The University of Queensland
Australia

UQ Genetics and Genomics Winter School 2023

Systems Genomics and Pharmacogenomics Module 6 Day 2

Gene expression signature matching for identifying drug candidates

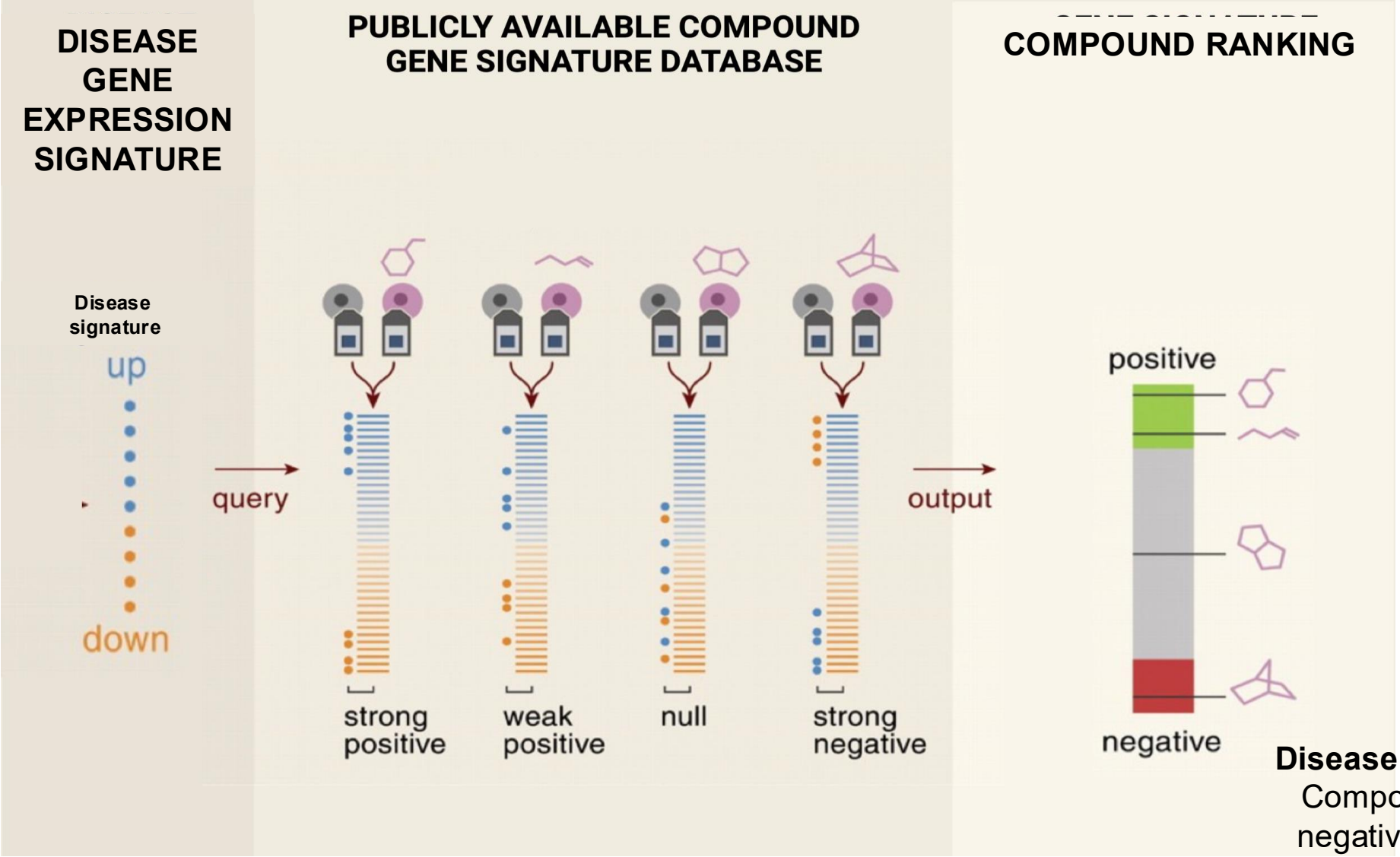
The University of Queensland
Australia

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

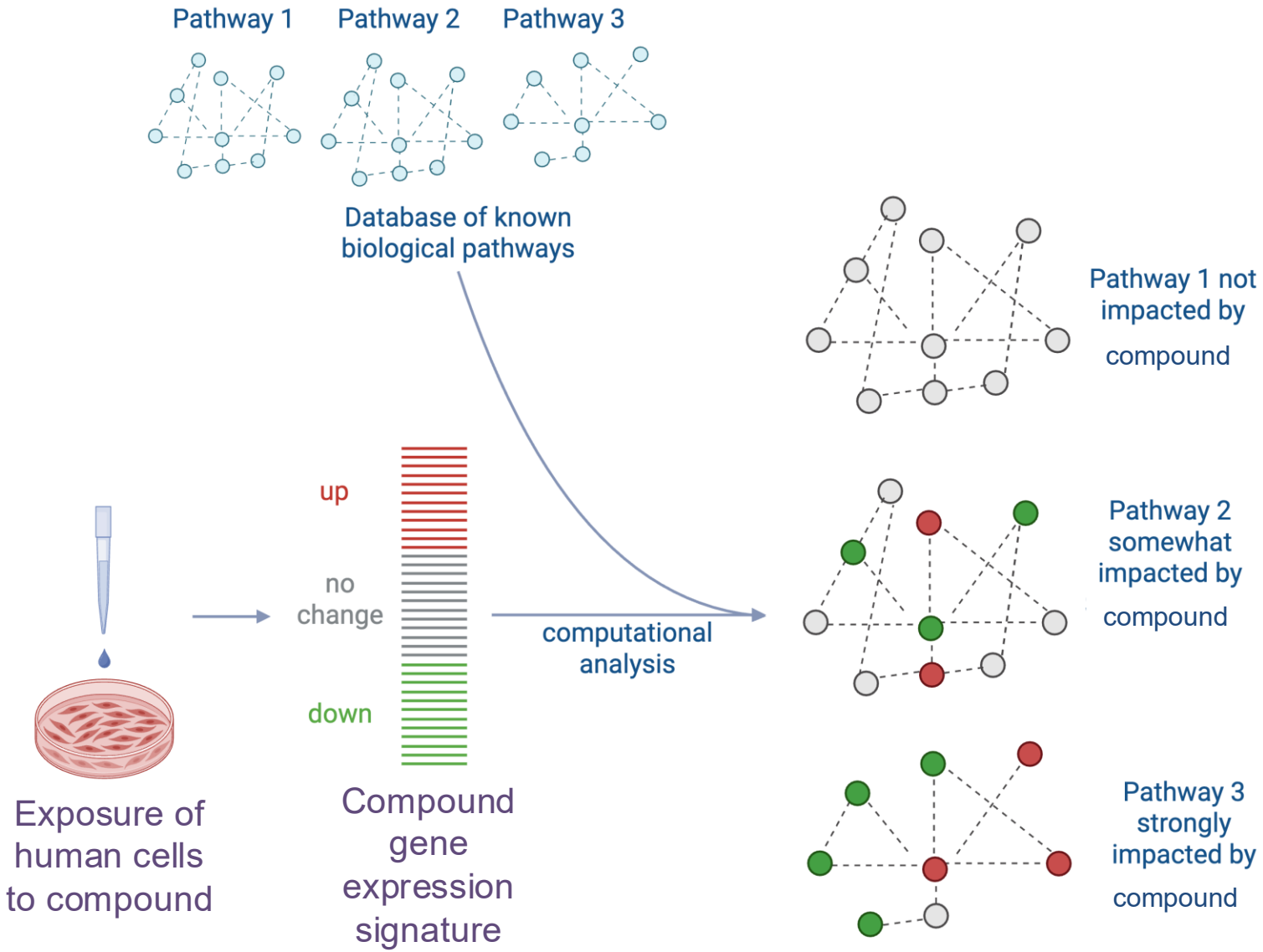
Gene expression signatures matching for understanding drug pharmacodynamics i.e. MoA

