High-Throughput Transcriptomics: From Screening to Pathways

Imran Shah
NCCT/ORD/EPA

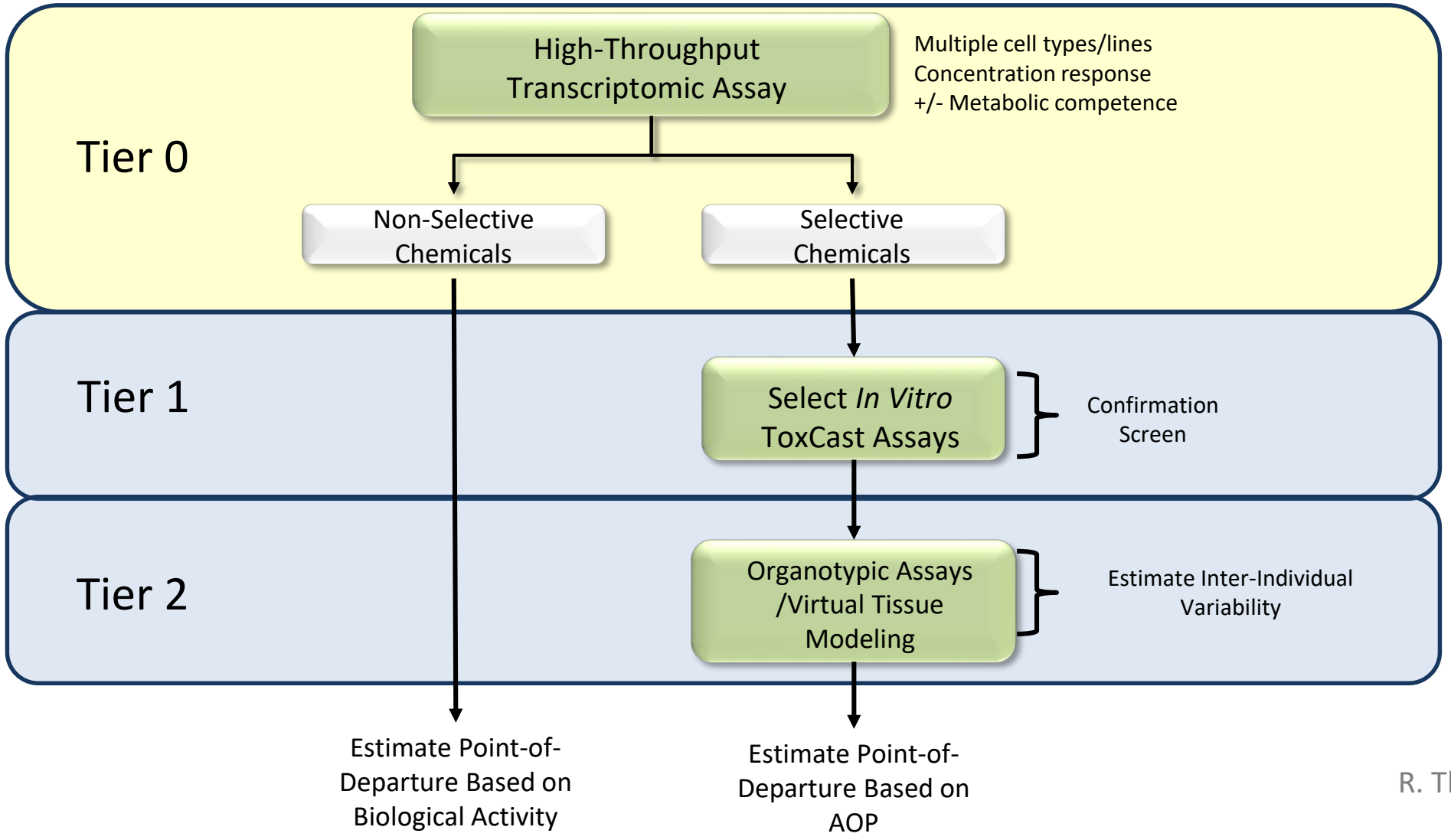
High-Throughput Transcript Profiling &
Functional Assessment

57th Annual Meeting of the Society of Toxicology (SOT)

March 12, 2018

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.

Background

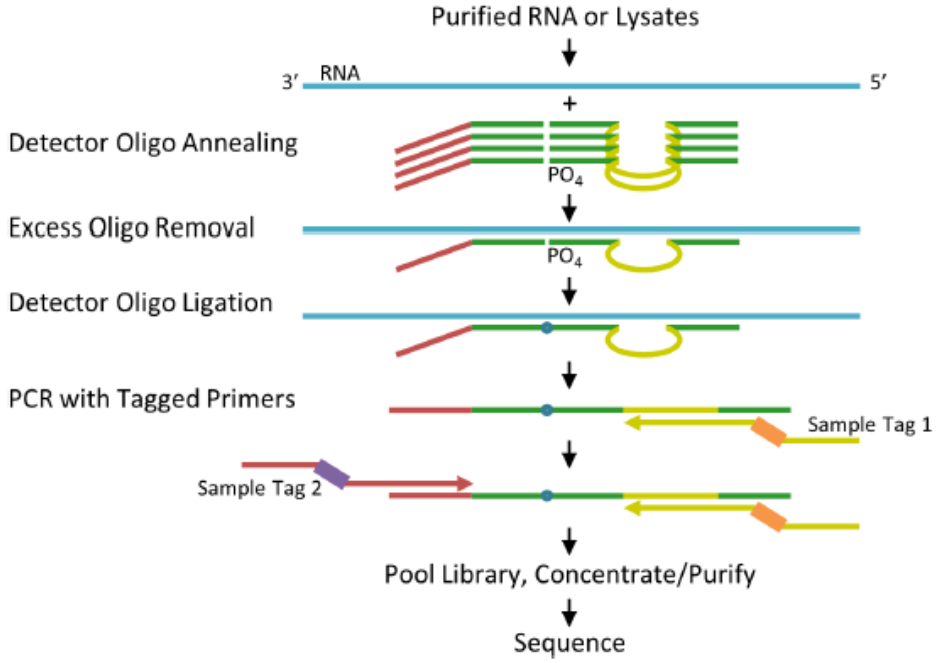


R. Thomas

TempO-Seq for HTTr

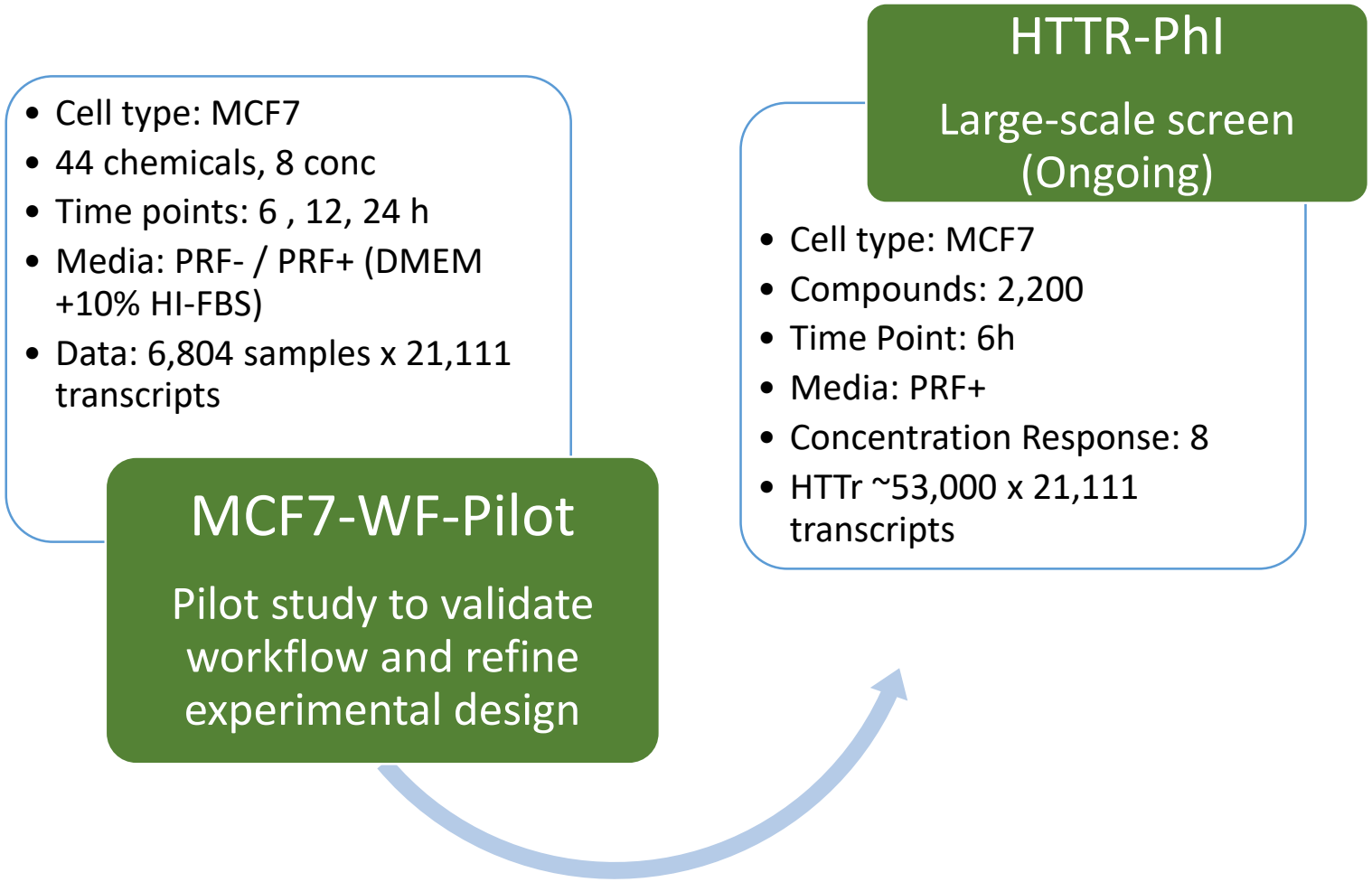
- The **TempO-Seq** human whole transcriptome assay measures the expression of ~21,100 transcripts.
- Requires only picogram amounts of total RNA per sample.
- Compatible with purified RNA samples or **cell lysates**.
- Transcripts in cell lysates generated in 384-well format barcoded to well position
- Scalable, targeted assay:
  - 1) Measures transcripts of interest
  - 2) Ignores intronic sequences or rRNA
  - 3) Much greater throughput than RNA-Seq
- Per sample **fastq files** are aligned to BioSpyder probe manifest to generate raw count data.

