Evaluating Chemical Safety Using Transcriptomics

In silico Toxicology Workshop
September 30, 2020

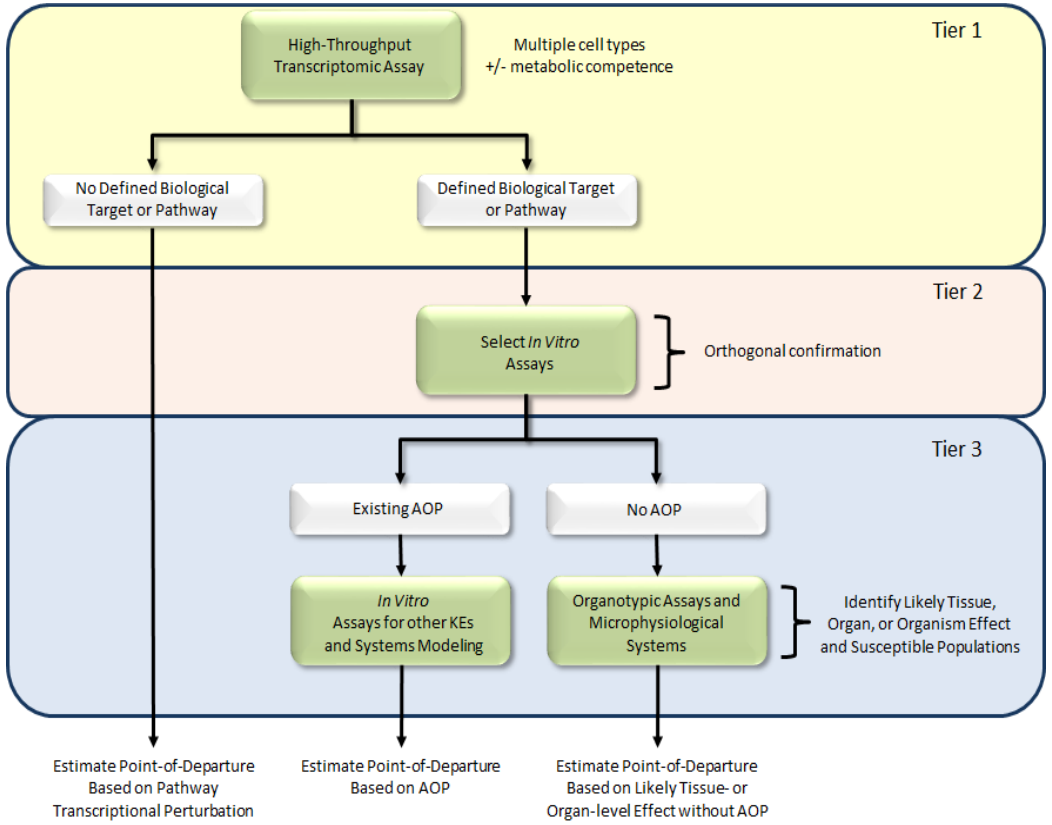
Imran Shah
Center for Computational Toxicology & Exposure

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.

EPA Context

- Need new approach methodologies (NAMs) to evaluate thousands of untested chemicals effectively
- NAMs are any technology, methodology, approach, or combination of methods that can provide information about chemical hazard and point of departure (POD) without using whole animals
- High-throughput transcriptomics (HTTr) is a flexible, portable and cost-efficient platform to comprehensively evaluate the potential biological pathways impacted by chemical exposure
  - NAMs for hazard: identify putative molecular initiating events (MIEs)/targets or pathways
  - NAMs for POD: estimate *in vitro* POD associated with pathways/MIEs

"Tiered testing framework for hazard characterization"



Thomas et al., 2018

Typical Transcriptomics Workflow

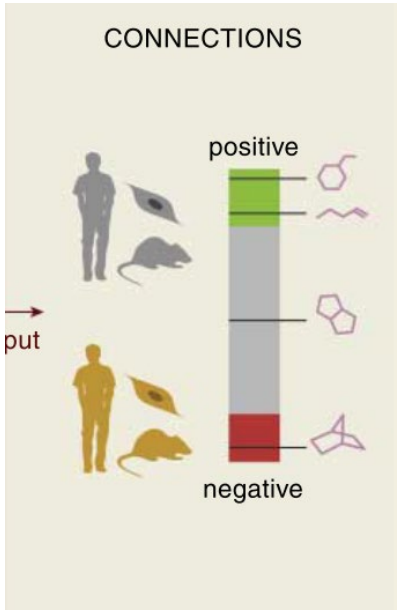
How do we infer hazard and estimate POD based on transcriptomic profiles?