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

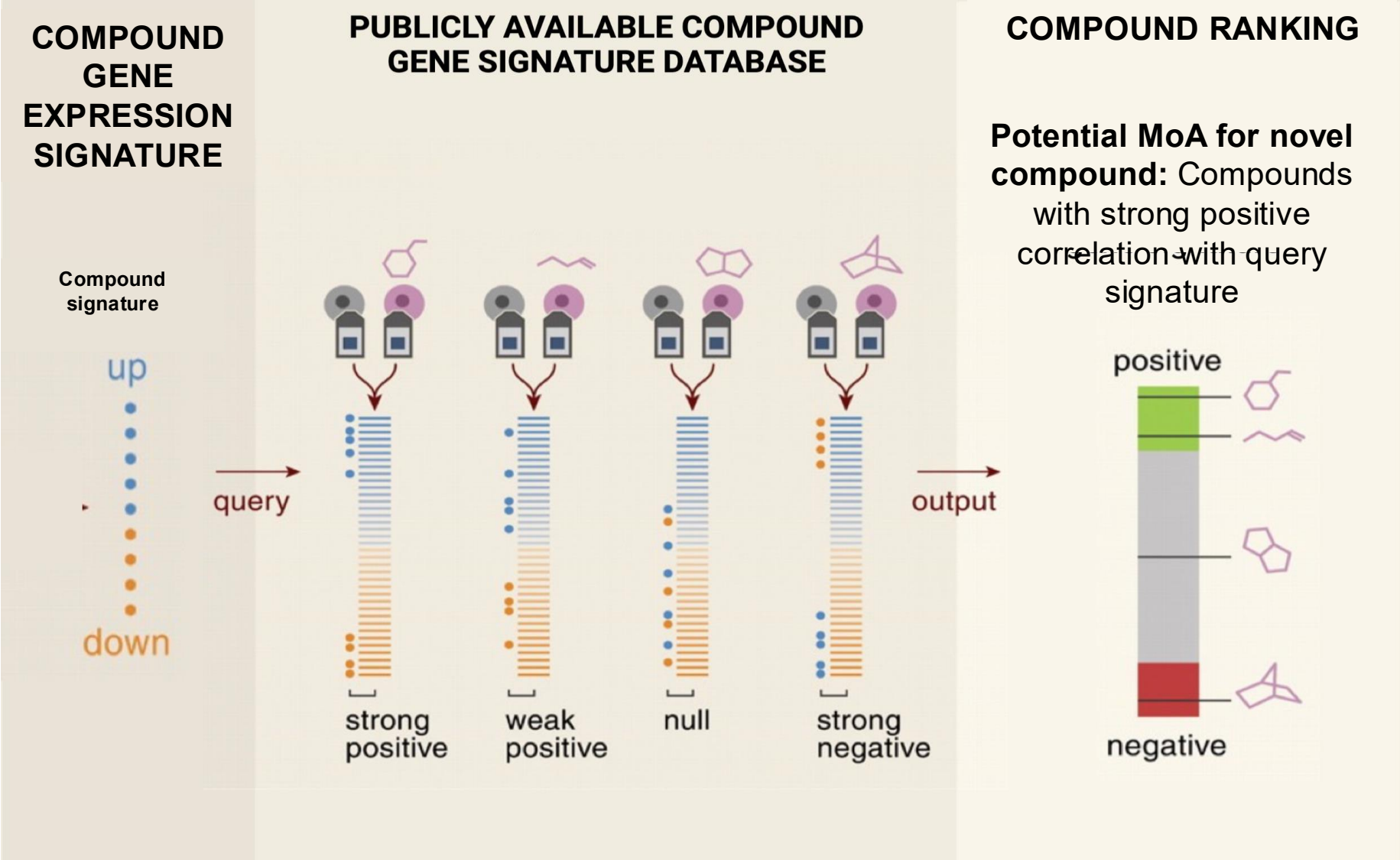


Gene expression signature matching to understand drug pharmacodynamics

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

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

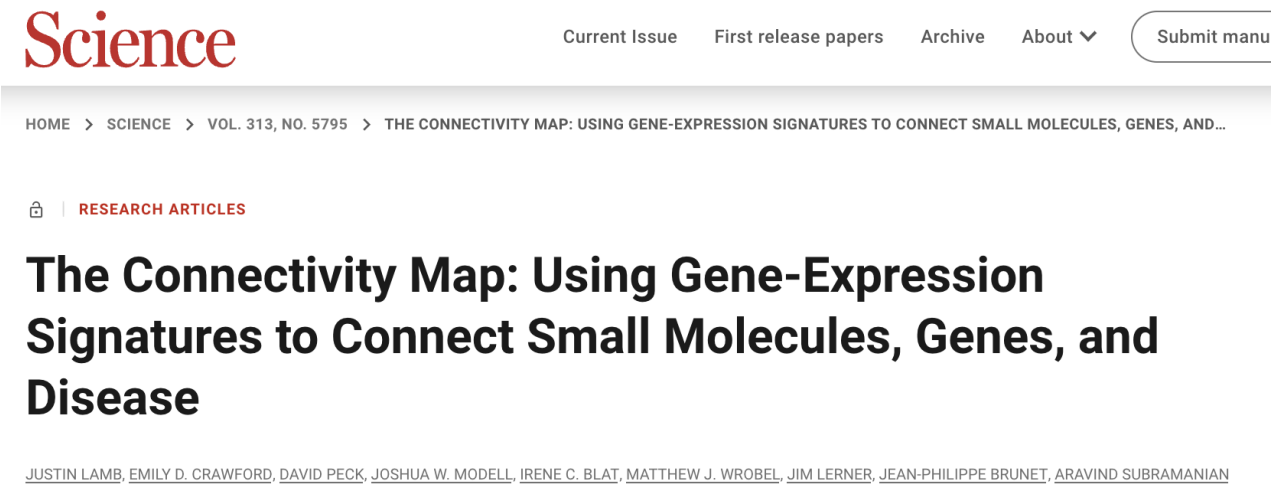
2) Novel compound gene expression signature – easily done using compound perturbation studies using cells.

