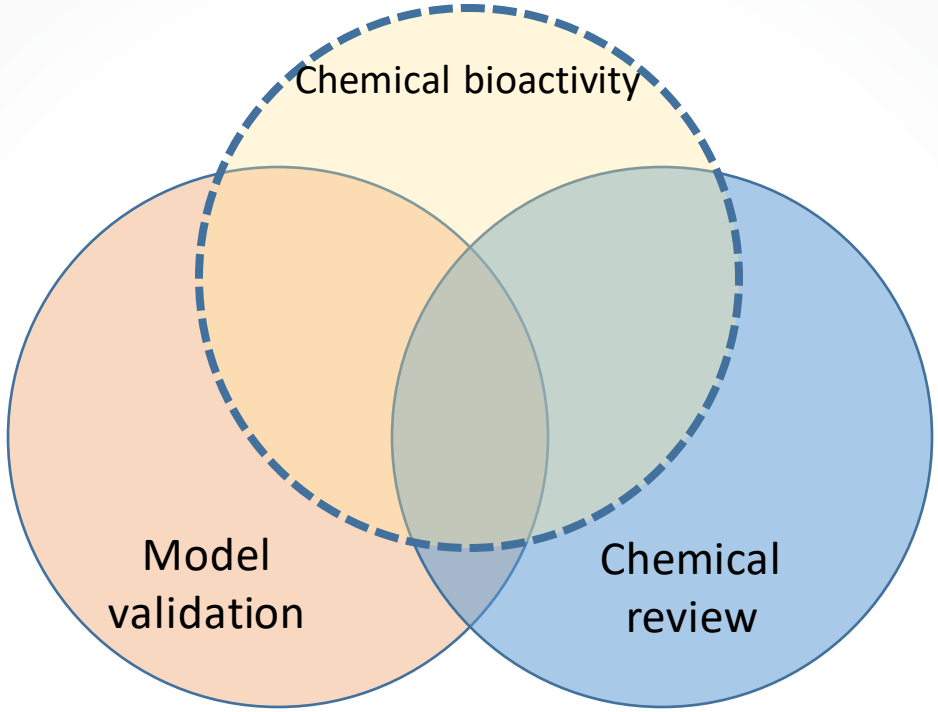
Mining chemical endpoints and gene targets from literature to automate hazard assignment for predictive toxicology

Systematic Review Communities of Practice

November 9th, 2021

Bryant A. Chambers, Ph.D.

A shared need for annotated chemical activity

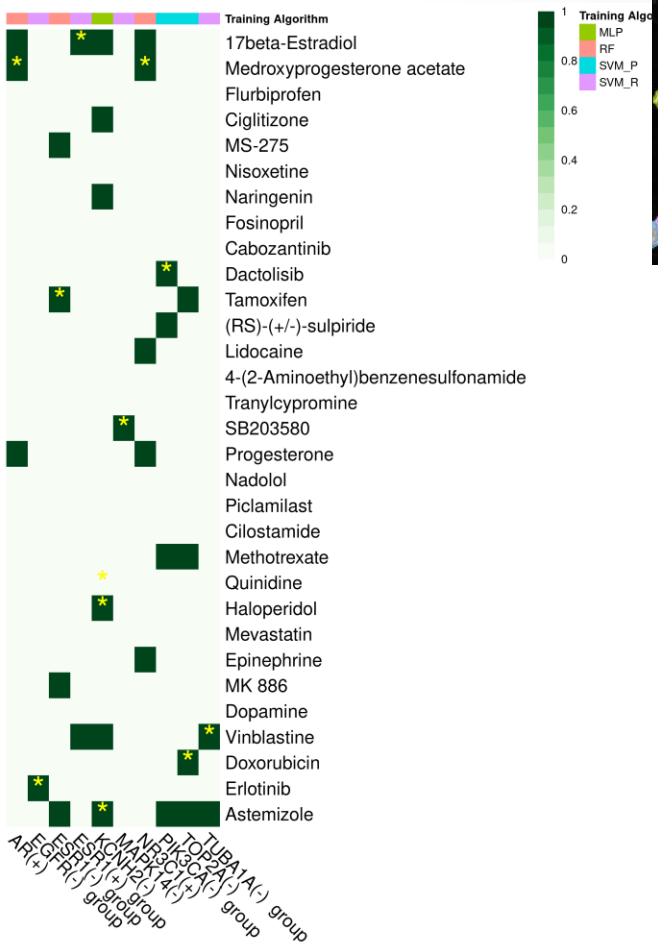


- Reviewers and modelers both seek accuracy to protect human health

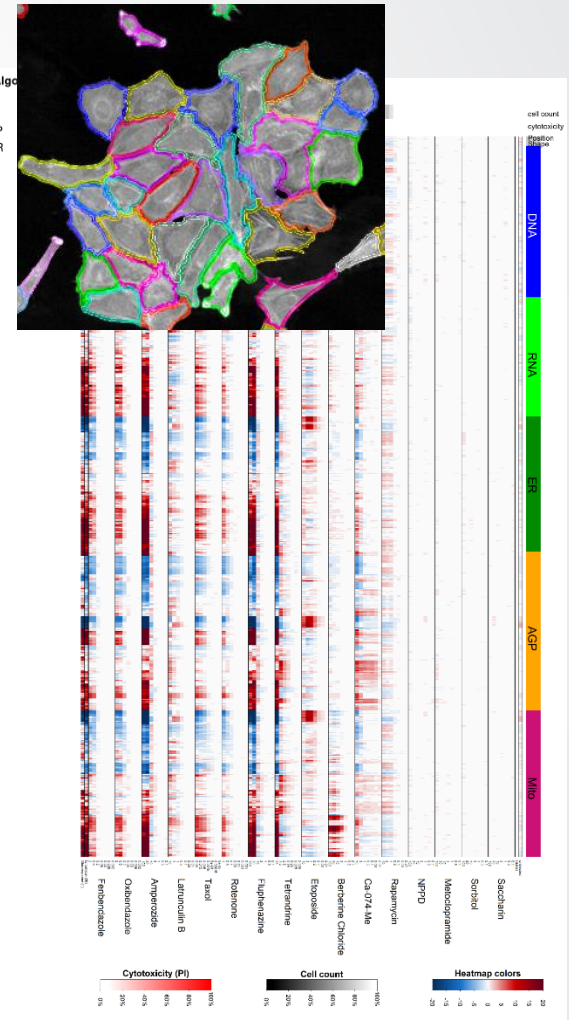
- We can use each other practices to build better models and provide data more useful to reviewers

… Just a little introduction

- Center for Computational Toxicology and Exposure
  - Develop New Approach Methods (NAMs)
    - High Throughput Screening – testing of hundreds – thousands of chemicals / screen
    - High Content Assays – measurement of hundreds-thousands of features within a single assay (cell painting/whole transcriptome)
- **Within CTBB:**
  - Tiered Testing
  - Molecular Initiating Event classification (Adverse Outcome Pathways)
  - Pathway and mechanistic classification
  - Pathway, gene, and cellular feature activity and response
  - Tipping Points



Credit: Joe Bundy



Credit: Johana Nyffeler

Identification of stress response active chemicals

1

The Major Adaptive Stress response pathways

Stress response pathway	Chemical inducers	TF	Activated gene promoters
OSR: Oxidative stress	Quinones, hydroperoxides, heavy metals, trivalent arsenicals	Nrf2	*HMOX1*, *NQO1*, *GST2A*
HSR: Heat shock response	Heat, Heavy Metals	HSF-1	*HSPA6*
DDR: DNA damage response	Etoposide, Methyl Methanesulfonate, N-Dimethylnitrosamine, Cyclophosphamide, UV radiation	p53	*CDKNIA*, *GADD45A*, *MDM2*, *BCL2*, *TP53I3*
HPX: Hypoxia	Hypoxia, Cobalt, Desferriozamine, Quercetin, Dimethyloxalylglycine	HIF-1	*VEGF*, *TF*, *EPO*
UPR: ER stress	Tunicamycin, Thapsigargin, Caplain, Brefeldin A	XBP-1, ATF6, ATF4	*HSP90B1*, *HSPA5*, *DNAJB9*
MTL: Metal stress	Heavy Metals	MTF-1	*MT1E*, *MT2A*
Inflammation	Metal, PCBs, Exhaust Particles, Smoke Particles	NF-κB	*IL1A*, *TNFA*
Osmotic stress	High salt, polyethylene glycol, mannitol	NFAT5	*AKR1B1*, *SLC6A12*, *SLC5A3*

Simmons et al., 2009

2

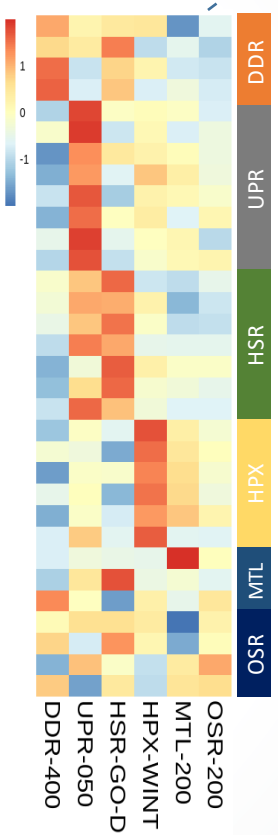
Published Signatures (MSigDB v7.2)

Overlapping Consensus Genes

Unique Consensus Signatures

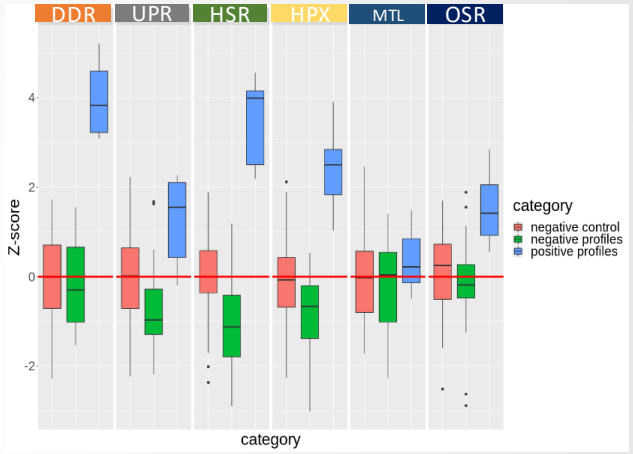
Connectivity Mapping with Transcriptomic Data for reference Chemicals (GSEA)

3



Performance

4



- Consensus signatures for most SRPs can identify reference perturbagens
- Can use SRP signatures to evaluate non-specific chemicals