Transcriptomics for Hazard and POD Estimation

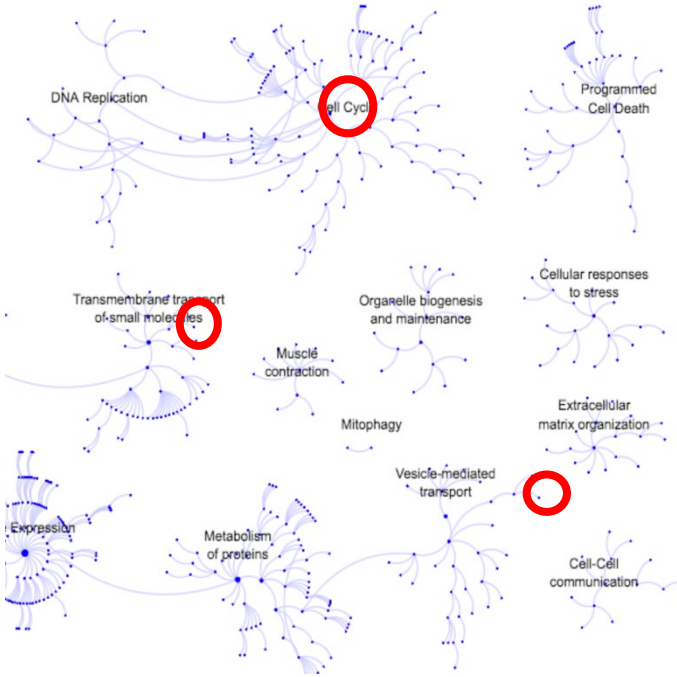
Connectivity Mapping

Pathway / Network analysis

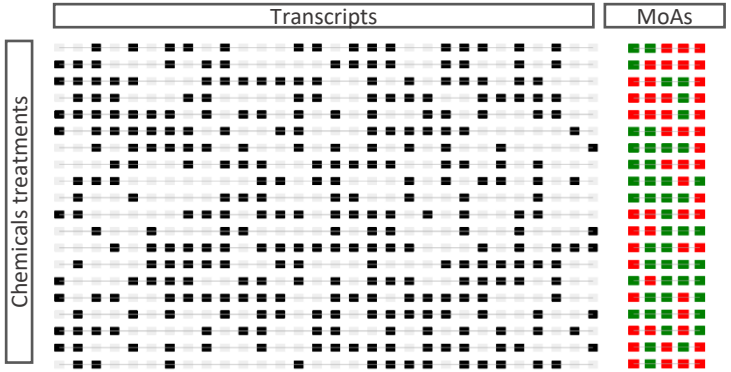
Machine Learning

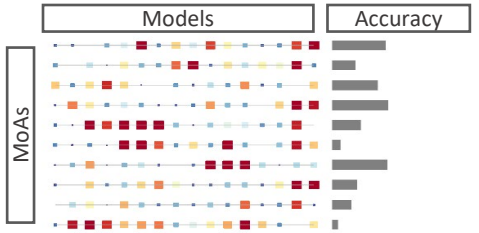


Lamb *et al.*, 2006



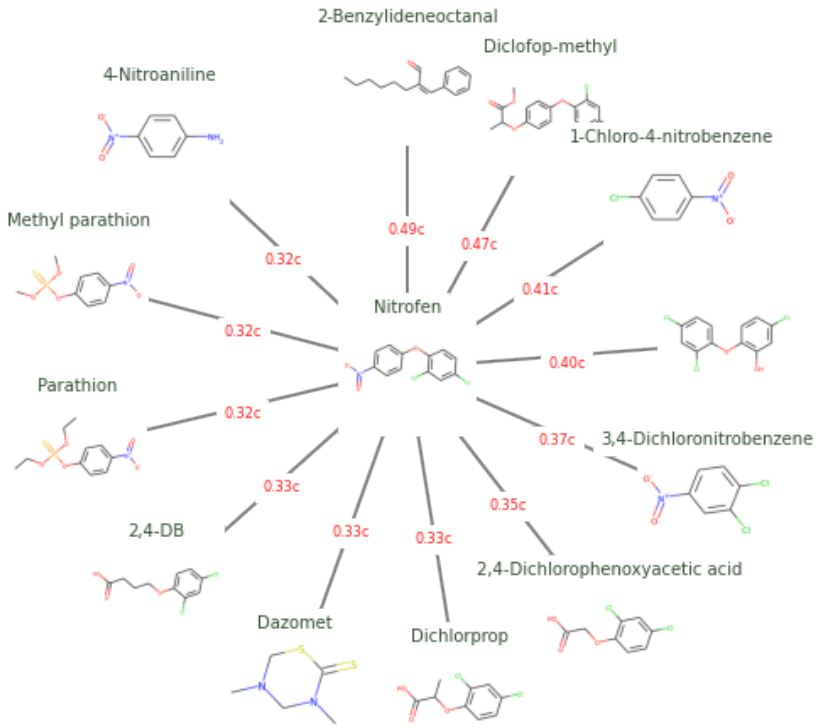
reactome.org





How is the risk of untested chemicals evaluated *today*?

- Generally, “read-across” techniques are used to fill data gaps by inference from a ‘similar’ substance or substances (OECD, 2017):
  - Identify analogues using between structure / physico-chemical similarity
  - Assign hazard and POD value based on analogue(s)
- However, many chemicals do not have substantial structural similarity to tested chemicals
- Can we use transcriptomic similarity to infer hazard and POD?

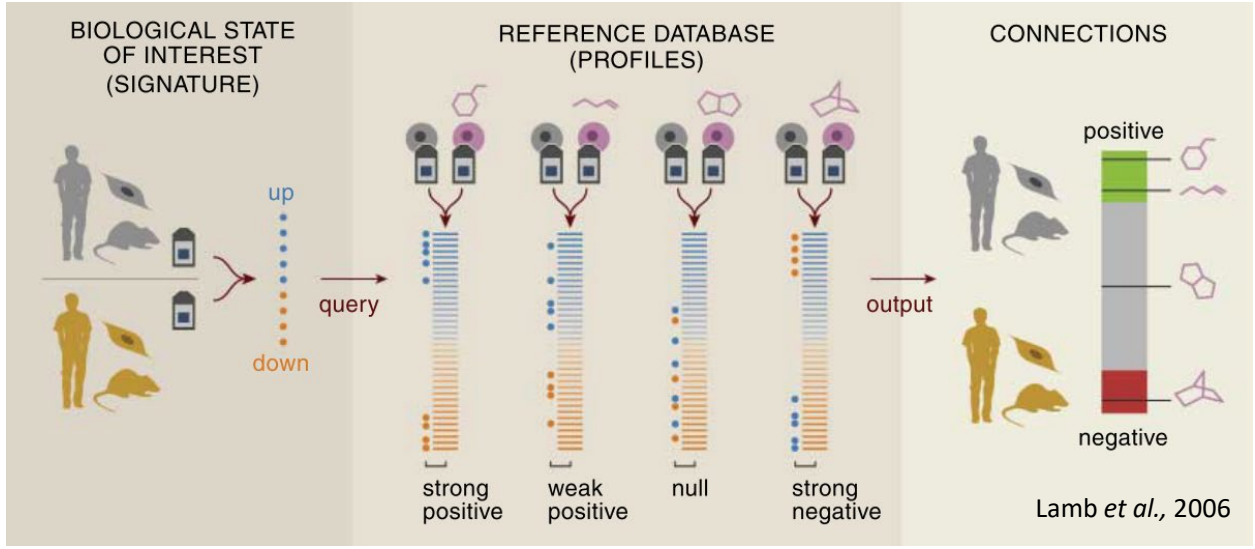


Generalized Read-across (GenRA)
Shah *et al.* 2016

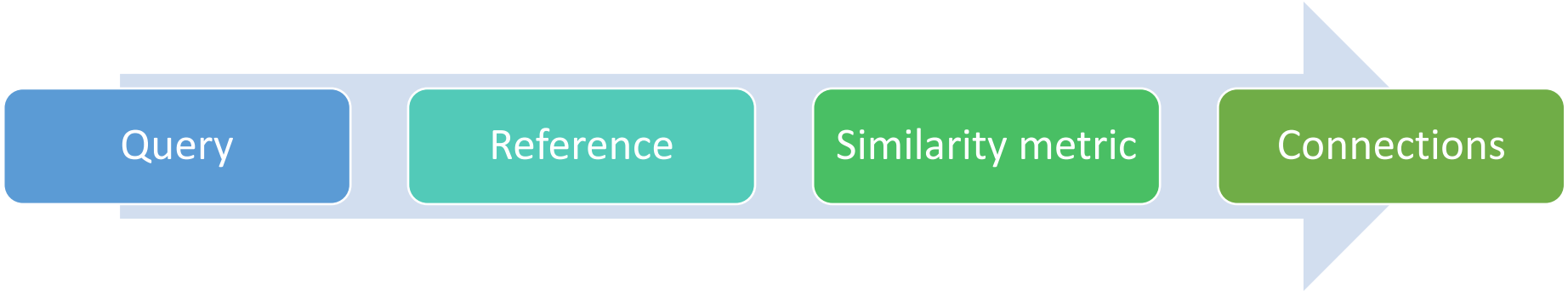
Inferring hazard based on transcriptomic similarity

- "Connectivity mapping" (Lamb *et al.,* 2006): If two chemicals produce similar transcriptomic profiles, then they could act via a common mechanism
- Connectivity mapping has been used in toxicology to find putative targets (DeAbrew *et al.,* 2016; Wang *et al.,* 2016).
- Many new approaches are being developed and there are a diverse set of tools
- Need to formalize connectivity analysis for putative target/pathway identification (hazard) and quantitative *in vitro* POD estimation