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)


Science

Current Issue | First release papers | Archive | About | Submit manu

HOME > SCIENCE > VOL. 313, NO. 5795 > THE CONNECTIVITY MAP: USING GENE-EXPRESSION SIGNATURES TO CONNECT SMALL MOLECULES, GENES, AND...

RESEARCH ARTICLES

The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease

JUSTIN LAMB, EMILY D. CRAWFORD, DAVID PECK, JOSHUA W. MODELL, IRENE C. BLAT, MATTHEW J. WROBEL, JIM LERNER, JEAN-PHILIPPE BRUNET, ARAVIND SUBRAMANIAN

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

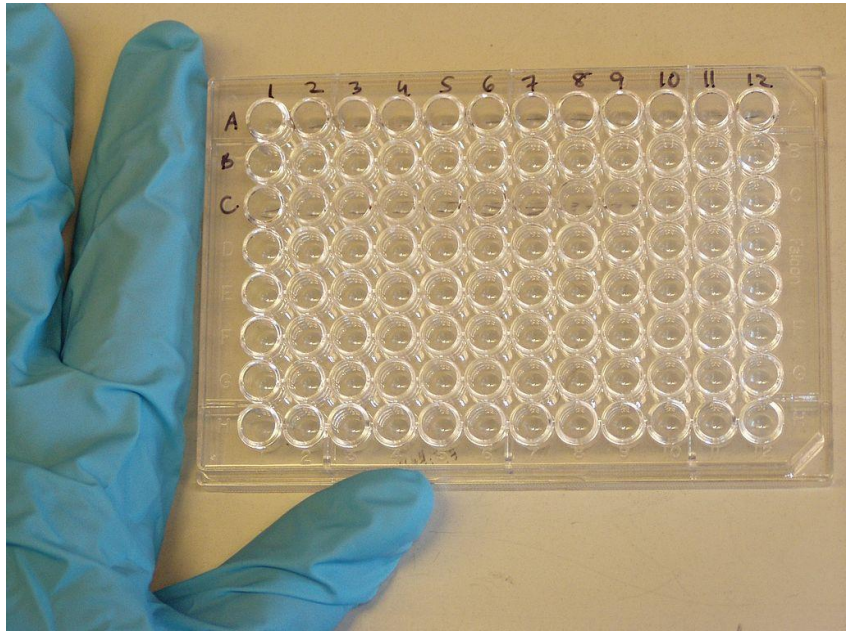
- Stably grown over long periods of time
- Amenable to culture in microtiter plates

- 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

- 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

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

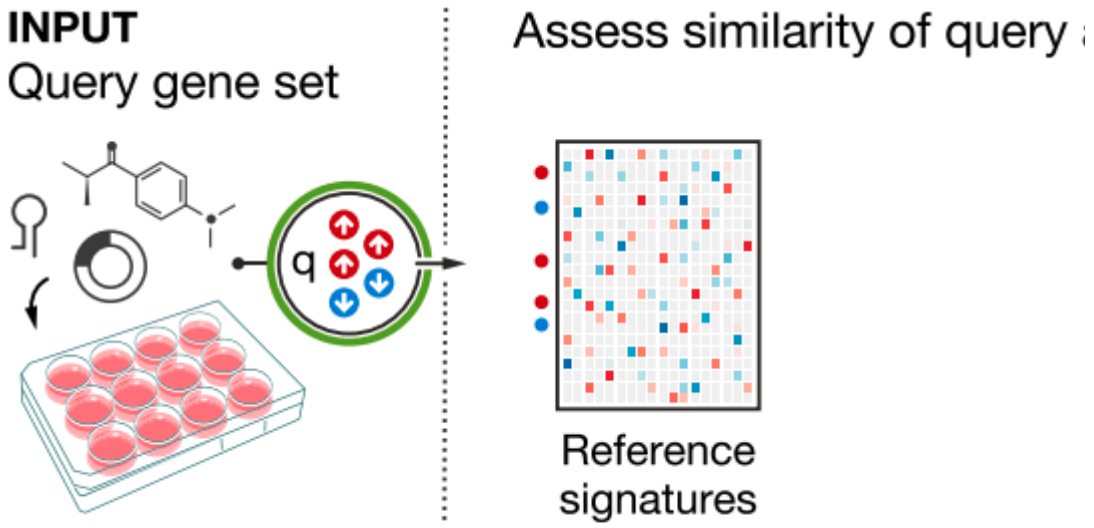
- Control perturbations for each treatment (cells grown on the same plate treated with vehicle only)
  - minimize the impact of batch-to-batch
  - biological and technical variation
- Replicates
- Data were collected in multiple batches over a period of 1 year by Affymetrix GeneChip microarrays.
- DEG 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

