

Evaluating adaptive stress consensus gene signatures using transcriptomics

Bryant A. Chambers, Ph.D.

Office of Research and Development

Center for Computational Toxicology and Exposure

OpenTox2020

September 22nd, 2020



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.

Building a transcriptomic basis for adaptive stress response

- Thousands of untested chemicals with little or no health effects data
- New approach methods (NAMs) can inform chemical hazard/risk without animal testing *e.g.*:-
 - High throughput screening (HTS)
 - High-throughput transcriptomics (HTTr)
- Half of the 1063 chemicals examined in an early ToxCast study do not act via specific mechanisms
 - Did activate stress response systems
- Overwhelming adaptive stress response systems beyond “tipping point” can lead to adverse outcomes (Shah et al. 2016)



Stress response (SR) pathways provide a “systems” basis for categorizing perturbagen action

- SR pathways maintain cellular homeostasis
- SR pathway activation

Sensor(SENS) -> TF -> Transducer (TRD) ->
Response Element (RE) -> Gene expression (EFF)

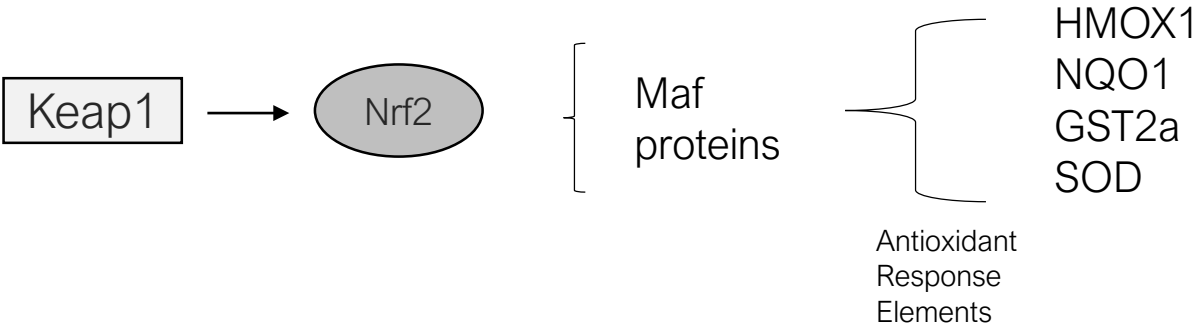
- Key questions:
 - How to develop signature sets?
 - Can we uniquely classify SRPs?
 - Can SR gene sets be used to evaluate tipping points and quantify cell stress?

Outcome: A NAM to classify non-specific chemicals using HTTr-SR activity

TABLE 1
The Major Adaptive Stress response pathways

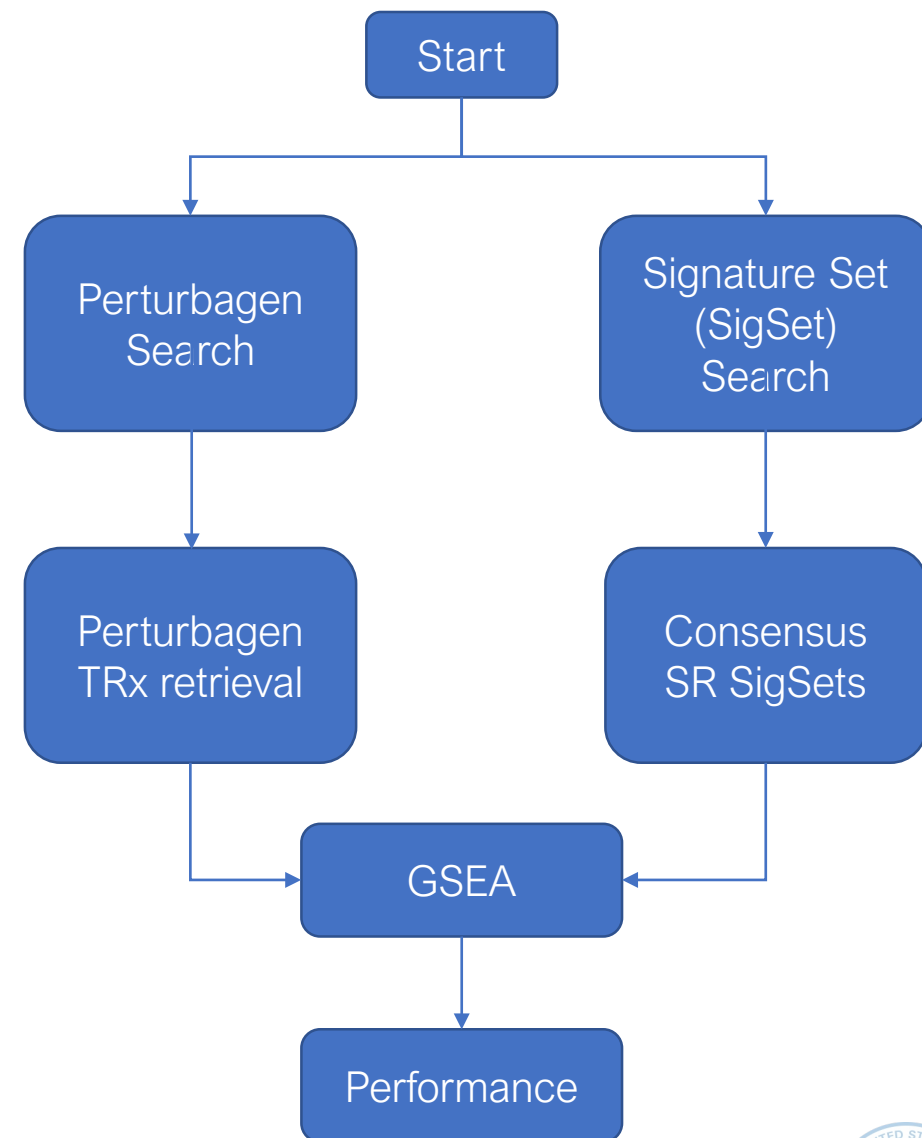
Stress response pathway	Chemical inducers	TF	Activated gene promoters
Oxidative stress	Quinones, hydroperoxides, heavy metals, trivalent arsenicals	Nrf2	<i>HMOX1, NQO1, GST2A</i>
Heat shock response	Heat, Heavy Metals	HSF-1	<i>HSPA6</i>
DNA damage response	Etoposide, Methyl Methanesulfonate, N-Dimethylnitrosamine,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

Simmons, S. O.; Fan, C.-Y.; Ramabhadran, R. Cellular Stress Response Pathway System as a Sentinel Ensemble in Toxicological Screening. *Toxicol. Sci.* 2009, 111 (2), 202–225. <https://doi.org/10.1093/toxsci/kfp140>.