Connectivity mapping



Elements of Connectivity Mapping



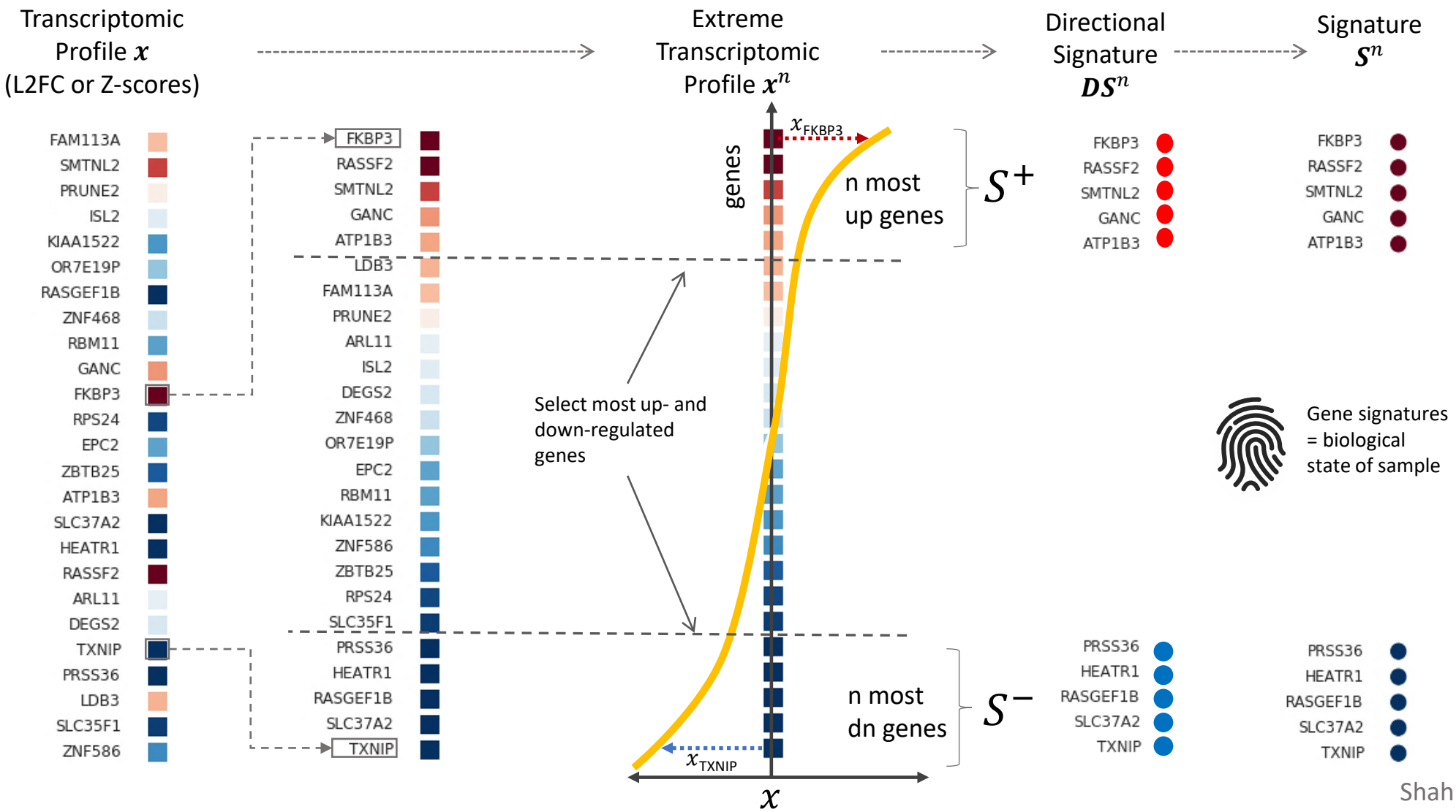
Find “Signature” of differential gene expression to define biological state associated with disease, chemical treatment, etc.

Database of differential gene expression profiles for reference chemicals with known biological activities

Pattern-matching algorithm to score the query signature and reference profiles

A set of high-scoring hits for the signature signifying connections with mechanisms, diseases or other phenotypes

Formalizing some concepts …



Shah *et al.* (in prep)

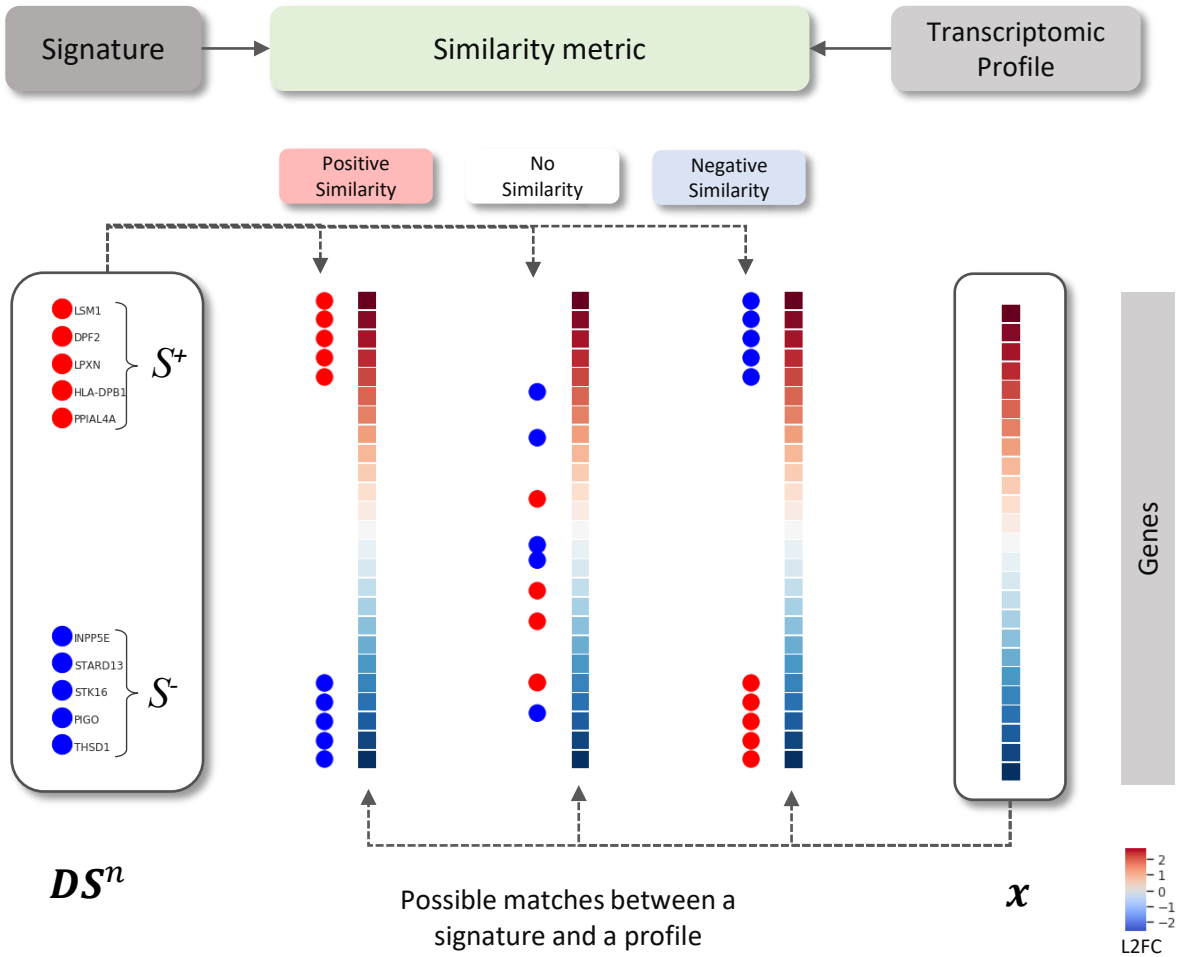
Generalizing Connectivity Mapping

- "Connectivity mapping" (Lamb *et al.,* 2006)
  - Query: directional signature ($\boldsymbol{DS}^n$)
  - Reference: transcriptomic profiles in Cmap v2 ($\boldsymbol{x}$)
  - Similarity metric: Gene set enrichment Analysis (GSEA) (Subramanian *et al.* 2005)
- Can generalize this as:

$$s = SM(O_q, O_r)$$

s = similarity / connectivity score
SM = similarity metric
O = gene set "Object"