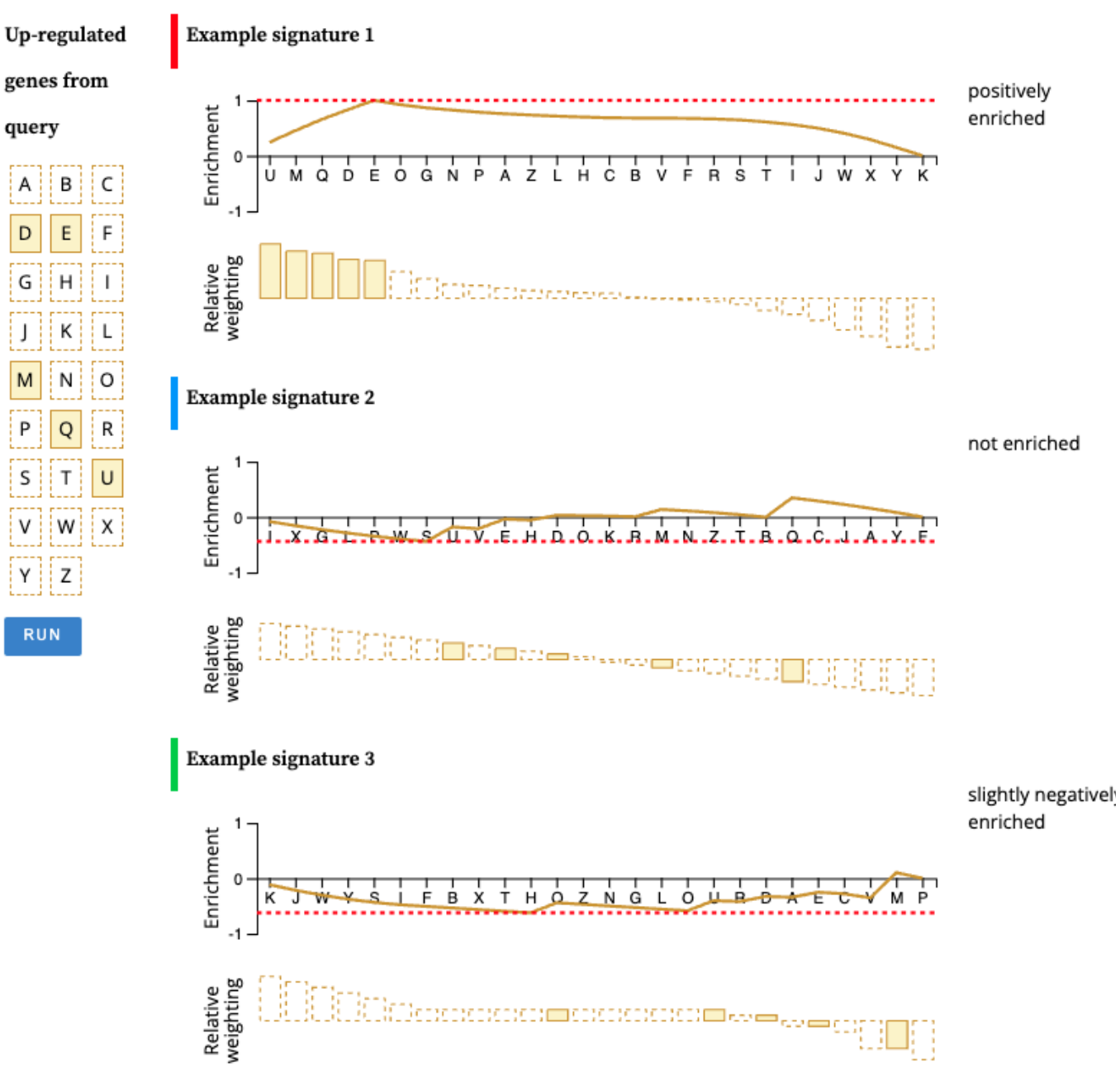
a) identify MoA of a compound?

b) identify drug candidates for disease?

Connectivity score - metric for signature similarity

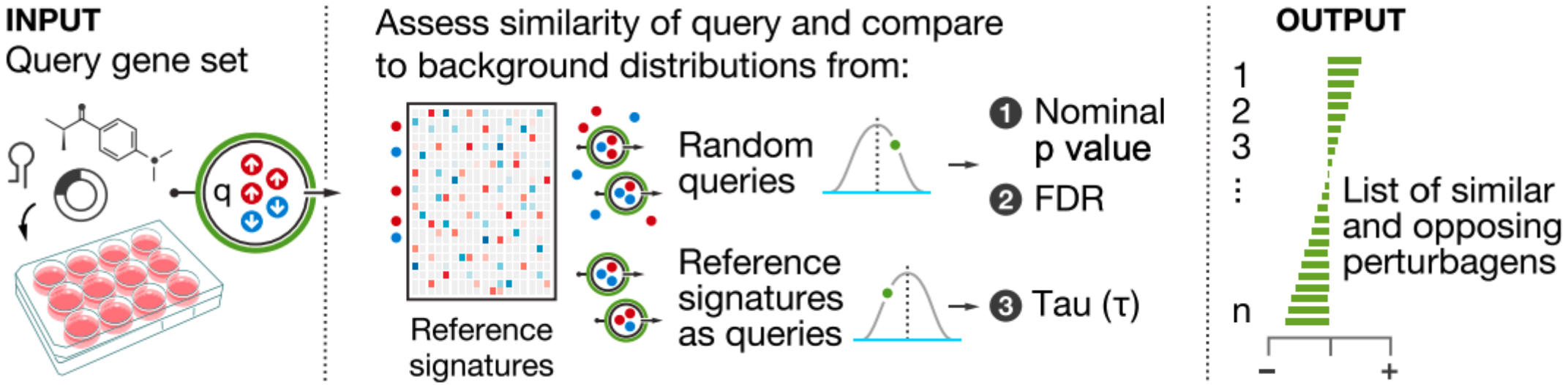


- Rank-based pattern-matching strategy based on the Kolmogorov-Smirnov statistic
- Determine if the most significant DEGs in query gene set are randomly distributed in the reference compound signature
- Enrichment score - reflects the degree to which your query gene set is overrepresented in the extremes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



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

- 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

- 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

B

rank	perturbagen	dose	cell	score
171	fulvestrant [704]	1 μM	PC3	0
261	fulvestrant [523]	1 μM	ssMCF7	0
447	fulvestrant [367]	1 μM	MCF7	-0.749
450	fulvestrant [310]	10 nM	MCF7	-0.843
451	fulvestrant [985]	1 μM	MCF7	-0.961
452	fulvestrant [1076]	10 nM	MCF7	-0.989
453	fulvestrant [1043]	1 μM	MCF7	-1

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

highest negative connectivity scores came from fulvestrant, a known anti-estrogenic drug