TempO-Seq Assay Illustration

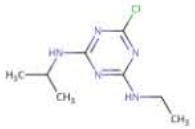


Wednesday, March 14th; 3-4PM; Room 217C
High Throughput Transcriptomics: Addressing the Human Risk Assessment Challenges of Chemical Coverage, Metabolism, and Populations Variation.

Experiments

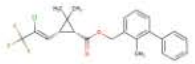


Pilot Chemicals (44) with known MoA



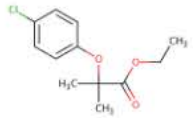
Atrazine
1912-24-9

cAMP inducer



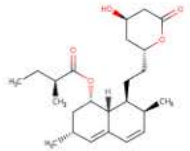
Bifenthrin
82657-04-3

Na+ channel



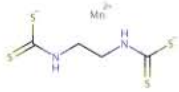
Clofibrate
637-07-0

Hypolipidemic



Lovastatin
75330-75-5

HMGCR



Maneb
12427-38-2

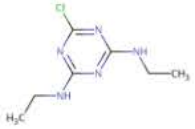
SH Reactive

Reserpine
50-55-5

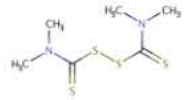
VMAT

Rotenone
83-79-4

Mito Complex I



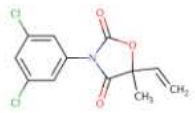
Simazine
122-34-9



Thiram
137-26-8

SH Reactive

Ziram
137-30-4

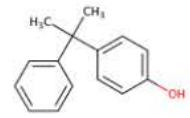


Vinclozolin
50471-44-8

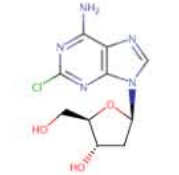
Antiandrogen

Fulvestrant
129453-61-8

(Z)-4-Hydroxytamoxifen
68047-06-3

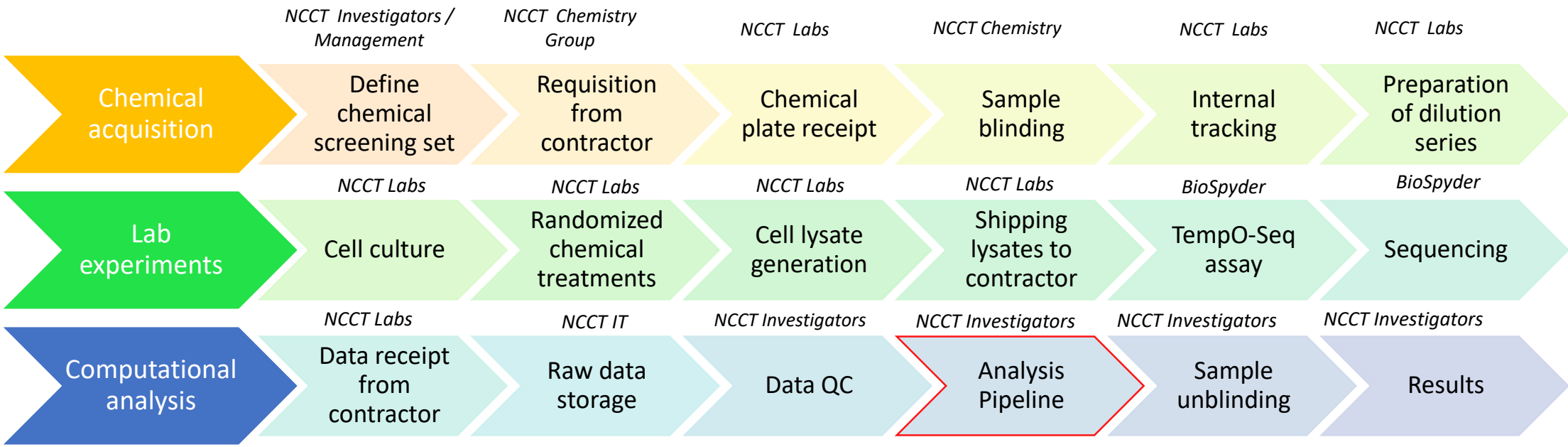


