UQ Genetics and Genomics Winter School 2026

Systems Genomics and Pharmacogenomics Module 6 Day 2

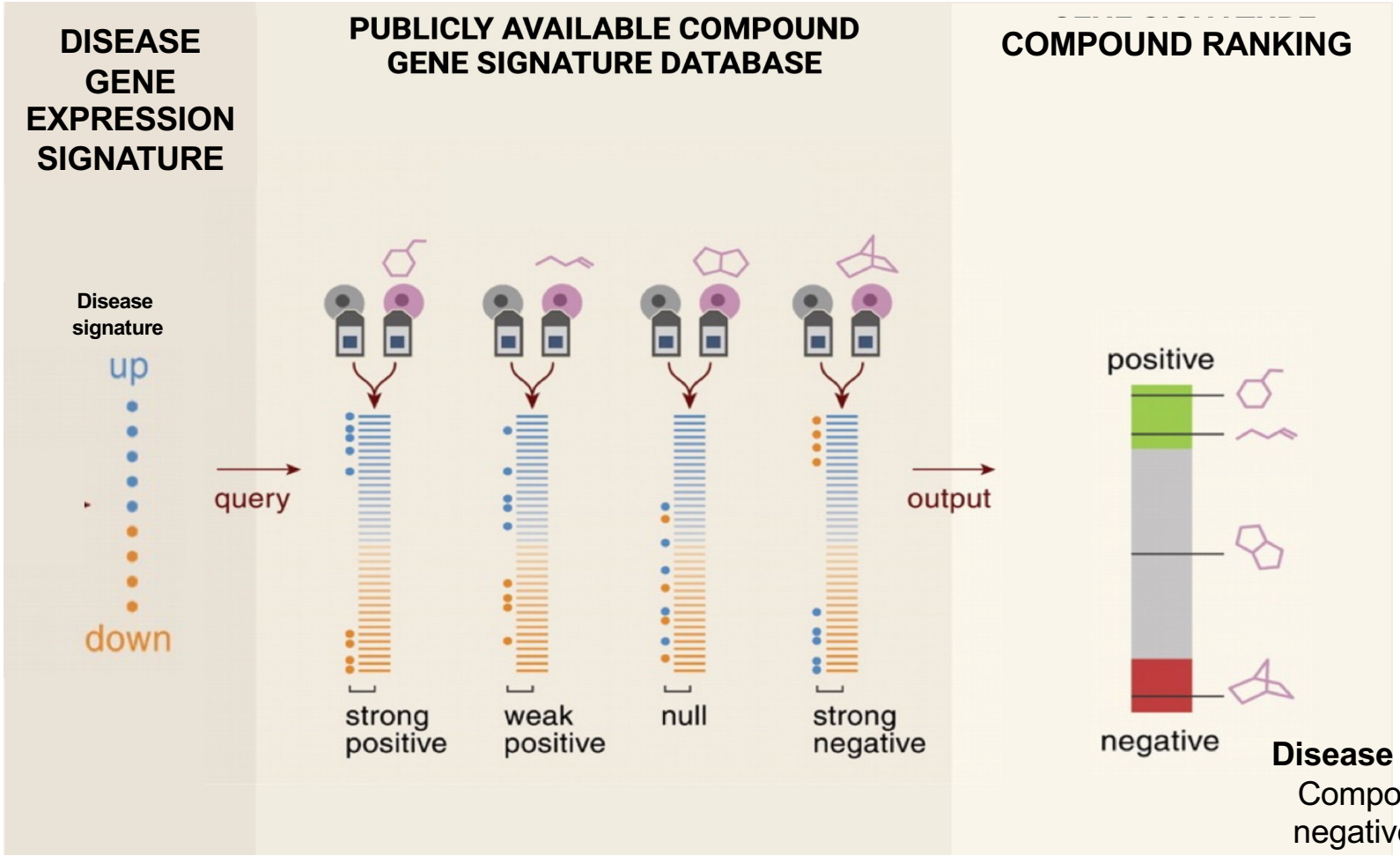
Gene expression signature matching for identifying drug candidates

GWAS to medicine

- Are GWAS-significant genes targets of existing drugs (identify drug repurposing candidates)
  - Repurposing FDA-approved compounds – better safety profile, lower risk, shortest path to approval
  - Can use MR approaches to prioritise genes targeted by existing drugs

- But…
  - Important disease biology may be lost under stringent p-value thresholds
  - Only considers a single gene target rather than a biological pathway
  - MR cannot be used for compounds with unknown mechanism of action (MoA)