Approach to evaluate SR pathway activity

1. Find reference chemicals
2. Reference chemical transcriptomics dataset (TRx)
3. Construct consensus SR signature sets
4. Evaluate performance of signature sets for characterizing reference chemicals



Three search methods used to compile hallmark perturbagen profiles

Three methods to search

- Literature

Known Chemicals identified by searching against stress response system induction



- Comparative Toxicogenomic Database (CTD)

Confirmed by reference association



- Library of Integrated Network-Based Cellular Signatures (LINCS)

Confirmed stress system activity by querying TF/SENS/TRD/EFF gene lists against for identified chemicals within the



Transcriptomic profiles sourced from public datasets on GEO



DNA	ERS	HTS	HYP	MTL	OXD
MDM2	HSPA5, BiP, Grp78	HSF1	VHL	MTF1	NFE2L2
ATM	PERK, CREB2	HSF2	PHD*	MT1E	MAPK
JNK	IRE1alph	HSF3	PSMA7	MT2A	ERK
Chk1	ATF4	HSF4	HIF1alph	PRNP	p38
Chk2	ATF6	HSPA6	HIF1bet	Zip10	PKC
p53	XBP1	DNAJB1, DNAJ1, HDJ1, HSPF1	p300, CBP	MT-I: MT1A, MT1B, MT1E, MT1F, MT1G1, MT1G2, MT1H, MT1HL1, MT1M, MT1X,	Keap1
GADD45a	DDIT3	HSP90AA1 HSP90A, HSPC1, HSPCA, HSP90AB1 HSP90B, HSPC2, HSPCB		MT-II: MT2A, CES1, MT2	HMOX1
BRCA1	IGFBP1	CK2	VEGF	AFP	MAF *
STRAP	IL8	CaMK2	EDN*	BiP	Roc1
DNA2	ATF3	HSPA4*	TFR*	CZBP	CUL*, CUL3
APE	HERPUD1, CHOP, GADD153	HSPA1A	ALDO*	Sepw1	SOD3, SOD*
RPA	HSP90	HSPA1B	GTR*	ZnT-1	CAT
LIG4	DNAJB9,	HSPA1L	VLP	yGCS	PRDX1 PAGA, PAGB, TDPX2
MRE1	ERdj4, ERdj5	HSPA2	VDU2	P1GF	GPX*
RAD51	p58	HSPA6	MT1	C/EBPbeta	TXN* TRDX, TRX, TRX1
PARP	PDI, MPD1, EUG1, MPD2, EPS1	HSPA7	HMOX*	PI3K	GSR* 1
EXO1	ERO1	HSPA8	ENO*	GSK-3	MT1, MT2
WRN	ERp57	HSP12A	HOGA*	HAMP	GST*
OGG1	Sec61	HSP12B		Fer2LCH	NQO1



Reference Perturbagen Dataset

Perturbagen	Profiles
<u>Hypoxia</u>	
0.1% Oxygen	2
VU-0418946-1 VU-0418946-2	2
<u>ERS/Unfolded Protein Response</u>	
Bredfeldin A	2
Tunicamycin	3
Thapsigargin	3
<u>Metal Stress</u>	
AgNO3	1
ZnO	3
<u>Heat Shock</u>	
Heat and Heat w/ recovery	1
Geldanimycin	3
Radcicol	2
<u>DNA Repair Stress</u>	
Lasicoarpine	1
Methylmethanesulfonate	1
Gylcidamide	1
Benzopyrene	1
<u>Oxidative Stress</u>	
hydrogen peroxide	3
tertbutylhydroperoxide	1

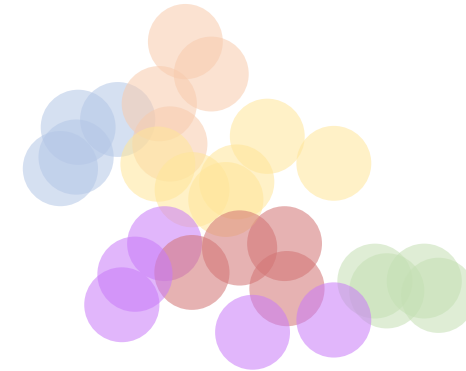


Process to find consensus SR Signature Sets

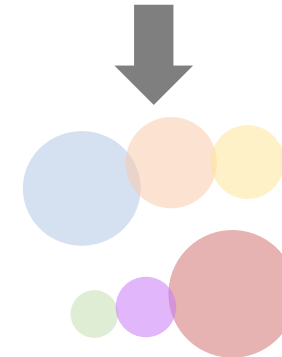
- SR systems are not formally defined
 - Incorporate many features
- Find existing signature sets
 - Use key SENS/TF/TRD/EFF genes from initial characterization to identify
 - Choose only those that are well associated with the SR in question
 - Significant cross-talk expected
- Prepare consensus signature sets
- Cut size to limit overlap while applying a filter to ensure only most central genes are included