$O \in \{\boldsymbol{x}, \boldsymbol{x}^n, DS^n, S^n\}$



Shah *et al.* (in prep)

Similarity metrics beyond GSEA …

Aggregation-based metrics

Type	Method	Similarity Metrics (SM)
Aggregation-based enrichment scoring	eXtreme Sum (XS)	$\sum_{i\in Q} \lvert \boldsymbol{x_{ri}} \rvert - \sum_{i\in Q'} \lvert \boldsymbol{x_{ri}} \rvert$
	eXtreme Mean (XM)	$\frac{1}{q}\sum_{i\in Q} \lvert \boldsymbol{x_{ri}} \rvert - \frac{1}{q'}\sum_{i\in Q'} \lvert \boldsymbol{x_{ri}} \rvert$
	T-statistic (TT-p)	$ts = \frac{\overline{x_r[Q]} - \overline{x_r[Q']}}{\sqrt{\frac{\sigma_q^2}{q}+\frac{\sigma_{q'}^2}{q'}}}$; $\sigma_q^2 = \frac{1}{q}\sum_{i\in Q}(x_{ri} - \overline{\boldsymbol{x_r}[Q]})^2$, $\sigma_{q'}^2 = \frac{1}{q}\sum_{i\in Q'}(x_{ri} - \overline{\boldsymbol{x_r}[Q']})^2$
	Ranksum statistic (RS)	$rs = \min\left(qq' + \frac{q(q+1)}{2} - \sum y_r \, , qq' + \frac{q'(q'+1)}{2} - \sum y'_r\right);\ y = rank(x)$
	Kolmogoror-Smirnov statistic (GSEA)	$\mathrm{ES} = max_{1\le j\le m}(S_i - S'_i)$; $S_i = \sum_{\substack{i\in R \\ j\le i}} \frac{\lvert x_j \rvert^b}{\sum_{i\in R} \lvert x_i \rvert^b}$, $S'_i = \sum_{\substack{i\in R' \\ j\le i}} \frac{\lvert x_j \rvert^b}{\sum_{i\in R'} \lvert x_i \rvert^b}$
	Total enrichment score (TES)	$TES = 1 - \frac{ES^+ - ES^-}{2}$

Vector-based metrics

Type	Method	Similarity Metrics (SM)
Vector-based similarity scoring	Extreme Pearson correlation (XCP)	$\frac{cov(\boldsymbol{x}_q, \boldsymbol{x}_r)}{\sigma_q \sigma_r}$
	Extreme Spearman Correlation (XCS)	$\frac{cov(\boldsymbol{y}_q, \boldsymbol{y}_r)}{\sigma_{y_q} \sigma_{y_r}}, y = rank(\boldsymbol{x})$
	Extreme Cosine (XC)	$\frac{\boldsymbol{x_q} \cdot \boldsymbol{x_r}}{\lvert \boldsymbol{x_q} \rvert \lvert \boldsymbol{x_r} \rvert}$
	Jaccard index (JI)	$J(Q,R) = \frac{Q \cap R}{Q \cup R}$
	Signed Jaccard (SJI)	$\frac{J(Q^+,R^+) + J(Q^-,R^-) - J(Q^+,R^-) - J(Q^-,R^+)}{2}$
	Szymkiewicz–Simpson index (SI)	$S(Q,R) = \frac{Q \cap R}{\min(n_q, n_r)}$
	Signed Szymkiewicz–Simpson index (SSI)	$\frac{S(Q^+,R^+) + S(Q^-,R^-) - S(Q^+,R^-) - S(Q^-,R^+)}{2}$

Shah *et al.* (in prep)

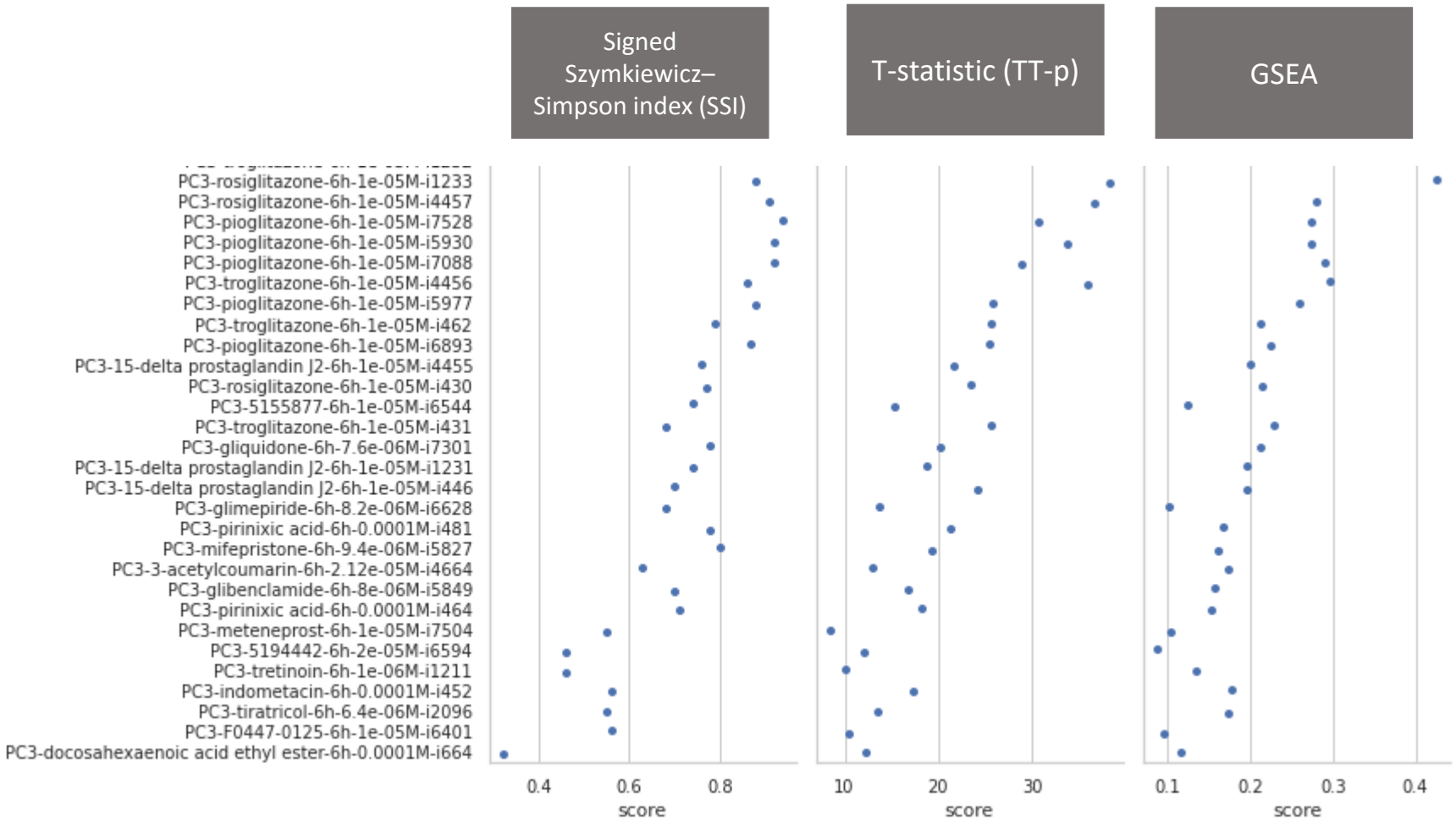
Example: Troglitazone Signature Connections in CMap