Chambers & Shah, 2021

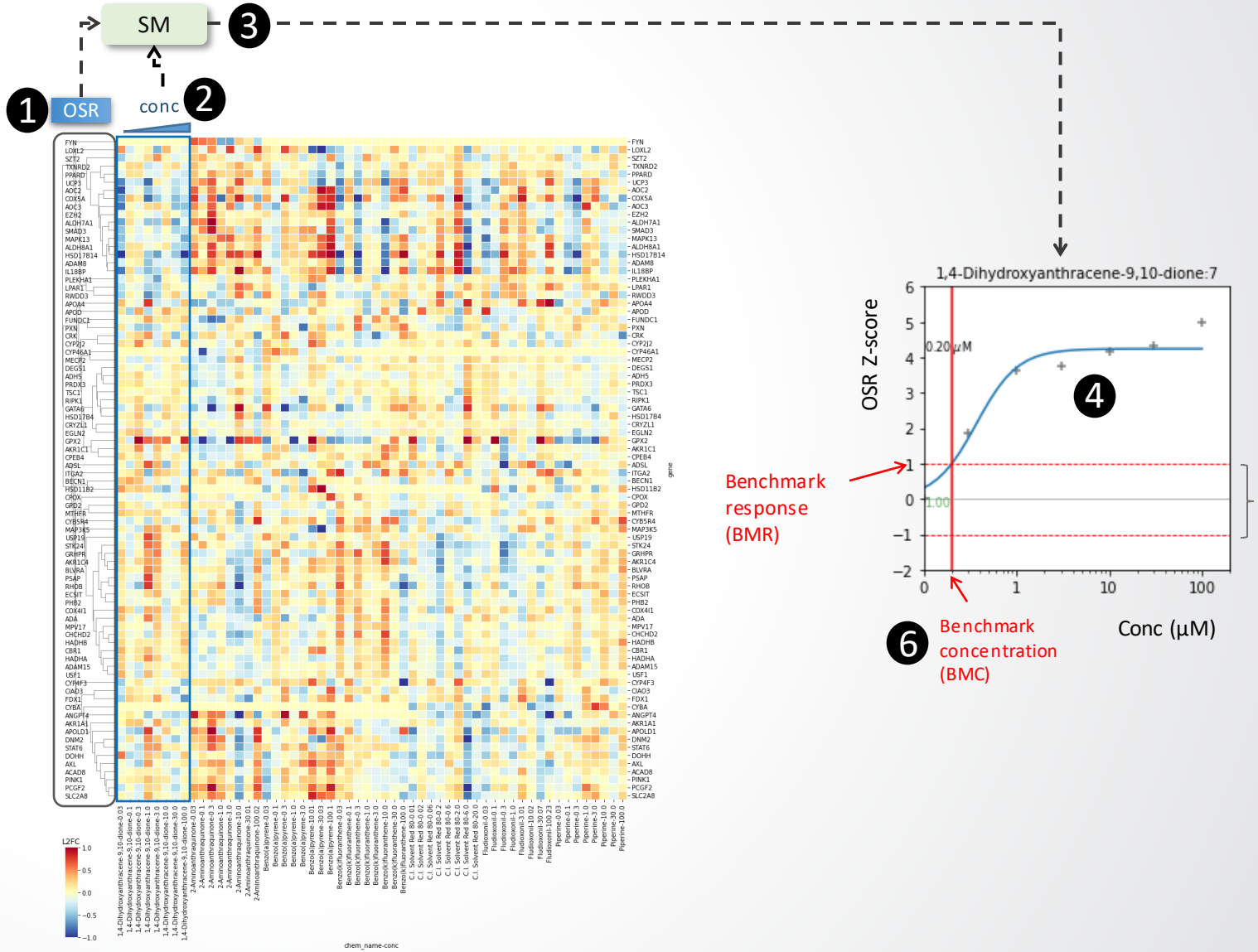
Slide: Imran Shah

Using gene signature scores to identify SRP "actives"

Signatures used to score large transcriptomic datasets
- Tuning of methodology necessary

Measure of pathway activity
- Known?
- Novel?

How to assess accuracy of score and chemical categorization
- Potential rouge signal?
- Background?
- True activity
- Co/cross activation



Slide Adapted from Imran Shah

Systems overlap drives need for specificity

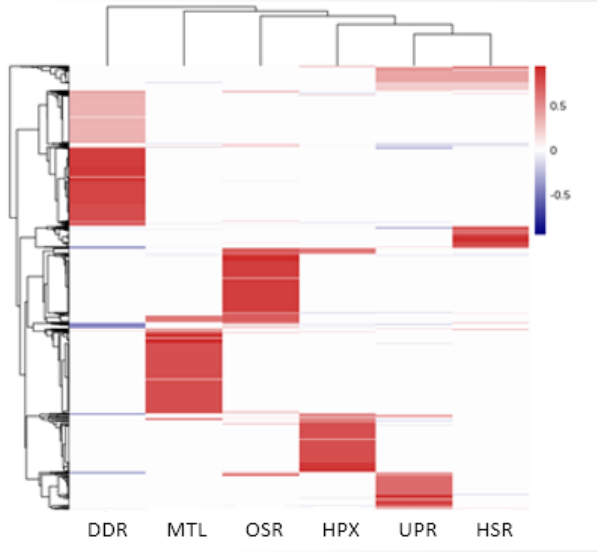
- Stress inducing reference chemicals are hard to classify

  This is in part due to:
  - Cross-talk between systems
  - Dose and Time factors
  - Cell specific effects
  - Overlapping system activation (e.g., oxidative stress and DNA damage; vascularization and hypoxia)

- SRP cross-talk necessitate the formulation of highly specific and highly sensitive bioactivity assays

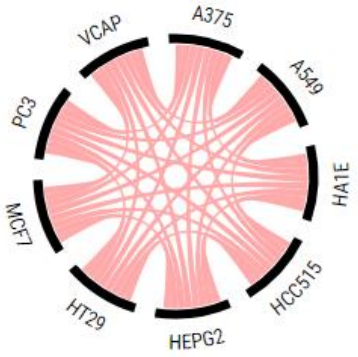
- **Can we use literature review and mining to validate transcriptomic chemical activity quickly?**

Almost 20% of genes overlapping in curated sets:

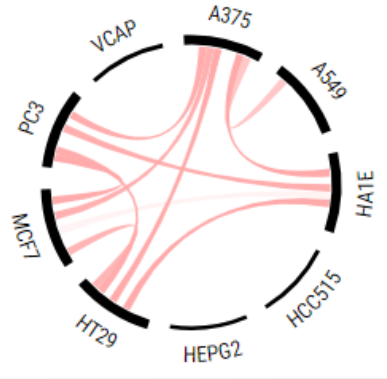


Variation in cell TRx

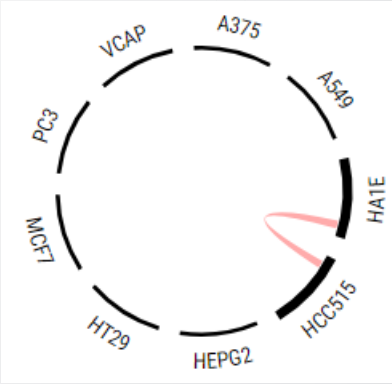
Brefeldin-a



Cobalt (II) Chloride



Benzo(a)pyrene



Automated systematic review to identify SRP actives

- Can we employ some of these systematic review process to help us score chemicals and predict transcriptomic activity?

- Question: Is a chemical SRP active?
  - Does the chemical activate one or more of the 6 canonical stress pathways?

- Method: Score chemicals by using literature query results and summarize output quantitatively by information content and text mining
  - Utilize expanded query terms
  - Link with additional DBs

- Result: Output chemical-SRP association rank with each possible activity

Rapid Measurement of Literature-SRP Association Strength

- Approach to quantify chemical-bioactivity associative strength in literature
  - Automated & High throughput
  - Ranked output for all chemical – target combinations
- Make use of two approaches
  - Pairwise Mutual Information (PMI)
    - Joint frequency vs independent frequency of two comparators
  - Text mining (gene extraction)
    - Extraction of key elements to match with prior information

❶ PMI scores joint frequency vs independent frequency

$$PMI(chemical, stress) = \log \frac{F(chemical, stress)}{F(chem)F(stress)}$$

❷ Text mining pulls entities from which relevance can be scored

Since Inhibitor of Apoptosis (IAP) proteins have been implicated in cellular adaptation to endoplasmic reticulum (ER) stress, we investigated the regulation of ER stress-induced apoptosis by small-molecule second mitochondria-derived activator of caspase (Smac) mimetics that antagonize IAP proteins. Here, we discover that Smac mimetic suppresses tunicamycin (TM)-induced apoptosis via resolution of the unfolded protein response (UPR) and ER stress. Smac mimetics such as BV6 selectively inhibit apoptosis triggered by pharmacological or genetic inhibition of protein N-glycosylation using TM or knockdown of DPAGT1, the enzyme that catalyzes the first step of protein N-glycosylation. In contrast, BV6 does not rescue cell death induced by other typical ER stressors (i.e., thapsigargin (TG), dithiothreitol, brefeldin A, bortezomib, or 2-deoxyglucose). The protection from TM-triggered apoptosis is found for structurally different Smac mimetics and for genetic knockdown of cellular IAP (cIAP) proteins in several cancer types, underlining the broader relevance. Interestingly, lectin microarray profiling reveals that BV6 counteracts TM-imposed inhibition of protein glycosylation. BV6 consistently abolishes TM-stimulated accumulation of ER stress markers such as glucose-regulated protein 78 (GRP78) and C/EBP homologous protein (CHOP) and reduces protein kinase RNA-like ER kinase (PERK) phosphorylation and X box-binding protein 1 (XBP1) splicing upon TM treatment. BV6-stimulated activation of nuclear factor-κB (NF-κB) contributes to the resolution of ER stress, since NF-κB inhibition by overexpression of dominant-negative IκBα superrepressor counteracts the suppression of TM-stimulated transcriptional activation of CHOP and GRP78 by BV6. Thus, our study is the first to show that Smac mimetic protects from TM-triggered apoptosis by resolving the UPR and ER stress. This provides new insights into the regulation of cellular stress responses by Smac mimetics.