Consensus SigSet approach

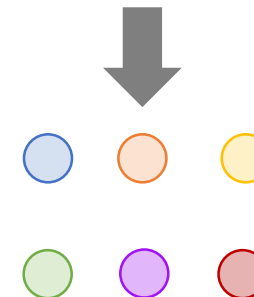
collect existing
SigSets



consolidate

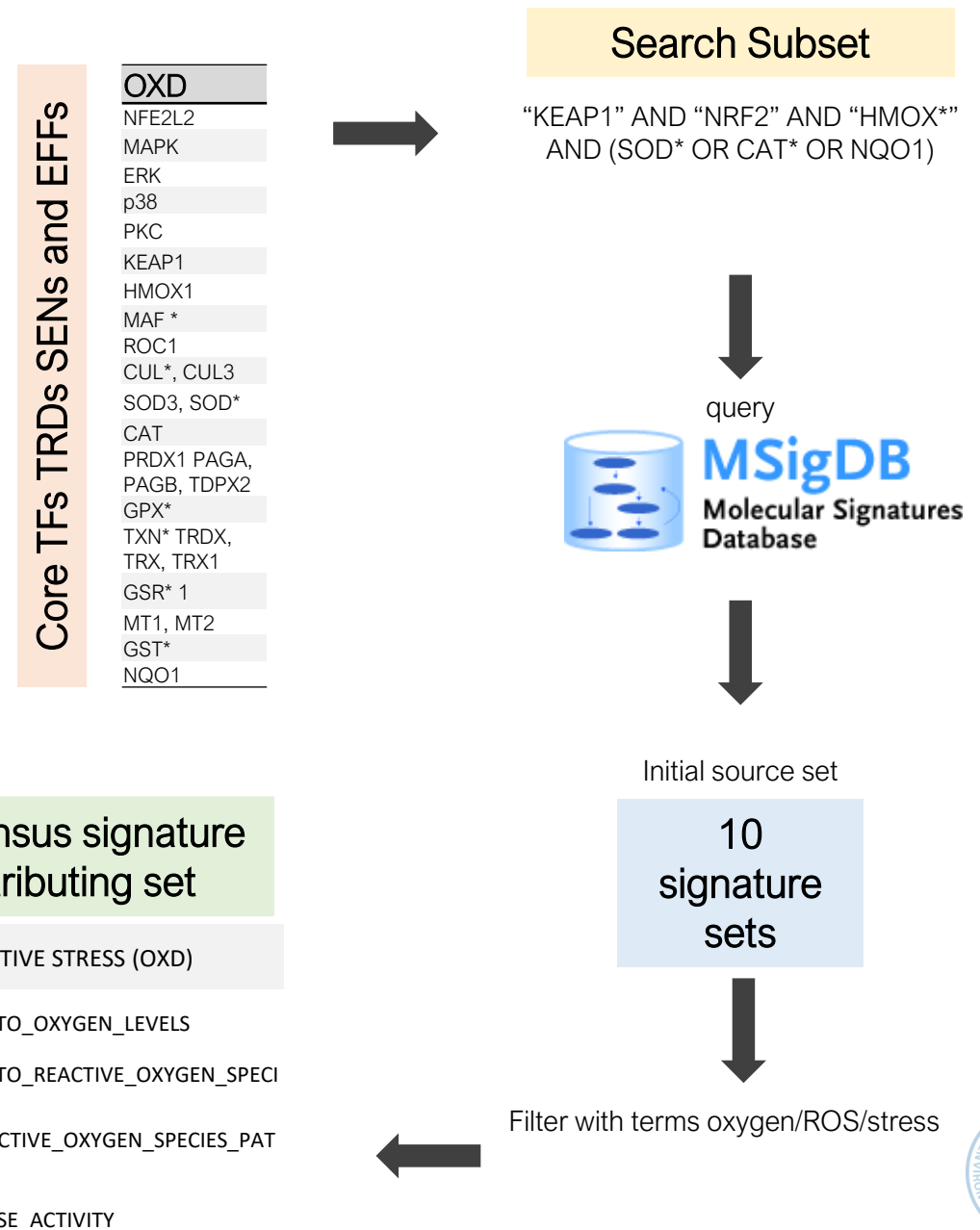


filter and
threshold



Identifying relevant signatures with SPR central elements

- Search for signature sets in MSigDB with combination of SR system TF, TRD, SENS and subset of key EFF
 - Similar to approach used to identify hallmark perturbagens
- Find EFF combination returning maximum and use keyword filter to return SigSets of interest
 - *i.e.*, “hypoxia”, “oxidative”, or “damage”
- Construct consensus signature sets



Consensus signature source sets vary in size and abundance

DNA DAMAGE (DNA)	ER STRESS (ERS)	HEAT SHOCK (HTS)	HYPOXIA (HYP)	METAL STRESS (MTL)	OXIDATIVE STRESS (OXD)
BIOCARTA_P53_PATHWAY	GO_CELLULAR_RESPONSE_TO_TOPOL OGICALLY_INCORRECT_PROTEIN	GO_CELLULAR_RESPONSE_TO_HEAT	GROSS_HYPOXIA_VIA_ELK3_AND_HIF1A _UP	GO_METAL_ION_TRANSMEMBRANE_TRA NSPORTER_ACTIVITY	GO_RESPONSE_TO_OXYGEN_LEVELS
GO_BASE_EXCISION_REPAIR	GO_CHAPERONE_COFACTOR_DEPENDE NT_PROTEIN_REFOLDING	GO_CHAPERONE_COFACTOR_DEPENDE NT_PROTEIN_REFOLDING	GROSS_HYPOXIA_VIA_HIF1A_DN	GO_RESPONSE_TO_CADMIUM_ION	GO_RESPONSE_TO_REACTIVE_OXYGEN _SPECIES
GO_CELLULAR_RESPONSE_TO_DNA_DA MAGE_STIMULUS	GO_CHAPERONE_MEDIATED_PROTEIN_ FOLDING	GO_CHAPERONE_MEDIATED_PROTEIN_ FOLDING	HARRIS_HYPOXIA	GO_RESPONSE_TO_METAL_ION	HALLMARK_REACTIVE_OXYGEN_SPECIE S_PATHWAY
GO_DNA_REPAIR	GO_DE_NOVO_PROTEIN_FOLDING	GO_PROTEIN_FOLDING	JIANG_HYPOXIA_NORMAL	GO_TRANSITION_METAL_ION_HOMEOST ASIS	OXIDOREDUCTASE_ACTIVITY
GO_DOUBLE_STRAND_BREAK_REPAIR	GO_PROTEIN_FOLDING			ON_METAL_ION_TRANSMET NSPORTER_ACTIVITY	
GO_G1_DNA_DAMAGE_CHECKPOINT	GO_RESPONSE_TO_TO NCORRECT_PROTEIN			RESPONSE_TO_METAL_ION	
GO_NUCLEASE_ACTIVITY	GO_UNFOLDED_PROTEIN				
GO_REGULATION_OF_DNA_DAMAGE_RE SPONSE_SIGNAL_TRANSDUCTION_BY_P 53_CLASS_MEDIATOR	HALLMARK_MTORC1_S				
GO_REGULATION_OF_RESPONSE_TO_D NA_DAMAGE_STIMULUS	HALLMARK_UNFOLDED ONSE				
GO_SIGNAL_TRANSDUCTION_IN_RESPO NSE_TO_DNA_DAMAGE					
HALLMARK_DNA_REPAIR					
KEGG_MISMATCH_REPAIR					
KEGG_NUCLEOTIDE_EXCISION_REPAIR					
KEGG_P53_SIGNALING_PATHWAY					
REACTOME_DNA_REPAIR					
REACTOME_G1_S_DNA_DAMAGE_CHEC KPOINTS					
REACTOME_TRANSCRIPTION_COUPLED _NER_TC_NER					

- Abundance of DNA is significantly greater than OXD in MSigDB
- Overlap in available sets for ERS and HTS



Frequency of occurrence determines consensus signature set

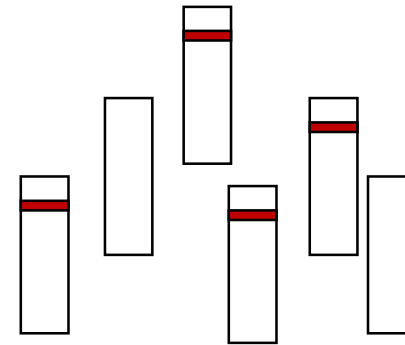
- Merge gene sets into consensus SR sig set by filtering genes based on...
- Occurrence Frequency factor (OFF)

$$\text{OFF} = \frac{f_{\text{in group}} - f_{\text{out group}}}{f_{\text{in group}} + f_{\text{out group}}}$$

Where f = frequency (count/#groups sampled)