Gene expression signatures matching for drug discovery



Disease drug candidates:
Compounds with strong negative correlation with disease signature

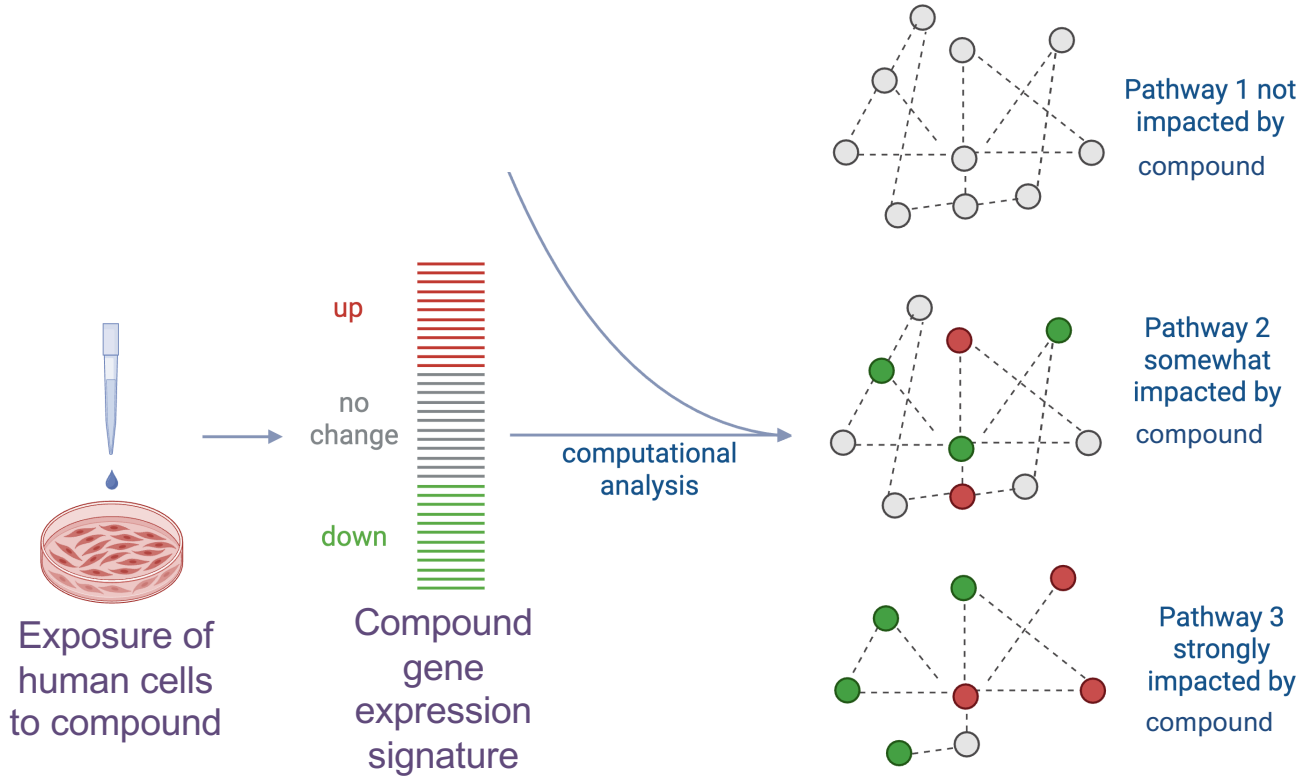
Gene expression signatures matching for understanding drug pharmacodynamics i.e. MoA

Gene expression signature matching to understand drug pharmacodynamics

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Approach 1: **Network analysis**
Which biological pathways are perturbed by your compound in human cells?