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

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

- **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

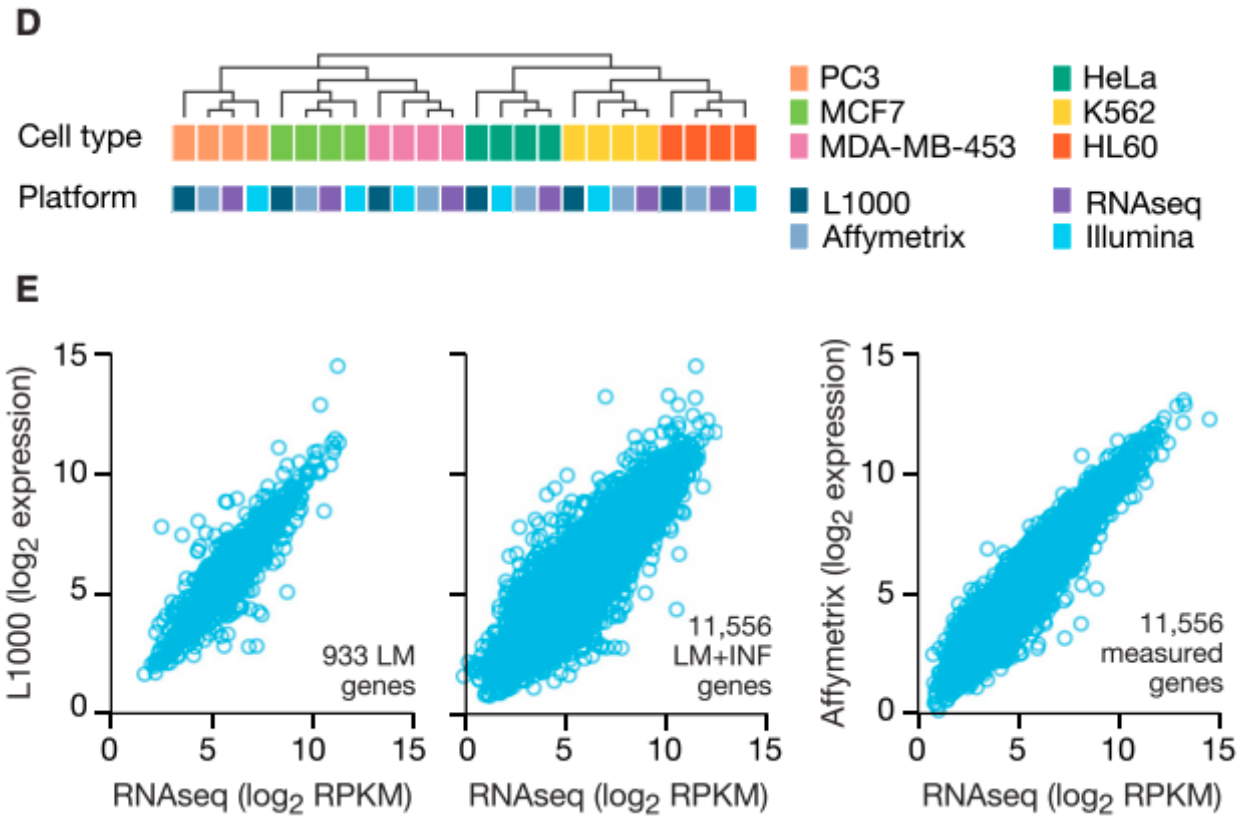
- Capture cellular state at low cost by measuring a reduced representation of the transcriptome.

- 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

Using CMap Data

CMAP – One dataset several names

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CMap-L1000 version 1:

- L1000-based compendium
- Phase 1
- *A Next Generation Connectivity Map: L1000 platform and the first 1,000,000 profiles (2017)*
- Available at GSE92742

Data:

19,811 compounds (drugs and small molecules):

- Different time (6 hours or 24 hours)
- Different concentration (0.04uM to 90uM)
- Different cell lines: (71 different cell lines)

<u>1,319,138 replicates measured.</u>

This dataset also include genetic perturbation

CMAP – One dataset several names

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CMap-L1000 version 2:

- Phase 2
- *No paper published on this dataset (but the data was included in clue.io).*
- Available at GSE70138

Data:

1,768 additional compounds:

- Different time (**3hours**, 6 hours or 24 hours)
- Different concentration (0.04uM to **40uM**)
- Different cell lines: (30 different cell lines)

354,123 replicates measured

This dataset also include genetic perturbation

CMAP – One dataset several names

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

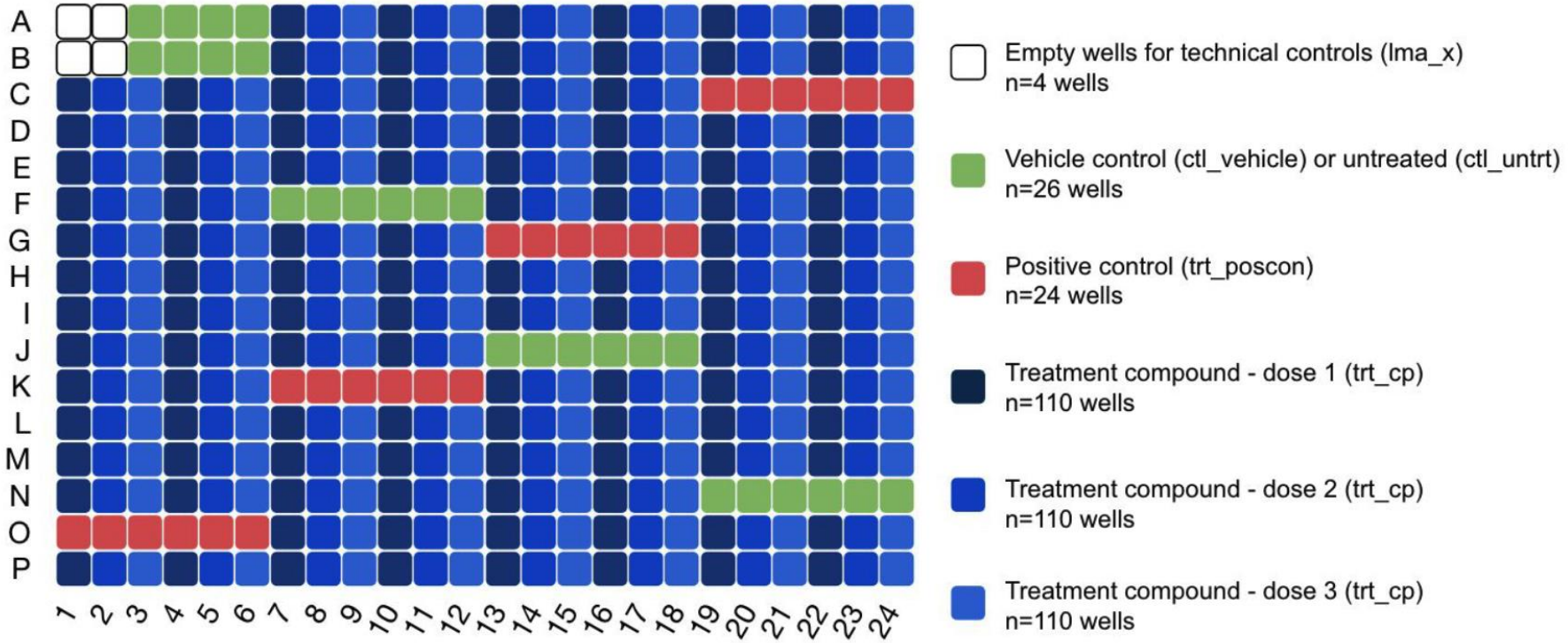
iLincs:

- Uses both GSE92742 and GSE70138.
- This is a meta-analysis combining phase 1 and phase 2.
- Paper: **Connecting omics signatures and revealing biological mechanisms with iLINCS (2022)**
- Data is accessed through the ilincs website

1,673,261 replicates measured

CMAP – What does it actually look like?