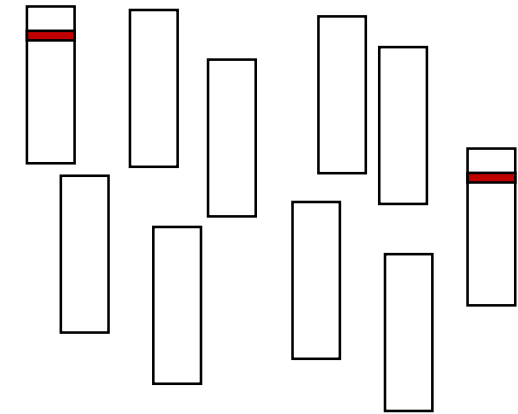
- All genes in set ranked by OFF and thresholded by SigSet sizes:
 - TH50 = 50 top genes
 - TH100 = 100 top genes,
 - TH200 = 200 top genes
 - ...
 - FULL = all genes

In group
(e.g., all HYP sig sets)

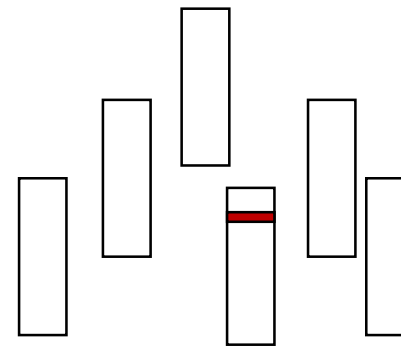


$$f_{\text{in group}} = 4/6 \quad f_{\text{out group}} = 2/10$$
$$\text{occurrence frequency factor} = (.667 - .2)/(0.867) = 0.54$$

Out group
(e.g., all non-HYP sig sets)

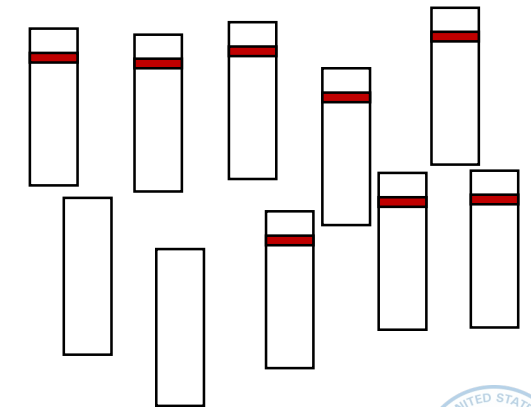


In group
(e.g., all HYP sig sets)



$$f_{\text{in group}} = 1/6 \quad f_{\text{out group}} = 8/10$$
$$\text{occurrence frequency factor} = (.167 - .8)/(0.967) = -0.66$$

Out group
(e.g., all non-HYP sig sets)

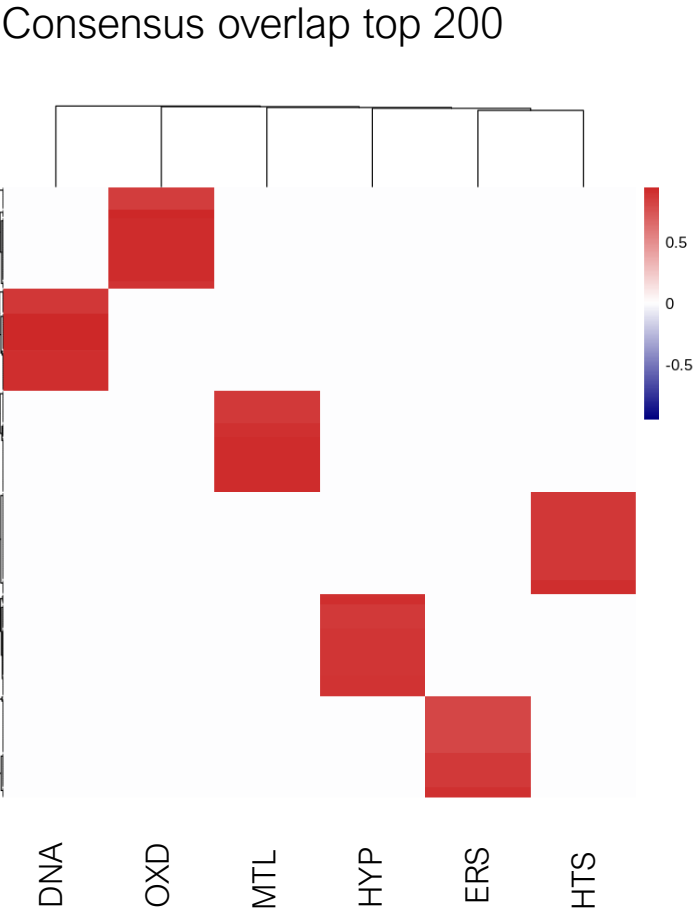
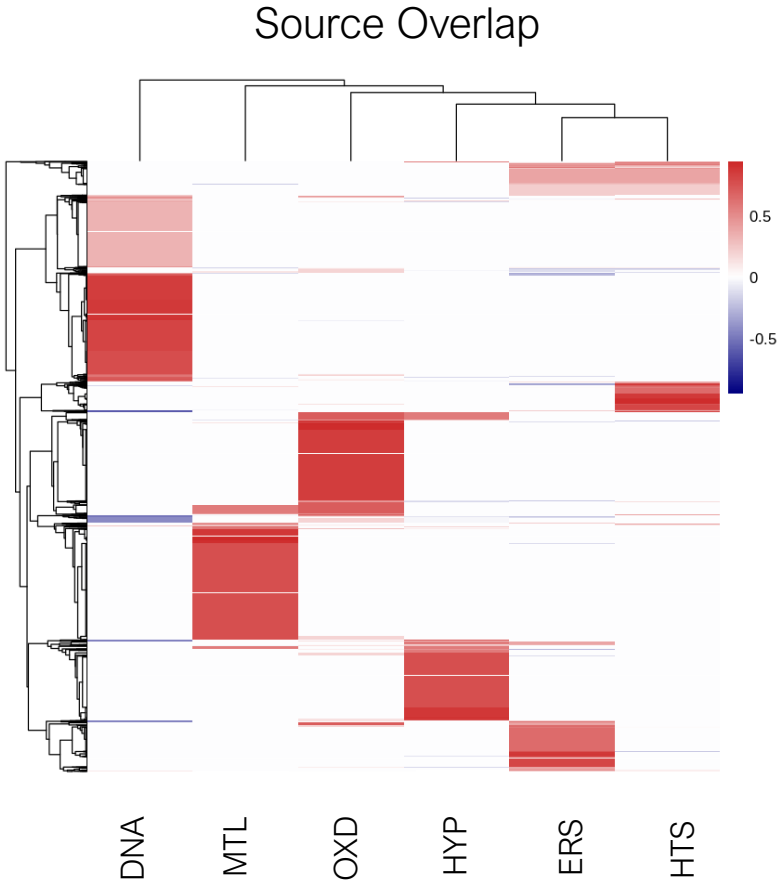


Consensus Signature construction reduces gene space overlap

All signature sets share approximately 50 or more genes

Similar signatures like ER and HS share many more

Thresholding reduces Signature set overlap by only 4 genes at a size of 200 total genes and overlap coefficients increase after this thresholding