Gene
Chemical
Cellular Process

Linking Rich TRx DBs with Literature

1

ConnectivityMap

Lit Database

PubMed

- 31M Articles
- 3.3B Searchers/y
- Query API / Programmatic Access

TRx Database

Library of Integrated Network-based Cellular Signatures (LINCS)

- 20K Chemicals
- 300K Transcriptomic Profiles
- Multi - Time/Conc/Cell
- 12K Genes

2

Subset a searchable list of chems

3

Information and Text Mining Approaches

Pointwise mutual information

Abstract GSEA

4

Transcriptomic Approaches

GSEA profiles

Similarity scoring

5

Evaluate Accuracy

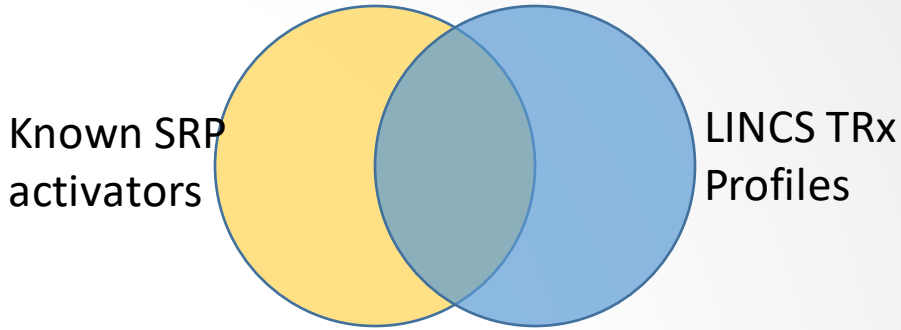
Filter and validate high scoring subset along side hallmark chems

6

1. Expanded automated profile clustering
2. Construct multilevel model explaining assignments and cell dependency

High confident set used to tune process

- Find overlap of high confidence reference chemicals
  - Well established literature and history



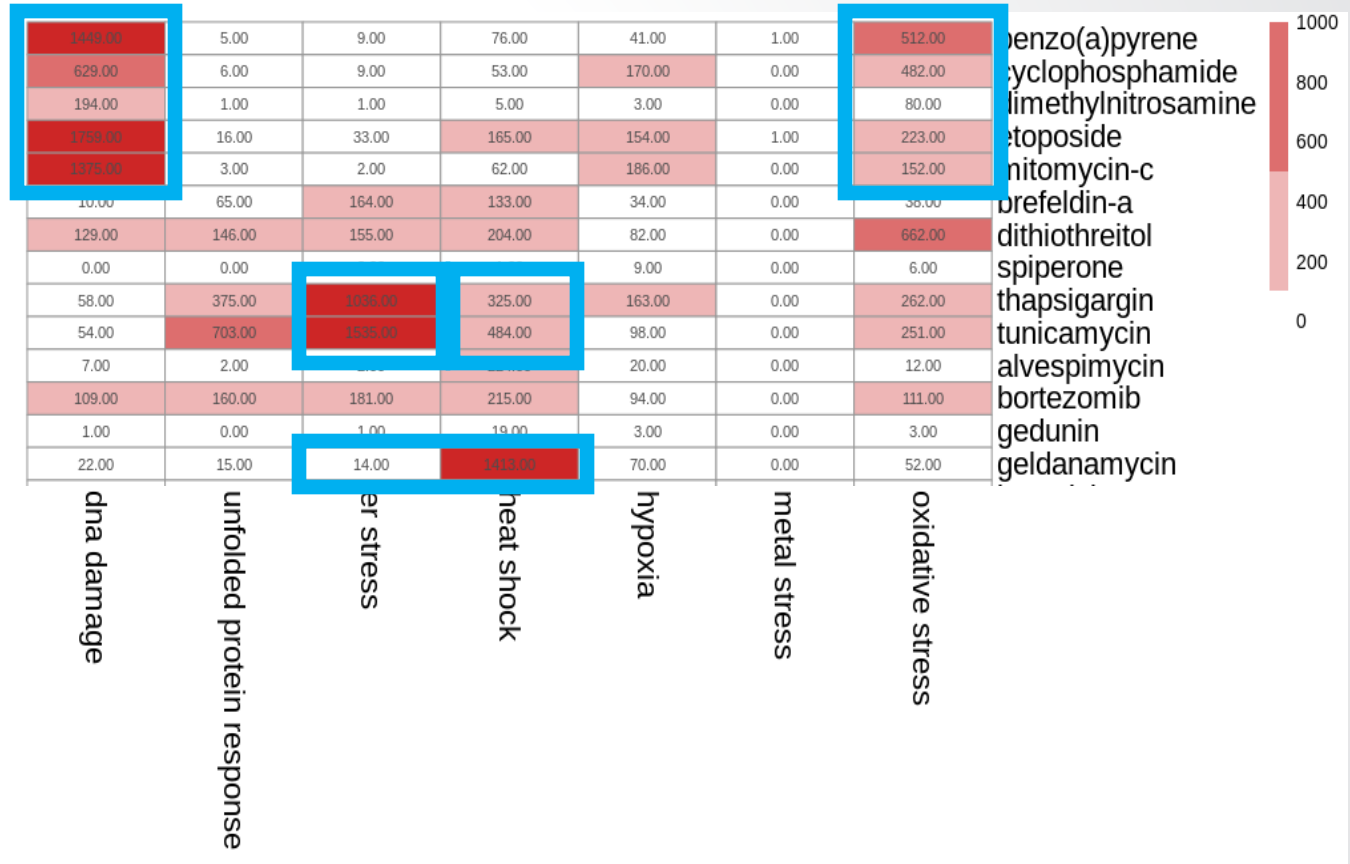
Stress Response Pathway	Chemical
DDR	benzo(a)pyrene, etoposide, mitomycin-c
HSR	radicicol, geldanamycin, bortezomib
HPX	cobalt II chloride, YC-1
MTL	cadmium chloride
OSR	tert butylhydroquinone, 1,2, dichlorobenzene, amodiaquine
UPR	brefeldin-a, thapsigargin, tunicamycin

- Look for the presence of these chemical in clusters and review additional chemicals

Search results associate with knowns

Source chemical profiles in LINCS

1. Identify list of chemicals perturbagens in LINCS (~20,547 chemicals)
2. Remove chemical only identified by Broad Institute ID (e.g., BRD-)
   - Reduced to 4671 chemicals
3. Search PubMed with all chemicals & a set of stress response pathways (SRPs)
   - Used EZ PubMED R utility for queries and XML parsing
   - 9 terms (e.g., 'dna damage', 'er stress', unfolded protein response')
   - Totaled 42,039 searches
   - 609,698 Abstracts



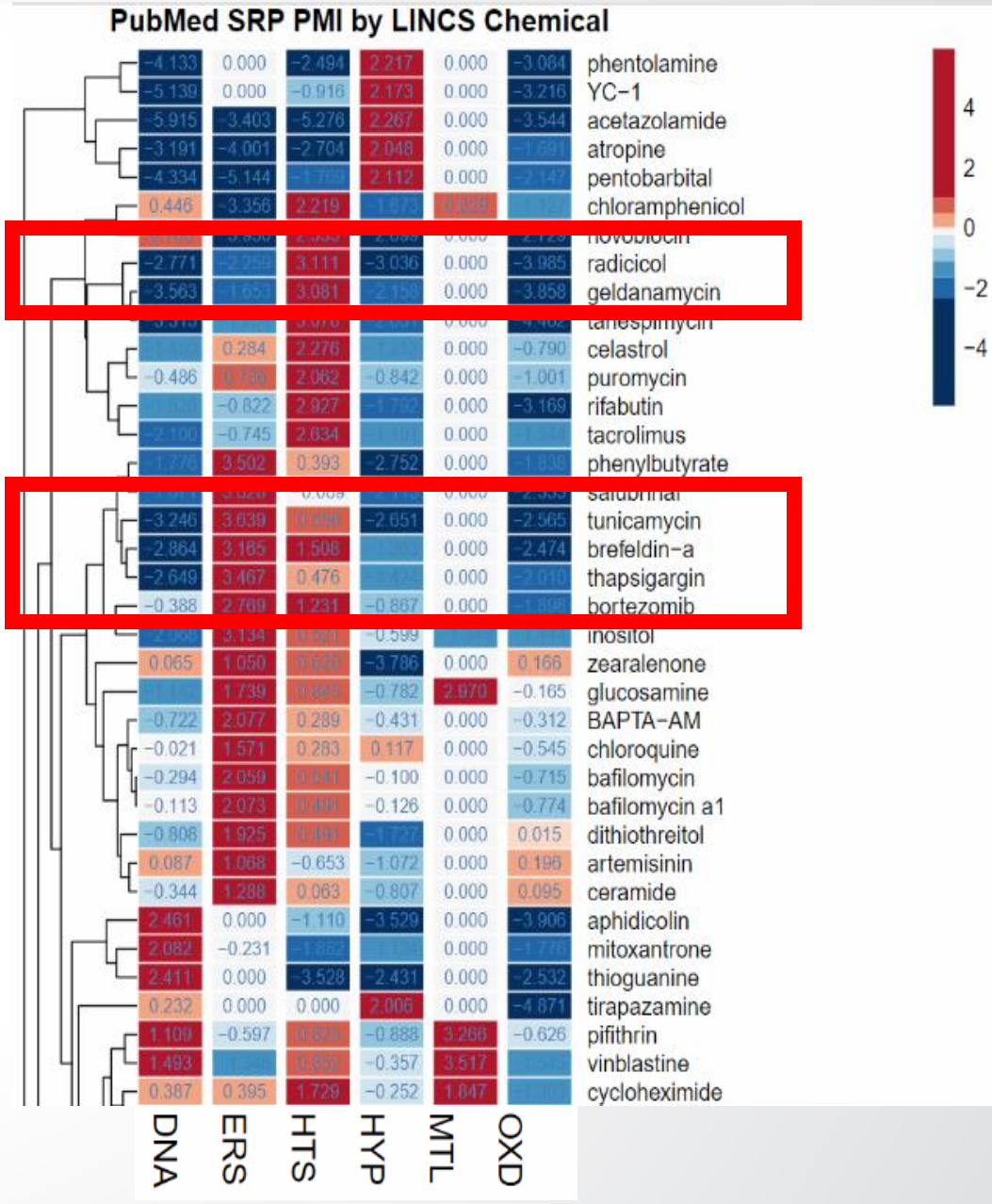
Example of reference counts for a common set of DNA damaging and ER stressing chemicals

Search pair information content normalizes outcome

- Calculated as:

$$PMI(chemical, stress) = \log \frac{F(chemical, stress)}{F\ (chem)F(stress)}$$

  - F(x,y) is the joint relative frequency and F(x) is the relative frequency
  - PMI aids in understanding the associated amount of information
  - Chemicals with low search returns but only in one stress category become important
  - PMI calculated on a subset with more than 5 returned abstracts in all stress categories