4-Cumylphenol
599-64-4

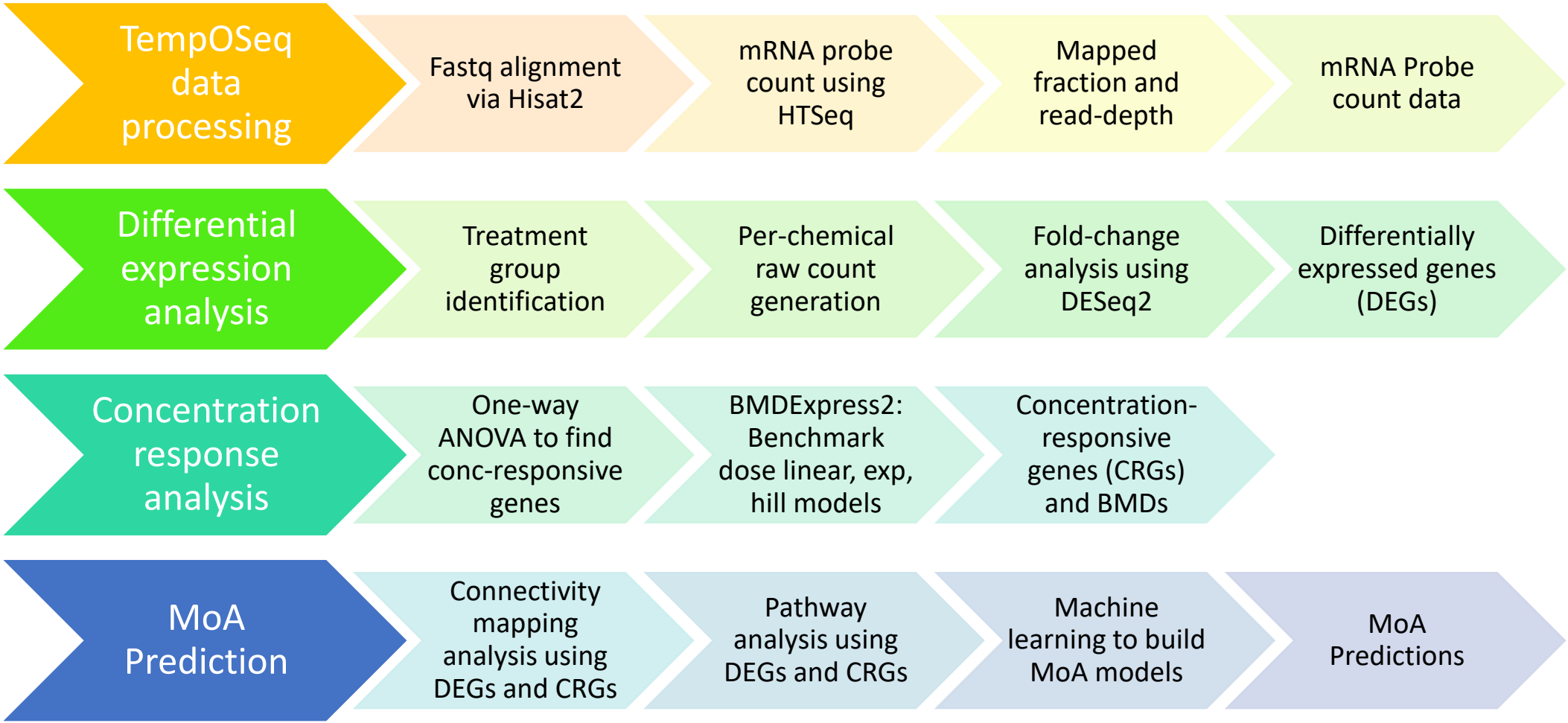


Cladribine
4291-63-8

HTTr Workflow



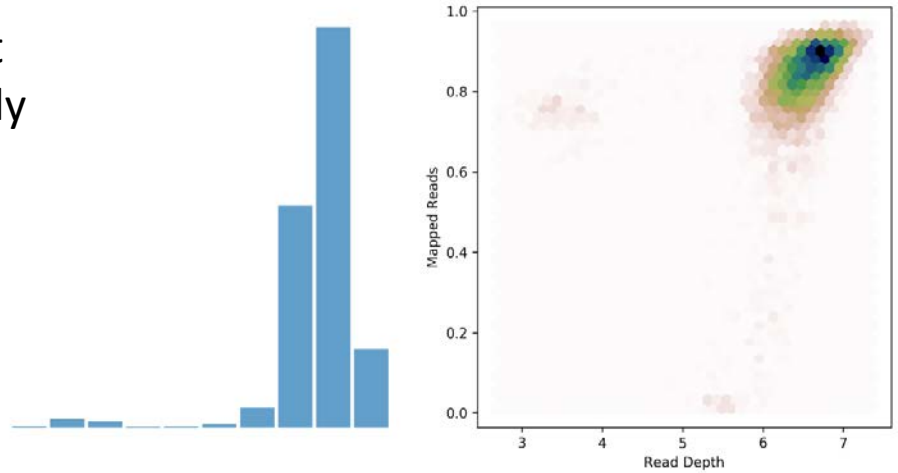
Analysis Pipeline

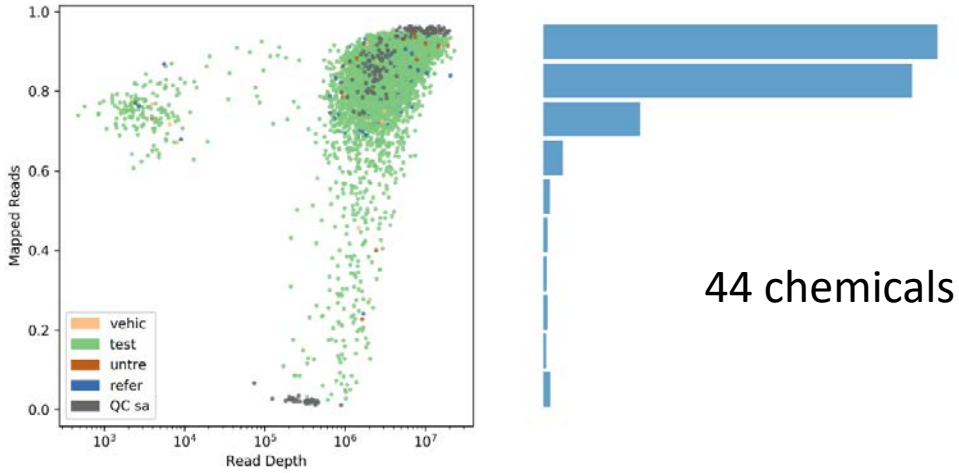


Python & R analysis pipeline will be available from EPA@GitHub after publication

TempO-Seq Quality

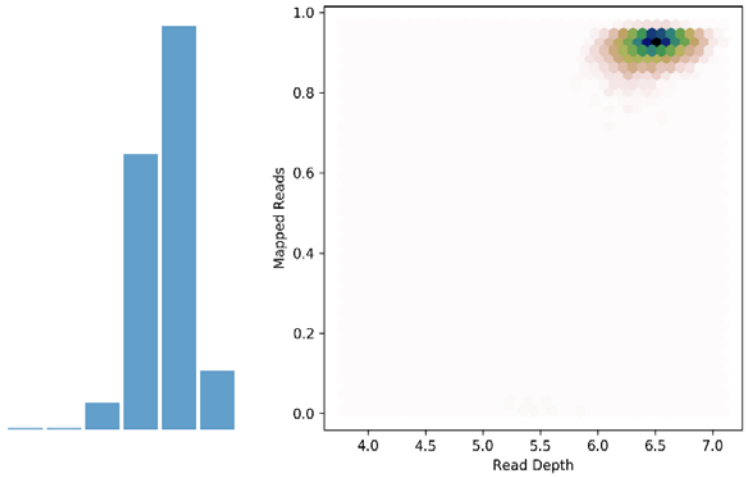
Pilot Study

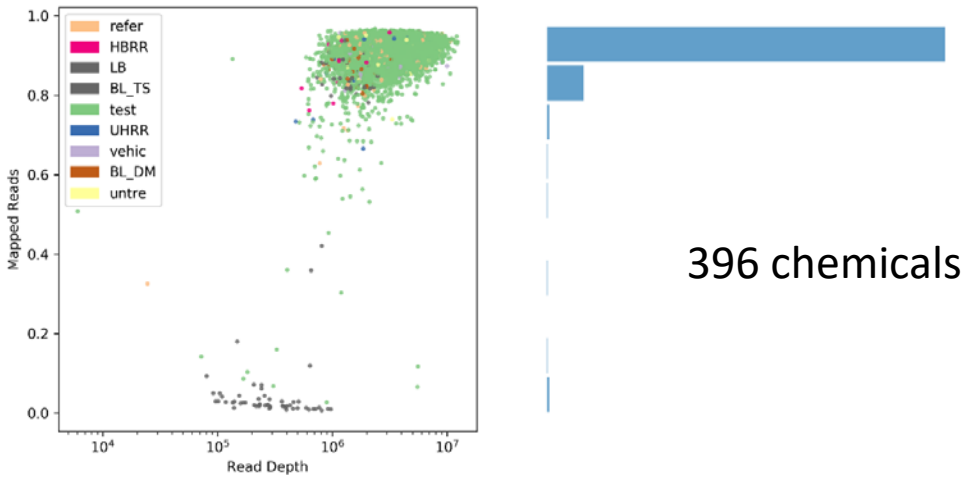




44 chemicals

Screening Study

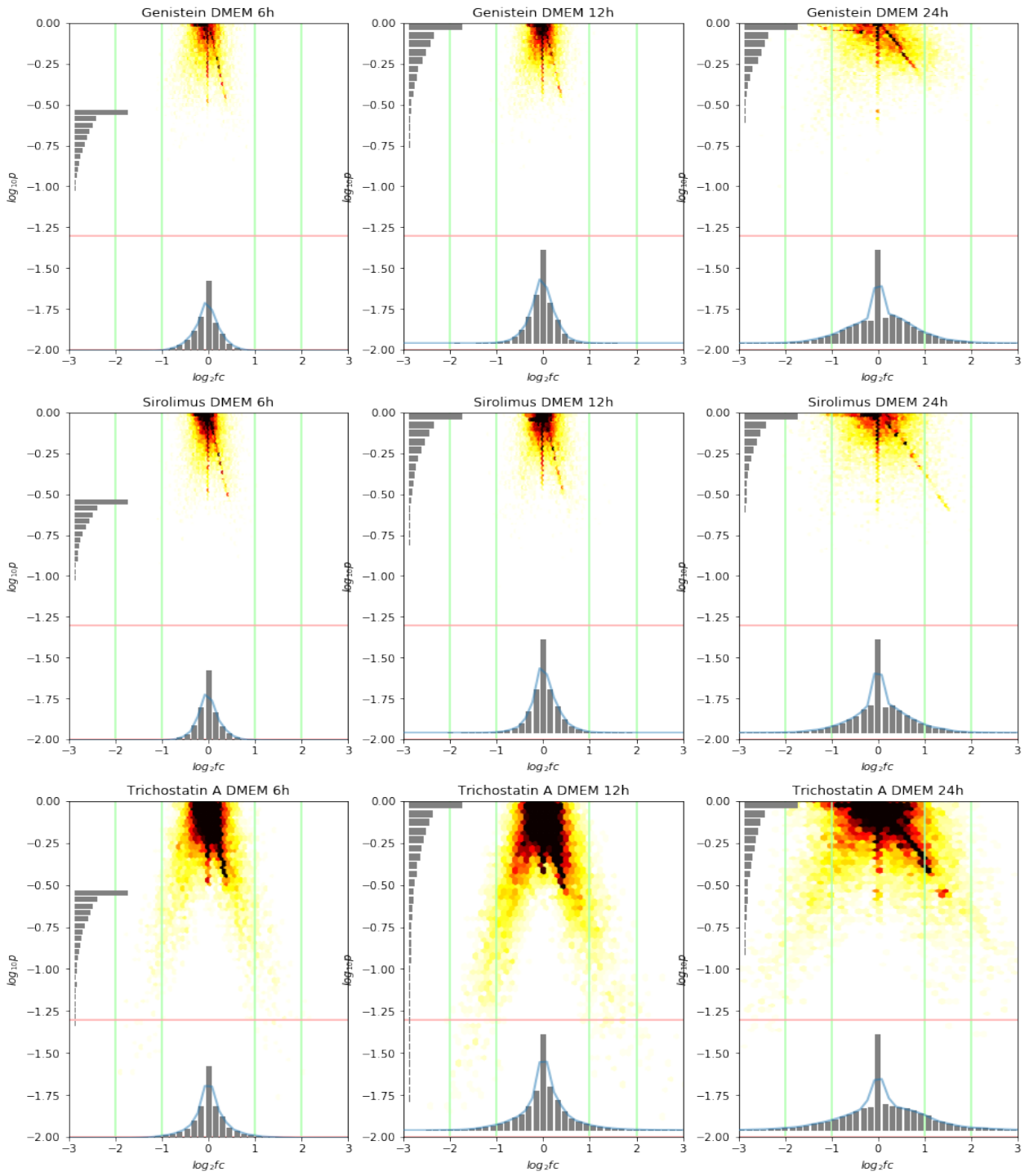


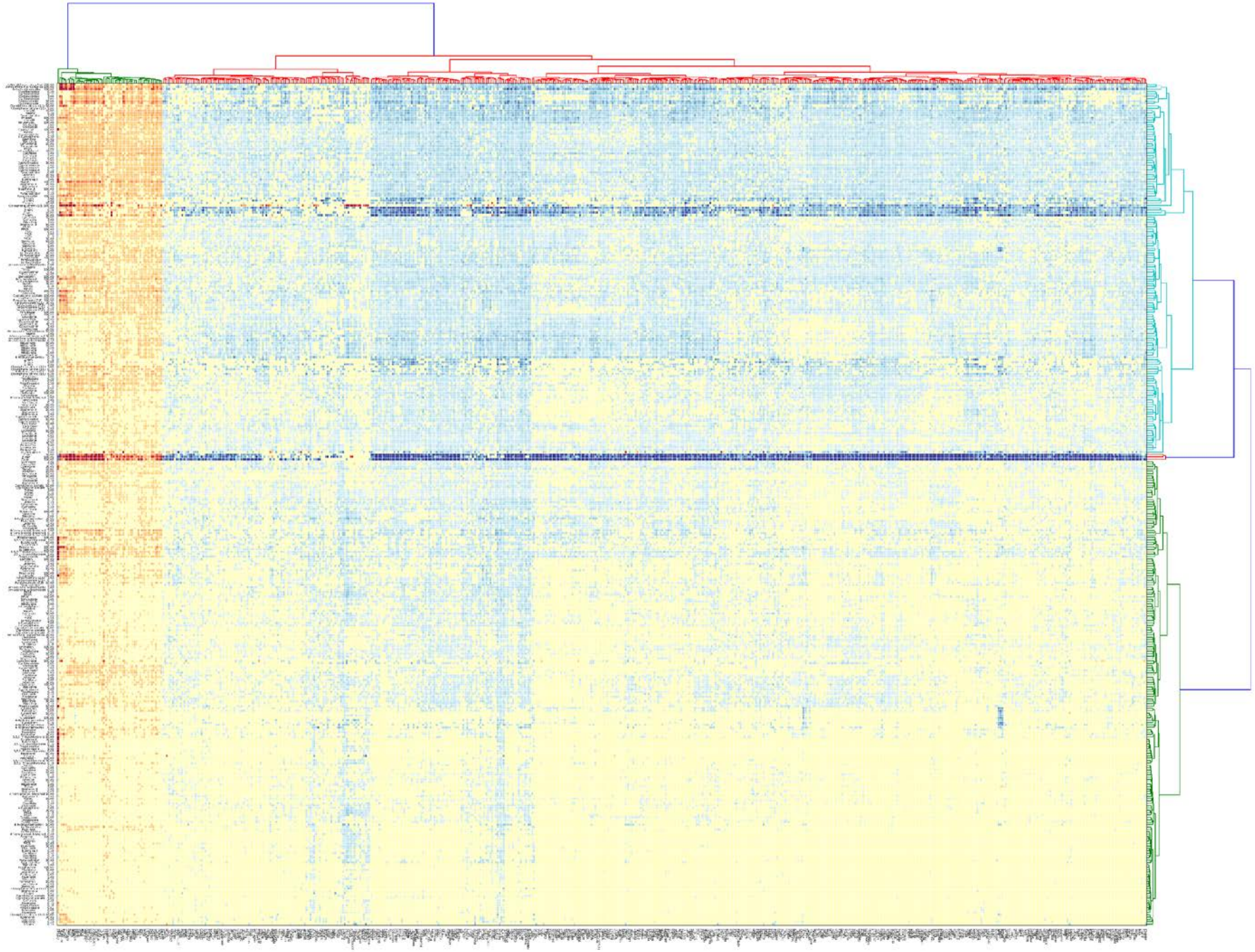


396 chemicals

DEGs

- RNA-Seq data have
  - Variance in total # of transcripts between samples
  - Heterogeneous transcript distributions
- A variety of techniques for analyzing differentially expressed genes in transcriptomic studies. DESeq2 has been shown to accurately identify transcriptional responses
- DESeq2 (R package) that uses the negative Binomial (NB) distribution to model count data and effects are “smoothed” over genes with similar variance
- For screening, we ran DESeq2 for count data for each chemical (all concentrations and same-plate DMSO controls)