Data set

- CMap v2 (Lamb 2017)
- 1,309 chemicals (6 h) and 3 cells: MCF7, PC3 and HL60
- Transcriptomics: Affymetrix U133A GeneChip (22,215 transcripts associated with 13,609 genes).
- Differential expression (L2FC) by comparison with DMSO controls
- 6,100 differential expression profiles ($\boldsymbol{x}$)

Example

- Query: troglitazone 10µM (PC3 cells) top 50-up and 50-dn genes
- Reference: 6,100 profiles $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots.\}$
- Best hits: Thiazoladinediones and PPAR-activators
- Multiple similarity metrics – similar results



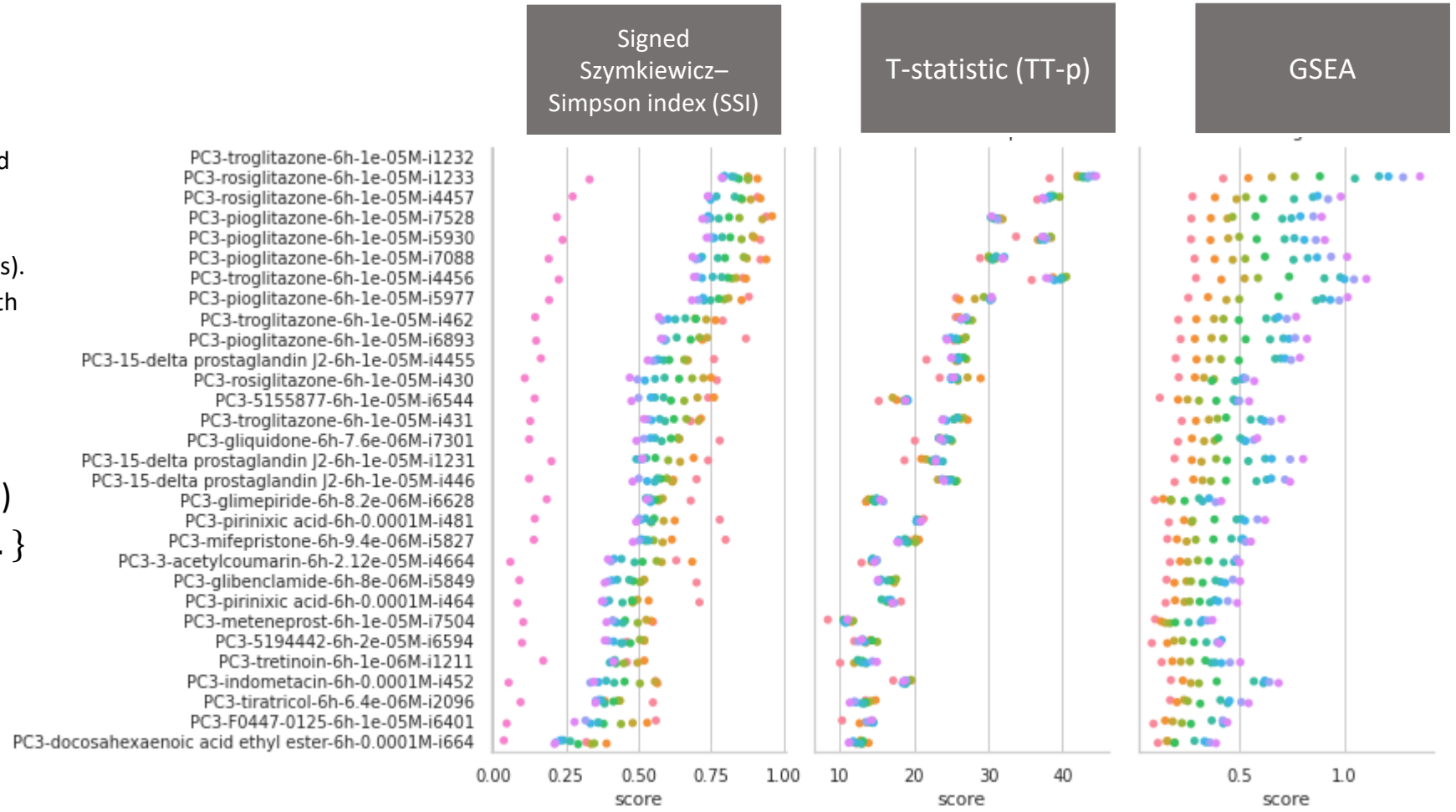
Troglitazone Signature Connections in CMap

Data set (Query and Reference)

- CMap v2 (Lamb 2017)
- 1,309 chemicals (6 h) and 3 cells: MCF7, PC3 and HL60
- Transcriptomics: Affymetrix U133A GeneChip (22,215 transcripts associated with 13,609 genes).
- Differential expression (L2FC) by comparison with DMSO controls
- 6,100 differential expression profiles ($\boldsymbol{x}$)

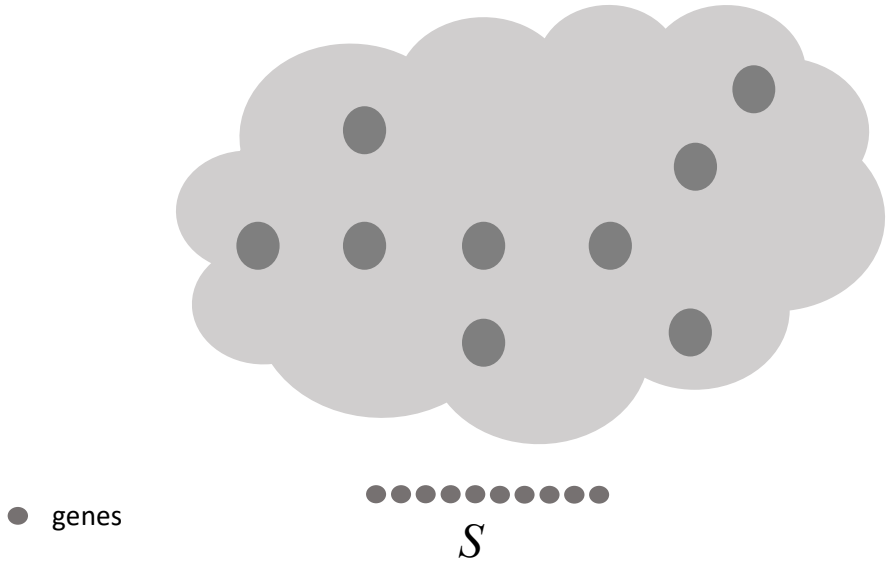
Example

- Query: troglitazone 10µM (PC3 cells)
- Reference: 6,100 profiles $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots.\}$
- Best hits: Thiazoladinediones and PPAR-activators
- Multiple similarity metrics – similar results