Expert Curated Reference Set

- Initial list of 93 chemicals were identified in from HC clusters
- The first 500 abstracts were pulled
- Hand Validated
- 68 chemicals (73%) were selected for further TRx analysis
- Issues
  - Hypoxia associated with organismal hypoxia/protection
  - Metal associated with protection chelation

stresscat	pert_iname	n_cell_types	cell_types	n_time_points	time_points	n_doses	dosing_conc	mber_of_profi	id	description	direction	PMID	Linked PMI	NOTES
DNA	azacitidine	22	A375, A549, .	2	6, 24	9	0.04, 0.12, 0.	181	BRD-K03406	DNA methyltr	positive	32676814	link	some potentia
DNA	benzo(a)pyre	4	HA1E, HCC5	2	6, 24	1	10	8	BRD-K09668	pro carcinoge	positive	32866872	link	NA
DNA	camptothecir	14	A375, A549, .	2	6, 24	12	0.001, 0.01, (	129	BRD-A30437	topoisomeras	positive	32846134	link	some potentia
DNA	cyclophospha	8	A375, A549,	2	6, 24	2	10, 20	29	BRD-A09722	alkylating agr	positive	32717509	link	some potentia
DNA	cytarabine	12	A375, A549,	2	6, 24	6	0.04, 0.12, 0.	61	BRD-K33106	antimetabolite	positive	32474729	link	NA
DNA	dacarbazine	7	A375, HA1E,	2	6, 24	6	0.04, 0.12, 0.	46	NA	NA	positive	25697728	link	NA
DNA	daunorubicin	21	A375, A549, .	2	6, 24	12	0.001, 0.01, (	171	BRD-K43389	RNA synthesi	positive	32554494	link	NA
DNA	dimethylnitros	1	HA1E	2	6, 24	1	10	3	NA	NA	positive	31092975	link	NA
DNA	DMBA	1	HA1E	2	6, 24	1	10	3	NA	NA	positive	32159784	link	NA
DNA	doxorubicin	14	A375, A549, .	2	6, 24	12	0.001, 0.01, (	192	BRD-K92093	topoisomeras	positive	27852227	link	NA
DNA	etoposide	14	A375, A549,	2	6, 24	9	0.04, 0.12, 0.	100	BRD-K37798	topoisomeras	positive	32866497	link	some potentia
DNA	gemcitabine	53	A375, A549, .	2	6, 24	12	0.001, 0.01, (	195	BRD-K15108	cell cycle inhi	positive	32688248	link	some potentia
DNA	hydroxyurea	7	A375, HA1E,	1	24	6	0.04, 0.12, 0.	42	NA	NA	positive	32541066	link	NA
DNA	irinotecan	15	A375, A549, .	2	6, 24	6	0.04, 0.12, 0.	76	BRD-K08547	topoisomeras	positive	27852227	link	NA
DNA	melphalan	1	HA1E	2	6, 24	1	10	3	NA	NA	positive	32717133	link	NA
DNA	mitomycin-c	15	A375, A549, .	2	6, 24	6	0.04, 0.12, 0.	76	BRD-A48237	DNA alkylatin	positive	32087850	link	NA
DNA	mitoxantrone	32	A375, A549, .	4	3, 6, 24, 48	13	0.001, 0.01, (	439	BRD-K21680	topoisomeras	positive	27852227	link	NA
DNA	nocodazole	15	A375, A549, .	2	6, 24	4	0.5, 1, 3, 10	40	BRD-K12539	tubulin inhibit	positive	30311985	link	NA
DNA	ochratoxin-a	14	A375, A549, .	2	6, 24	1	10	18	BRD-K39944	phenylalanyl t	positive	32454083	link	NA
DNA	olaparib	69	A375, A549, .	2	6, 24	14	0.04, 0.1, 0.1	463	BRD-K02113	PARP inhibito	positive	32215876	link	NA
DNA	paclitaxel	18	A375, A549, .	2	6, 24	12	0.001, 0.01, (	147	BRD-A28746	tubulin inhibit	positive	32651820	link	NA
DNA	temozolomide	56	A375, A549, .	2	6, 24	12	0.001, 0.01, (	176	BRD-K32107	DNA alkylatin	positive	32708903	link	NA
DNA	topotecan	13	A375, A549, .	2	6, 24	1	10	31	BRD-A59985	topoisomeras	positive	32200233	link	some potentia
DNA	vorinostat	71	A375, A549, .	3	6, 24, 48	15	0.001, 0.01, (	1533	BRD-K81418	HDAC inhibit	positive	32512035	link	NA
ERS	brefeldin-a	56	A375, A549, .	2	6, 24	2	0.1, 10	84	BRD-A17065	antibiotic that	positive	32815086	link	NA
ERS	chloroquine	58	A375, A549, .	2	6, 24	12	0.001, 0.01, (	169	BRD-A91699	antimalarial a	positive	30451364	link	NA
ERS	dithiothreitol	3	FIBRNPC, NE	2	6, 24	1	10	6	NA	NA	positive	32850859	link	NA
ERS	phenylbutyrat	13	A375, A549,	2	6, 24	6	0.04, 0.12, 0.	59	BRD-K67102	HDAC inhibit	negative	32793713	link	ANTAGONIST
ERS	salubrinal	8	A375, A549,	2	6, 24	1	10	15	BRD-A77299	eukaryotic tra	negative	17198682	link	ANTAGONIST
ERS	spiperone	16	A375, A549, .	2	6, 24	2	3, 10	44	BRD-K55468	dopamine rec	positive	25925857	link	NA
ERS	thapsigargin	53	A375, A549, .	2	6, 24	3	0.5, 1, 10	84	BRD-A62809	ATPase inhib	positive	32846985	link	NA
ERS	tunicamycin	15	A375, A549, .	3	6, 24, 144	2	10	24	BRD-K10573	GLCNAC phc	positive	32815086	link	NA
HTS	alvespimycin	16	A375, A549, .	3	6, 24, 48	11	0.04, 0.1, 0.1	84	BRD-K83988	HSP inhibitor,	positive	24471734	link	NA
HTS	bortezomib	53	A375, A549, .	2	6, 24	15	0.001, 0.01, (	2459	NA	NA	positive	32026316	link	NA
HTS	gedunin	8	A375, HA1E,	2	6, 24	1	10	12	BRD-A17819	HSP inhibitor	positive	24969439	link	some potentia
HTS	geldanamycir	69	A375, A549, .	4	3, 6, 24, 48	15	0.001, 0.01, (	1200	BRD-A19500	HSP inhibitor	positive	32497199	link	NA
HTS	hypericin	15	A375, A549, .	2	6, 24	6	0.04, 0.12, 0.	65	BRD-K29673	PKC inhibitor,	positive	21949677	link	double check
HTS	parthenolide	58	A375, A549, .	2	6, 24	5	0.5, 1, 3, 10,	174	BRD-K98548	NFkB pathwa	positive	30023543	link	NA
HTS	puromycin	14	A375, A549, .	2	6, 24	4	0.5, 1, 3, 10	43	BRD-A28970	adenosine re	positive	24178293	link	NA
HTS	radicicol	21	A375, A549, .	4	3, 6, 24, 48	11	0.04, 0.1, 0.1	315	BRD-K33551	HSP inhibitor,	positive	31782811	link	NA
HTS	tanespimycin	65	A375, A549, .	3	6, 24, 48	9	0.001, 0.01, (	257	NA	NA	positive	32585150	link	NA

Context is important

- Most common mistake is protective treatment
- Here phentolamine acts to restore basal conditions
- Metals associate with chelators

The goldfish (Carassius auratus) exhibits a remarkable capacity to survive and remain active under prolonged and severe hypoxia, making it a good model for studying cardiac function when oxygen availability is a limiting factor. Under hypoxia, the goldfish heart increases its performance, representing a putative component of hypoxia tolerance; however, the underlying mechanisms have not yet been elucidated. Here, we aimed to investigate the role of β3-adrenoreceptors(ARs) in the mechanisms that modulate goldfish heart performance along with the impact of oxygen levels. By western blotting analysis, we found that the goldfish heart expresses β3-ARs, and this expression increases under hypoxia. The effects of β3-AR stimulation were analyzed by using an ex vivo working heart preparation. Under normoxia, the β3-AR-selective agonist BRL37344 (10-12 to 10-7 mol l-1)elicited a concentration-dependent increase of contractility that was abolished by a specific β3-AR antagonist (SR59230A; 10-8 mol l-1), but not by α/β1/β2-ARinhibitors (phentolamine, nadolol and ICI118,551; 10-7 mol l-1). Under acute hypoxia, BRL37344 did not affect goldfish heart performance. However, SR59230A,but not phentolamine, nadolol or ICI118,551, abolished the time-dependent enhancement of contractility that characterizes the hypoxic goldfish heart. Under both normoxia and hypoxia, adenylate cyclase and cAMP were found to be involved in the β3-AR-dependent downstream transduction pathway. In summary, we show the presence of functional β3-ARs in the goldfish heart, whose activation modulates basal performance and contributes to a hypoxia-dependent increase of contractility.

Summary of Expert Curated

Chemical passing review = 68/93 (73%)

Transcriptomic profiles = 11,185

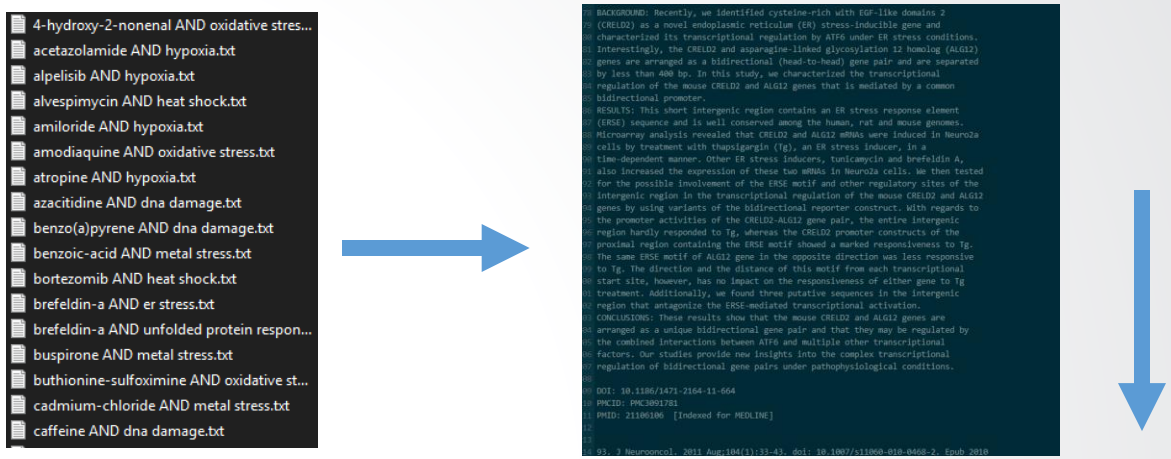
Weakest (profiles; ranked): MTL, HPX, UPR

SRP	Chemicals	Profiles
DDR	24	4162
HPX	7	158
HSR	11	5135
MTL	1	9
OSR	16	1405
UPR	8	316

Expert	Chemicals	PMI (unfilt)	PMI (filt)
DDR	24	0.96	1
HPX	7	1	1
HSR	11	0.64	0.81
MTL	1	0	1
OSR	16	0.94	0.75
UPR	8	0.5	0.88