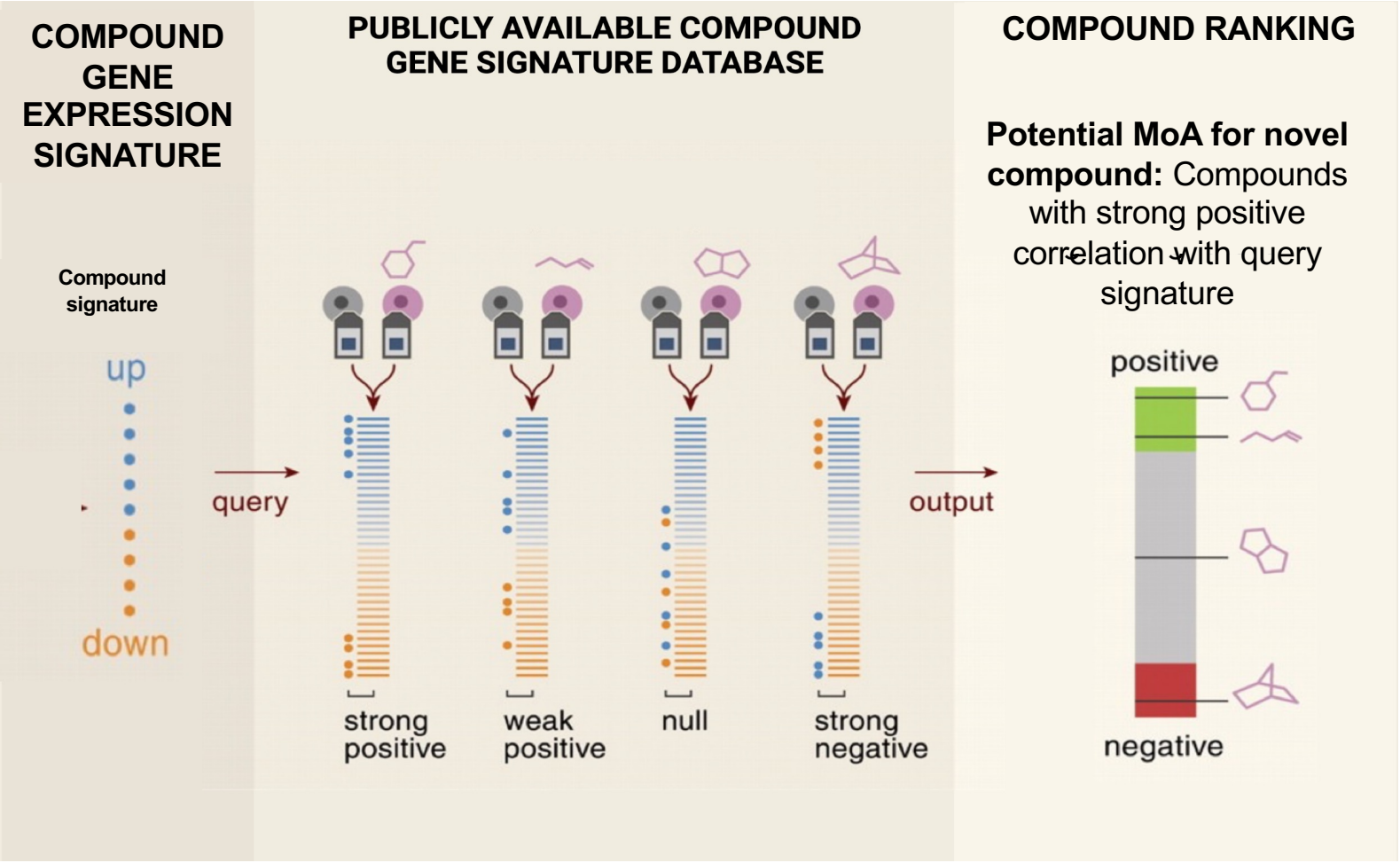
- Map these genes to their biological networks/pathways to understand which pathways are strongly impacted by the compound
- Identify “hub” genes which play a crucial role in these biological processes



Gene expression signature matching to understand drug pharmacodynamics

Approach 2:
Comparative analysis

Which compounds with known MoA have similar signatures to your compound?



Gene expression signature matching for drug discovery

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

1) A database of gene expression signatures for drugs

2) A disease gene expression signature

3) Query the signature database using the disease gene expression signature to identify compounds that ‘reverse’ disease gene expression changes.

Does not require knowledge of the drug’s MoA
Does not require an understanding of disease pathophysiology

