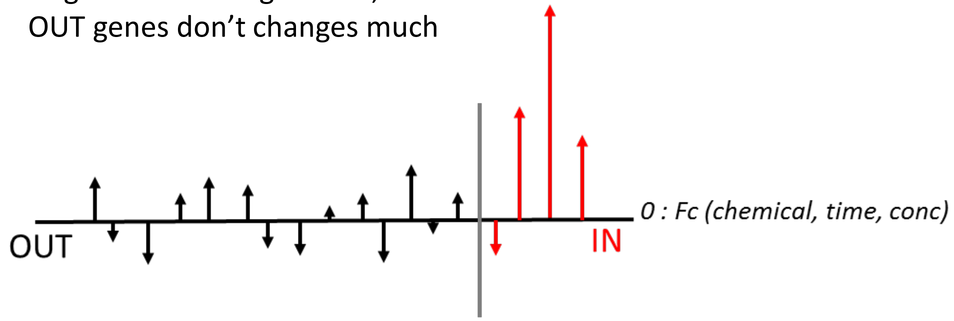
$$M = \sum_{i=1}^{ngene} fc_i$$

$$R = M_{in} - M_{out}$$

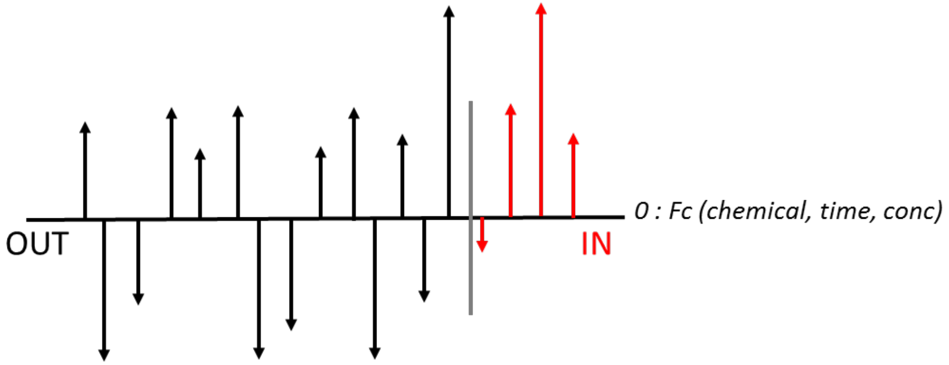
Step 2: CR Modeling

- For each chemical, fit using tcplFit
  - Constant, Hill , Gain-Loss methods
  - BMAD(pathway) = MAD of response for the pathway across the two lowest concentrations across all chemicals and times
- Hitcall:
  - tcplFit calls a hit
  - Top > 3*BMAD

- IN genes are changed a lot, and are coherent in direction
- OUT genes don’t changes much



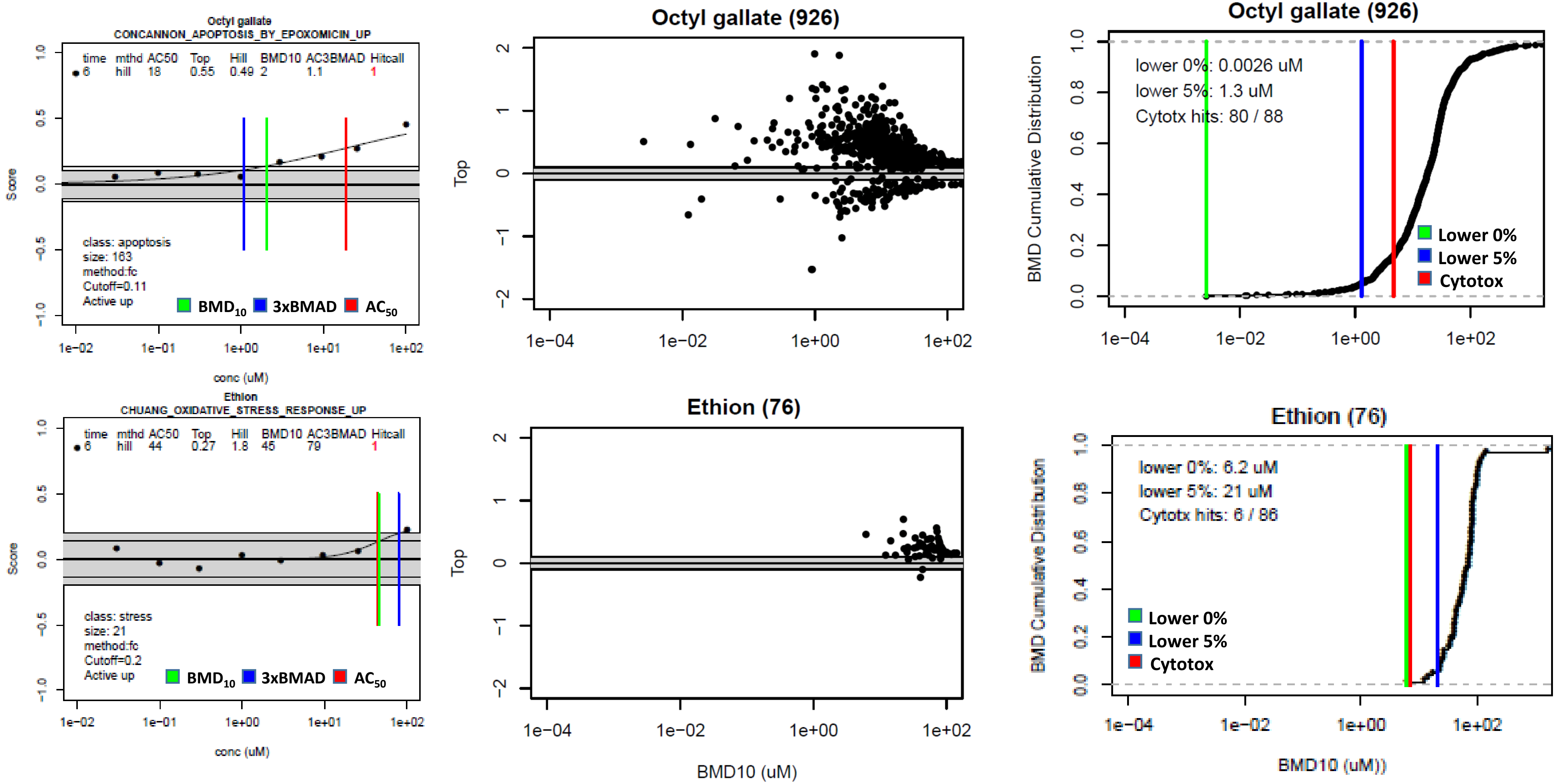
- IN genes are changed a lot, and are coherent in direction
- OUT genes change a lot but are not coherent (mean ~ 0)



Gene Set Collections:

- **MSigDB_C2:** Curated gene sets from online pathway databases, publications and knowledge of domain experts (n = 4738).
- **BioPlanet:** Curated pathway set developed by National Toxicology Program (n = 1700).