Using abstract referenced genes to clarify SRP association

- Significant amount of information captured in abstracts
  - Chemical
  - Gene
  - Cell process
  - Pathway
- Gene abundance reflects important regulators and targets of chemical activity identified in literature
- Aggregate all abstracts for chemical ~ stress search (200-6000 abstracts)
- mine human/mouse genes
- assign quantitative score to each



Since Inhibitor of Apoptosis (IAP) proteins have been implicated in cellular adaptation to endoplasmic reticulum (ER) stress, we investigated the regulation of ER stress-induced apoptosis by small-molecule second mitochondria-derived activator of caspase (Smac) mimetics that antagonize IAP proteins. Here, we discover that Smac mimetic suppresses tunicamycin (TM)-induced apoptosis via resolution of the unfolded protein response (UPR) and ER stress. Smac mimetics such as BV6 selectively inhibit apoptosis triggered by pharmacological or genetic inhibition of protein N-glycosylation using TM or knockdown of DPAGT1, the enzyme that catalyzes the first step of protein N-glycosylation. In contrast, BV6 does not rescue cell death induced by other typical ER stressors (i.e., thapsigargin (TG), dithiothreitol, brefeldin A, bortezomib, or 2-deoxyglucose). The protection from TM-triggered apoptosis is found for structurally different Smac mimetics and for genetic knockdown of cellular IAP (cIAP) proteins in several cancer types, underlining the broader relevance. Interestingly, lectin microarray profiling reveals that BV6 counteracts TM-imposed inhibition of protein glycosylation. BV6 consistently abolishes TM-stimulated accumulation of ER stress markers such as glucose-regulated protein 78 (GRP78) and C/EBP homologous protein (CHOP) and reduces protein kinase RNA-like ER kinase (PERK) phosphorylation and X box-binding protein 1 (XBP1) splicing upon TM treatment. BV6-stimulated activation of nuclear factor-κB (NF-κB) contributes to the resolution of ER stress, since NF-κB inhibition by overexpression of dominant-negative IκBα superrepressor counteracts the suppression of TM-stimulated transcriptional activation of CHOP and GRP78 by BV6. Thus, our study is the first to show that Smac mimetic protects from TM-triggered apoptosis by resolving the UPR and ER stress. This provides new insights into the regulation of cellular stress responses by Smac mimetics.

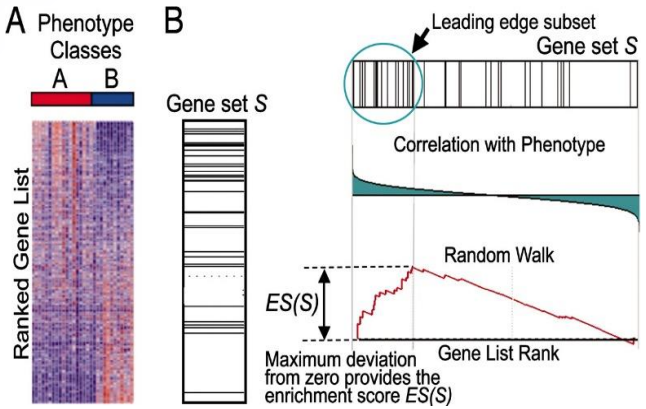
Score abstract gene enrichment

Reduce resulting abstracts into genes listed in associated literature

Use existing enrichment analysis tools to score gene enrichment

- single sample gene set enrichment analysis
- Log10 counts used to rank genes
- Scored with published signature set
- TF only increases performance

GSEA workflow



Aravind Subramanian DOI: 10.1073/pnas.0506580102

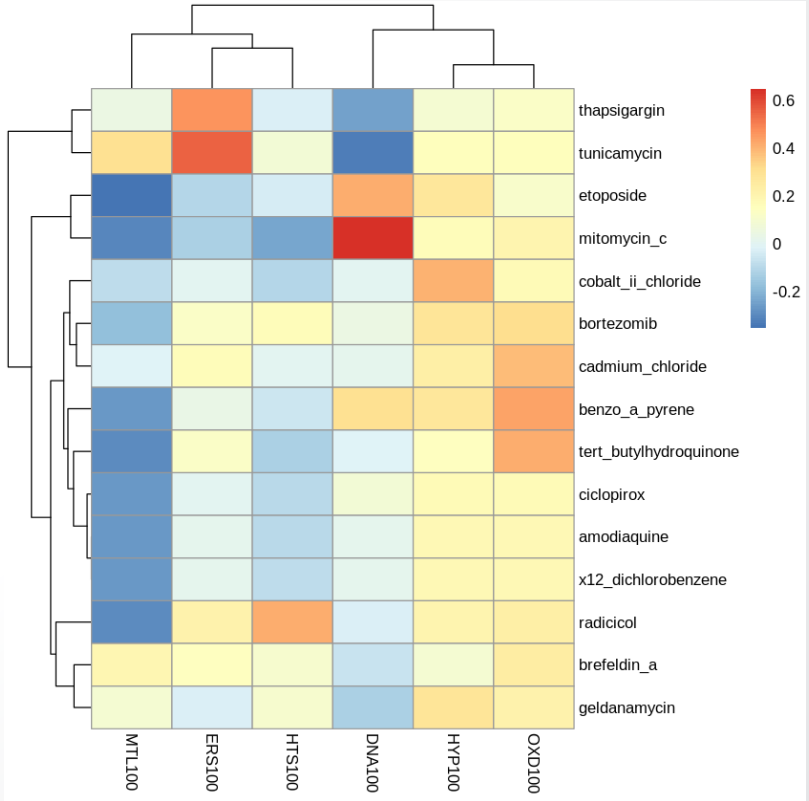
tert-butylhydroquinone		mitomycin-c		etoposide		brefeldin-a		geldanamycin		cobalt(II) chloride	
NQO1	3.00	FANCD2	5.46	ATM	5.81	ATF6	4.28	EGFR	4.91	RPE	2.89
KEAP1	2.89	BRCA1	5.14	BRCA1	4.96	XBP1	4.22	HSF1	4.80	AQP1	2.30
SLC4A11	2.89	BRCA2	5.12	E2F1	4.90	APP	4.20	AR	4.06	ARNT	2.30
GC	2.48	HR	5.02	H2AX	4.87	GBF1	3.91	RPE	4.03	BNIP3	1.95
GSTA2	2.48	RAD51	4.73	HR	4.56	CRELD2	3.87	MDM2	3.91	ENO1	1.95
SOD2	2.30	PALB2	4.11	SIRT1	4.50	PRNP	3.71	IL6	3.89	MMP2	1.95
SQSTM1	2.30	FANCA	4.01	PARP1	4.33	ATF4	3.66	TPR	3.71	KLF13	1.79
ATF4	2.20	BLM	3.95	ATR	4.29	TG	3.58	CYP2E1	3.66	VEGFC	1.79
GCLC	2.20	FANCG	3.91	PTEN	4.28	MANF	3.47	STAT3	3.66	PDCD4	1.39
CFTR	2.08	ATM	3.87	RAD51	4.26	ADM2	3.40	GDA	3.64	HK2	1.10
CRBN	2.08	ATR	3.87	WRN	4.14	CFTR	3.37	HR	3.61	IL6	1.10
CS	2.08	NQO1	3.87	BCL2	4.09	FGF1	2.89	ARNT	3.53	CA9	0.69
GCLM	2.08	WRN	3.81	BRCA2	4.03	HSPA6	2.89	NQO1	3.53	EPO	0.69
PTEN	2.08	ERCC1	3.78	HIC1	3.89	OSR1	2.89	PDK1	3.43	HMOX1	0.69
SFN	2.08	PCNA	3.69	MDM2	3.83	CD4	2.83	ERBB2	3.14	IL18	0.69
NLRP3	1.79	CP	3.58	BLM	3.81	RPE	2.83	MMP9	3.14	MMP9	0.69
NRF1	1.79	FANCI	3.43	PDCD5	3.76	BOK	2.71	HGF	3.09	ULK1	0.69
VCAM1	1.79	XRCC2	3.40	TP53	3.69	CENPF	2.71	GC	3.04		
				EGFR	3.61			HSPA6	3.04		

Example gene ranks from chemical search profiles

Example HTS signature:

'BAG1' · 'CCT2' · 'CD74' · 'CHCHD4' · 'DNAJB1' · 'DNAJB12' · 'DNAJB13' · 'DNAJB14' · 'DNAJB4' · 'DNAJB5' · 'DNAJC18' · 'DNAJC2' · 'DNAJC7' · 'ENTPD5' · 'ERO1A' · 'FKBP1A' · 'FKBP1B' · 'GAK' · 'HSPA13' · 'HSPA14' · 'HSPA1A' · 'HSPA1B' · 'HSPA1L' · 'HSPA2' · 'HSPA5' · 'HSPA6' · 'HSPA7' · 'HSPA8' · 'HSPA9' · 'HSPD1' · 'HSPE1' · 'HSPH1' · 'PTGES3' · 'SELENOF' · 'ST13' · 'ST13P4' · 'ST13P5' · 'TOR1A' · 'TOR1B' · 'TOR2A' · 'UGGT1'

Extracted abstract genes GSEA scores



Abstract GSEA can add more resolution

Example DDR abstract gene scoring

Benzo(a)pyrene is a lower scoring DDR but has significant OSR activity (matches review)

Chemical	DDR	UPR	HSR	HPX	MTL	OSR	SRP
hydroxyurea	2.58	-0.87	0.99	0.24	-0.62	-0.40	DDR
mitomycin-c	2.32	-0.82	-0.46	0.14	-0.58	-0.46	DDR
olaparib	2.24	-0.55	0.16	-0.11	-1.08	-1.24	DDR
camptothecin	2.24	0.46	-0.12	0.54	0.90	-0.43	DDR
etoposide	1.74	0.53	0.69	1.53	0.67	0.24	DDR
doxorubicin	1.71	0.78	0.94	2.66	1.54	2.14	DDR
gemcitabine	1.61	1.51	1.50	-0.78	-1.13	0.33	DDR
paclitaxel	1.47	0.95	1.63	1.17	0.33	0.52	DDR
irinotecan	1.43	-0.46	-1.02	-0.94	0.26	-0.63	DDR
melphalan	1.27	-0.11	-0.11	-1.26	-0.98	-1.25	DDR
temozolomide	1.25	0.49	0.07	2.17	0.02	0.01	DDR
benzo(a)pyrene	1.03	0.18	1.70	0.44	-0.58	0.58	DDR

Example UPR abstract gene scoring

Brefeldin A mentions less core UPR genes but indicates more HSR genes

Chemical	DDR	UPR	HSR	HPX	MTL	OSR	SRP
tunicamycin	-0.62	4.13	2.31	1.29	2.76	0.63	UPR
thapsigargin	-0.73	3.40	0.75	0.35	2.37	0.18	UPR
phenylbutyrate	-0.80	1.99	1.30	-0.45	1.25	0.82	UPR
gemcitabine	1.61	1.51	1.50	-0.78	-1.13	0.33	DDR
4-hydroxy-2-nonenal	-0.28	1.29	-1.22	0.27	1.58	2.06	OSR
dithiothreitol	-0.53	1.05	-0.44	0.08	0.64	0.15	UPR
paclitaxel	1.47	0.95	1.63	1.17	0.33	0.52	DDR
radicicol	-0.70	0.82	1.83	-0.76	-0.96	-1.00	HSR
chloroquine	-0.46	0.79	0.82	1.39	1.15	1.58	UPR
doxorubicin	1.71	0.78	0.94	2.66	1.54	2.14	DDR
salubrinal	-0.80	0.76	-0.28	-0.37	-0.21	0.84	UPR
buthionine-sulfoximine	-0.02	0.57	2.61	0.37	0.60	1.66	OSR
bortezomib	0.08	0.54	1.73	1.41	0.40	1.02	HSR
etoposide	1.74	0.53	0.69	1.53	0.67	0.24	DDR
cadmium-chloride	-0.31	0.51	-0.19	0.56	0.80	0.75	MTL
temozolomide	1.25	0.49	0.07	2.17	0.02	0.01	DDR
brefeldin-a	-0.95	0.47	0.52	-1.26	2.17	-0.30	UPR

Abs GSEA vs Transcriptomic approaches

Strength is in number of documents and breadth of studies

Abstracts tend to focus on a smaller number key regulatory genes

Signatures can have much finer grained resolution

Abs Abstract scoring can be improved

- Small TF based signatures?