Szymkiewicz-Simpson Overlap

Coefficient overlap

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

~ 100 x less overlap by thresholding ranked OFF signatures sets

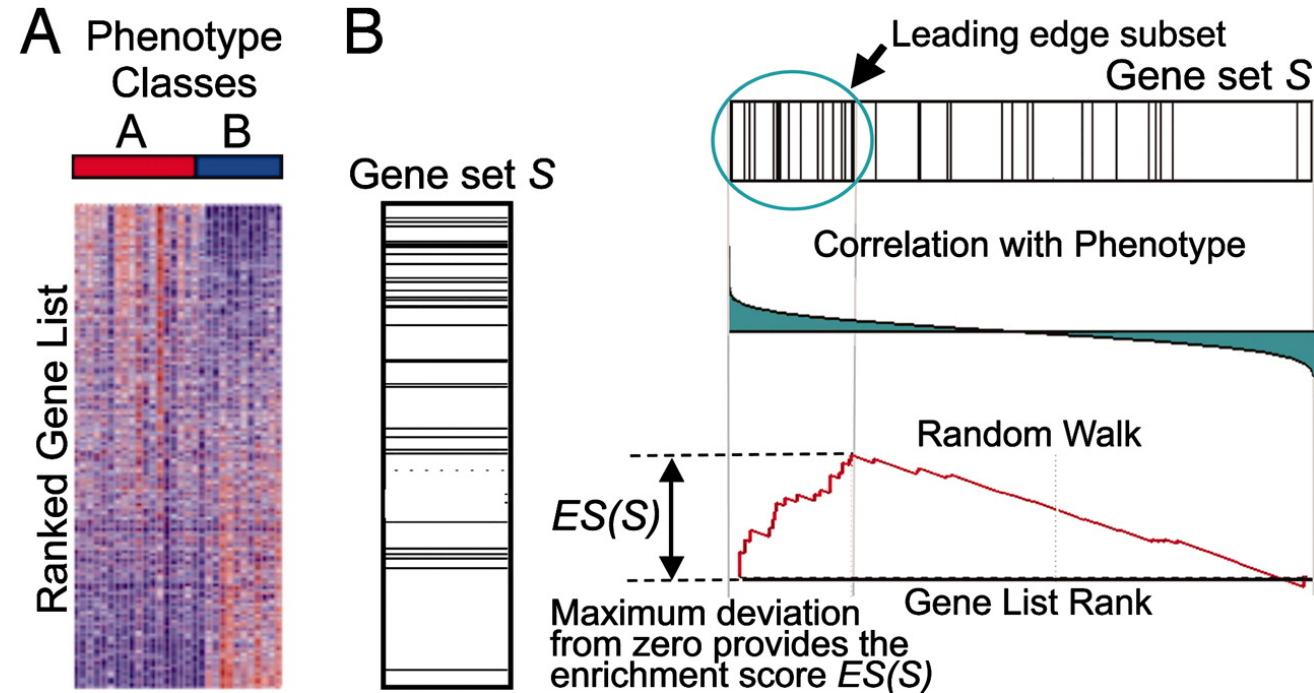
	DNA	ERS	HTS	HYP	MTL	OXD
DNA	1	0.13	0.10	0.16	0.06	0.17
ERS	0.13	1	0.58	0.13	0.05	0.13
HTS	0.10	0.58	1	0.07	0.10	0.13
HYP	0.16	0.13	0.07	1	0.08	0.28
MTL	0.06	0.05	0.10	0.08	1	0.17
OXD	0.17	0.13	0.13	0.28	0.17	1

	DNA	ERS	HTS	HYP	MTL	OXD
DNA	1	0	0	0	0	0
ERS	0	1	0.005	0	0	0
HTS	0	0.005	1	0.005	0.005	0
HYP	0	0	0.005	1	0.005	0
MTL	0	0	0.005	0.005	1	0
OXD	0	0	0	0	0	1



GSEA scores activity of a pathway in transcriptomic data

- Gene signature enrichment analysis
 1. Rank order genes by expression
 2. KS random walk through SigSet
 3. Count genes in SigSet with up score
 4. Count genes in SigSet with down score
 5. Calculate total
- Calculate GSEA NES
 - Used myGSEA implementation of vGSEA (Richard Judson at the EPA)
 - No Up/Down characteristic included



Aravind Subramanian *et al.* PNAS Oct 2005, 102 (43) 15545-15550; DOI: 10.1073/pnas.0506580102

GSEA scores to assess SR activity and assign stress class to perturbagens

preliminary run scored with
TH200 SigSet

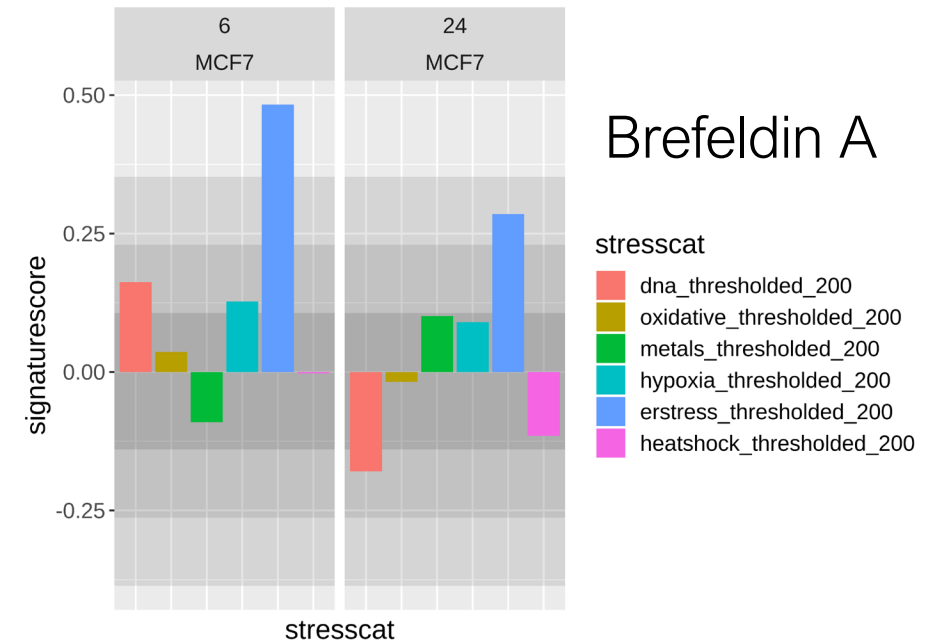
- *Brefeldin as ERS case study*
- *Decrease in DNA at 24 may indicate cell death*

“Accuracy by” is the depth into the ranked GSEA scores at which the reference categorical assignment is met

- e.g.:

1/4 correctly classified by GSEA score at top ranked score depth is 25% Accuracy by 1st