Example plate layout: 3 dose

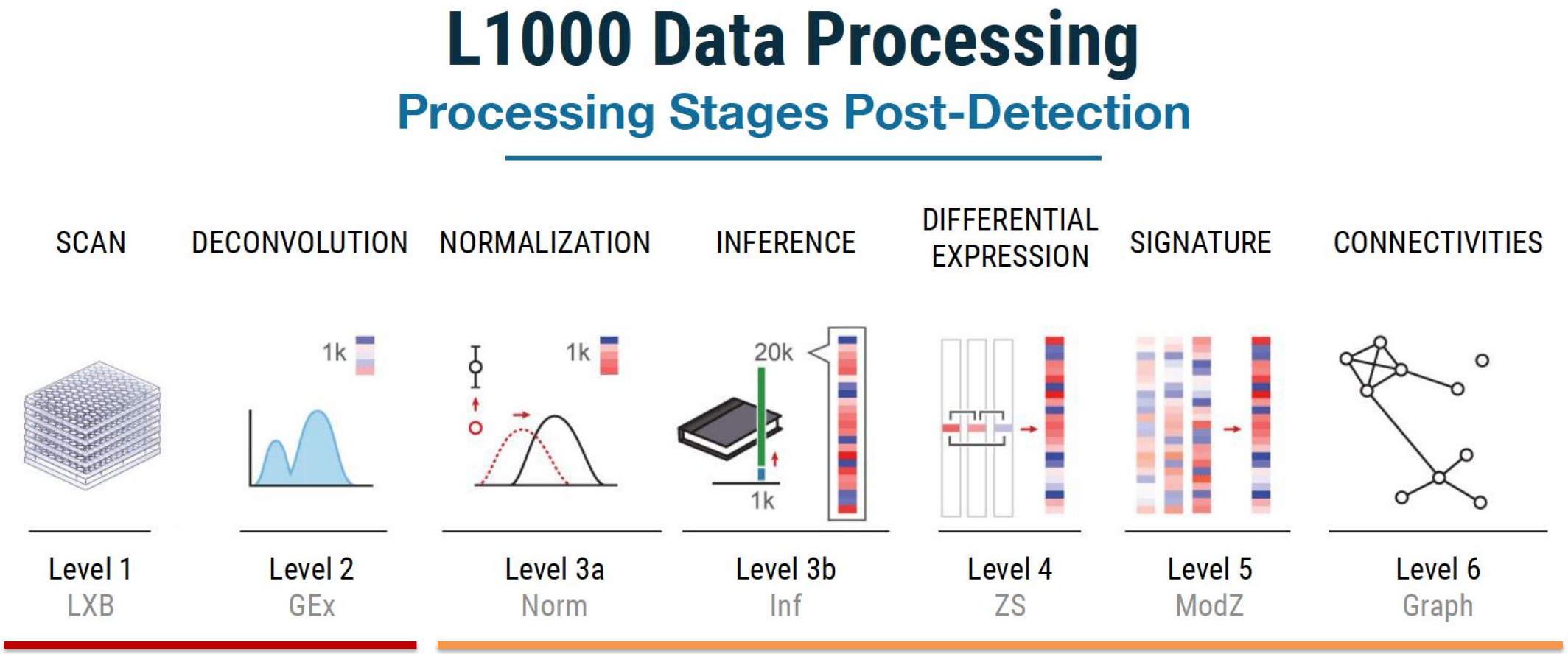


CMAP – Microarray.

Definition of the landmark genes:

- 12,063 gene expression samples profiles using Affymetrix HG-U133A microarrays from the Gene Expression Omnibus (GEO) called DS_{GEO}.

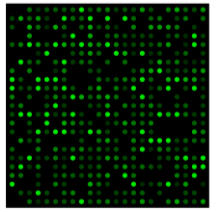
CMAP – Data processing

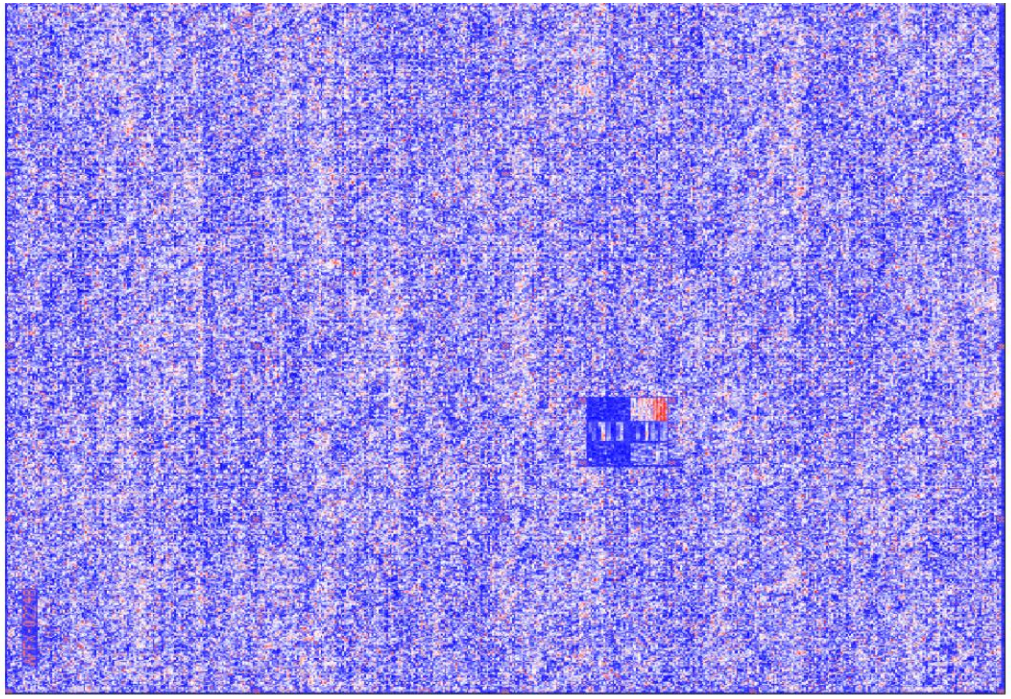


Default microarray processing

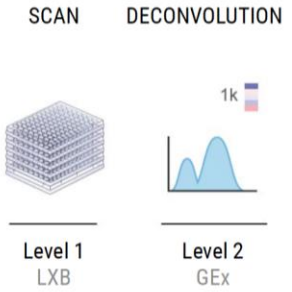
CMAP specific processing

Microarray processing:





Example Slide Affymetrix microarray (~18,000 genes)

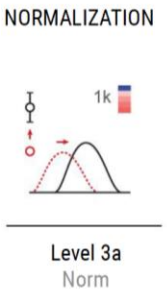


Data generation for each of the replicates:

- Each dot on the slide corresponds to a different gene.
- The intensity and color of the dot represent:
  - **Blue** indicates **low signal intensity** (low expression or hybridization).
  - **Red** indicates **high signal intensity** (high expression or hybridization).
  - **Intermediate shades** (purple, magenta) indicate **medium signal.**

The centre of the image corresponds to control genes of known intensity. This allows to generate a numerical value for each gene measured.