Search and Text Mining Approaches

Expert	Chemicals	PMI (unfilt)	PMI (filt)	Abs GSEA
DDR	24	0.96	1	0.67
HPX	7	1	1	0.57
HSR	11	0.64	0.81	0.81
MTL	1	0	1	1
OSR	16	0.94	0.75	0.5
UPR	8	0.5	0.88	0.88

Extending assessment with TRx based approaches

- Ensemble scoring approaches:

Similarity Scoring

Selected top 100 up and down genes from median of all profiles of small high confidence case

Gene	Rank
Gene A	1
Gene B	2
Gene C	3
Gene D	4
Gene E	5
Gene F	6
Gene G	7
Gene …	…
Gene n	n

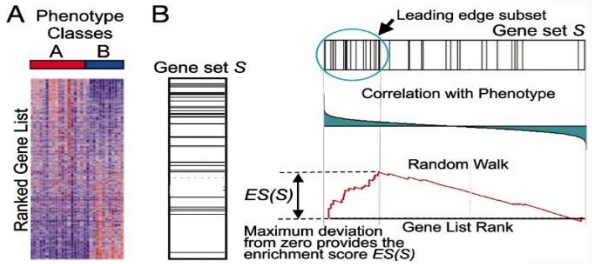
Test Set

$$\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

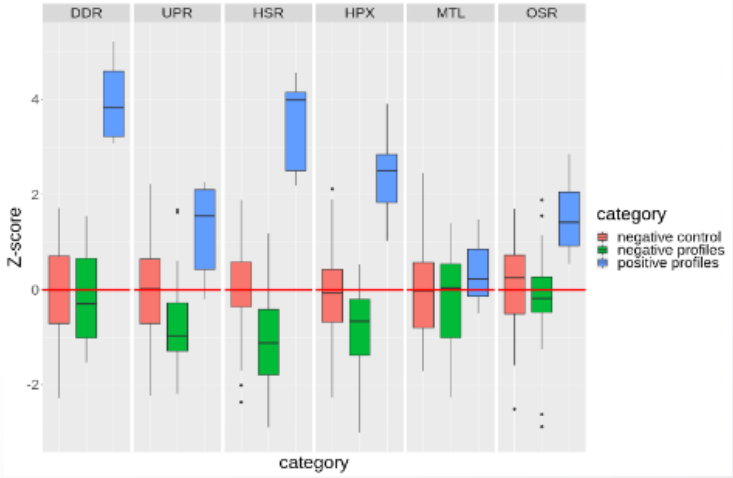
Examine all top up and down for each prospective chemical

Gene set enrichment analysis

1. Rank order genes by expression
2. KS random walk-through GS
3. Count genes in GS with up score
4. Count genes in GS with down score
5. Calculate total



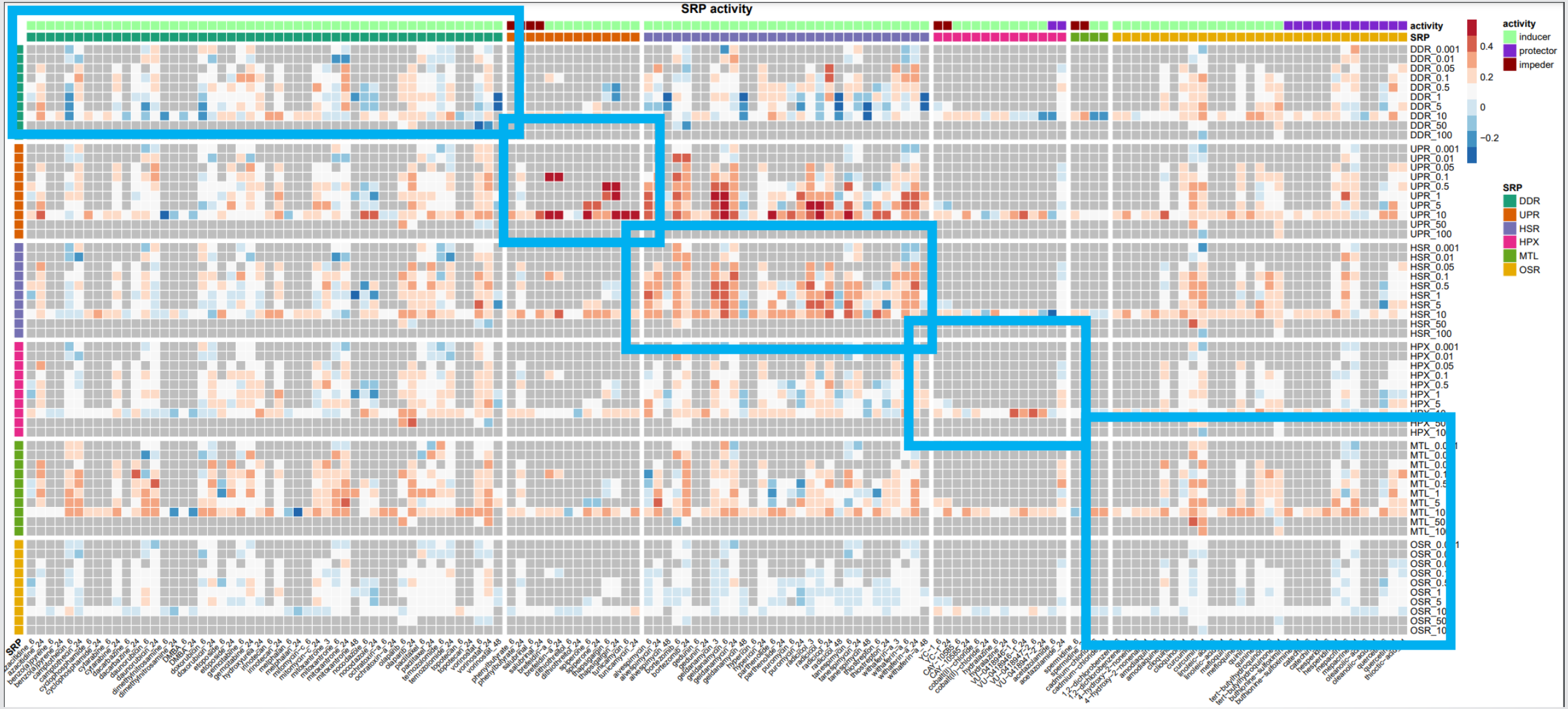
Aravind Subramanian
DOI: 10.1073/pnas.0506580102



GSEA predicts assigned chemical activity

Chemical and Time →

Signature and Dose →



Transcriptomic Scoring Approaches

		Search and Text Mining Approaches			Transcriptomic Approaches	
Expert	Chemicals	PMI (unfilt)	PMI (filt)	Abs GSEA	Similarity	TRx GSEA
DDR	24	0.96	1	0.67	0.50	0.21
HPX	7	1	1	0.57	0.57	0.50
HSR	11	0.64	0.81	0.81	0.81	0.81
MTL	1	0	1	1	1	1
OSR	16	0.94	0.75	0.5	1	0.13
UPR	8	0.5	0.88	0.88	0.88	0.63