3/4 correctly classified by GSEA score at the second highest score rank depth is 75% accuracy by 2nd



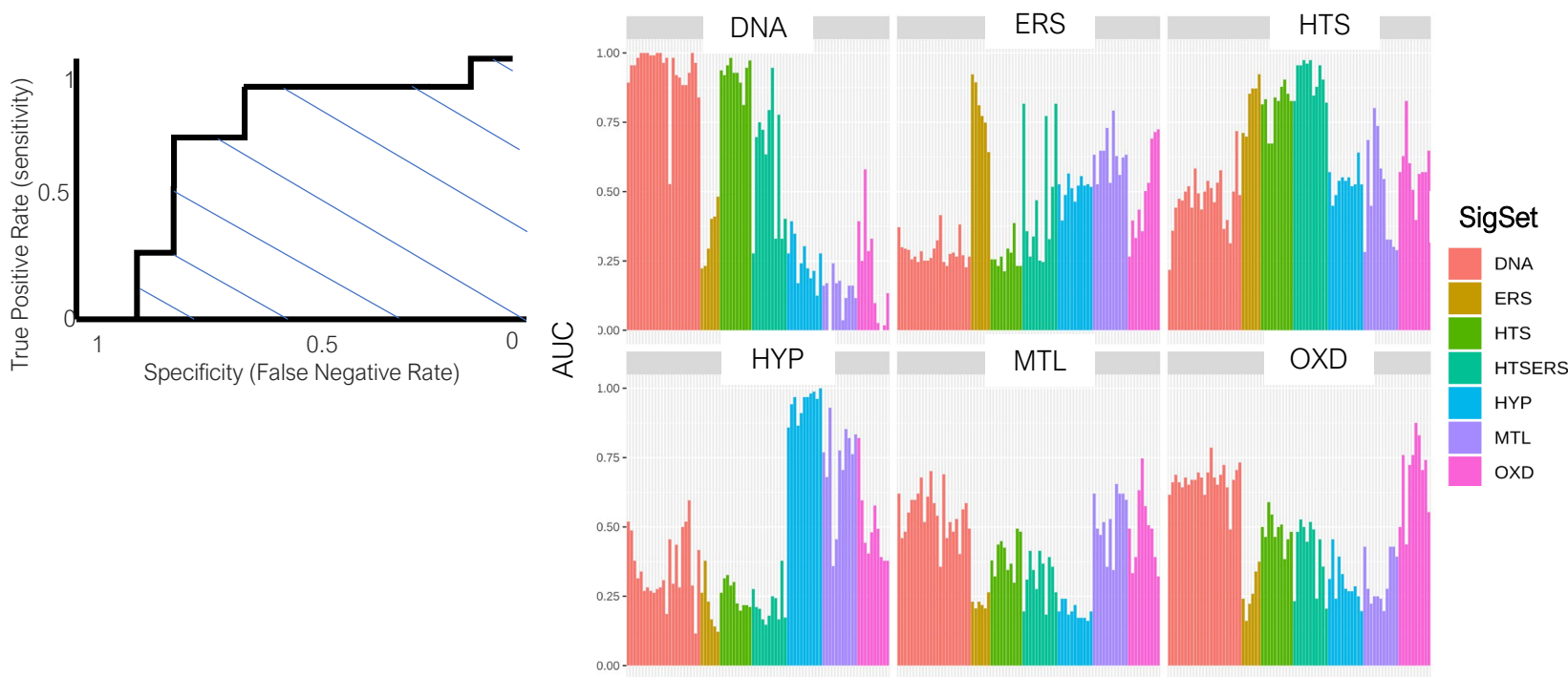
SR pathway	Accuracy by 1 st	Accuracy by 2 nd	Accuracy by 3 rd
DNA	50%	100%	100%
ERS	74%	100%	100%
HTS	0%	50%	89%
HYP	100%	100%	100%
MTL	25%	25%	25%
OXD	25%	50%	100%
Mean	46%	71%	86%

Accuracy by score depth for TH200 Consensus Set



Assess diagnostic ability of all consensus signature sets with receiver operating characteristics(ROC)/Area Under Curve (AUC) analysis

- Calculate the receiver operating characteristic (ROC) area under the curve (AUC) as a measure of performance for each SR category
- Assess the ability of all signature sets to diagnose each SR pathway
- The winning set was chosen by selecting on two parameters:
 - Maximized $AUC_{in\ group} / (Total\ AUC_{out\ group})$
 - Signatures set Size
- A group of the 6 most diagnostic sig sets were chosen
 - MTL has some overlap with HYP
 - OXD is not very selective and is overpowered by DNA



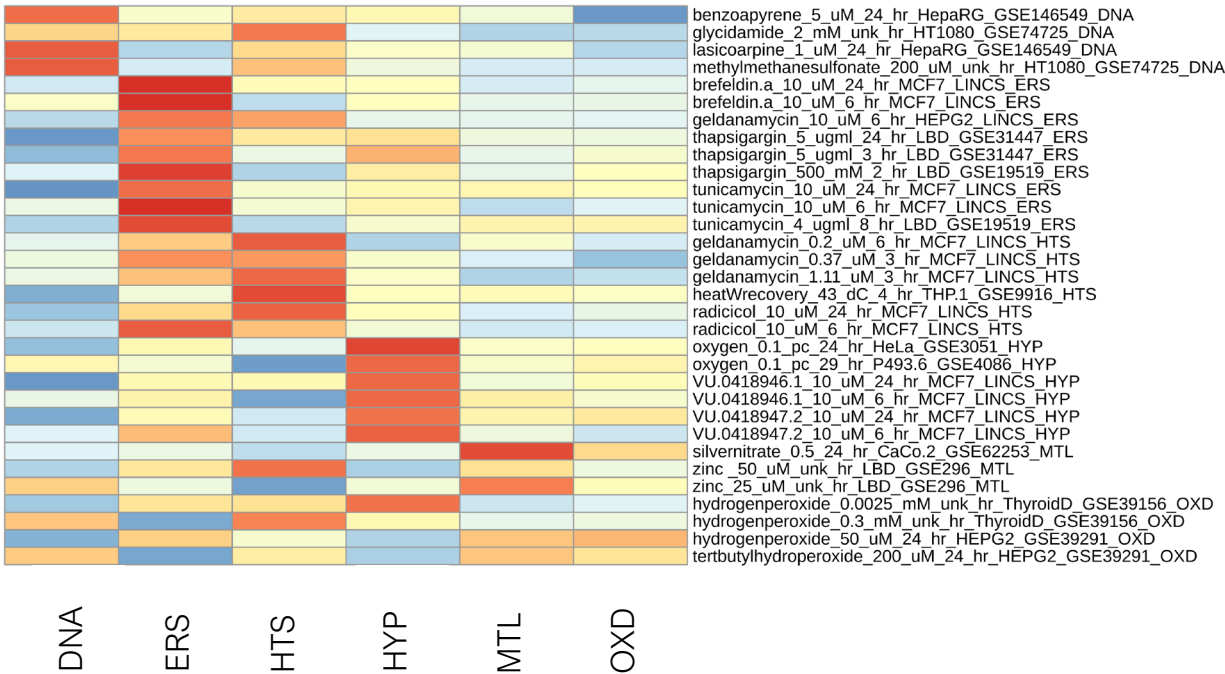
names	DNA	ERS	HTS	HYP	MTL	OXD
DNA_thresholded_300	100	25.60	46.79	33.97	59.77	64.28
DNA_thresholded_400	100	26.57	50	26.92	59.77	67.85
DNA_thresholded_477	100	24.63	51.92	28.20	62.06	65.17
GO_DNA_REPAIR	100	25.12	49.35	27.56	60.91	69.64
GO_DOUBLE_STRAND_BREAK_REPAIR	100	26.08	43.58	28.20	70.11	67.85
REACTOME_DNA_REPAIR	100	27.05	50	28.84	56.32	66.96
GO_BASE_EXCISION_REPAIR	99.10	28.50	44.23	26.92	67.81	66.96
GO_CELLULAR_RESPONSE_T O_DNA_DAMAGE_STIMULUS	99.10	25.12	58.33	26.28	51.72	66.96