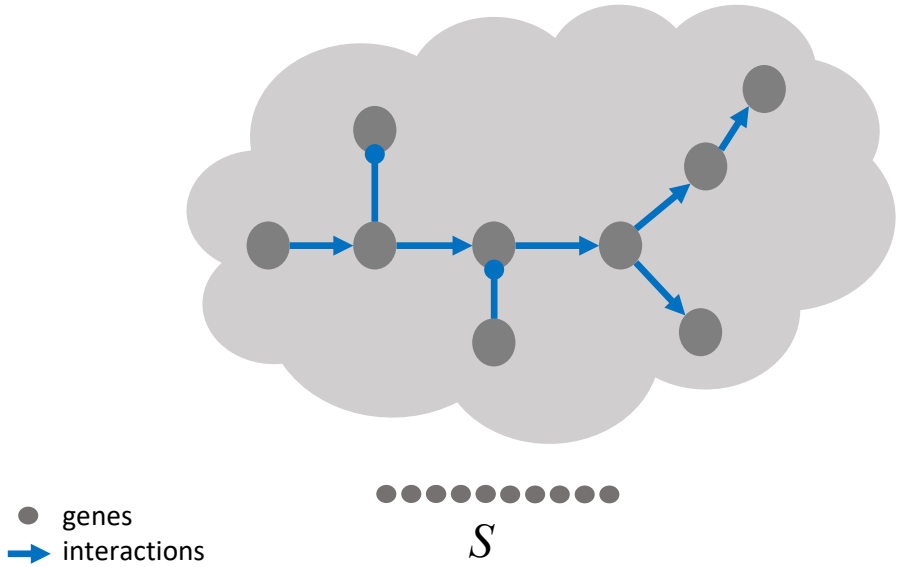
Other gene sets: canonical pathways

MSigDB pathways and signatures



Signature = S

Reactome/KEGG/etc. fancy network

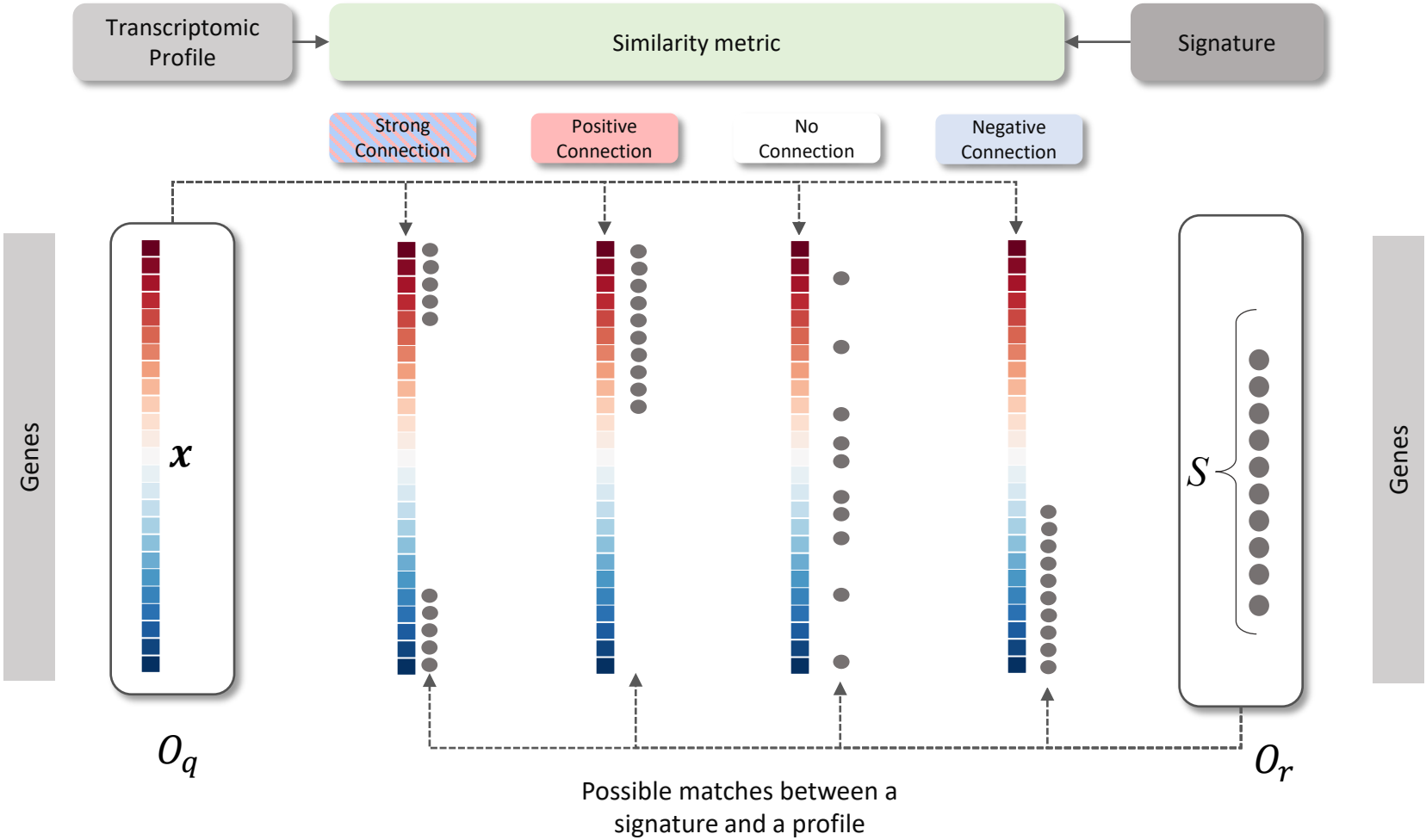


Signature = S

MSigDB: Molecular Signatures Database

Shah *et al.* (in prep)

Matching profiles to disease or pathway signatures



Shah *et al.* (in prep)