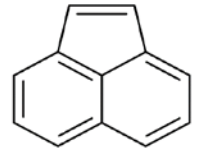


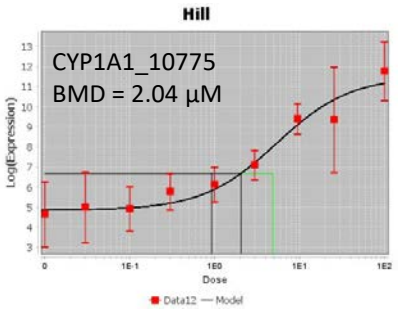


Concentration-Response Analysis

- Filter transcripts by ANOVA p<0.05 & |LFC2|$_{once}$>1
- BMDS is standard approach for concentration-response analysis
- BMD 2.0 Java GUI-based interactive wrapper for BMDS, currently maintained by Scott Auerbach@NIH
- Scott Auerbach facilitated the implementation of BMD 2.0 command-line version (currently alpha)
  - BMD2 –input data –config bmd2.json –out bmd.json
- All model fits and BMD values exported for storage

Acenaphthylene
208-96-8 | DTXSID3023845





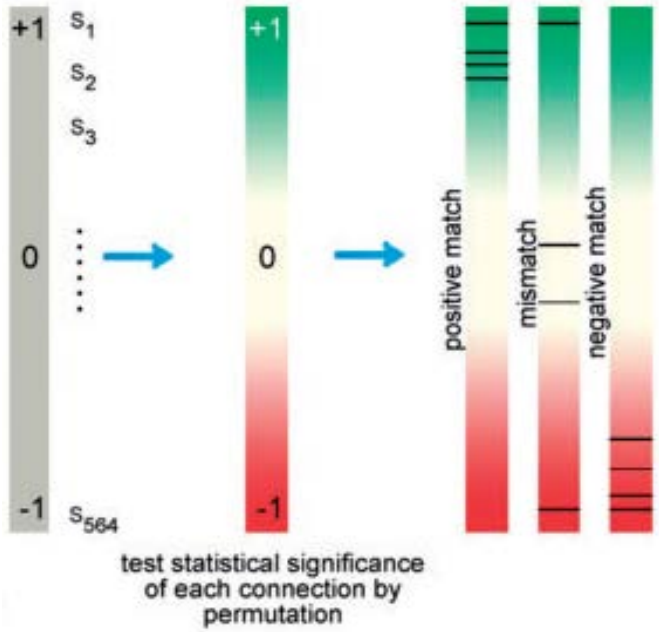
Tuesday 10:45-12:15 CC Exhibit Hall **P453**

Joshua Harrill

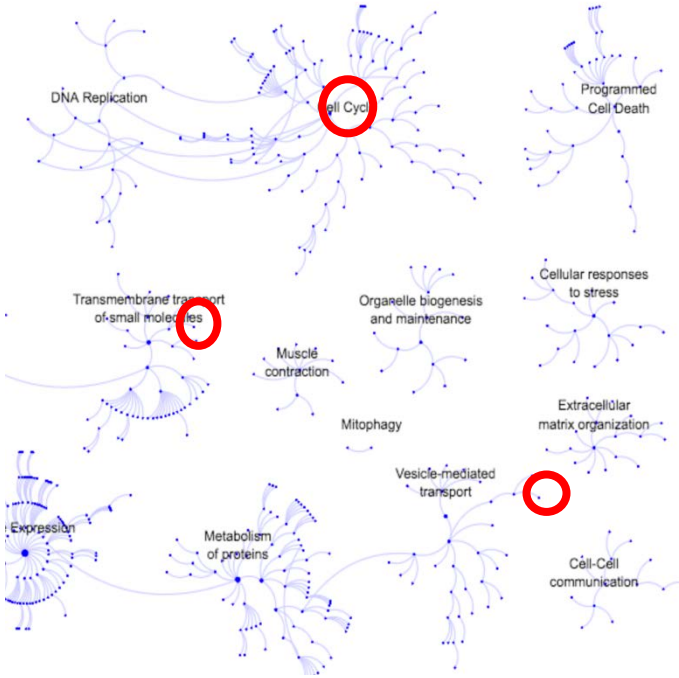
Differential Gene Expression and Concentration-Response Modeling Workflow for High-Throughput Transcriptomic (HTTr) Data: Results from MCF7 Cells

MoA Prediction

Connectivity mapping analysis using DEGs and CRGs

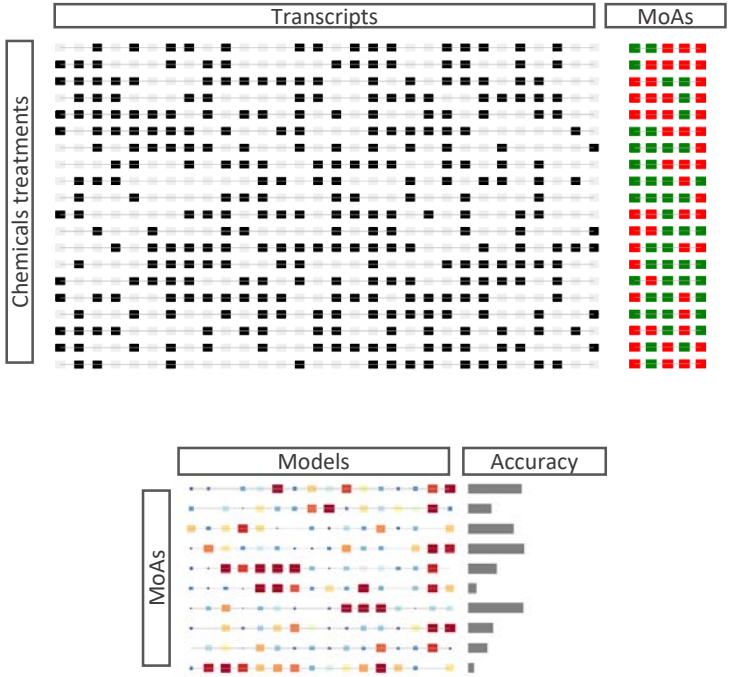


Pathway analysis using DEGs and CRGs



reactome.org

Machine learning to build MoA models

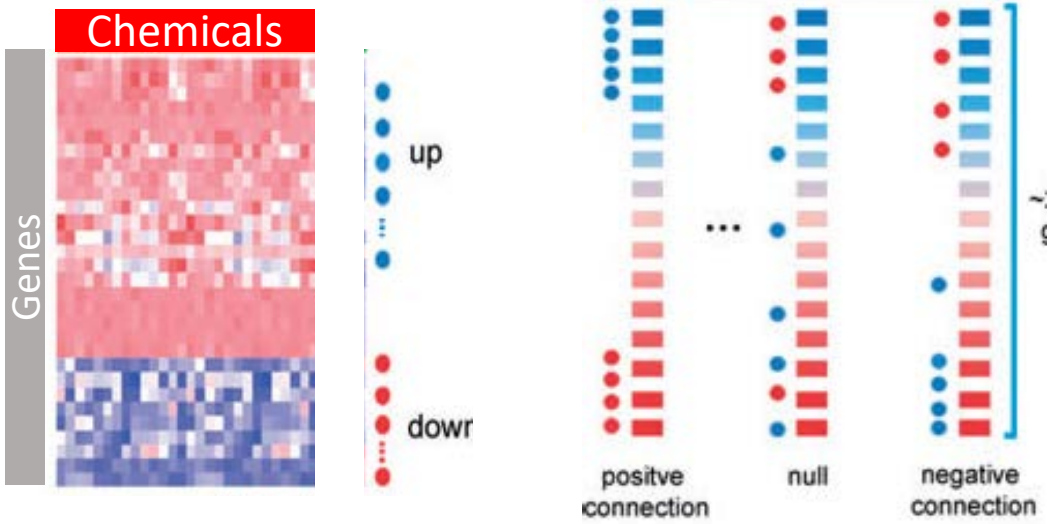


MoA by Connectivity Mapping

Input DEGs or CRGs

Query Signature DB CMap or BSP

Find best positive matches



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

Infer MIE/Target 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 MoAs – how do we distinguish signal from noise ?

CMap Signatures

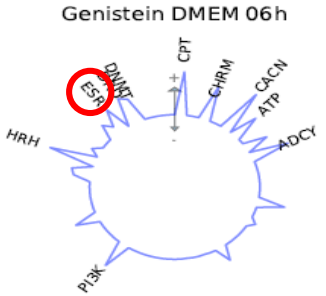
- Use CMap v2 database: Affymetrix data on 1176 chemicals, 5 cell lines
- Translate FC profiles to “signatures”
- Convert L2FC data to standardized Z vector
- For z0=1,2,3 create discrete Z where value = 1 if Z>z0 and -1 where Z<z0
- Store signatures in MongoDB database for rapid searching

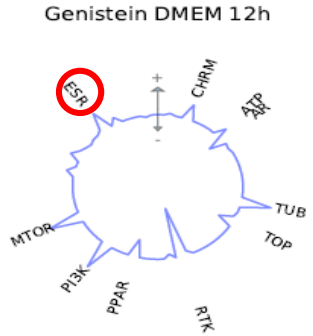
HTTr Signatures creation

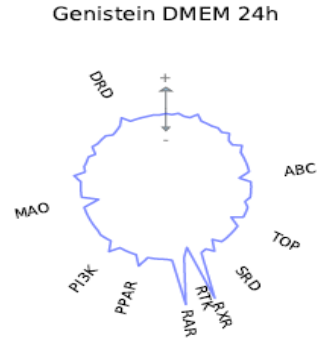
- Calculate FC using DESeq2
- Filter probes with |L2FC|>=0.6 & p<0.05
- Construct Z-score based profiles

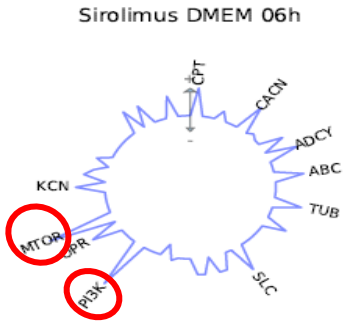
Search all query profiles against database (MCF7 only)