Full method results in a rankable output

CHEMICAL	SPID	SUPER_CAT	SPID_MEDIAN_AC50	SPID_SD_AC50	SPID_MEDIAN_TPCOF	SPID_SD_TPCOF	CHEM_MEDIAN_AC50	CHEM_SD_AC50	CHEM_MEDIAN_TPCOF	CHEM_SD_TPCOF	SCALED_CHEM_MEDIAN_AC50	CHEM_SUPERCAT_MEDIAN_AC50_PCNTILE	CHEM_SUPERCAT_MEDIAN_TPCOF_PCNTILE	PMI_CAT	PMI	PMI_PCNTILE	MAX_PMI	MAX_PMI_CAT	SEARCH_HITS	TOTAL_SEARCH_HITS
Butachlor	TP0000077A01	DDR	18.49	NA	1.18	NA	18.49	NA	1.18	NA	0.06	26.45	2.81	DDR	1.54	16.05	1.54	dna damage	6	14
Chloroacetaldehyde	TP0000719G11	DDR	3.29	NA	1.22	NA	3.29	NA	1.22	NA	0.01	7.5	4.69	DDR	2.23	3.91	2.23	dna damage	45	65
Azathioprine	TP0000422E07	DDR	90.00	NA	1.23	NA	90.00	NA	1.23	NA	0.28	80.68	5.63	DDR	0.95	32.68	0.95	dna damage	37	130
Aminopterin	TP0000456H05	DDR	46.40	NA	1.25	NA	46.40	NA	1.25	NA	0.15	49.34	7.32	DDR	2.31	3.13	2.31	dna damage	11	15
Chlorothalonil	TP0000077G06	DDR	1.28	NA	1.27	NA	1.28	NA	1.27	NA	0.00	4.69	7.69	DDR	1.50	17.42	1.50	dna damage	15	36
Quinoline	TX002359	DDR	43.24	NA	1.29	NA	43.24	NA	1.29	NA	0.14	46.9	8.44	DDR	1.68	12.92	1.68	dna damage	185	391
Pirimiphos-methyl	TX002825	DDR	30.41	NA	1.32	NA	30.41	NA	1.32	NA	0.10	38.09	10.69	DDR	1.44	18.4	1.44	dna damage	2	5
Thidiazuron	TX000704	DDR	51.67	NA	1.33	NA	51.67	NA	1.33	NA	0.16	52.16	10.88	DDR	1.44	18.4	1.44	dna damage	2	5
Captafol	TP0000077F04	DDR	6.31	NA	1.35	NA	6.31	NA	1.35	NA	0.02	12.01	11.82	DDR	1.76	9.78	1.76	dna damage	3	6
2,4,6-Tribromophenol	TP0000722F09	DDR	50.85	NA	1.37	NA	50.85	NA	1.37	NA	0.16	51.41	13.32	DDR	1.54	16.05	1.54	dna damage	3	7
Hydroquinone	TP0000131H07	DDR	28.00	NA	1.05	NA	28.00	40.43543	1.39	0.651327	0.09	35.65	14.26	DDR	1.33	23.09	1.33	dna damage	187	503
Hydroquinone	TP0000371F02	DDR	53.93	53.12	1.85	0.6521	28.00	40.43543	1.39	0.651327	0.09	35.65	14.26	DDR	1.33	23.09	1.33	dna damage	187	503
Benzyl butyl phthalate	TP0000458H06	DDR	99.21	NA	1.41	NA	99.21	NA	1.41	NA	0.31	85.37	16.32	DDR	1.50	17.22	1.50	dna damage	5	12
Chlorambucil	TP0001079B03	DDR	55.21	NA	1.41	NA	55.21	NA	1.41	NA	0.17	54.03	16.7	DDR	2.30	3.33	2.30	dna damage	97	134
9-Nitroanthracene	TP0001323A04	DDR	14.54	NA	1.42	NA	14.54	NA	1.42	NA	0.05	21.58	17.07	DDR	1.44	18.4	1.44	dna damage	4	10
all-trans-Retinoic acid	TX008810	DDR	24.19	NA	1.44	NA	24.19	NA	1.44	NA	0.08	32.65	18.76	DDR	0.71	39.53	0.71	dna damage	69	286
Carbendazim	TP0000454D03	DDR	3.32	0.26	1.44	0.20344	3.32	0.256725	1.44	0.203436	0.01	7.69	18.95	DDR	1.62	15.07	1.62	dna damage	25	55
Digoxigenin	TP0001180G11	DDR	3.60	NA	1.47	NA	3.60	NA	1.47	NA	0.01	8.82	19.7	DDR	1.02	30.33	1.02	dna damage	47	157
4-Aminoazobenzene	TP0000385E05	DDR	63.25	NA	1.50	NA	63.25	NA	1.50	NA	0.20	61.16	20.64	DDR	2.57	1.17	2.57	dna damage	7	8
2,4,5-Trichlorophenol	TP0000386E01	DDR	41.28	NA	1.51	NA	41.28	NA	1.51	NA	0.13	45.22	21.01	DDR	1.54	16.05	1.54	dna damage	3	7
Phenolphthalein	TP0000392C01	DDR	38.82	19.91	1.51	0.55954	38.82	19.9146	1.51	0.559538	0.12	44.28	21.2	DDR	1.91	7.83	1.91	dna damage	5	9
6-Thioguanine	TP0000392E03	DDR	24.11	NA	1.62	NA	14.69	13.32761	1.52	0.150176	0.05	21.95	21.95	DDR	2.57	0.98	2.57	dna damage	151	172
6-Thioguanine	TX006867	DDR	5.26	NA	1.41	NA	14.69	13.32761	1.52	0.150176	0.05	21.95	21.95	DDR	2.57	0.98	2.57	dna damage	151	172
p,p'-DDE	TP0000386B03	DDR	58.91	NA	1.56	NA	58.91	NA	1.56	NA	0.19	57.04	24.77	DDR	1.67	13.11	1.67	dna damage	16	34
Pyrimethamine	TP0000372C06	DDR	3.75	NA	1.33	NA	33.22	41.68395	1.57	0.330496	0.10	39.59	25.89	DDR	0.65	41.1	0.65	dna damage	9	39
Pyrimethamine	TX003331	DDR	62.70	NA	1.80	NA	33.22	41.68395	1.57	0.330496	0.10	39.59	25.89	DDR	0.65	41.1	0.65	dna damage	9	39

Automated TRx profile annotation similarity

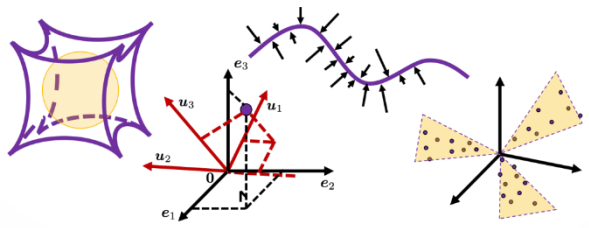
Qualitatively - how similar are the TRX from top ranked PMI and abstract GSEA

Evaluate similarity of profiles via dimensional reduction

- Take 1000s of features and reduce to a plot-able subset
- Unsupervised

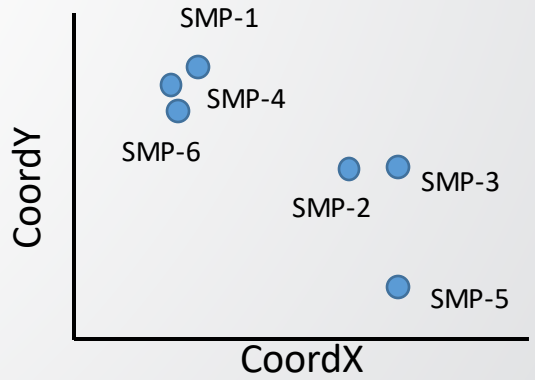
Sample	GeneA	GeneB	…	Gene ZZZZ
SMP-1	-0.024	1.32	…	0.013
SMP-2	4.231	0.002	…	-0.073
…	…	…	…	…
SMP-100	0.327	1.32	…	-2.823

local distance



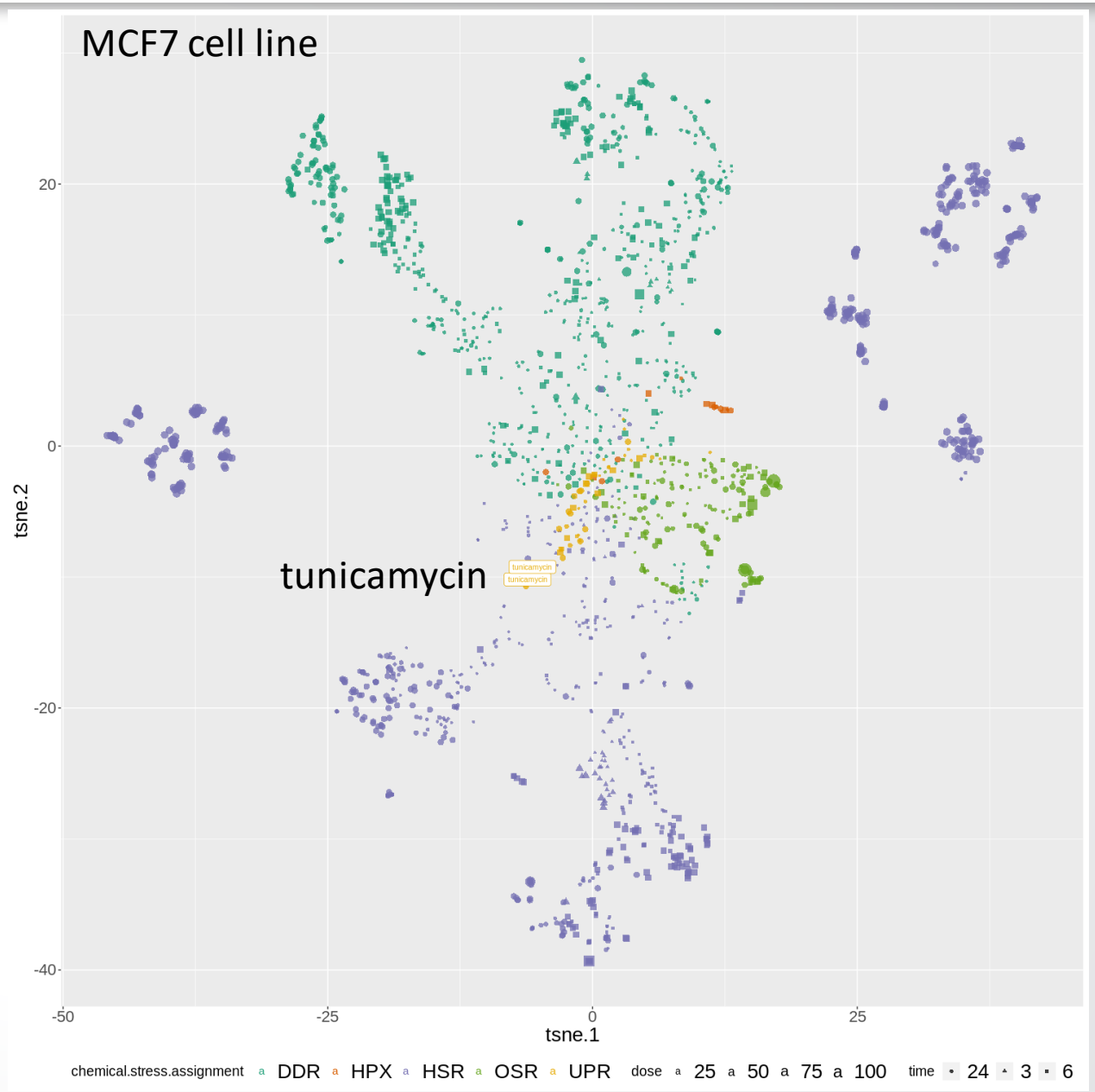
https://data-science-blog.com/blog/2020/10/02/illustrative-introductions-on-dimension-reduction/

Sample	CoordX	CoordY
SMP-1	-0.024	1.32
SMP-2	4.231	0.002
…	…	…
SMP-100	0.327	1.32



Clustering reflects SRPs

- Reference assignment guided annotation process reflects SRP bioactivity
- Unique groups are visible, clustering seems attributable to SRP assignment
- HSR; DDR show most obvious unique clusters
- UPR, OSR, and HPX overlap
- OSR and low conc SRPs group overlap
  - Either:
    - Most central and similar
    - Weakest response



Expanding the curated set

Method:

Expand Reference chemicals using meta-analysis score relative to other SRPs

1. Chose only those with search results (2580 chemicals)
2. Filter to those with > 5 references
3. PMI > 1

375 chemicals (of ~ 4761 named) with clearly associated stress activity

25014 TRx profiles

SRP	Perts	Profiles
DDR	74	6231
HPX	69	2425
HSR	39	4349
MTL	31	927
OSR	101	4516
UPR	61	6252