Gene Set Level CR Modeling Examples

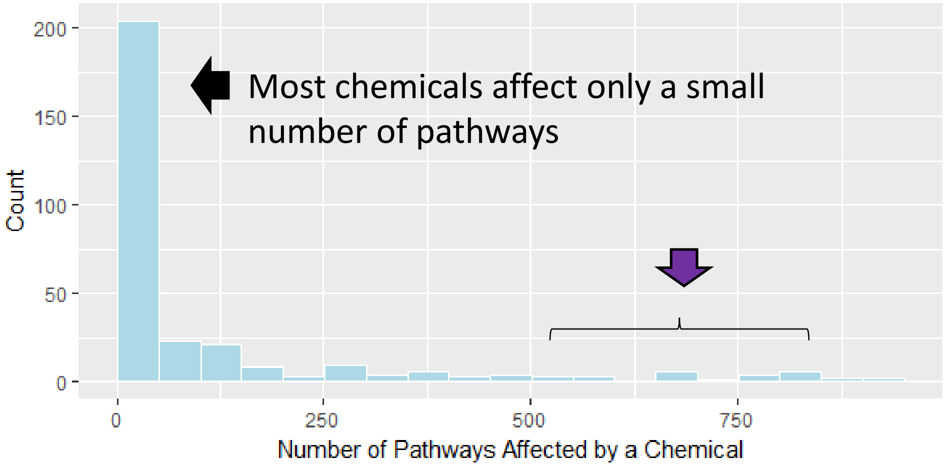


time	mthd	AC50	Top	Hill	BMD10	AC3BMAD	Hitcall
6	hill	18	0.55	0.49	2	1.1	1

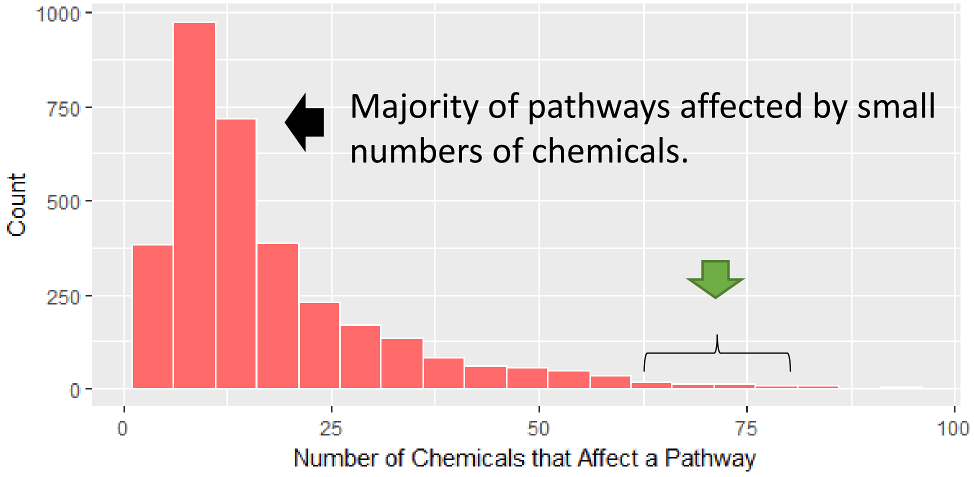
time	mthd	AC50	Top	Hill	BMD10	AC3BMAD	Hitcall
6	hill	44	0.27	1.8	45	79	1

- **Top Row:** Chemical produced effects on biological pathways at concentrations **below** cytotoxicity.
- **Bottom Row:** Chemical produced effects on biological pathways at or **above** the cytotoxicity threshold.

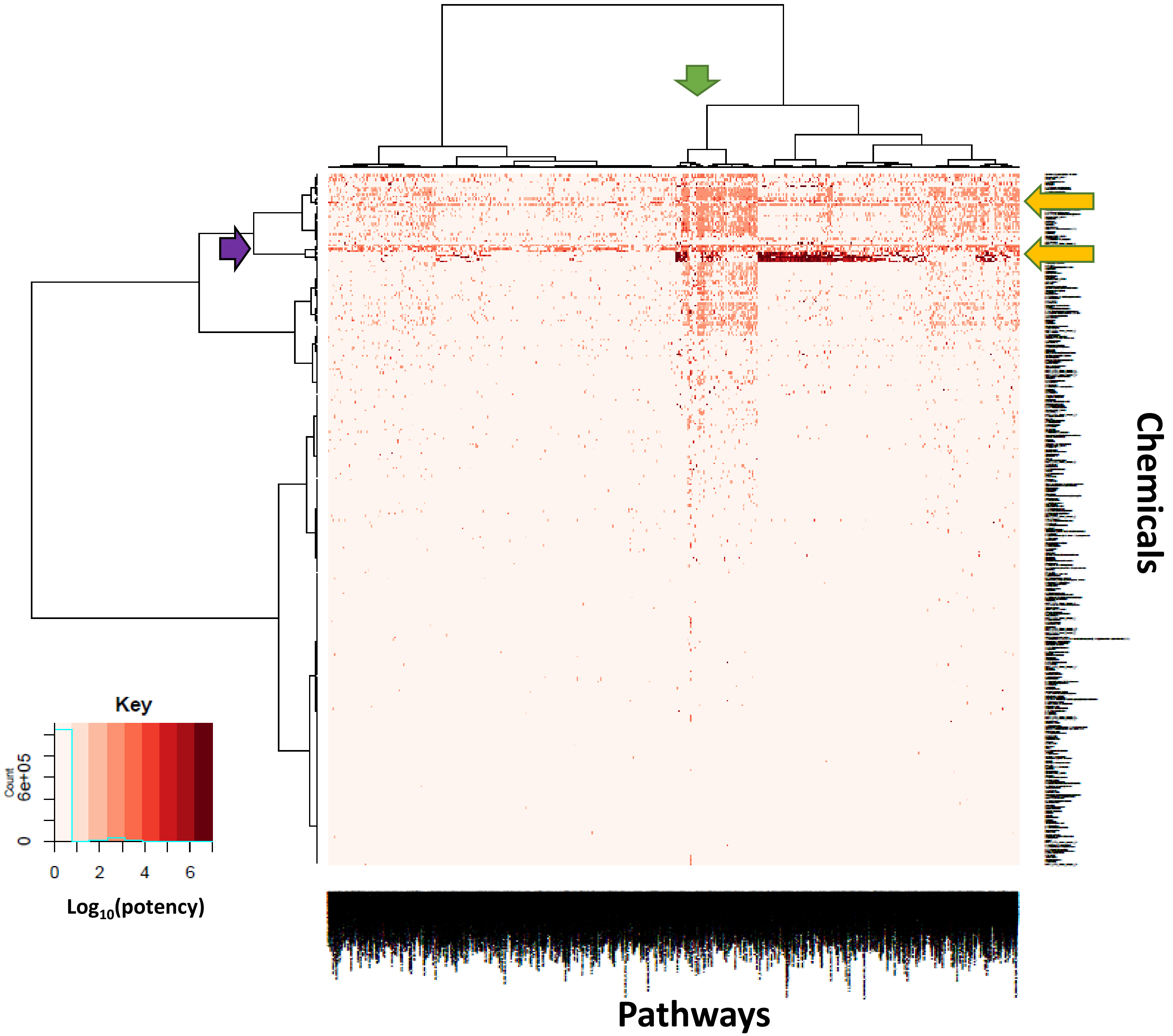
Gene Set Level CR Summary



Most chemicals affect only a small number of pathways



Majority of pathways affected by small numbers of chemicals.