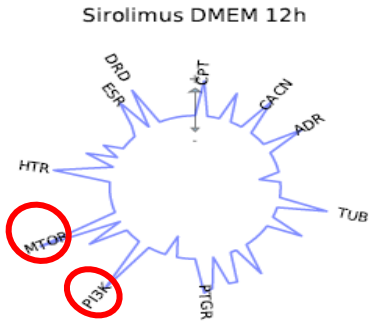
- Score hits using different GSEA, Jiang & Gentleman (JG),
- Assign MoA(s) based on hit

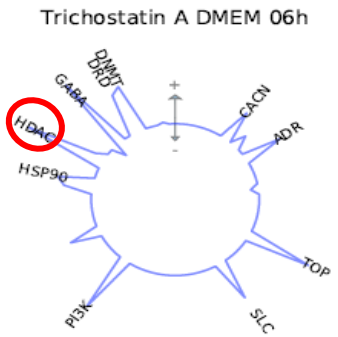


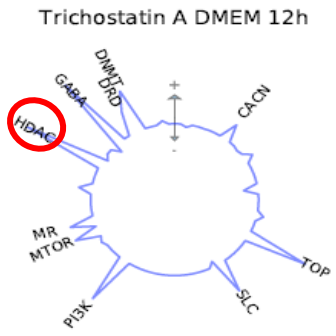


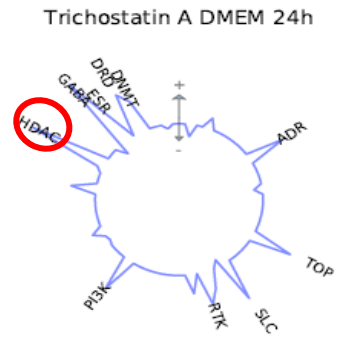


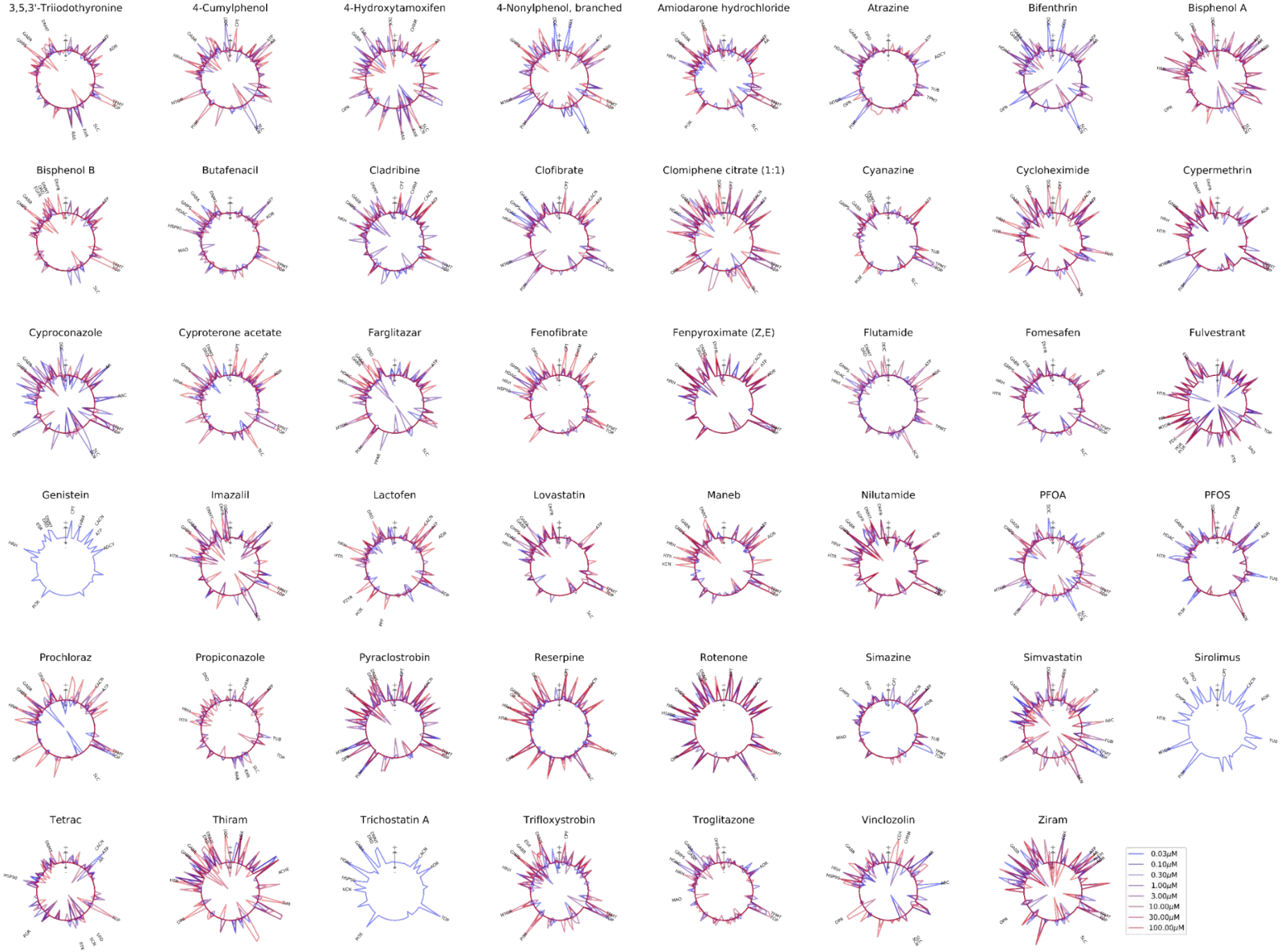











3,5,3'-Triiodothyronine
4-Cumylphenol
4-Hydroxytamoxifen
4-Nonylphenol, branched
Amiodarone hydrochloride
Atrazine
Bifenthrin
Bisphenol A
Bisphenol B
Butafenacil
Cladribine
Clofibrate
Clomiphene citrate (1:1)
Cyanazine
Cycloheximide
Cypermethrin
Cyproconazole
Cyproterone acetate
Farglitazar
Fenofibrate
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

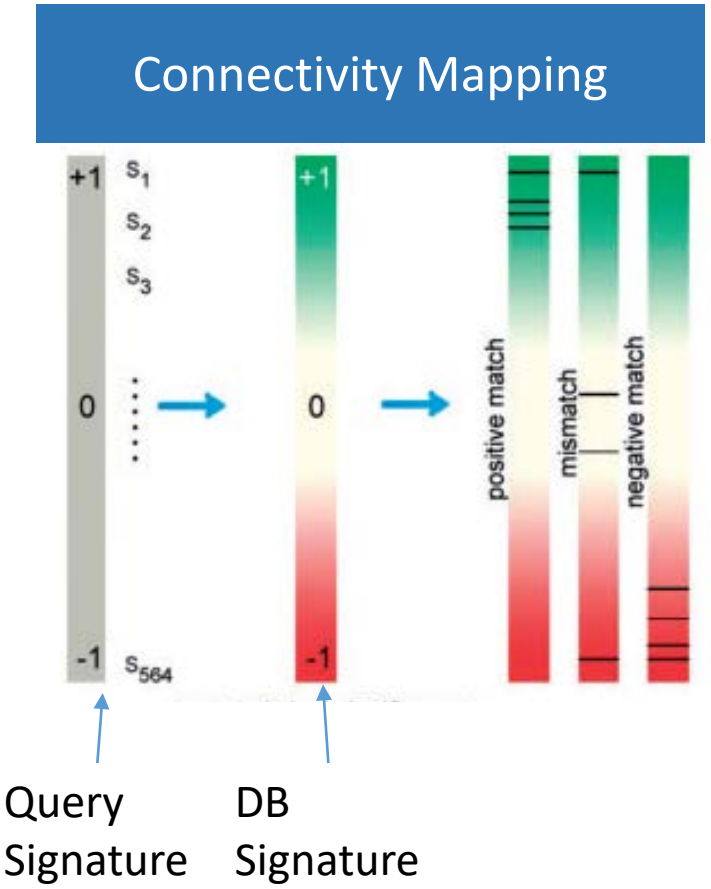
Quantifying Performance

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

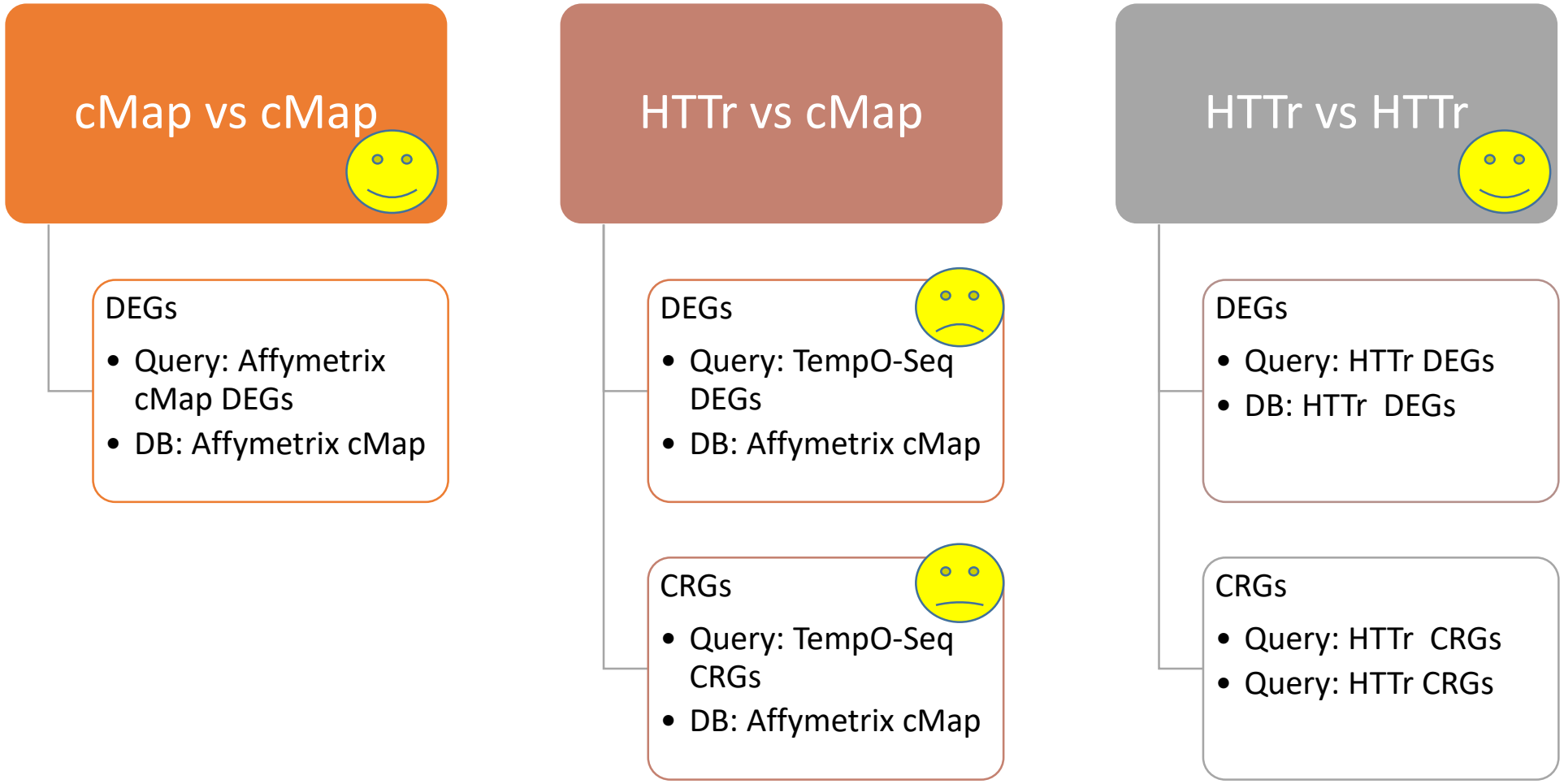
1. Annotate chemicals with known MoAs
   - MoAs/Targets: 143
   - Chemicals: 614
2. Use to predict MoA
3. Search "hits" by connectivity with score= ϑ
   - If $\vartheta > \vartheta_0$