SR consensus SigSet mean accuracy improved by selecting top AUC

- The six most diagnostic sig sets were primarily defined by consensus SigSets
- Improvement to 78% using new sigsets at top GSEA score and 91% by second
- OXD is least diagnostic – perhaps due to centrality of system

Signatures Set	DNA	ERS	HTS	HYP	MTL	OXD
DNA_thresholded_400 (DNA)	100	26.6	50	26.9	59.8	67.9
ERS_thresholded_200 (ERS)	23.2	89.4	69.9	37.8	20.7	16.1
GO_DE_NOVO_PROTEIN_FOLDING (HTS)	72.3	33.8	97.4	16.7	34.5	50
HYP_thresholded_400 (HYP)	27.7	51.7	52.6	100	19.5	19.6
MTL_thresholded_200 (MTL)	11.6	62.8	32.7	85.3	65.5	27.7
OXD_thresholded_200 (OXD)	9.82	50.2	39.7	57.7	57.5	87.5



Category	Accuracy in 1st	Accuracy by 2nd	Accuracy by 3rd
DNA	75	100	100
ERS	100	100	100
HTS	66.7	100	100
HYP	100	100	100
MTL	66.7	100	100
OXD	25	25	50
Mean Accuracy	78%	91%	93%



Null targets and negative scores do not overlap with positive scores

Find chemicals and TRx profiles from LINCS for chemicals acting on targets (based on Tau scores):-

Estrogen
Dopamine/Serotonin
Antipsychotics
Etc.

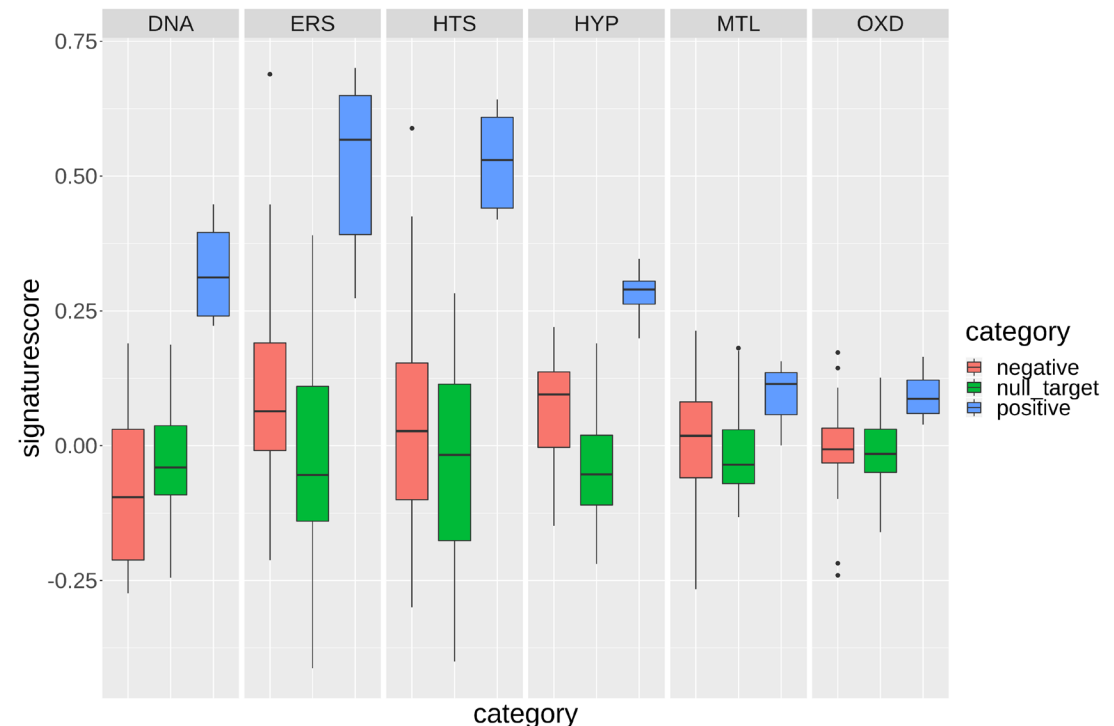
- 31 Chemicals / 64 profiles
- Defined as null targets

Categories of comparison

- Positive: hallmark assignment and pathway assignment should match
- Negative: hallmark assignment and pathway assignment should **not** match
- Null target: chemical should not have any pathway activity

- ERS shows some broader activation but this is expected
- While OXD and MTL are less sensitive they still exhibit good separation from negative and null scores in their class

Name	Description	Median Tau
azaperone	dopamine receptor antagonist	2.62
tetrahydropalmatine	serotonin release inhibitor	2.55
elvitegravir	HIV integrase inhibitor, HIV inhibitor	2.51
midazolam	benzodiazepine receptor agonist, GABA benzodiazepine site receptor agonist	2.47
androstanol	constitutive androstane receptor (CAR) inhibitor	2.3
bicalutamide	androgen receptor antagonist	2.26
tibolone	estrogen receptor agonist, androgen receptor agonist, progesterone receptor agonist, selective estrogen receptor modulator (SERM), sterol sulfatase inhibitor	2.22
GBR-13069	dopamine uptake inhibitor	2.18
estriol	estrogen receptor agonist, estrogen receptor antagonist	-1.36



Key Outcomes

- We can build gene sets to represent adaptive stress response pathways
 - While existing gene sets did exist, consensus SigSets were more accurate
- These stress response gene sets can accurately classify the 18 hallmark perturbagens
 - Based on ROC AUC analysis of GSEA scores
 - ERS tends to be less specific with null targets
- Limitations of this approach
 - The transcriptional fingerprint only as good as the starting gene sets
 - Evaluation performance as good as the set of perturbagen set – need more profiles
- SR gene set analysis can potentially characterize non-specific stress responses and enable identification of hazard and potency (using concentration response modeling)
 - Not shown: Failed Pharmaceutical, Concentration response mapping