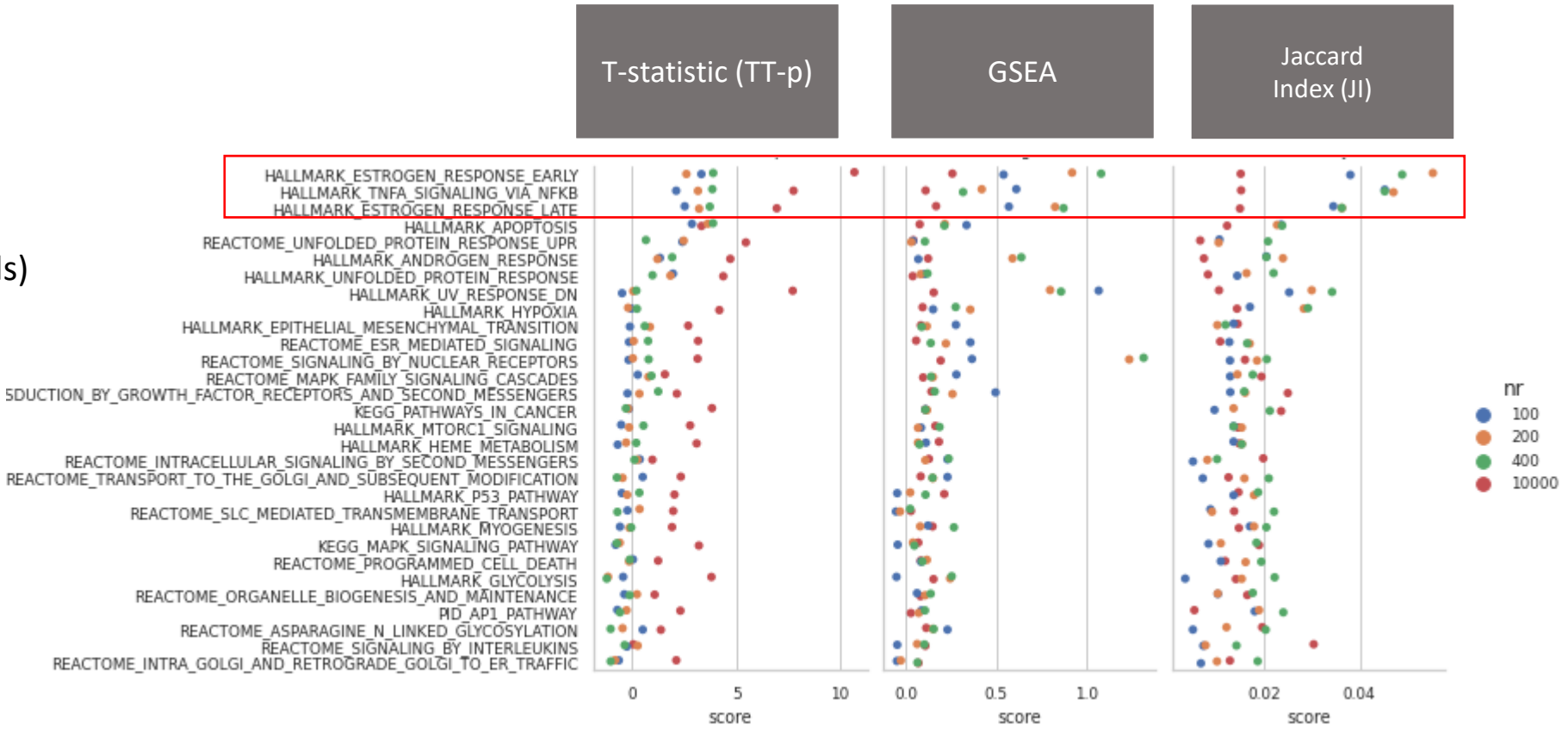
Estrogen Signature Connections in MSigDB

Data set

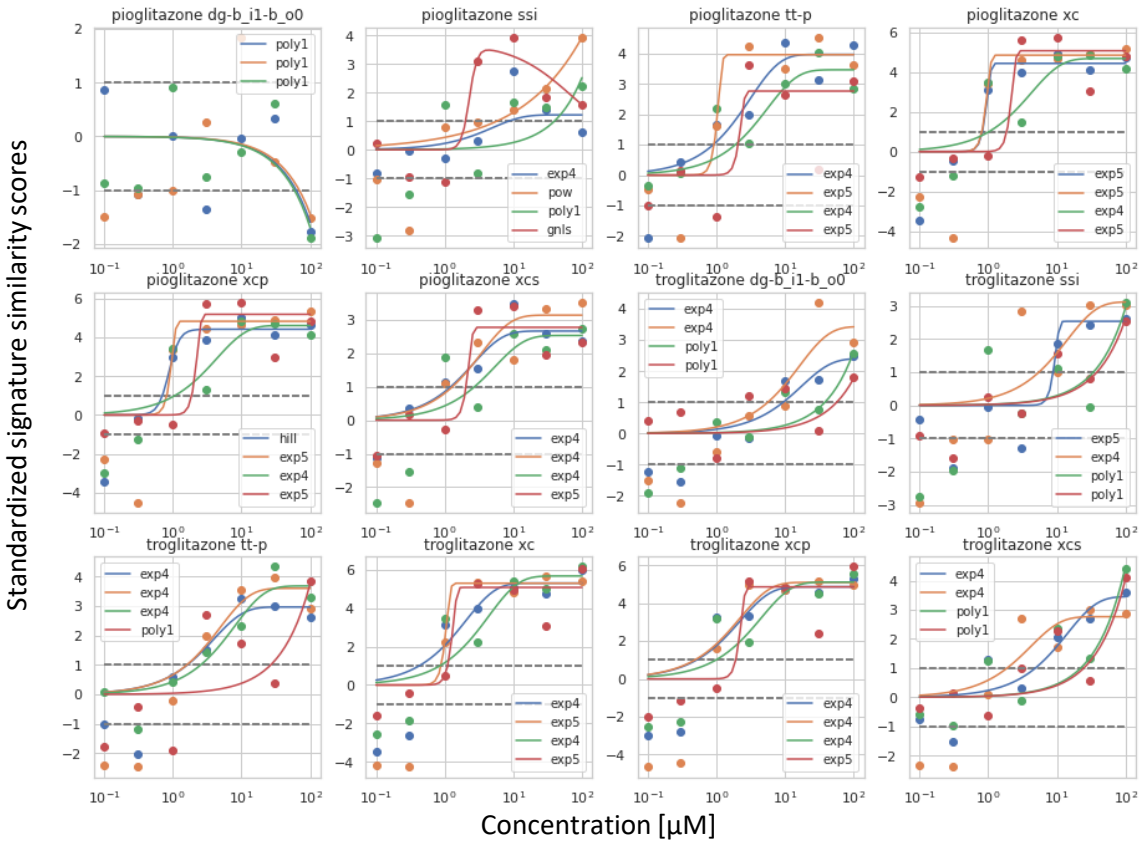
- Reference: MSigDB v7 (26,860 signatures)
- Query: CMap v2