     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
4. Measure sensitivity, specificity, BA



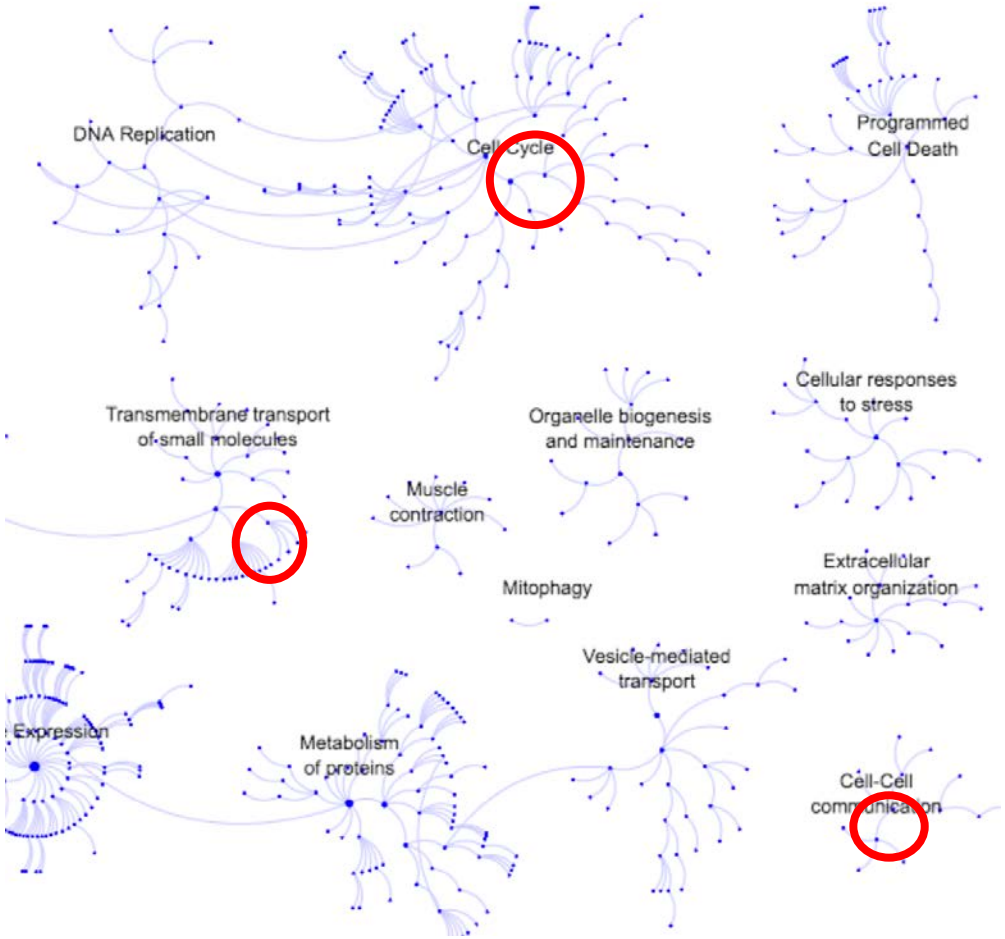
Evaluating Different Databases



media	timeh	moa	pos	neg	BA	Sn	Sp	th0	gene	httr-av	httr-md	httr-mx	httr-p75	httr-p95	Exp
DMEM	6	Protein-Reactive	3	41	0.82	1	0.63	0.216							
		Vmat	2	42	0.81	1	0.62	0.348	SLC18A2	1	0	4	0	4	off
					0.81	1	0.62	0.348	SLC18A3	0	0	1	0	1	off
					0.81	1	0.62	0.348	SLC18A1	0	0	0	0	0	off
					0.81	1	0.62	0.348	SLC18B1	0	0	0	0	0	off
		Estrogen	4	40	0.78	0.92	0.64	0.342	ESR1	6001	3060	14283	10365	14029	on
					0.78	0.92	0.64	0.342	ESR2	5	2	17	6	16	off
		Mitochondria	4	40	0.64	0.67	0.62	0.375	NDUFB9	8298	4191	36884	6012	27380	on
					0.64	0.67	0.62	0.375	NDUFB1	1629	611	8532	865	6097	on
					0.64	0.67	0.62	0.375	NDUFS6	1536	1059	5441	1279	4134	on
					0.64	0.67	0.62	0.375	CYB5R3	428	375	1083	444	893	on
					0.64	0.67	0.62	0.375	CYB5R1	423	309	1278	341	1044	on
					0.64	0.67	0.62	0.375	CYBA	294	198	804	293	682	on
		Anti-Estrogen	3	41	0.55	0.5	0.61	0.212	ESR1	6001	3060	14283	10365	14029	on
					0.55	0.5	0.61	0.212	ESR2	5	2	17	6	16	off
		Thyroid	2	42	0.55	0.5	0.6	0.296	THRB	47	44	80	68	76	off
					0.55	0.5	0.6	0.296	THRA	20	21	44	24	36	off
		Ppo	2	42	0.55	0.5	0.6	0.328							
		Anti-Androgen	4	40	0.47	0.33	0.6	0.374	AR	56	29	198	78	158	on
		Steroidogenesis	4	40	0.46	0.33	0.59	0.372							
		Ppara	5	39	0.32	0.06	0.57	0.317	PPARA	24	16	75	24	61	off
		Triazene	2	41		0	0.59	0.287	PDE7A	1284	848	3810	1842	3037	on
		Cholesterol	2	41		0	0.59	0.389	HMGCR	313	188	897	421	744	on
		Ion-Channel	2	42		0	0.59	0.341	SCN1B	272	183	857	368	665	on
		Pparg	2	42		0	0.61	0.298	PPARG	107	56	392	123	322	on
		Triazene	2	41		0	0.59	0.287	PDE6D	68	46	135	116	128	on
						0	0.59	0.287	PDE3B	68	52	170	69	162	on
		Pparg	2	42		0	0.61	0.298	PPARGC1B	46	46	77	61	71	off
						0	0.61	0.298	PPARGC1A	0	0	0	0	0	off
		Cytotoxicity	2	42		0	0.59	0.124							

Predicting MoA via Pathways

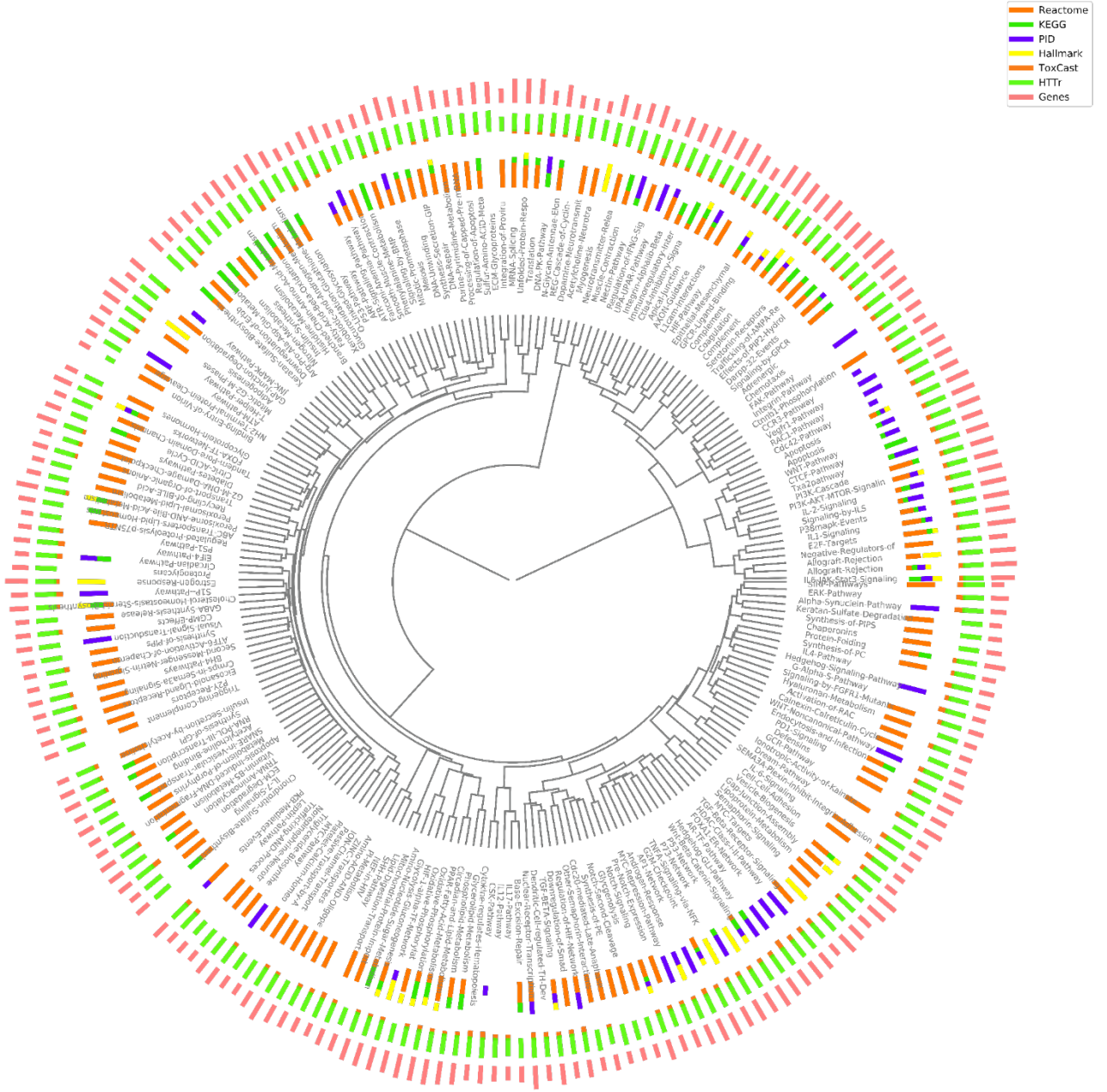
- Transcriptional perturbations of key pathways predicts MoA
- Pathway analysis
  - Select DEGs or CRGs to identify enriched pathways
  - Link enriched pathways to MoA
- Issues
  - Choice of pathway database
  - Scoring pathway enrichment
  - How do we objectively evaluate predictive accuracy
  - Linking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