Gene expression signature matching for drug pharmacodynamics

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

1) A database of gene expression signatures for drugs

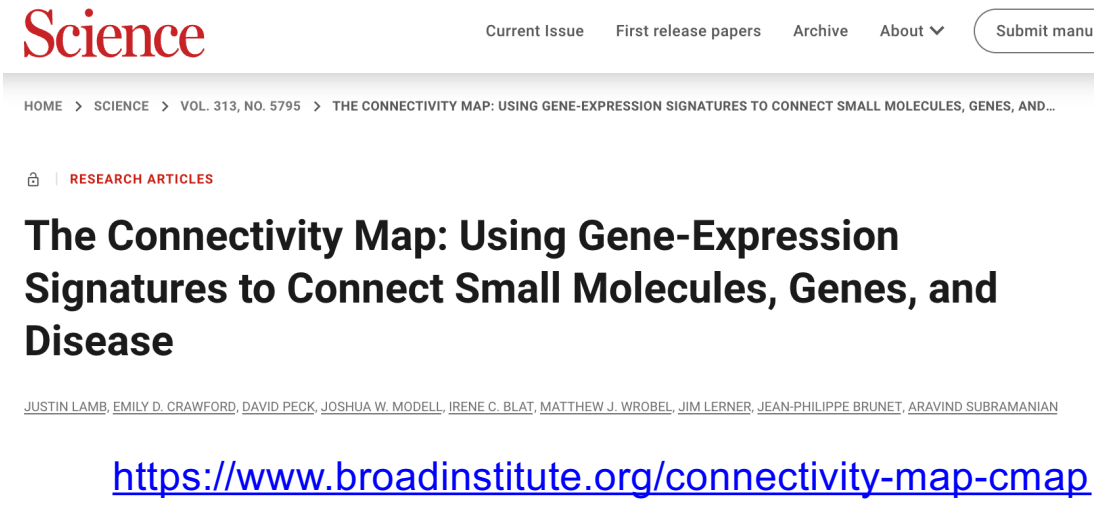
2) Novel compound gene expression signature – easily done using compound perturbation studies using cells.

3) Use network or comparative analysis (latter requires database of compound signatures)

Connectivity Map (CMap)

Library of gene expression signatures in response to chemical and genetic perturbation.

- >1 million gene expression profiles
- ~50 different cell lines
- ~20,000 compounds (chemical perturbation)
- ~5,000 knockdown/overexpression (genetic perturbations)



https://www.broadinstitute.org/connectivity-map-cmap

1st Generation CMap - Lamb et al Science 2006

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

- Need to establish the relation among diseases, physiological processes, and the action of small-molecule therapeutics.

- Previous compound and genetic perturbation studies in yeast and rats
  - Translation to humans
  - High cost of animal studies

- Mammalian cells
  - Generalisable, systematic and biologically relevant
  - BUT…a large number of parameters would need to be optimized for each perturbation – cell type, dose, duration

- Pilot study demonstrated the feasibility of this approach

1st Generation CMap - compounds

164 distinct small-molecule perturbagens, selected to represent a broad range of activities:

- FDA–approved drugs
- nondrug bioactive ‘‘tool’’ compounds
- multiple compounds sharing molecular targets (test if they share gene signatures e.g. HDAC inhibitors)
- compounds with the same clinical indication (test whether compounds with different MoA that treat the same disease generate similar gene signatures e.g. antidiabetics)
- Molecules that are proximal (e.g. selective estrogen receptor modulators) and distal to gene expression
- Molecules whose targets are not expressed in the cell types being tested (COX2 inhibitors)

1st Generation CMAP – cell lines

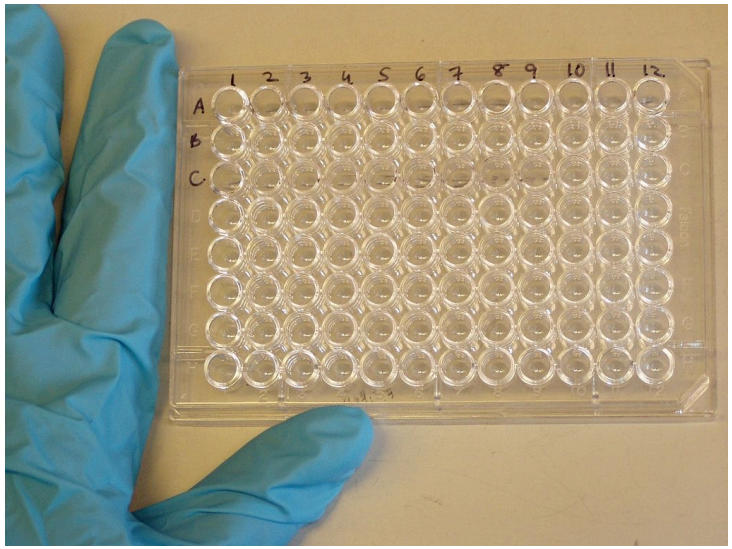
- Stably grown over long periods of time
- Amenable to culture in microtiter plates