➡ Similarity in the pattern of chemical responses

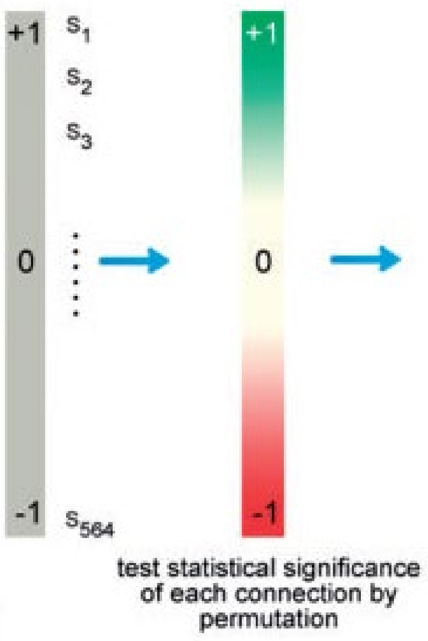
Putative Target Prediction

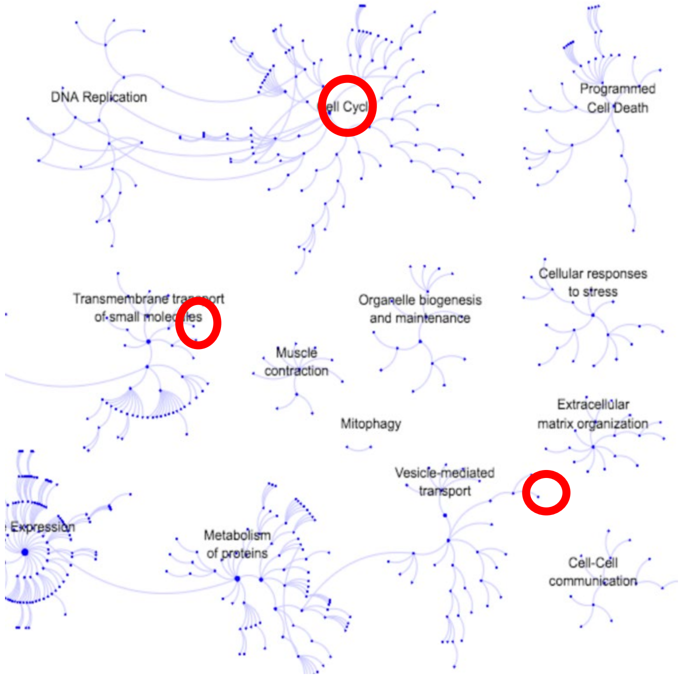
Putative Molecular Target Prediction

Connectivity mapping analysis using DEGs and CRGs

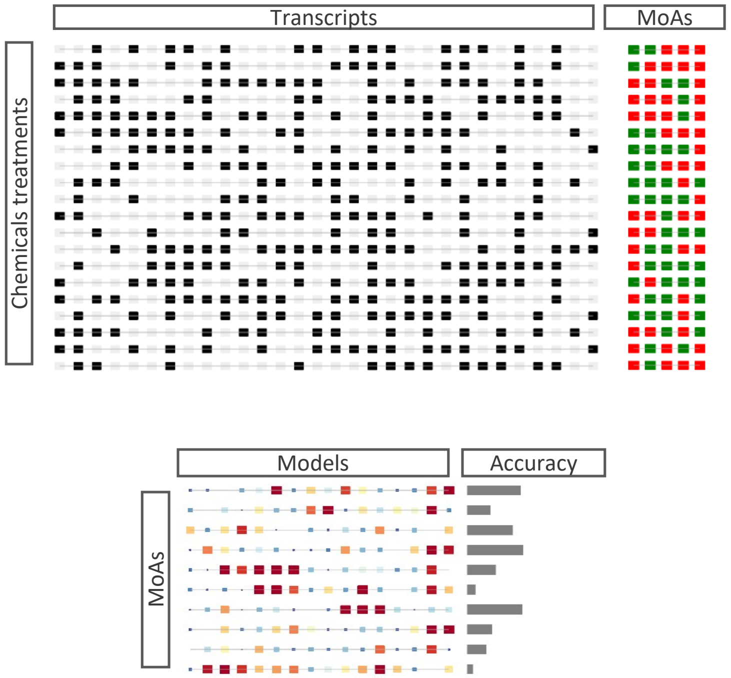
Pathway / Network analysis using DEGs and CRGs

Machine learning to build Target-specific models




DNA Replication
ell Cycl
Programmed Cell Death
Transmembrane tra port of small mole les
Organelle biogenesis and maintenance
Cellular responses to stress
Muscle contraction
Mitophagy
Extracellular matrix organization
Vesicle-mediated transport
Expression
Metabolism of proteins
Cell-Cell communication

reactome.org

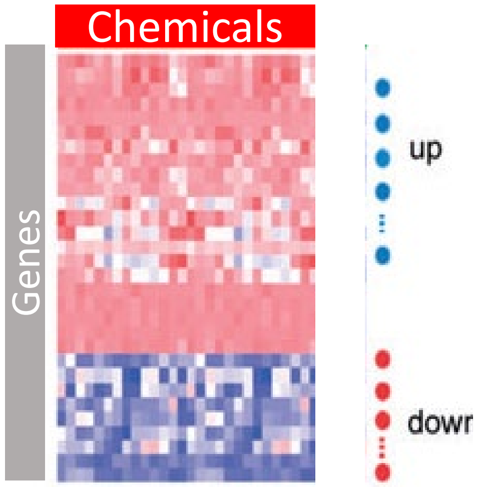


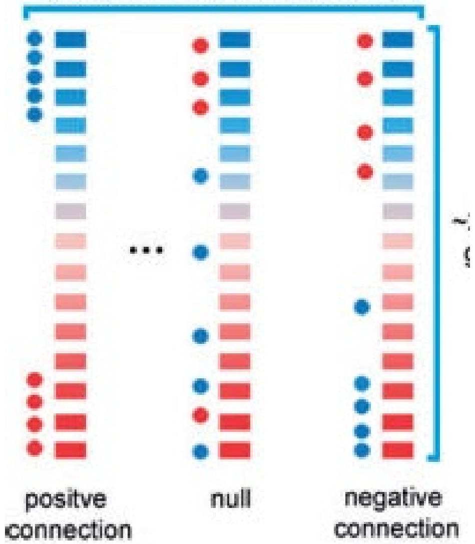
Connectivity Mapping

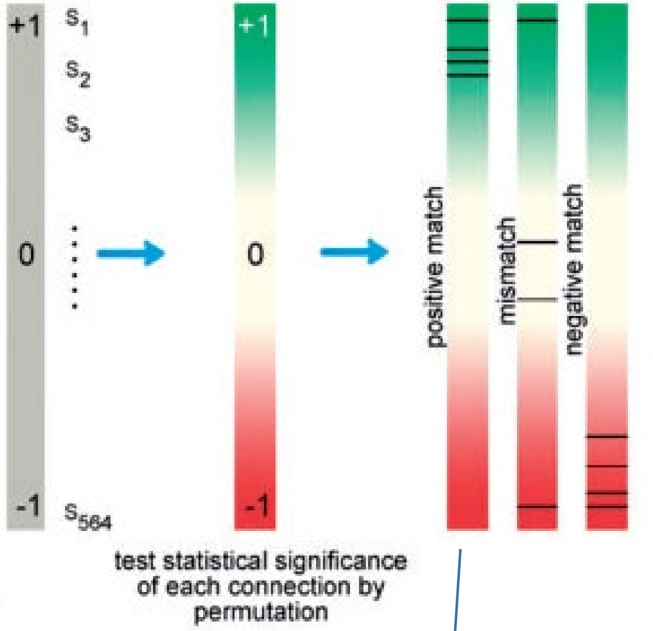
Input DEGs or CRGs

Query Signature DB CMap or BSP

Find best positive matches







BioSpyder HTTr (BSP)

Infer Tox/MoA by best match

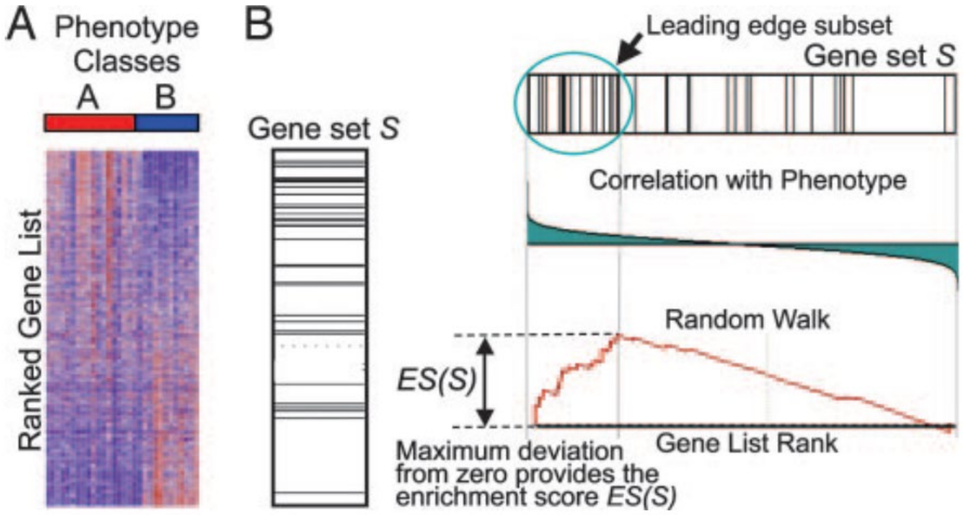
Issues

- Translating DEG/CRG to signature
- Many measures of similarity
- Only as good as reference chemical MoA annotation
- Highly sensitive but not very specific
- Chemicals that cause global perturbations “hit” all classes – how do we distinguish signal from noise ?