reactome.org

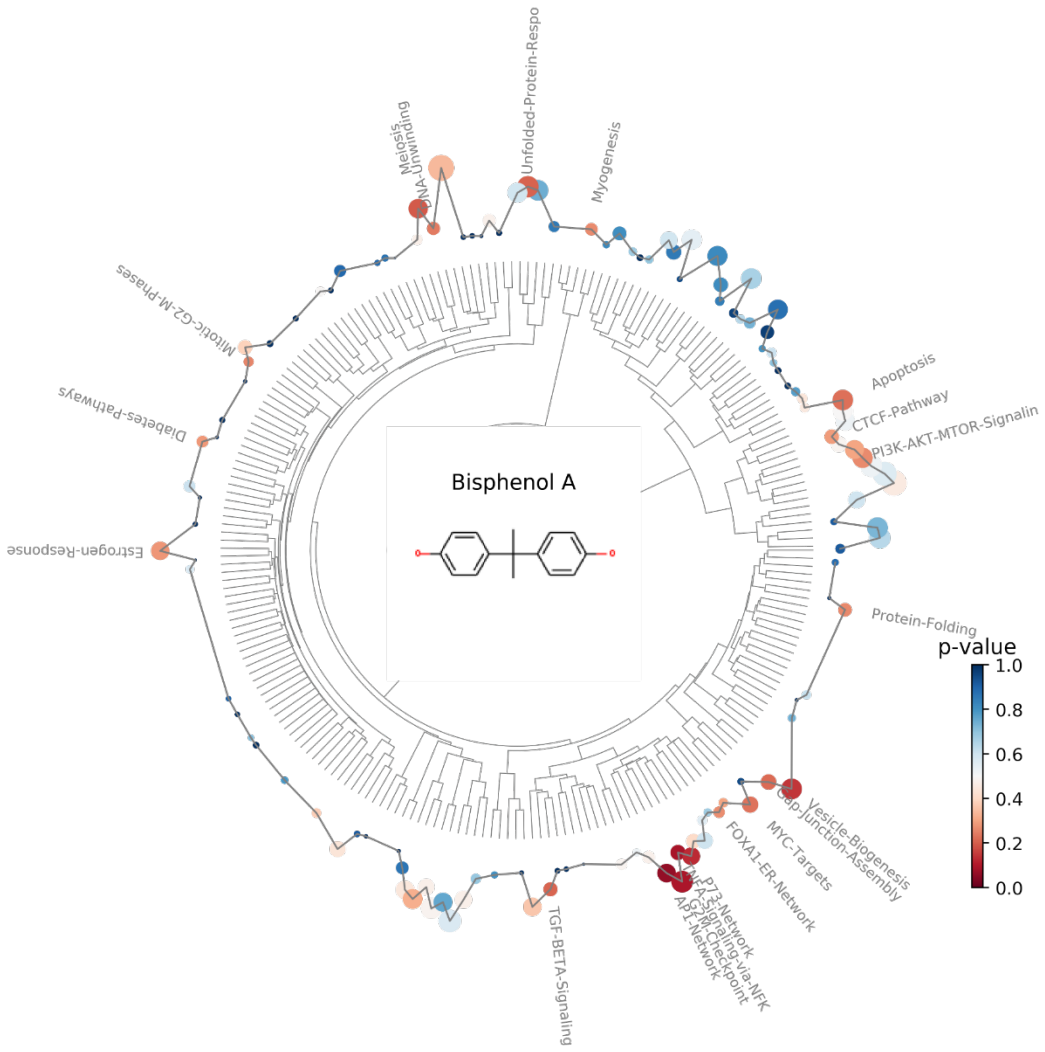
“Super-Pathways”

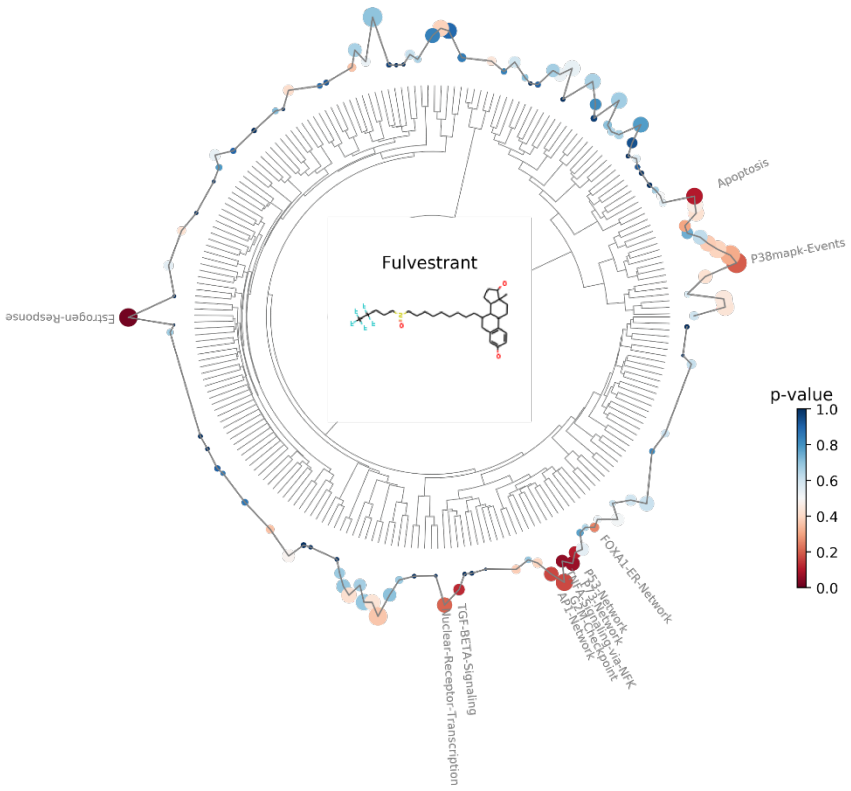
- Cluster Hallmark and canonical pathways (Reactome, KEGG, PID and BioCarta) from MSigDB V6 using genes
- Use hierarchical agglomerative clustering to organize super-pathways by similarity
- Each clade in the dendrogram shows groups of functionally related pathways
- Concentric rings show information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