- breast cancer epithelial cell line MCF7
  - extensively molecularly characterised,
  - used as a reference cell line

- prostate cancer epithelial cell line PC3

- nonepithelial lines HL60 (leukemia) and SKMEL5 (melanoma)

- Assess degree to which gene signatures are context-dependent



1st Generation CMAP – dose and duration

The University Of Queensland
AUSTRALIA

- 10uM – optimal concentration is not known for many compounds
  - Toxicity studies required for proper optimisation of dose

- 6 and 12 hrs post-treatment
  - Profiles obtained too early might not yield robust signals—esp for perturbations that do not directly modulate transcription
  - Profiles obtained too late may reflect secondary and tertiary responses
  - obtain signatures related to direct mechanisms of action

- Dose and duration dependent on question of interest, but difficult to optimise in such high-throughput experiments.

Compound gene signature generation

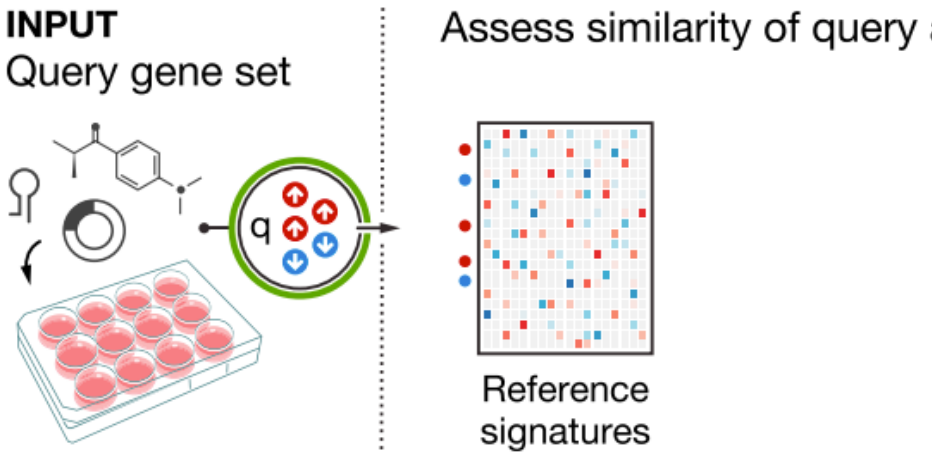
- Control perturbations for each treatment (cells grown on the same plate treated with vehicle only)
  - minimize the impact of batch-to-batch
  - biological and technical variation
- Replicates
- Data were collected in multiple batches over a period of 1 year by Affymetrix GeneChip microarrays.
- DEG analysis – compound-treated gene expression vs intra-batch vehicle-treated control
- For each treatment ~22,000 genes rank-ordered according to differential expression

Can gene expression signature matching

a) identify MoA of a compound?

b) identify drug candidates for disease?

Connectivity score - metric for signature similarity



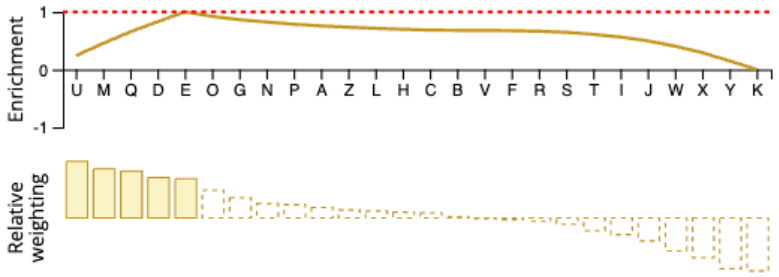
- Rank-based pattern-matching strategy based on the Kolmogorov-Smirnov statistic
- Determine if the most significant DEGs in query gene set are randomly distributed in the reference compound signature
- Enrichment score - reflects the degree to which your query gene set is overrepresented in the extremes of the ranked reference gene signature