Lamb *et al* (2006)
Musa *el al* (2017)

“Connectivity” Scoring

- Connectivity mapping is a similarity metric based on transcriptional descriptors
- Gene Set Enrichment Analysis (GSEA): Calculate score of signature with highly up or down regulated genes in reference profiles using KS statistics
- Many alternatives
  - ssCMap: subspace connectivity mapping based on DEGs
  - ProbCMap: probabilistic scoring based on latent factors
  - XCos: Cosine similarity based on overlapping genes
- We used GSEA in this analysis



Signature

Database Profile

Enrich. Score

Subramanian et al. 2005

Reference Database and Signatures

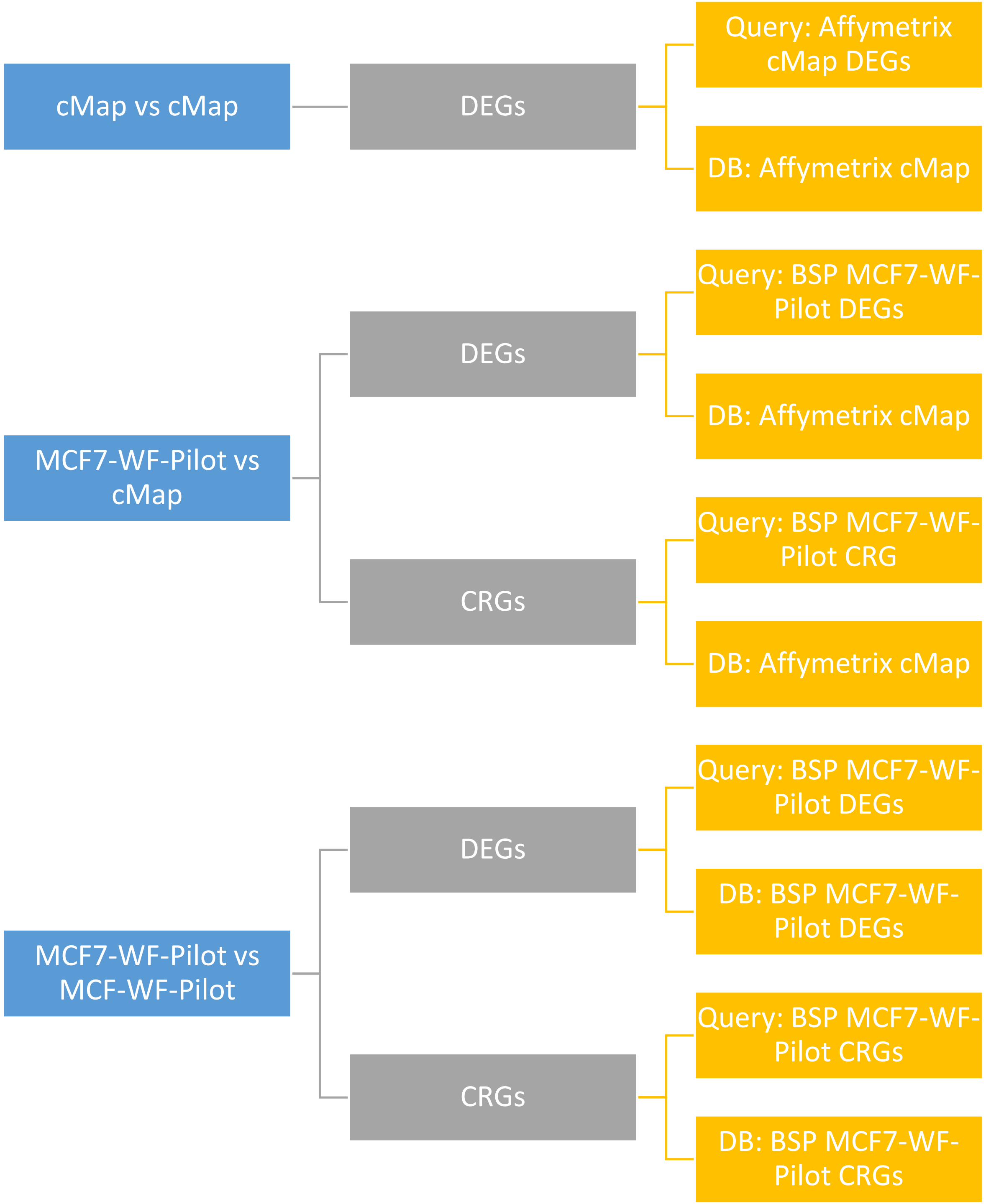
CMap Build 02

- CMap DB
  - Use CMap v2 database: Affymetrix data on 1176 chemicals, 5 cell lines
  - RMA Normalize CEL files
  - L2FC using treatment vs. matched DMSO
- Signatures (DEG)
  - Translate FC profiles in up/down profiles (signatures)
  - Convert L2FC data to Z-scores
  - DEG: For z0=1,2,3 create discrete Z where value = 1 if Z>z0 and -1 where Z<z0

MCF7-WF-Pilot BSP

- BSP DB
  - Use 44 chemicals x 8 conc x 3 times x 2 media combinations
  - Exclude probes with ave count<5
  - L2FC using DESeq2 (by chemical x 8 conc, time, media vs matched DMSO
- Signatures (DEG & CRG)
  - Convert L2FC data to Z-scores
  - |L2FC|>=0.6 & p<0.05 for at least one conc
  - DEG: For z0=1,2,3 create discrete Z where value = 1 if Z>z0 and -1 where Z<z0
  - CRG: Calc 1-way ANOVA on L2FC p<0.05