PMI based – completely automated

Clear clustering for many SRPs

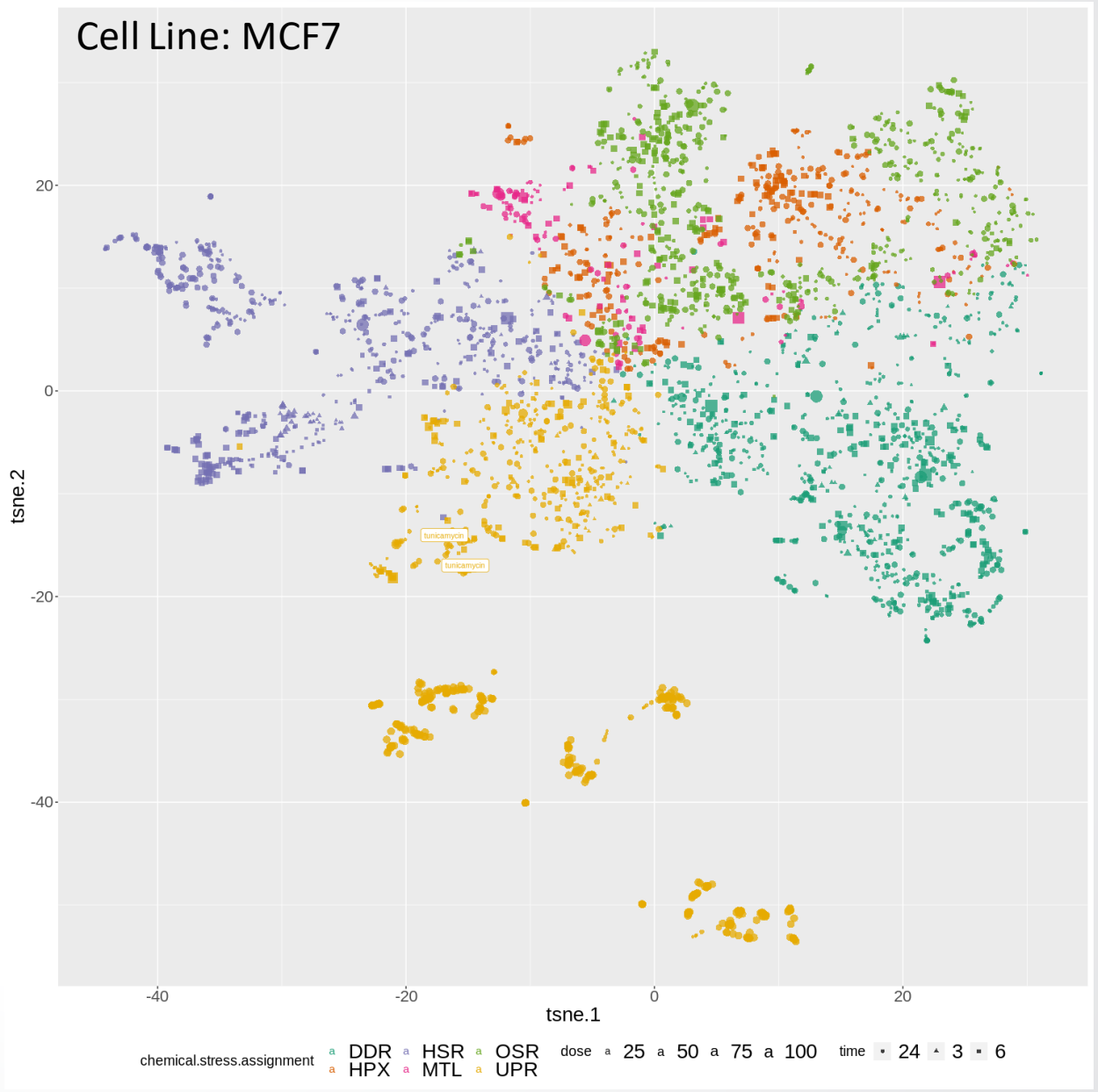
HPX might be incorporated into OSR region because of NLP issues

MSR situated near protein misfolding and OSR

HSR and UPR have overlapping region

Differentially expressed genes reflect SRP specific genes and could be used as signatures

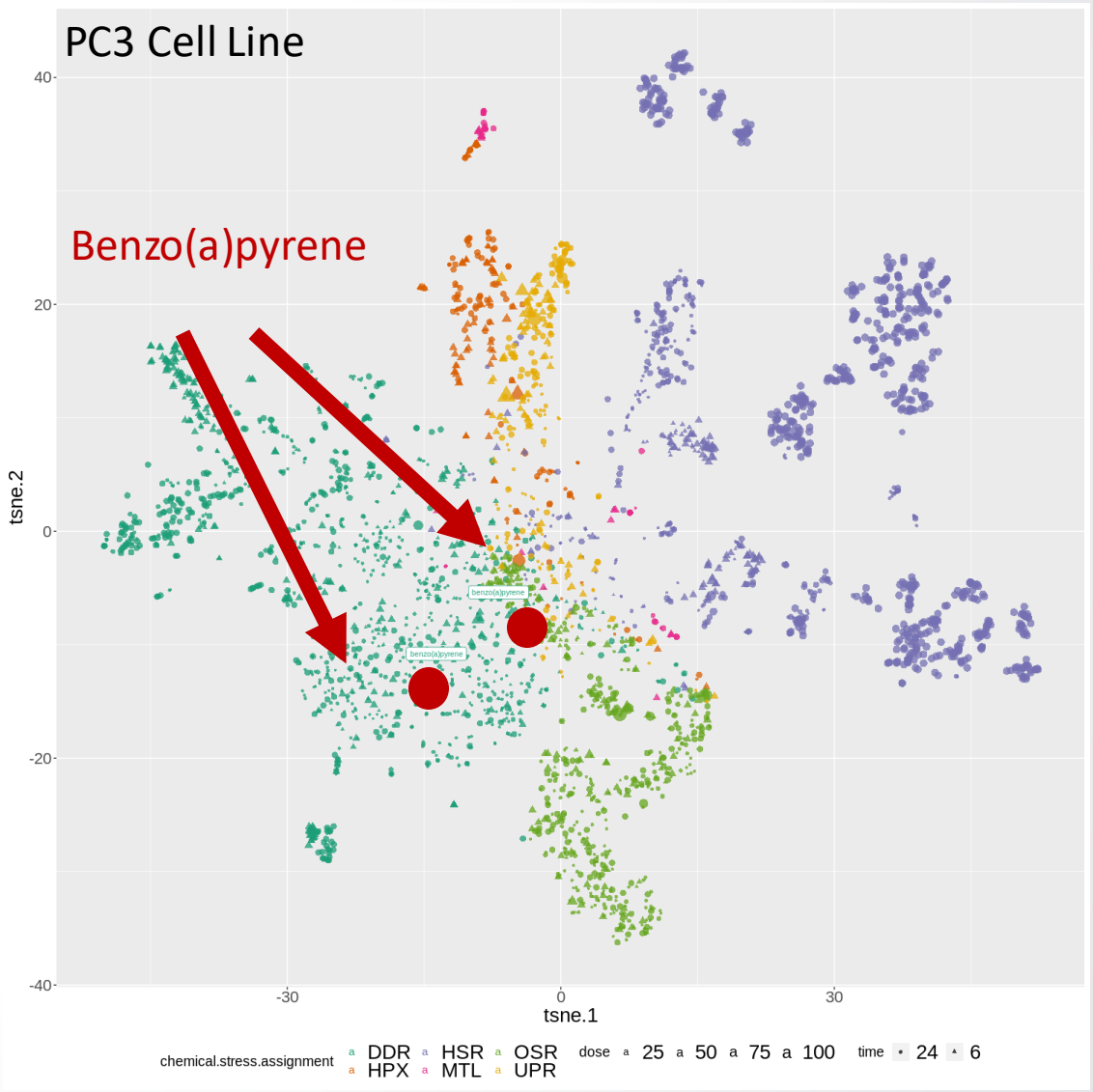
SRP	Profilen
DDR	821
HPX	385
HSR	660
MTL	139
OSR	701
UPR	803



Hazard / Pathway activity by dose

- Benzo(a)pyrene replicates cluster with DNA damage agents and with OSR inducers

- Overlap of low dose samples indicates a low signal area



Conclusions

- Literature mining is an effective tool for quickly finding and assigning SRP activity
  - Works well in a highly overlapping environment (nearly 80% recall)
  - Performs better or equal to transcriptomic approaches
- Abstract associated gene profiles contain much information
  - Best mined with a smaller signature using a similarity score?
  - Provide more information about PMI-Target/Pathway agreement
- Dimensional reduction of SRP associated chemicals highlight unique profile space
  - PMI automated assignment does yield extraction of similar transcriptomic signals
  - Cell type, time, and dose variance to be analyzed

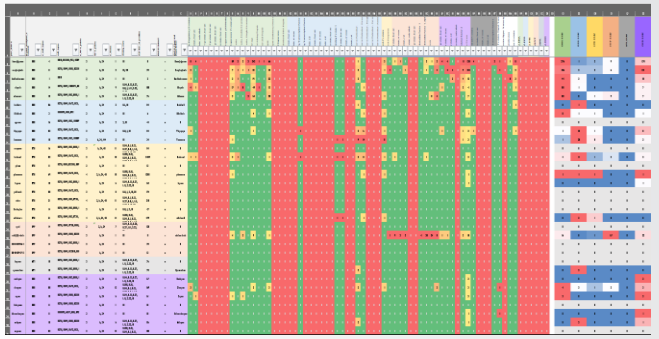
Applications to Systematic Review

- Simple informatics single metric approach scales easily to score/rank literature and chemical combination
  - 1000s of chemicals and 1000s of abstracts
  - Minutes runtime
  - Prioritize further evaluation
- Can further screen abstract relevancy
- Clusters with like chemicals
  - Chemical A's activity can be inferred from Chemical B's **literature/Reference Status**
- **Implementing abstract GSEA score individually can help to rank review priority**

Future Work

- Text mining
  - Use newer features of abstract sifter to pare chemical/disease/use context to a focused initial chemical list
  - Improve Abstract GSEA with small transcription factor centric signatures and switch to similarity scoring
  - Potentially use of gene ontology term enrichment for curated gene list as additional association

  - This scalable ranked scoring approach can interface with abstract sifter's MeSH and chemical



Abstract sifter, Credit: Nancy Baker

- Null and inactive
  - Mine targets and stress inactive compounds and compare clustering patterns

- Full characterization of cell, dose, and time dependency for stress response activity

- Combine with consensus signature methods to identify how to capture most important information in signatures

Acknowledgements

USEPA

Imran Shah
Nancy Baker
Laura Taylor
Joe Bundy
Johanna Nyffeler
Josh Harrill
Richard Judson
Logan Everett
Derik Haggard

UK

Danilo Basili
Andreas Bender (Germany)
Alistair Middleton