Up-regulated genes from query

A B C
D E F
G H I
J K L
M N O
P Q R
S T U
V W X
Y Z

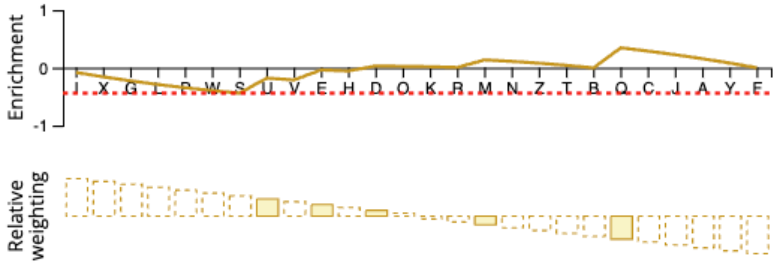
RUN

Example signature 1



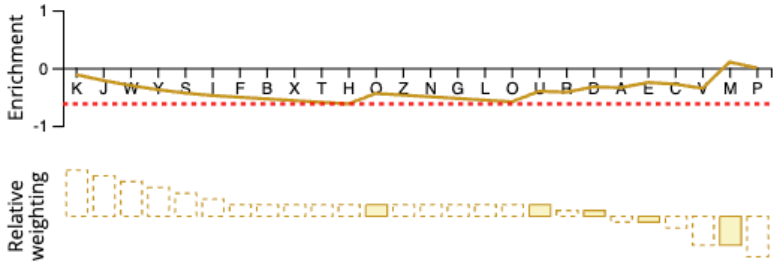
positively enriched

Example signature 2



not enriched

Example signature 3

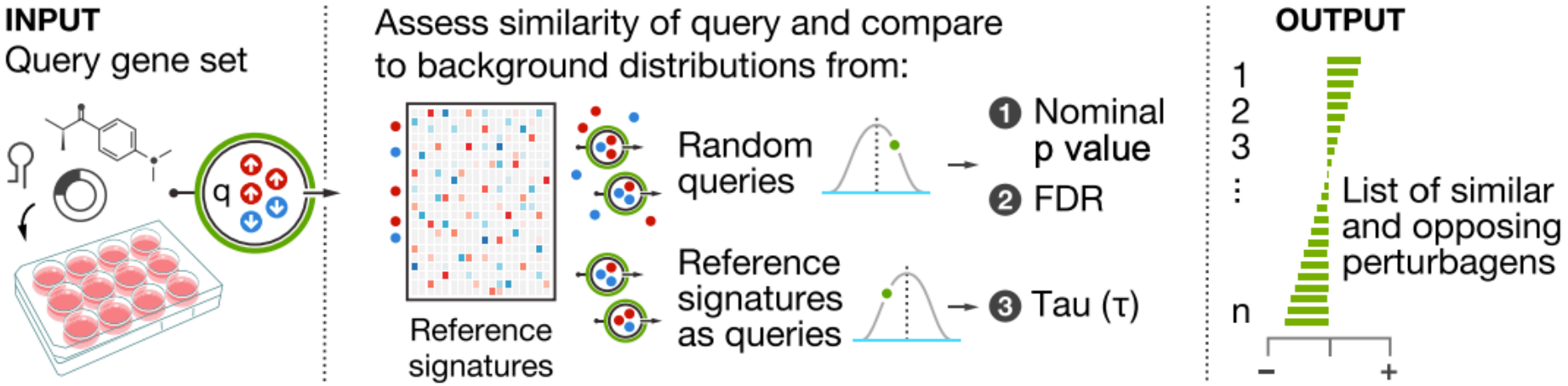


slightly negativel enriched

The University Of Queensland
AUSTRALIA

https://clue.io/connectopedia/cmap_algorithms

Connectivity score - metric for signature similarity



Measures of confidence:

- Nominal p-value - comparing similarity of query and reference signature to null distribution of random queries, using KS enrichment statistic
- Tau score – compares an observed enrichment score to all others in the database - a standardized measure ranging from -100 to 100. A Tau of 90 indicates that only 10% of signatures in the database had a stronger connectivity to the query than the compound in question.

https://clue.io/connectopedia/cmap_algorithms

Example results – HDAC inhibitors

- HDACs – remove acetyl groups on histones and regulate gene expression

- Query HDAC signature derived from independent study:
  - response of bladder and breast cancer cells treated with 3 HDAC inhibitors (vorinostat, MS-27-275, trichostatin)
  - 13-gene (8 up and 5 down-regulated) signature

- Determine if a query signature can recover compounds from the same class (same MoA).