Connectivity Mapping Comparisons



cMap vs cMap

DEGs

Query: Affymetrix cMap DEGs

DB: Affymetrix cMap

MCF7-WF-Pilot vs cMap

DEGs

Query: BSP MCF7-WF-Pilot DEGs

DB: Affymetrix cMap

CRGs

Query: BSP MCF7-WF-Pilot CRG

DB: Affymetrix cMap

MCF7-WF-Pilot vs MCF-WF-Pilot

DEGs

Query: BSP MCF7-WF-Pilot DEGs

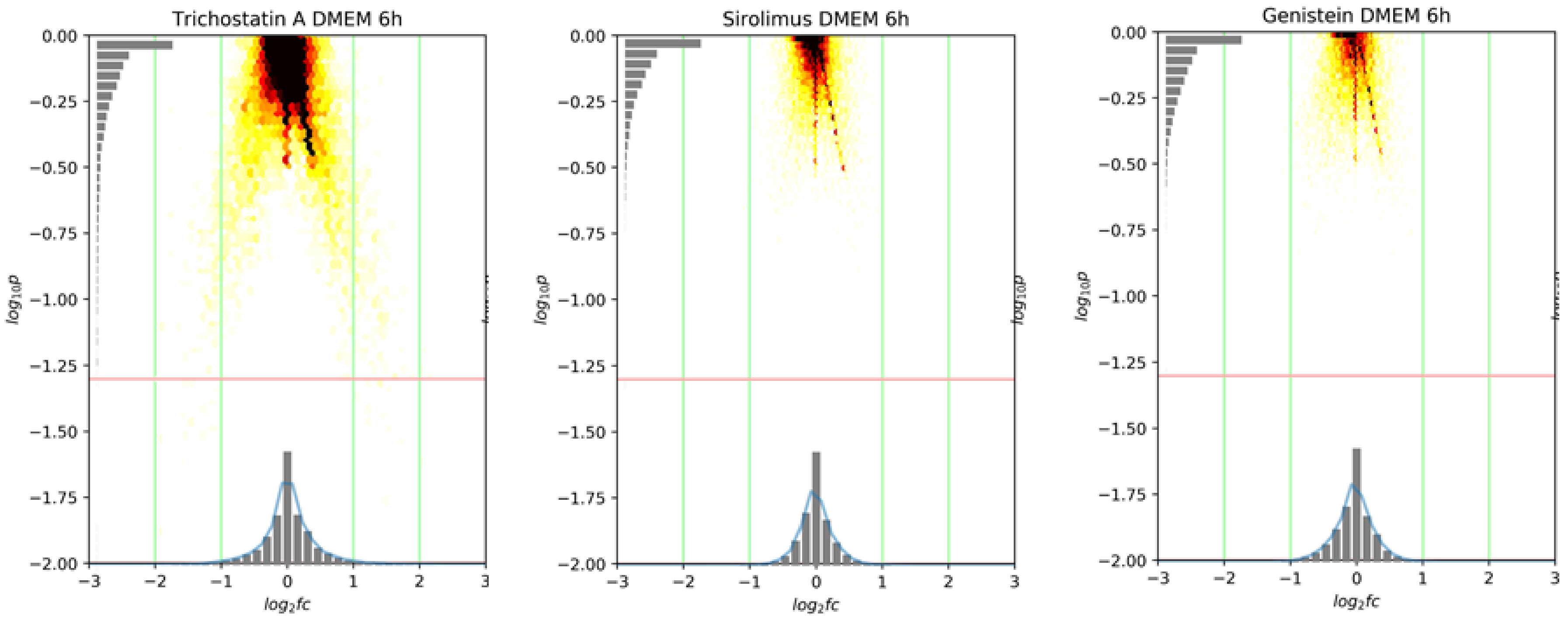
DB: BSP MCF7-WF-Pilot DEGs

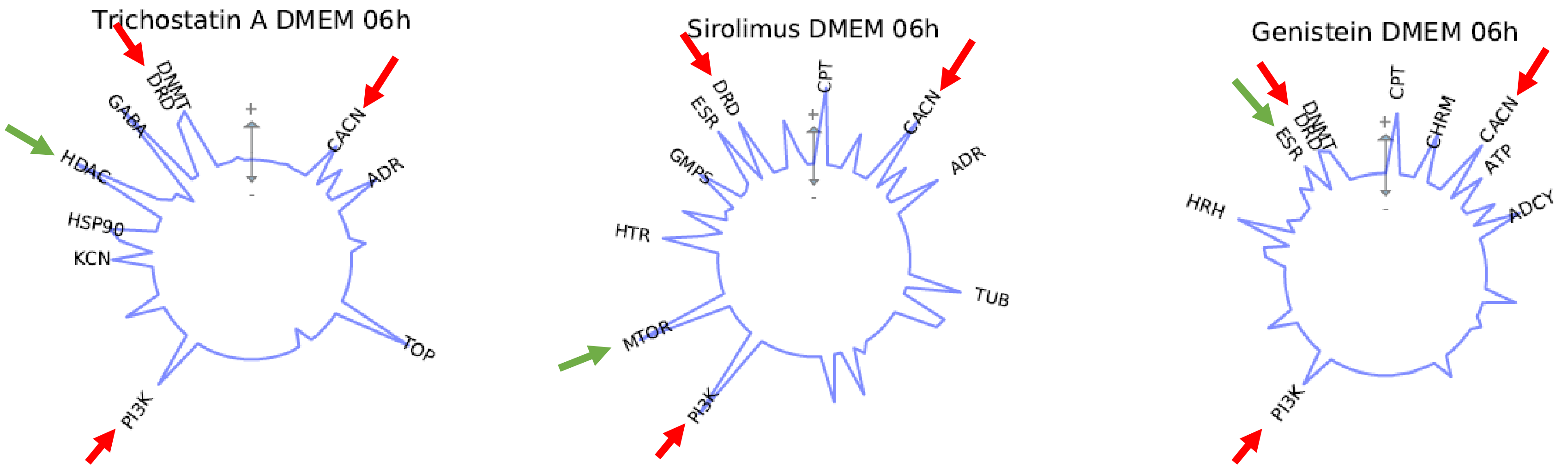
CRGs

Query: BSP MCF7-WF-Pilot CRGs

DB: BSP MCF7-WF-Pilot CRGs

Connectivity Mapping (MCF7-Pilot vs CMap)

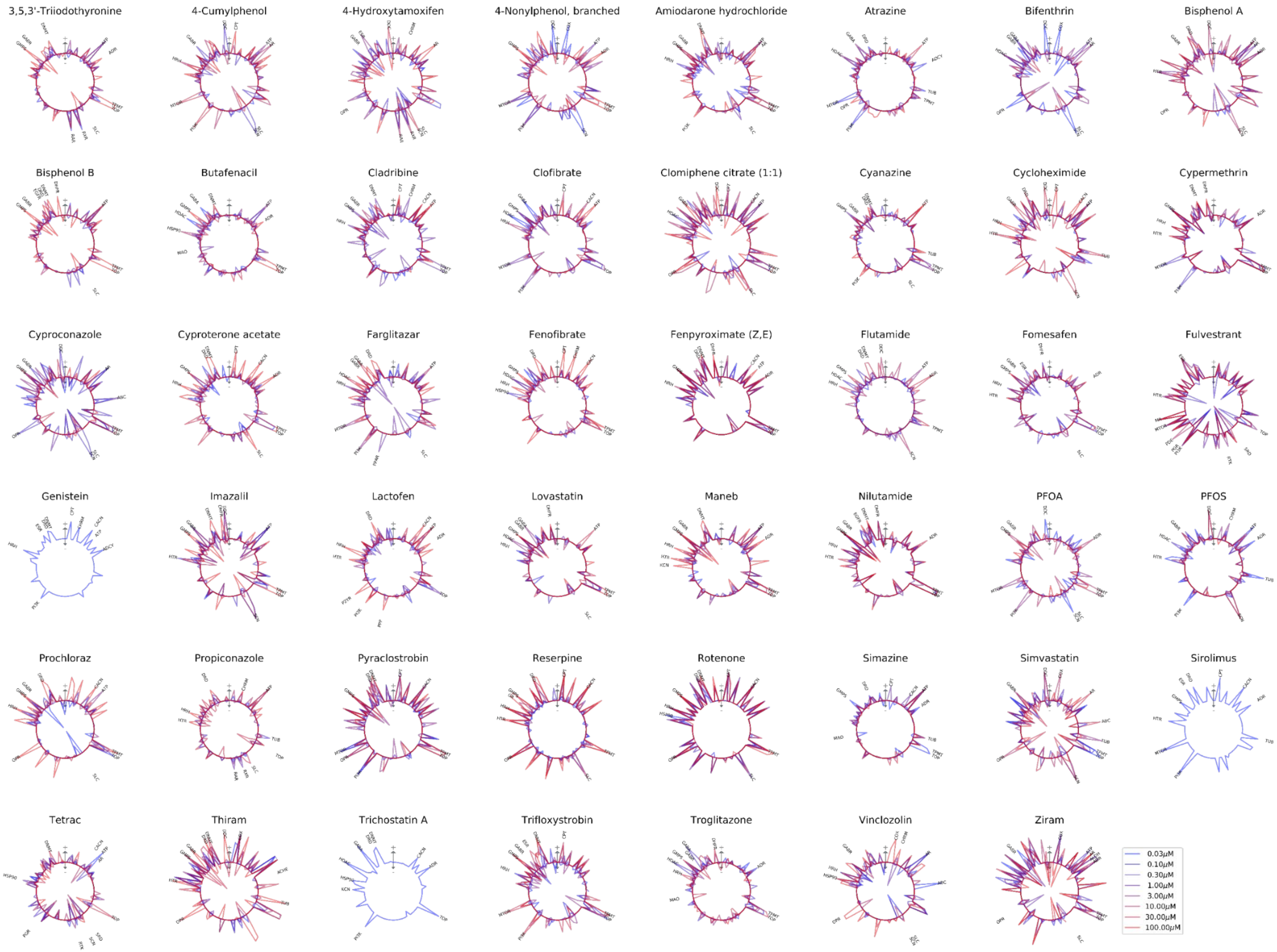




- Differential gene expression observed with reference chemicals.
- Putative targets identified using Connectivity Mapping
- Large degree of promiscuity of predicted targets observed.
- Currently evaluating additional methods for MIE prediction

→ *Putative target*

→ *Promiscuous Target Mapping*


3,5,3'-Triiodothyronine
4-Cumylphenol
4-Hydroxytamoxifen
4-Nonylphenol, branched
Amiodarone hydrochloride
Atrazine
Bifenthrin
Bisphenol A
Bisphenol B
Butafenacil
Cladribine
Clofibrate
Clomiphene citrate (1:1)
Cyanazine
Cycloheximide
Cypermethrin
Cyproconazole
Cyproterone acetate
Farglitazar
Fenofibrate
Fenpyroximate (Z,E)
Flutamide
Fomesafen
Fulvestrant
Genistein
Imazalil
Lactofen
Lovastatin
Maneb
Nilutamide
PFOA
PFOS
Prochloraz
Propiconazole
Pyraclostrobin
Reserpine
Rotenone
Simazine
Simvastatin
Sirolimus
Tetrac
Thiram
Trichostatin A
Trifloxystrobin
Troglitazone
Vinclozolin
Ziram
0.03μM
0.10μM
0.30μM
1.00μM
3.00μM
10.00μM
30.00μM
100.00μM

Quantifying Performance

Conduct Leave-one-out (LOO) evaluation of hits:

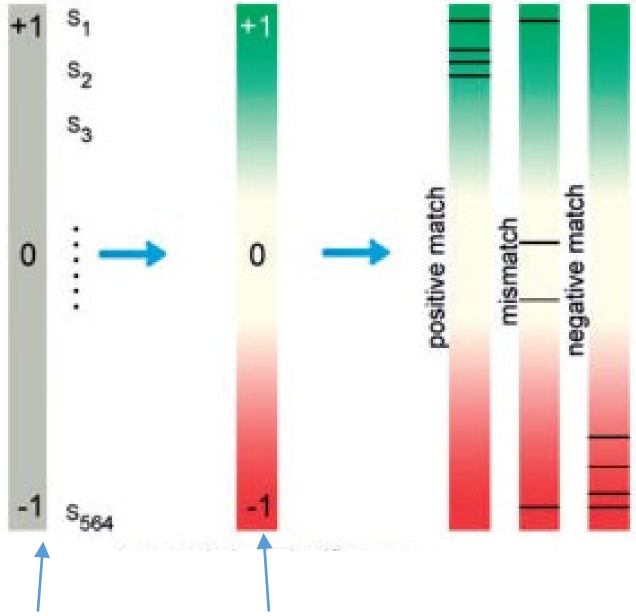
1. Annotate CMap chemicals with classes
   - Classes: 143 (Putative targets)
   - Chemicals: 614
2. Search "hits" by connectivity with score= ϑ
   - If $\vartheta > \vartheta_0$

     ```
       if query.target== hit.target:
                pred=TP
     elif query.target!= hit.target:
                pred=FP
     ```

   - If hit $\vartheta < \vartheta_0$

     ```
       if query.target== hit.target:
                pred=FN
     elif query.target!= hit.target:
                pred=TN
     ```
3. Measure sensitivity, specificity, BA

Connectivity Mapping



Query Signature

DB Signature

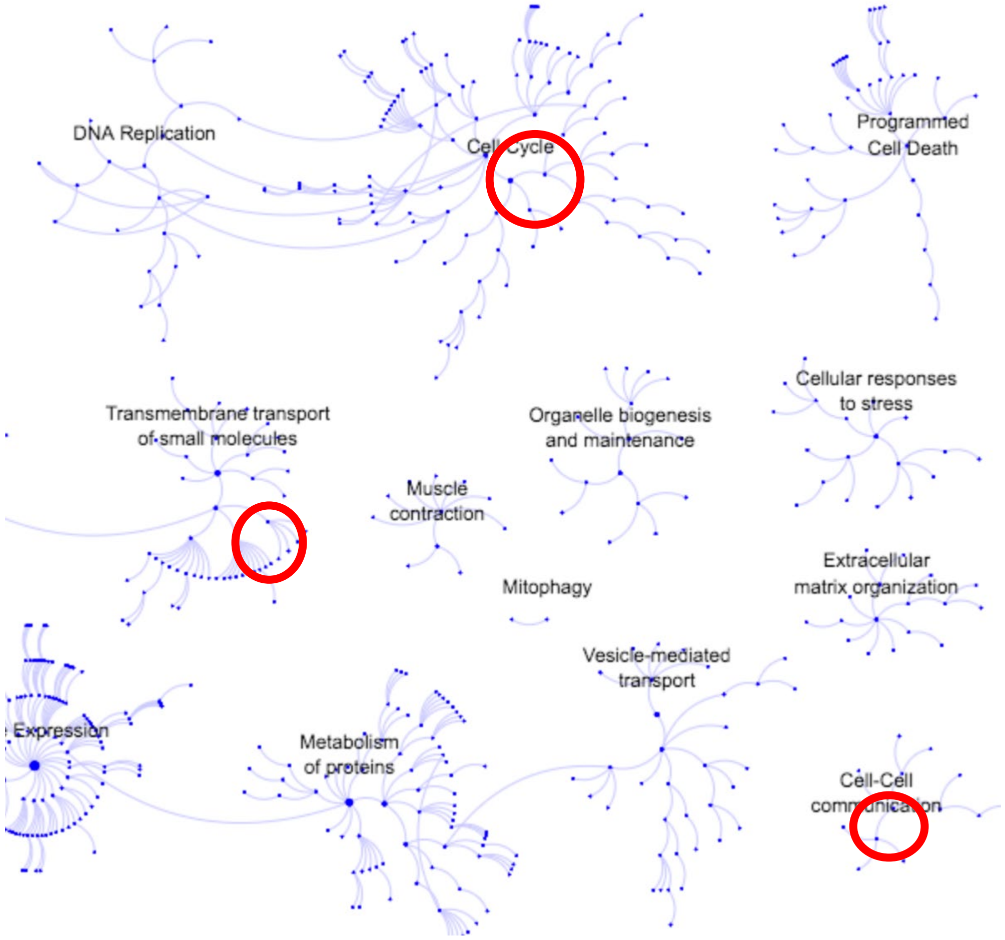
cMap 2.0 vs cMap 2.0

MoA	pos	neg	pos_annot	BA	Sn	Sp	th0
GABAT	2	117	2	0.85	1.00	0.71	0.19
HDAC	3	144	6	0.84	1.00	0.69	0.16
RAR	2	63	2	0.83	1.00	0.66	0.13
TUB	5	172	5	0.83	1.00	0.65	0.14
FKBP	2	41	2	0.82	1.00	0.63	0.33
HPRT	2	77	2	0.81	1.00	0.63	0.09
OPR	5	157	6	0.81	1.00	0.63	0.23
DNMT	2	32	2	0.81	1.00	0.63	0.28
DDC	2	84	2	0.81	1.00	0.62	0.17
TPO	2	78	3	0.81	1.00	0.62	0.04
DAT	2	73	3	0.81	1.00	0.62	0.03
PLG	2	71	3	0.81	1.00	0.62	0.13
DHFR	3	97	3	0.81	1.00	0.62	0.20
PTGER	4	113	4	0.81	1.00	0.62	0.07
NFKB	2	104	2	0.81	1.00	0.62	0.03
TR	2	82	2	0.81	1.00	0.62	0.14
ADORA	5	165	5	0.81	1.00	0.62	0.10
CHRN	4	139	6	0.81	1.00	0.62	0.06
TYMS	3	101	3	0.81	1.00	0.61	0.10
SRD	2	88	2	0.81	1.00	0.61	0.09