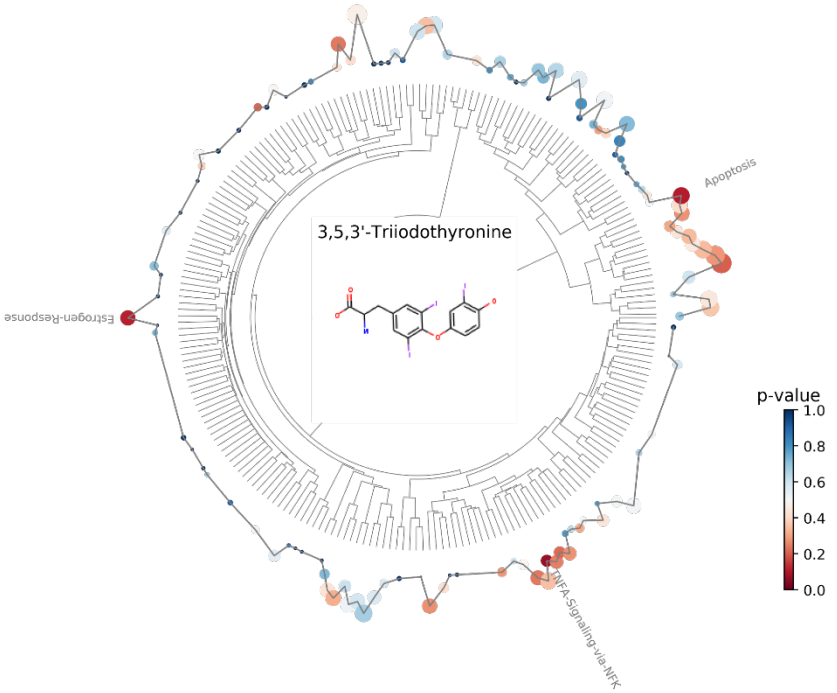


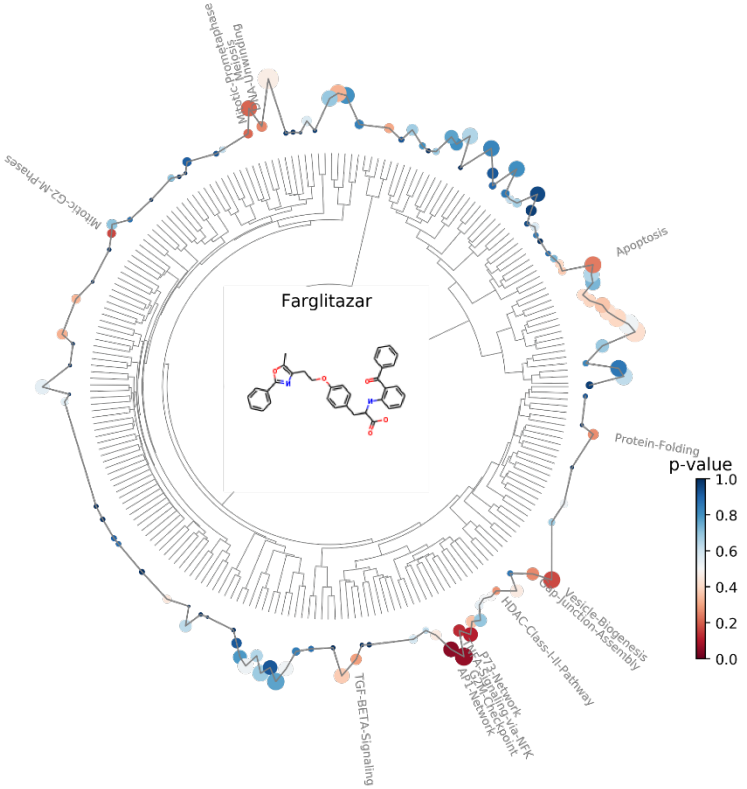
Pathway Analysis

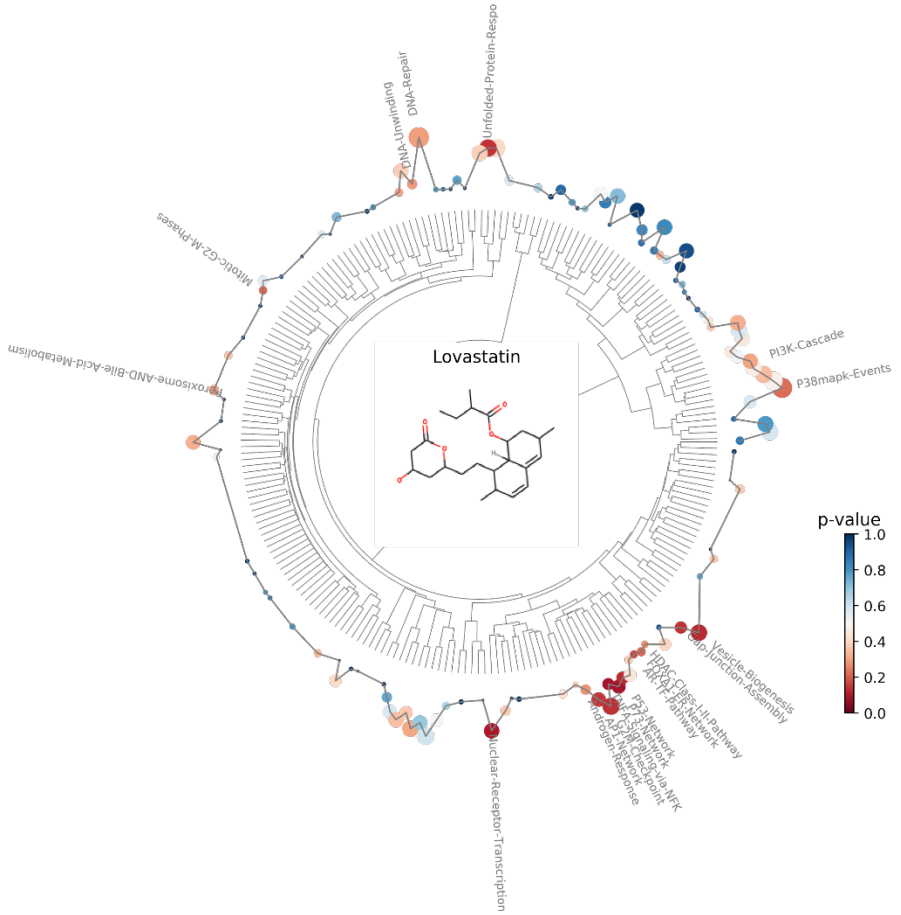
- The HTTr profiles for chemical treatments were searched against 224 super-pathways.
- Pathways were scored using different metrics that used the entire HTTr profile (e.g. enrichment scores), and just DEGs.
- The significance of scores was estimated by simulation.




Fulvestrant
Apoptosis
P38mapk-Events
Estrogen-Response
FOXA1-ER-Network
P53-Network
P53-Pathway
G1-S-Checkpoint
ATR-Network
TGF-BETA-Signaling
Nuclear-Receptor-Transcription
p-value
1.0
0.8
0.6
0.4
0.2
0.0

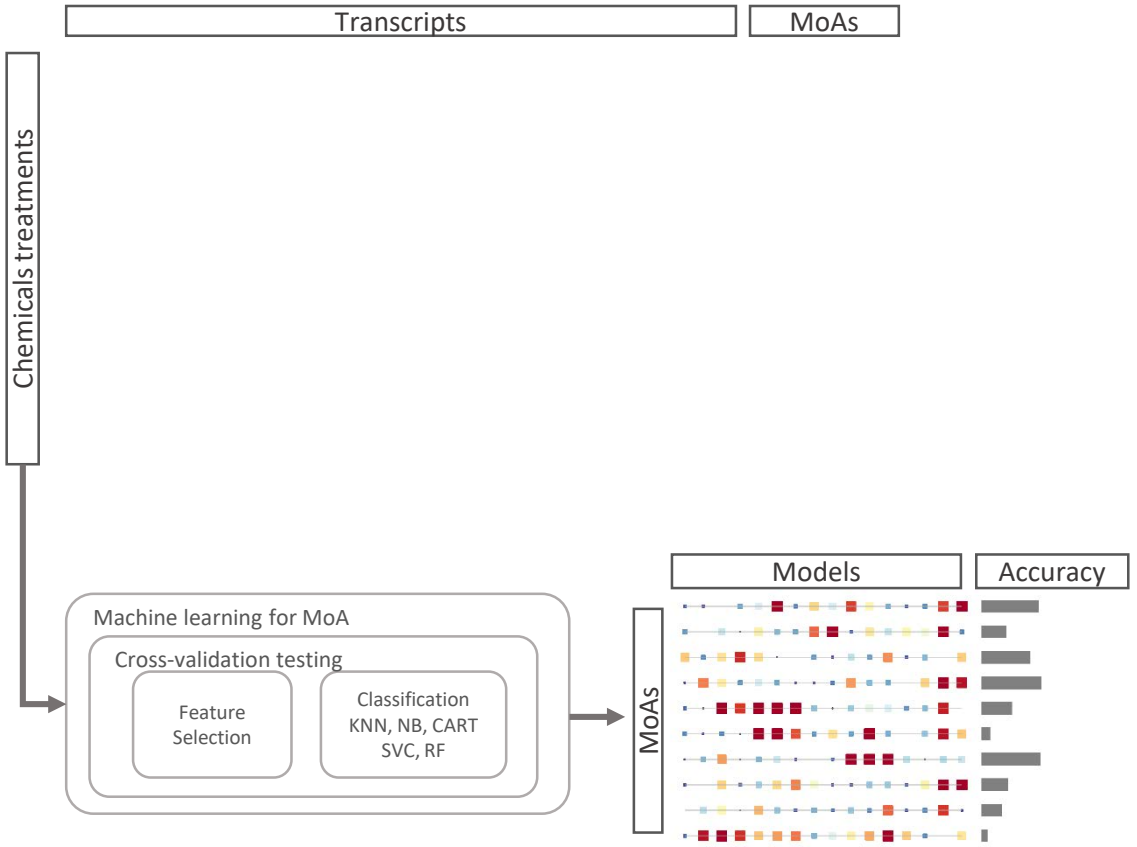

3,5,3'-Triiodothyronine
Apoptosis
Estrogen-Response
TNFA-Signaling-via-NFK
p-value
1.0
0.8
0.6
0.4
0.2
0.0


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


Lovastatin
PI3K-Cascade
P38mapk-Events
Vesicle-Biogenesis
Gap-Junction-Assembly
HDAC-Class-I-II-Pathway
ATR-Network
P73-Network
Cell-Cycle-Checkpoint
ATR-Signaling-via-NFK
ATR-Network
Nitrogen-Response
Nuclear-Receptor-Transcription
Peroxisome-AND-Bile-Acid-Metabolism
Mitotic-G2-M-Phases
DNA-Unwinding
DNA-Repair
Unfolded-Protein-Respo
p-value
1.0
0.8
0.6
0.4
0.2
0.0

Predicting MoA by Machine Learning

- Claim: Classifiers can be learned from transcriptional profiles labeled with MoA
- Machine Learning
  - Train models using DEGs for chemicals with known MoA
  - Predict MoA using classifiers
  - Added curated data improves performance
- Issues
  - Finding reproducible models difficult (too many genes too few chemicals)
  - Need new approaches for reducing dimensionality and increasing signal to noise



Summary

- Technology: Targeted RNA-Seq based HTTr is a powerful tool for assaying global chemical-induced perturbations.
- Workflow: We have developed a standardized, scalable workflow to generate large-scale HTTr data for thousands of chemicals.
- Connectivity mapping: Performance baseline for MoA prediction. Like most nearest neighbor prediction techniques, it is highly sensitive but not specific. Cross-platform connectivity mapping currently shows limited accuracy.
- Pathway analysis: Aggregated pathway analysis could provide a more biologically-relevant approach for MoA interpretation. Need new approaches to distinguish primary mechanisms from non-specific secondary/tertiary transcriptional cascades.
- Machine learning: We are developing machine learning approaches to improve specificity of MoA predictions. Challenges remain in curating information about primary MoA of chemicals, and dealing with paucity of data for MoA classes.
- Further evaluation of MoA Prediction approaches is being conducted and will be published soon. All computational workflows and data will be disseminated publicly.

Additional Information

10:45-12:15 CC Exhibit Hall P866
Imran Shah
Predicting Chemical Mechanisms-of-Action Using High-Throughput Transcriptomic Data

Tuesday 10:45-12:15 CC Exhibit Hall **P453**
Joshua Harrill
Differential Gene Expression and Concentration-Response Modeling Workflow for High-Throughput Transcriptomic (HTTr) Data: Results from MCF7 Cells

Wednesday 3-4PM Room 217C
High Throughput Transcriptomics: Addressing the Human Risk Assessment Challenges of Chemical Coverage, Metabolism, and Populations Variation.

Acknowledgements

Additional Information

10:45-12:15 CC Exhibit Hall P866
Imran Shah
Predicting Chemical Mechanisms-of-Action Using High-Throughput Transcriptomic Data

Tuesday 10:45-12:15 CC Exhibit Hall **P453**
Joshua Harrill
Differential Gene Expression and Concentration-Response Modeling Workflow for High-Throughput Transcriptomic (HTTr) Data: Results from MCF7 Cells

Wednesday 3-4PM Room 217C
High Throughput Transcriptomics: Addressing the Human Risk Assessment Challenges of Chemical Coverage, Metabolism, and Populations Variation.