A

rank	perturbagen	dose	cell	score
1	vorinostat [1000]	10 μM	MCF7	1
2	trichostatin A [873]	1 μM	MCF7	0.969
3	trichostatin A [992]	100 nM	MCF7	0.931
4	trichostatin A [1050]	100 nM	MCF7	0.929
5	vorinostat [1058]	10 μM	MCF7	0.917
6	trichostatin A [981]	1 μM	MCF7	0.915
7	HC toxin [909]	100 nM	MCF7	0.914
8	trichostatin A [1112]	100 nM	MCF7	0.908
9	trichostatin A [1072]	1 μM	MCF7	0.906
10	trichostatin A [1014]	1 μM	MCF7	0.893
11	trichostatin A [332]	100 nM	MCF7	0.882
12	trichostatin A [331]	100 nM	MCF7	0.846
13	trichostatin A [448]	100 nM	PC3	0.788
14	valproic acid [345]	10 mM	MCF7	0.743
15	valproic acid [23]	1 mM	MCF7	0.735
16	valproic acid [1047]	1 mM	MCF7	0.733
17	trichostatin A [413]	100 nM	ssMCF7	0.725
18	valproic acid [410]	10 mM	HL60	0.725
19	valproic acid [458]	1 mM	PC3	0.680
33	valproic acid [409]	1 mM	HL60	0.634
39	valproic acid [1020]	500 μM	MCF7	0.619
52	valproic acid [346]	2 mM	MCF7	0.582
61	valproic acid [1078]	500 μM	MCF7	0.563
71	valproic acid [629]	1 mM	SKMEL5	0.539
72	valproic acid [347]	500 μM	MCF7	0.539
73	valproic acid [989]	1 mM	MCF7	0.538
76	valproic acid [433]	1 mM	PC3	0.528
89	trichostatin A [364]	100 nM	HL60	0.507
92	valproic acid [497]	1 mM	ssMCF7	0.501
297	valproic acid [348]	50 μM	MCF7	0
388	valproic acid [994]	200 μM	MCF7	0
403	valproic acid [1002]	50 μM	MCF7	0
419	valproic acid [1060]	50 μM	MCF7	-0.537

IVERSITY
ENSLAND
ITA

- Compounds with HDAC inhibitory effects shown by black lines
- Despite differences in cell lines used to generate query signature, the approach identifies HDAC inhibitors as the top scoring compounds.
- Not highly sensitive to concentrations
- Strong connectivity with two structurally distinct compounds, valproic acid (developed as an antiseizure drug) and HC toxin, both now known to have HDAC-inhibitory activity

(green, positive; gray, null; red, negative

Example - Estrogens

The University Of Queensland
AUSTRALIA

- Estrogen – modulates nuclear hormone signaling by binding to estrogen receptor.

- Query signature from an independent experiment – MCF7 cells treated with 17beta-estradiol
  - 129-gene signature (40 up and 89 down-regulated)

A

rank	perturbagen	dose	cell	score
2	estradiol [988]	100 nM	MCF7	0.936
3	estradiol [373]	10 nM	ssMCF7	0.918
4	genistein [1015]	10 μM	MCF7	0.913
5	estradiol [1079]	10 nM	MCF7	0.899
6	estradiol [1021]	10 nM	MCF7	0.813
8	alpha-estradiol [990]	10 nM	MCF7	0.809
9	alpha-estradiol [403]	10 nM	ssMCF7	0.807
10	estradiol [414]	10 nM	ssMCF7	0.794
11	estradiol [121]	10 nM	MCF7	0.758
12	genistein [1073]	10 μM	MCF7	0.753
13	genistein [638]	10 μM	MCF7	0.730
17	alpha-estradiol [1048]	10 nM	MCF7	0.646
20	genistein [268]	1 μM	MCF7	0.619
21	estradiol [365]	100 nM	MCF7	0.610
25	genistein [382]	10 μM	MCF7	0.561
27	genistein [267]	1 μM	MCF7	0.552
46	alpha-estradiol [122]	10 nM	MCF7	0.435
51	estradiol [387]	10 nM	HL60	0.421
64	estradiol [782]	10 nM	HL60	0.376
148	alpha-estradiol [702]	10 nM	PC3	0
152	genistein [703]	10 μM	PC3	0
162	alpha-estradiol [762]	10 nM	MCF7	0
278	estradiol [665]	10 nM	PC3	0

no robust connections recovered in PC3 or HL60 cells, neither of which expresses ER.