Pathway Analysis

Predicting Tox/MoA via Networks & Pathways

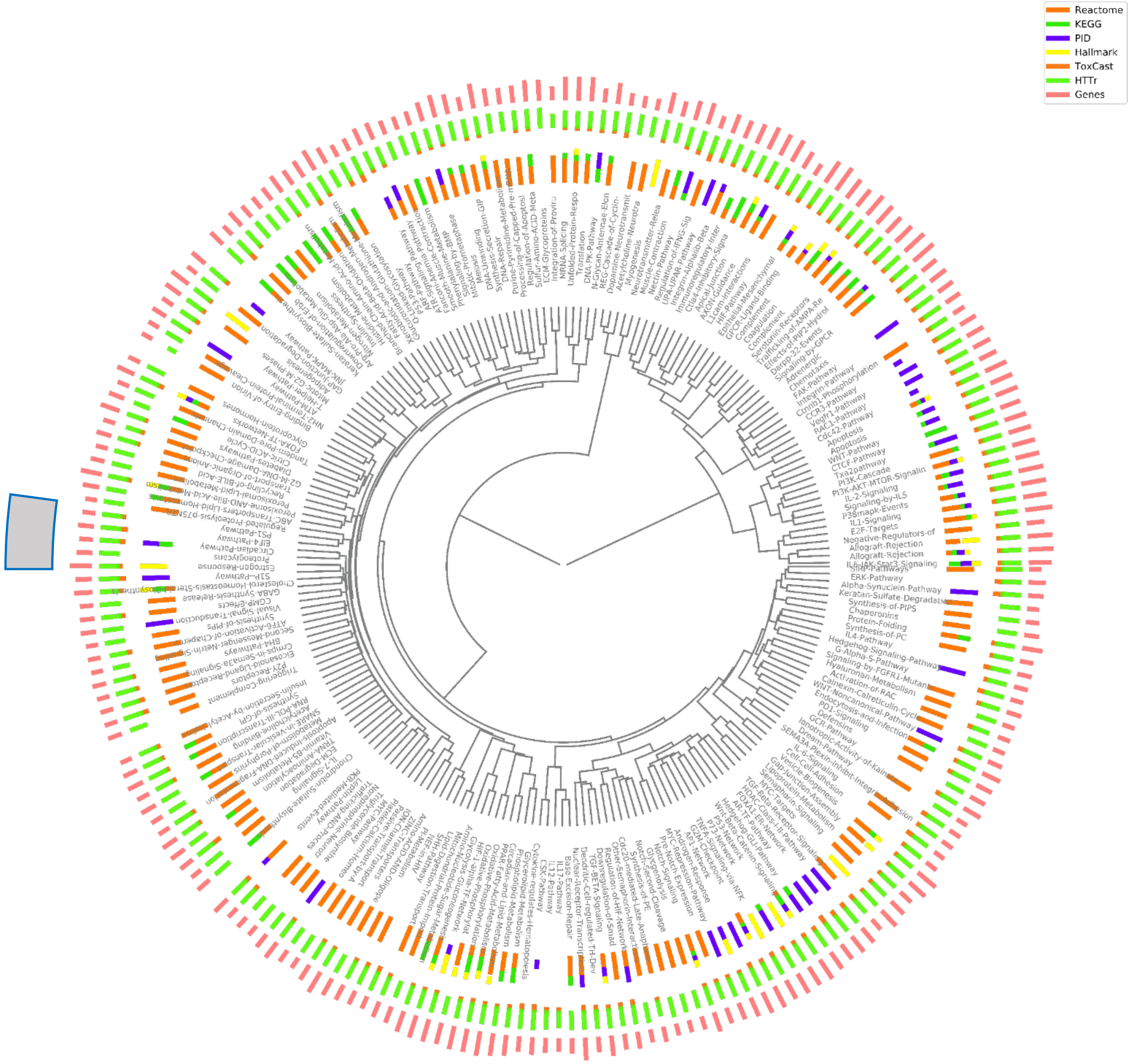
- Transcriptional perturbations of key pathways/interactions predicts Tox/MoA
- Pathway analysis
  - Select DEGs or CRGs to identify enriched pathways
  - Link enriched pathways to Tox/MoA
- Network analysis
  - Select DEGs or CRG to identify critical interactions
  - Link interactions to upstream or downstream targets
- Issues
  - Choice of pathway database
  - Scoring pathway/interaction enrichment
  - How do we objectively evaluate predictive accuracy
  - Effectively using signaling and genetic-regulatory network information
  - Linking pathways/interactions → MoA?


DNA Replication
Cell Cycle
Programmed Cell Death
Transmembrane transport of small molecules
Organelle biogenesis and maintenance
Cellular responses to stress
Muscle contraction
Mitophagy
Extracellular matrix organization
Vesicle-mediated transport
Expression
Metabolism of proteins
Cell-Cell communication

reactome.org

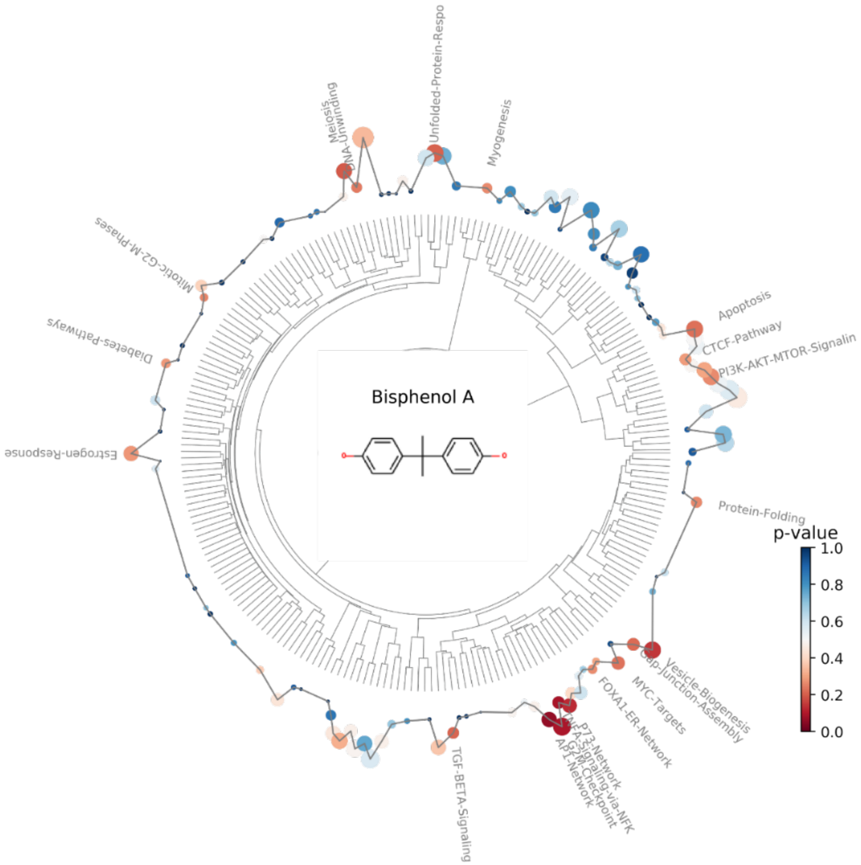
“Super-Pathways”

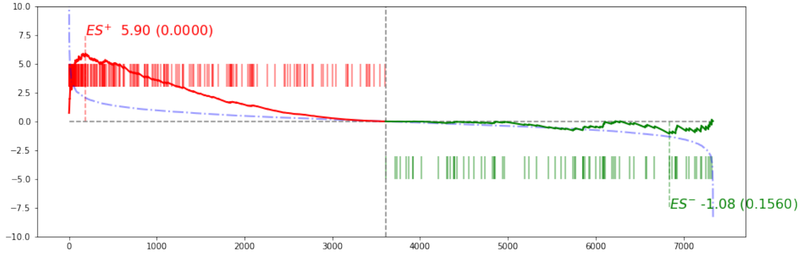
- Cluster Hallmark and canonical pathways (Reactome, KEGG, PID and BioCarta) from MSigDB V6 using genes
- Use hierarchical agglomerative clustering to organize super-pathways by similarity
- Each clade in the dendrogram shows groups of functionally related pathways
- Concentric rings show information about the source of information, HTTr coverage, and # of genes in each super-pathway