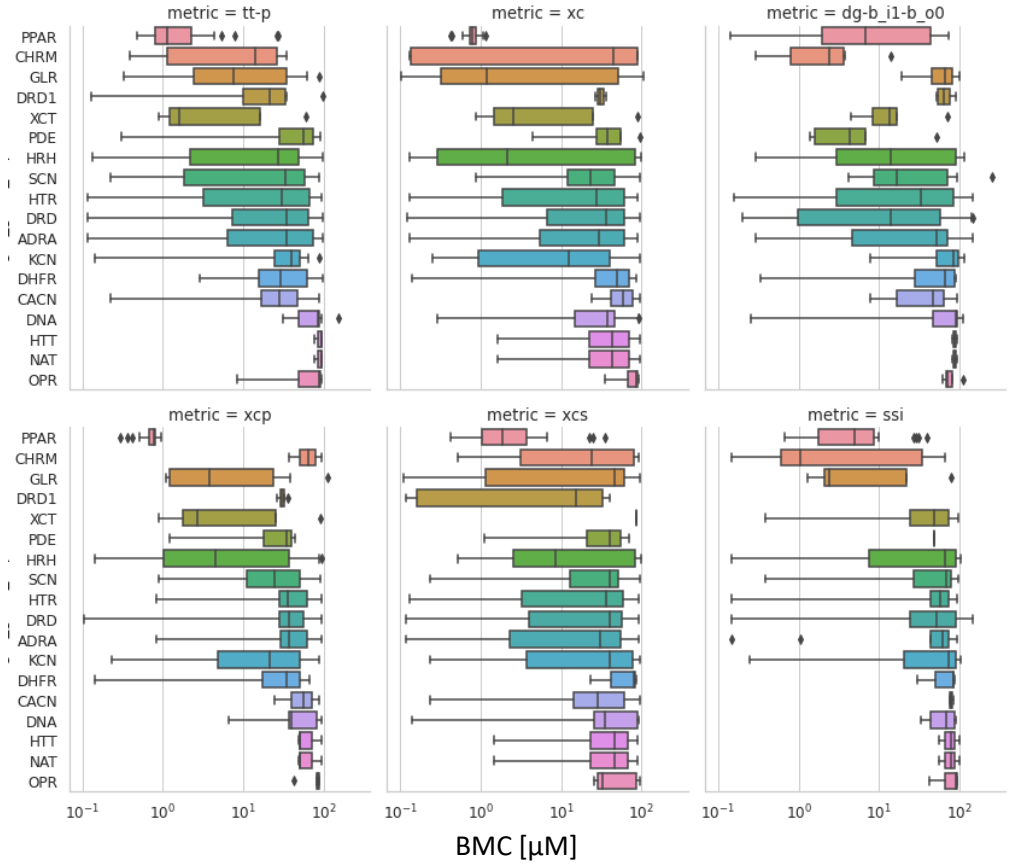
Example

- Query: Estrogen 14.6 µM (MCF7 cells)
- Reference: 2,253 canonical and hallmark pathways
- Best hits: Estrogen responses
- Multiple similarity metrics – similar results
- Difference signature sizes impact scores and depend on metric used



Signature similarity scores to PODs: Troglitazone



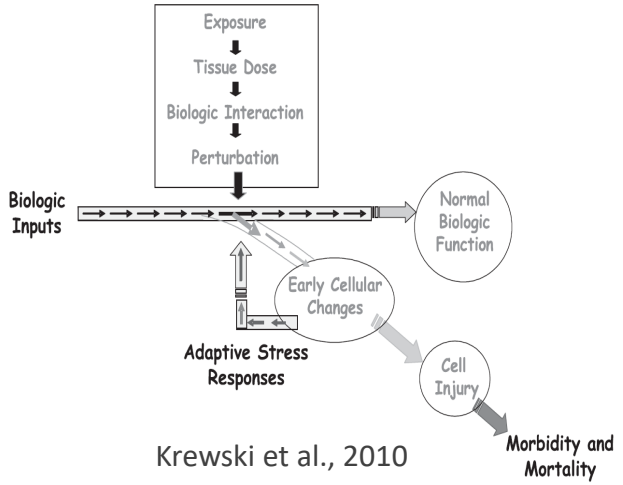


- Data: HepaRG cells treated with 1,366 chemicals 0.01-100μM
- Query: Troglitazone profiles replicated across 29 batches
- Reference: CMap v2 Affymetrix data
- Connectivity analysis using multiple signatures and metrics
- Score signatures against random profiles to estimate background

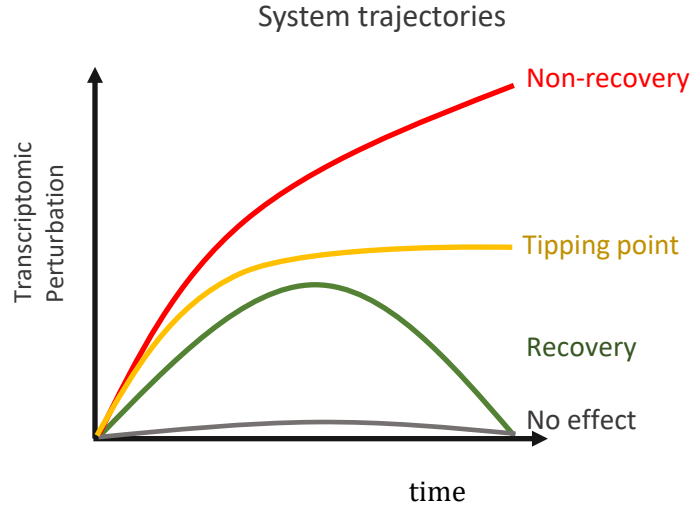
- Standardize similarity scores using background distribution (Z)
- Concentration-response modeling using tcplFit2 (cnst, hill, gnls, poly1, poly2, pow, exp2, exp3, exp4, exp5)
- Estimate benchmark concentration (BMC) using benchmark response (BMR) of Z=1

Challenges & Future directions

- Chemicals cause toxicity via complex pathways that are poorly defined. Two main conceptual approaches to map adverse outcome pathways (AOPs) using transcriptomics:-
  - Specific receptor-mediated mechanisms (e.g. nuclear receptor-mediated developmental or reproductive effects)
  - Non-specific adaptive stress response pathways (e.g. oxidative stress, unfolded protein response, etc.)
- We are developing transcriptomic signatures of receptor-mediated and non-specific adaptive stress response pathways
- Hypothesis: Increasing the level of chemical(s) beyond "tipping point" can overwhelm the adaptive stress responses and result in adverse outcomes
  - HepG2 cells using high-content imaging (HCI) to measure time-course cell phenotypic data (Shah et al. 2016)
  - Developing rat neuronal networks and time-course microelectrode array data on electrophysiological activity (Franks et al. 2018)
  - Induced pluripotent stem cells and time-course transcriptomic data during endodermal differentiation linked to ATRA signaling and toxicity (Saili et al. 2020)
- We are also developing NAMs to estimate tipping points from transcriptomic and other data streams using systems modeling



Krewski et al., 2010



Shah et al. 2016

Summary

1. High-throughput transcriptomics is promising for NAM development

   We are using TempO-Seq technology (targeted RNA-Seq) to evaluate thousands of chemicals in multiple cell lines and have developed a high-throughput pipeline to process and analyze transcriptomic concentration-response data.

2. Feasible to identify hazard and estimate POD using gene signature "similarity"

   We are systematically evaluating gene signature-based connectivity mapping and other approaches for identifying putative targets, AOPs and *in vitro* POD values. Gene signature-based approaches are more sensitive than single gene-based techniques.

3. Connectivity mapping, read-across and risk assessment

   Transcriptomic nearest-neighbor techniques are conceptually like expert read-across approaches, which are widely used to fill data gaps for untested chemicals. Could be easier "sell" than more sophisticated network inference and AI/ML/DL.

4. Future directions

   Systems biology of adaptive stress response pathways using transcriptomics to investigate the molecular basis of cellular resilience and tipping points; streamline the development of NAMs for evaluating untested chemicals based on adaptive stress responses.

Acknowledgements

US EPA

Joshua Harrill

Logan Everett

Richard Judson

Bryant Chambers

Woodrow Setzer

Joe Bundy

Grace Patlewicz

Derik Haggard

Tia Tate

Chris Corton

Beena Valanat

Thomas Knudsen

Tim Shafer

Brian Chorley

John Cowden

Russell Thomas

UniLever, UK

Alistair Middleton

Andy White

Paul Carmichael

Joe Reynolds

University of Cambridge, UK

Andreas Bender

Danilo Basili

A*Star