Connections with Disease States

- Query – DEGs from a rat model of diet-induced obesity
- Several differences in exp design:
  - Species: Rat vs human,
  - Exposure duration: 65 days vs 6 hrs
  - Tissue: adipose vs cancer cell line
- 3 PPAR-gamma agonists identified
- PPAR-gamma agonists are known potent inducers of adipogenesis in vitro
- Troglitazone and rosiglitazone are anti-diabetic treatments, with weight-gain as a known major side effect
- BUT…null or negative scores in non-PC3 cell lines, (only PC3 expresses PPAR-gamma)

rank	perturbagen	dose	cell	score
3	indometacin [452]	100 µM	PC3	0.874
4	rosiglitazone [430]	10 µM	PC3	0.838
11	troglitazone [462]	10 µM	PC3	0.737
20	troglitazone [431]	10 µM	PC3	0.696
116	15-delta prostaglandin J2 [446]	10 µM	PC3	0

Findings from CMap pilot study

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Gene expression signatures can

1. Identify drugs with common MoA
2. Identify unknown MoA of drugs
3. Identify potential new therapeutics for disease
4. Are often conserved across diverse cell types and settings
   - Drug target needs to be expressed in that cell line
5. Not highly sensitive to the precise concentration of drug

2nd Generation CMAP - LINCS1000

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

- **L**ibrary of **I**ntegrated **N**etwork-**B**ased **C**ellular **S**ignatures

- 1000-fold scale up of the CMAP – more compounds and cell lines plus genetic perturbations.

- Gene arrays and RNAseq not suitable for large-scale profiling
  - High cost
  - RNAseq cannot detect low abundant transcripts without deep sequencing which is costly

2nd Generation CMAP - LINCS1000

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

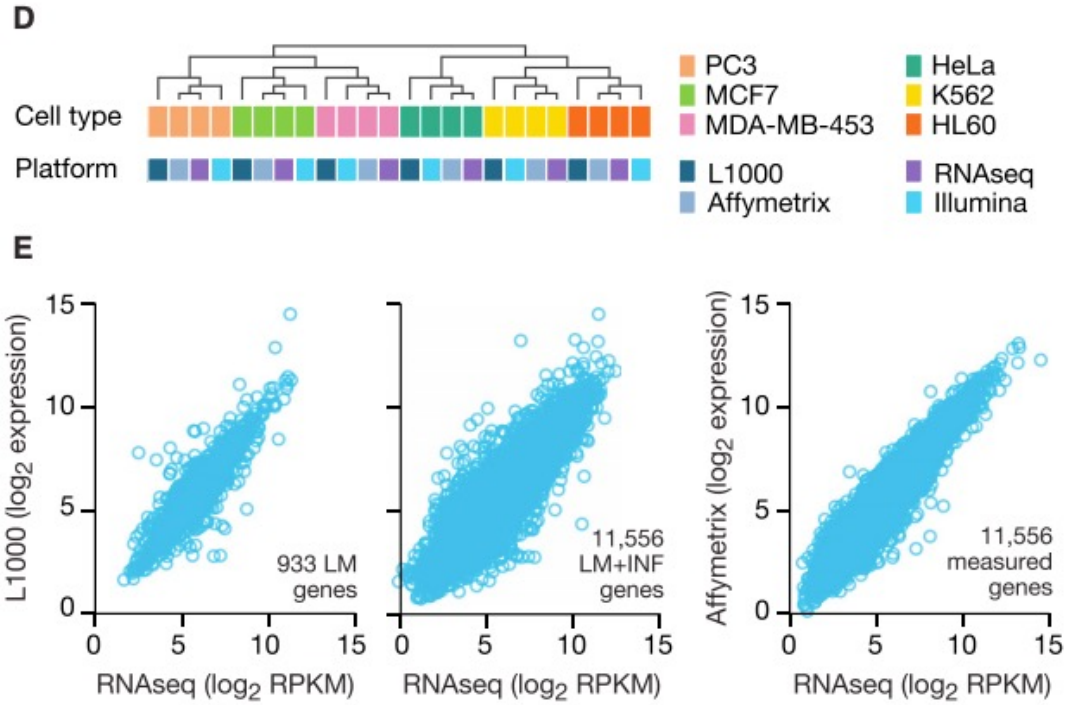
- Capture cellular state at low cost by measuring a reduced representation of the transcriptome.

- Analysed 12K Affy HGU133A expression profiles in GEO
  - Identified the optimal number of informative transcripts (“landmark” transcripts)
  - Cost vs information captured
  - 1000 landmarks enough to capture 82% of full transcriptome

- Tested ability of different number of landmark genes to recover connections observed in pilot data (for 25 signatures)

- No substantial enrichment of particular protein class or developmental lineage in landmark list (some generic classes enriched e.g. enzyme binding, ATP binding).

Comparison of L1000 with RNAseq



strong degree of similarity of profiles across L1000 and RNA-seq platforms