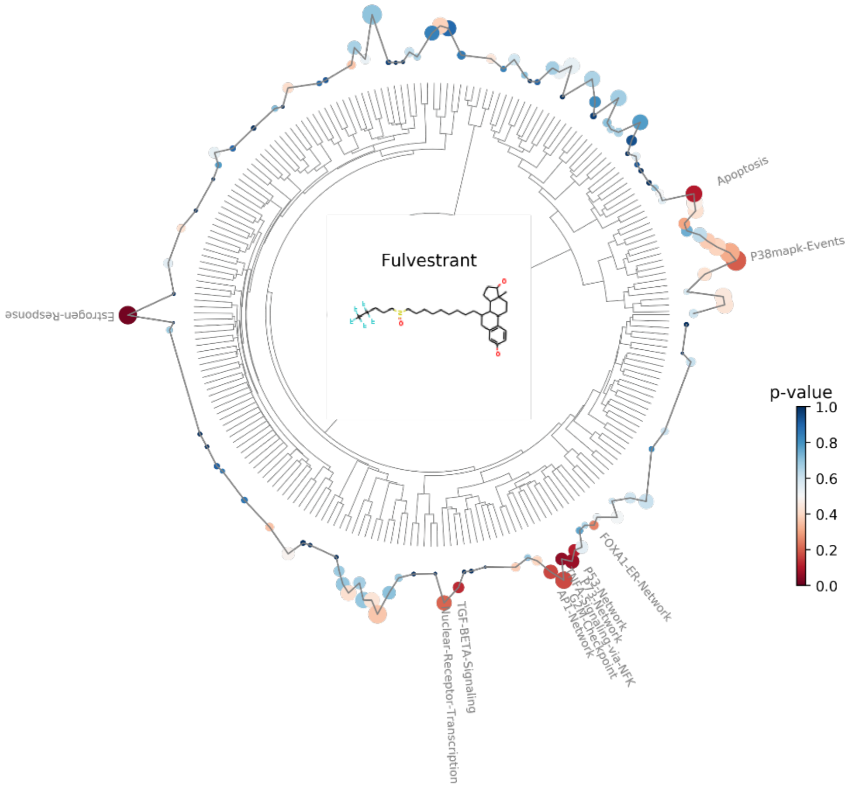


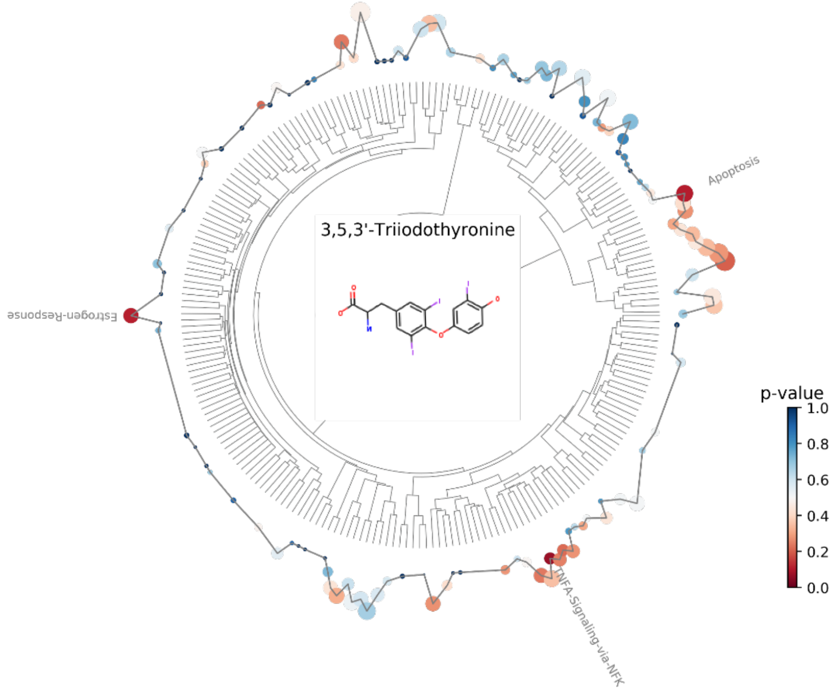
Pathway Analysis

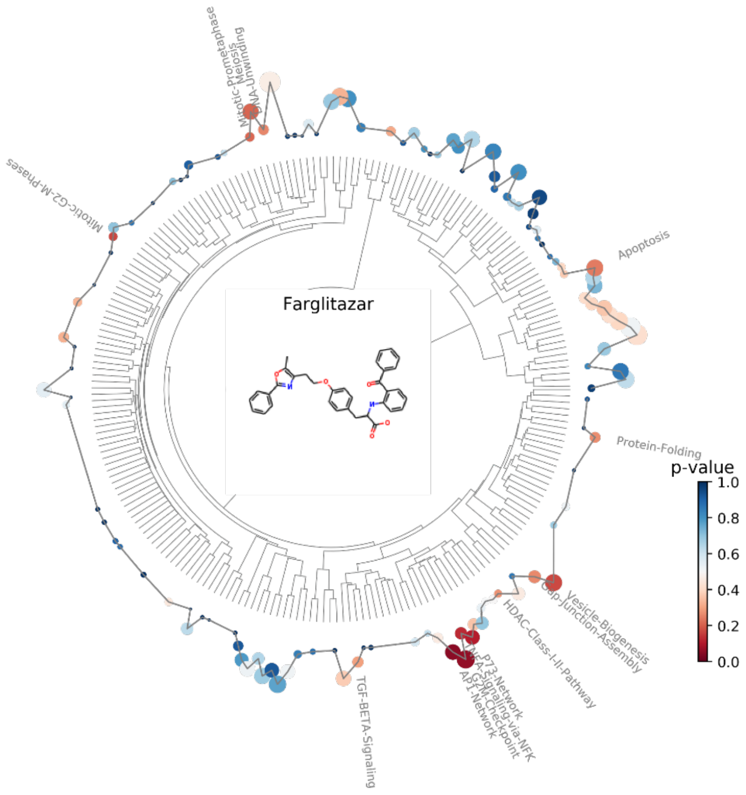
- The HTTr profiles for chemical treatments were searched against 224 super-pathways.
- Pathways were scored using different metrics that used the entire HTTr profile (e.g. enrichment scores), and just DEGs.
- The significance of scores was estimated by simulation.

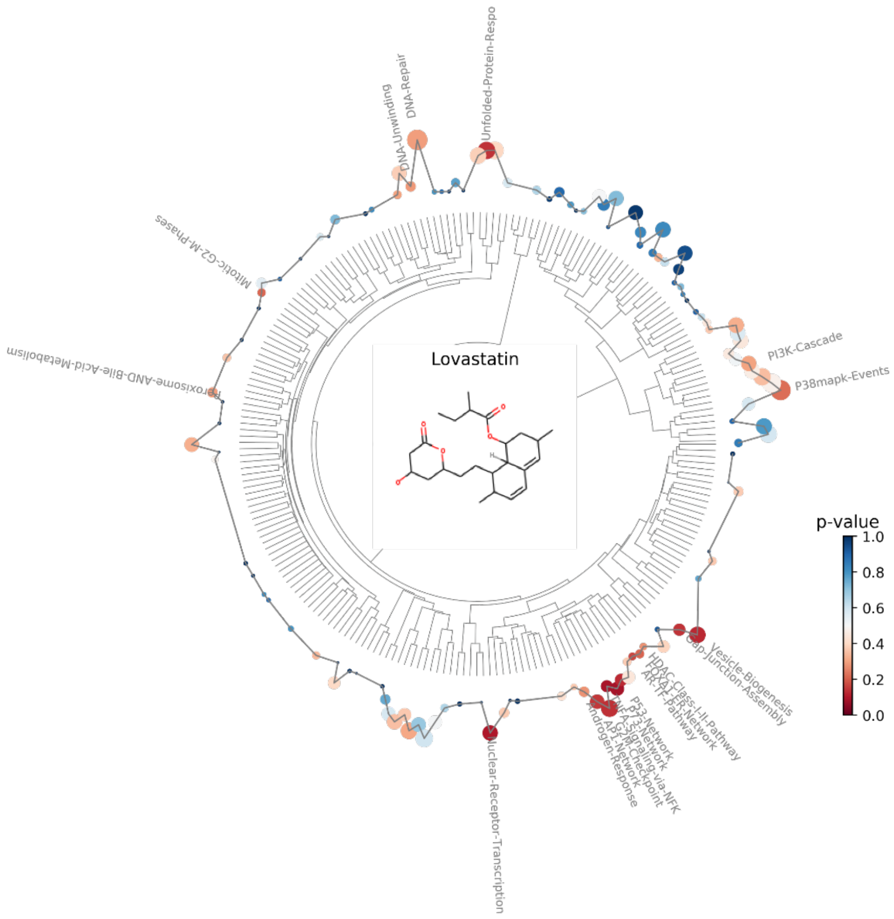





Fulvestrant
Apoptosis
P38mapk-Events
Estrogen-Response
FOXA1-ER-Network
P53-Network
P73-Network
TNFA-Signaling-via-NFK
G2M-Checkpoint
AP1-Network
TGF-BETA-Signaling
Nuclear-Receptor-Transcription
p-value
1.0
0.8
0.6
0.4
0.2
0.0


3,5,3'-Triiodothyronine
Apoptosis
Estrogen-Response
TNFA-Signaling-via-NFK
p-value
1.0
0.8
0.6
0.4
0.2
0.0


Farglitazar
Mitotic-Prometaphase
Meiosis
DNA-Unwinding
Mitotic-G2-M-Phases
Apoptosis
Protein-Folding
Vesicle-Biogenesis
Gap-Junction-Assembly
HDAC-Class-I-II-Pathway
P73-Network
TNFA-Signaling-via-NFK
G2M-Checkpoint
AP1-Network
TGF-BETA-Signaling
p-value
1.0
0.8
0.6
0.4
0.2
0.0


Lovastatin
DNA-Unwinding
DNA-Repair
Unfolded-Protein-Respo
Mitotic-G2-M-Phases
Peroxisome-AND-Bile-Acid-Metabolism
PI3K-Cascade
P38mapk-Events
Vesicle-Biogenesis
Gap-Junction-Assembly
HDAC-Class-I-II-Pathway
FOXA1-ER-Network
AR-TF-Pathway
P53-Network
P73-Network
TNFA-Signaling-via-NFK
G2M-Checkpoint
AP1-Network
Androgen-Response
Nuclear-Receptor-Transcription
p-value
1.0
0.8
0.6
0.4
0.2
0.0