Finding additional SR pathway inducers – a taste of what is to come

Identify list of chemicals perturbagens in LINCS
(~20,547 chemicals)

- Remove chemical only identified by Broad Institute ID (e.g., BRD-)
- Reduced to 4671 chemicals

Search PubMed with all chemicals & a set of stress response pathways (SRPs)

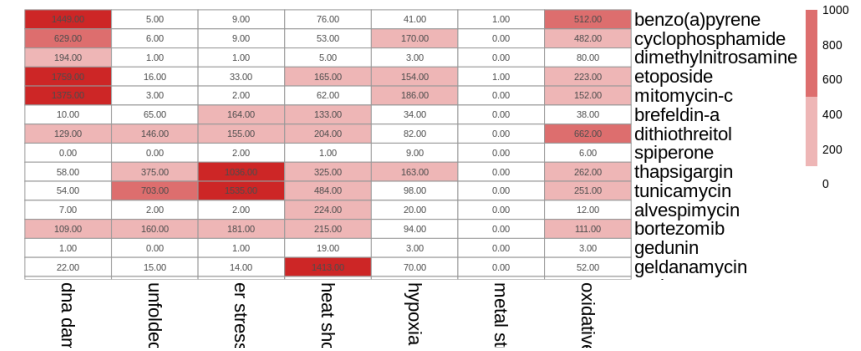
- 7 terms (e.g., 'dna damage', 'er stress', unfolded protein response')
- Totaled 32,679 searches

PMI calculated for all returned chemicals exceeding threshold abstracts to guide perturbagen selection and assignment

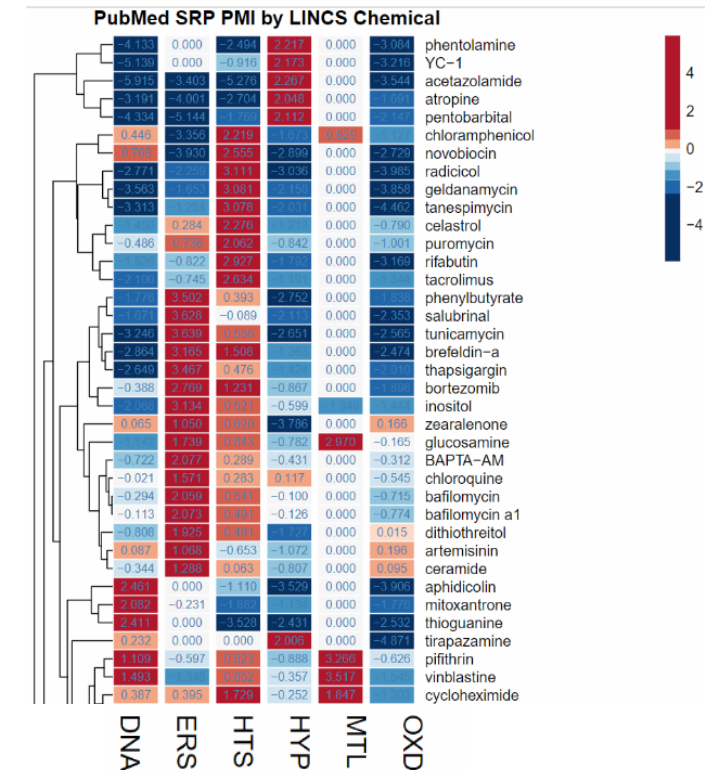
$$PMI(chemical, stress) = \log \frac{F(chemical, stress)}{F(chem)F(stress)}$$

Pulled first 500 abstracts for each selected perturbagen and validated assignment by hand for 97 chemicals

Inclusion of genetic perturbations (TR/SENS/EFF gene KD and OEs) currently in progress



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



Clustering of PMI scored LINCS perturbagens

Breakdown of chemical SR pathway inducing perturbagens TRx profiles

Total result is 11,364 transcriptomic profiles characterized by:

- 6 SRPs (DNA, ERS, HTS, HYP, MTL, and OXD)
- 83 cells types
- 26 doses
- 5 time points

stresscat	nchems	n_cell_types	n_time_point	time_points	n_doses	number_of_profiles
DNA	24	81	4	3, 6, 24, 48	17	4162
ERS	8	61	3	6, 24, 144	13	485
HTS	11	79	4	3, 6, 24, 48	16	5135
HYP	7	17	2	6, 24	8	168
MTL	2	4	2	6, 24	3	21
OXD	17	62	2	6, 24	19	1405

stresscat	pert_iname	n_cell_types	cell_types	n_time_points	time_points	n_doses	dosing_conc	nber_of_prof	id	description	direction	PMID	Linked PMI	NOTES	additional_PMIDS
DNA	azacitidine	22	A375, A549, .	2	6, 24	9	0.04, 0.12, 0.	181	BRD-K03406	DNA methyltr	positive	32676814	link	some potent	31208284, 29167115, 28851
DNA	benzo(a)pyre	4	HA1E, HCC5	2	6, 24	1	10	8	BRD-K09668	pro carcinoge	positive	32866872	link	NA	32609278, 32470682, 32234
DNA	camptothecin	14	A375, A549, .	2	6, 24	12	0.001, 0.01, 0	129	BRD-A30437	topoisomeras	positive	32846134	link	some potent	32753484
DNA	cyclophosphaz	8	A375, A549, .	2	6, 24	2	10, 20	29	BRD-A09722	alkylating agr	positive	32717509	link	some potent	32609954, 32534006, 32346
DNA	cytarabine	12	A375, A549, .	2	6, 24	6	0.04, 0.12, 0.	61	BRD-K33106	antimetabolite	positive	32474729	link	NA	32460231
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, 0	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	27482301, 25410580, 2426
DNA	DMBA	1	HA1E	2	6, 24	1	10	3	NA	NA	positive	32159784	link	NA	31070092, 31037472, 30311
DNA	doxorubicin	14	A375, A549, .	2	6, 24	12	0.001, 0.01, 0	192	BRD-K92093	topoisomeras	positive	27852227	link	NA	32866497, 32855734, 32825
DNA	etoposide	14	A375, A549, .	2	6, 24	9	0.04, 0.12, 0.	100	BRD-K37798	topoisomeras	positive	32866497	link	some potent	32846134, 32786121, 3252
DNA	gemcitabine	53	A375, A549, .	2	6, 24	12	0.001, 0.01, 0	195	BRD-K15108	cell cycle inhi	positive	32688248	link	some potential	ER stress
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, 0	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	



Questions & Comments?

Acknowledgements

- Imran Shah
- Nancy Baker
- Richard Judson
- Derik Haggard
- The HTTr Team
- Danilo Basili and Team
and Cambridge