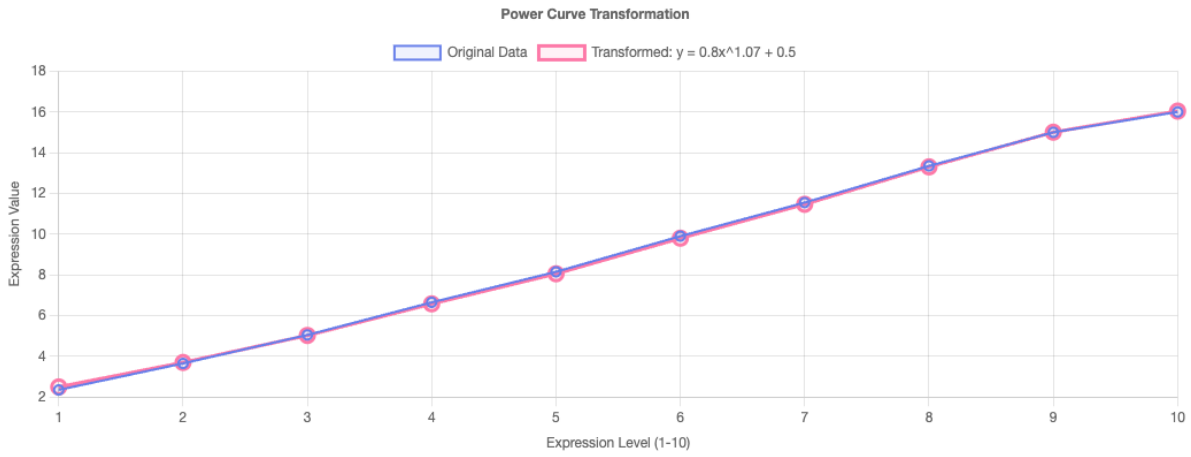
CMAP processing - Level 3 normalisation



At this stage, the data contain only 1,058 genes measured for each replicate (1,673,261)

Normalisation is called L1000 Invariant Set Scaling.

For each sample the expression of 80 invariable gene is used to generate a “calibration curve”



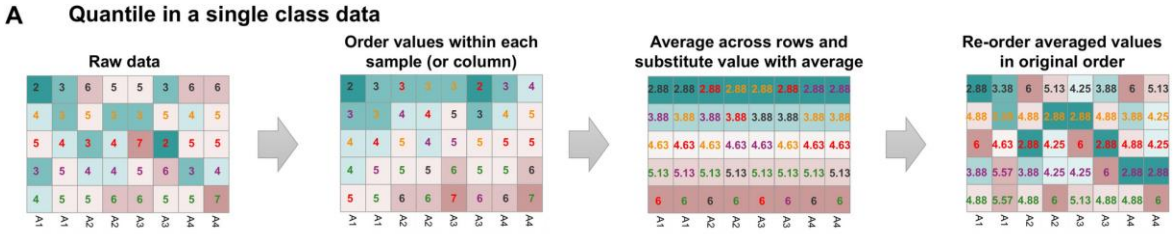
The data is then recalibrated using the following equation:

$$y_{scaled} = a y_{raw}^{b} + c$$

Where:

- a,b and c are estimated within each sample using a least square approach.
- y_{raw} is the unscaled data.
- y_{scaled} is the scaled data used for further analysis.

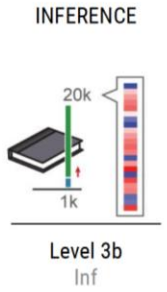
y_{scaled} is the normalized using a quantile normalization.



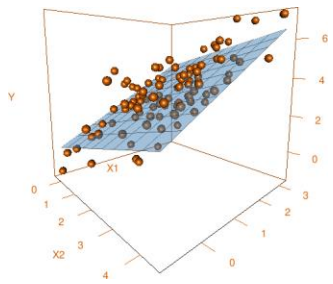
CMAP processing - Level 3 gene inference

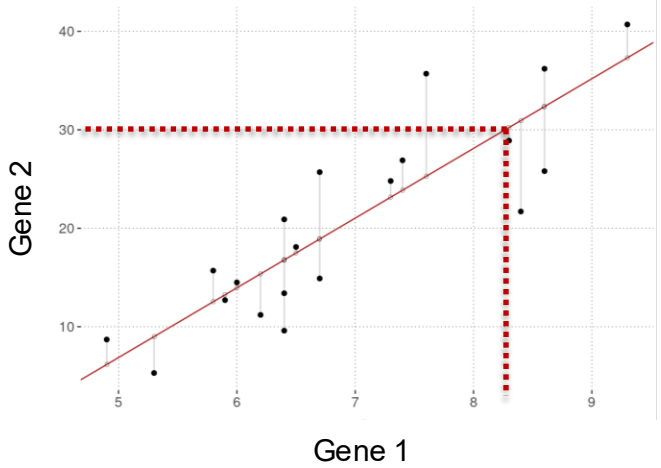
To infer the value of the missing genes, they assume that unmeasured genes can be predicted from the measured landmark genes using the following linear combination:

$$x = w_0 + \sum_{i=1}^{978} w_i y_i$$



2 response variables:

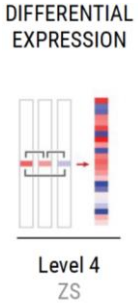




Example:
Gene 1: Measured value of 8.2
Gene 2: Predicted value of 30

Gene Symbol	Gene Title	Self-Correlation	Feature_set
ESRRA	estrogen related receptor alpha	0.89	BING
EIF3D	eukaryotic translation initiation factor 3 subunit D	0.90	BING
HAUS2	HAUS augmin like complex subunit 2	-0.38	Inferred

CMAP processing – Level4: Z-score scaling

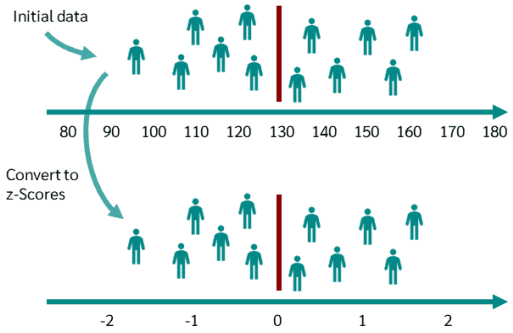


To make genes comparable, they are changed to a z-score scale using the following formula:

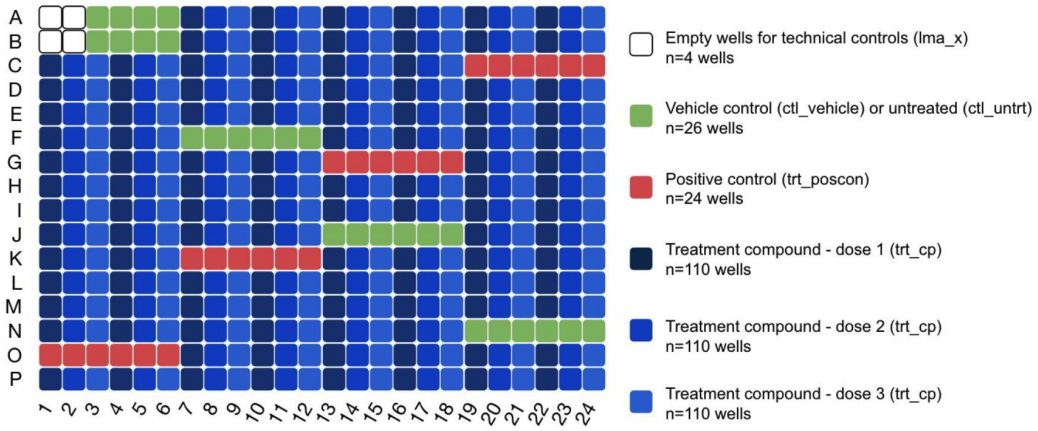
$$z_i = \frac{y_{norm} - \mu}{\sigma}$$

Where:

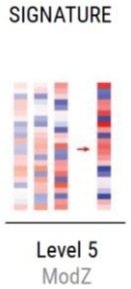
- z_i is the z-score transformed expression value.
- y_{norm} is the normalized expression (measured or inferred)
- μ is the mean normalized expression *on the plate*
- σ is the standard deviation of the normalized expression *on the plate*



Example plate layout: 3 dose

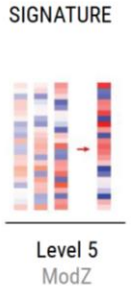


CMAP processing – Level 5: consensus signatures



- Pairwise correlation is calculated between each replicates of a signature
  - Drug
  - Time of expose
  - Dose
  - Cell line
- The consensus signature is then calculated as the linear combination of the replicates gene expression.
  - The coefficients are the sum of its correlation to the other replicates normalized to sum to 1.

CMAP processing – Level 5: consensus signatures



Example 3 genes, 3 replicates:

$$\begin{array}{c|c} & \begin{array}{ccc} \text{Rep1} & \text{Rep2} & \text{Rep3} \end{array} \\ \begin{array}{c} \text{Gene 1} \\ \text{Gene 2} \\ \text{Gene 3} \end{array} & \begin{bmatrix} 10 & 13 & 9 \\ 8 & 6 & 0 \\ 2 & 2 & 2 \end{bmatrix} \end{array}$$

Step 1: correlation matrix:

$$\begin{bmatrix} 1 & 0.9 & 0.5 \\ 0.9 & 1 & 0.82 \\ 0.5 & 0.82 & 1 \end{bmatrix}$$

Step 2: Set self-correlation to 0:

$$\begin{bmatrix} 0 & 0.9 & 0.5 \\ 0.9 & 0 & 0.82 \\ 0.5 & 0.82 & 0 \end{bmatrix}$$

Step 3: Raw weights:

$$\begin{bmatrix} 1.4 & 1.72 & 1.32 \end{bmatrix}$$

Normalizing factor: 4.44

Step 4: Normalized weights:

$$\begin{bmatrix} 0.31 & 0.39 & 0.30 \end{bmatrix}$$

Step 5: Linear combination

$$\begin{bmatrix} 0.31 * 10 + 0.39 * 13 + 0.30 * 9 \\ 0.31 * 8 + 0.39 * 6 + 0.30 * 0 \\ 0.31 * 2 + 0.39 * 2 + 0.30 * 2 \end{bmatrix} = \boxed{\begin{bmatrix} 10.87 \\ 4.82 \\ 2 \end{bmatrix}}$$

Consensus Signature

Generating disease gene expression signatures for querying CMap

1. Gene Expression Omnibus

- https://www.ncbi.nlm.nih.gov/geo/
- Public repository of microarray, next-generation sequencing, and other forms of high-throughput functional genomic data
- Allows differential gene analysis of data
  - Select significance threshold, fold change threshold, multiple correction method
- Provides R-script for analysis

Browse Content

Repository Browser

DataSets:	4348
Series:	202182
Platforms:	25116
Samples:	5887793

Series GSE11393

Query DataSets for GSE11393

Status	Public on Jun 03, 2008
Title	Monocyte gene expression profiling in familial combined hyperlipidemia and its modification by atorvastatin treatment
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Introduction: The genetic origin of familial combined hyperlipidemia (FCH) is not well understood. We used microarray profiling of peripheral blood monocytes to search novel genes and pathways involved in FCH. Methods: Fasting plasma for determination of lipid profiles, inflammatory molecules, and adipokines was obtained and peripheral blood monocytes were isolated from male FCH patients basally and after 4 weeks of atorvastatin treatment. Sex-, age- and adiposity-matched controls were also studied. Gene expression profile was analyzed using Affymetrix Human Genome U133A 2.0 GeneChip arrays. Results: Analysis of gene expression by cDNA microarrays showed that 82 genes were differentially expressed in FCH monocytes compared to controls. Atorvastatin treatment modified the expression of 87 genes. Changes in the expression of some genes, confirmed by real time RT-PCR, (CD36, leucine-rich repeats and immunoglobulin-like domains-1, tissue factor pathway inhibitor 2, myeloid cell nuclear differentiation antigen tumor necrosis factor receptor superfamily, member 25 and CD96) may be related to a proinflammatory environment in FCH monocytes, which is partially reversed by atorvastatin. Higher plasma levels of triglycerides and free fatty acids and lower levels of adiponectin in FCH patients could also trigger changes in gene expression that atorvastatin cannot modify. Conclusions: Our results demonstrate clear differences in gene expression in FCH monocytes compared with those of matched healthy controls, some of which are influenced by atorvastatin treatment. Keywords: comparative study differential gene expression
Overall design	Peripheral blood monocytes were isolated from male FCH patients basally and after 4 weeks of atorvastatin treatment. Sex-, age- and adiposity-matched

Platforms (1) GPL571 [HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array

Samples (9)
More...

GSM287664 Monocyte control rep 1
GSM287665 Monocyte control rep 2
GSM287666 Monocyte control rep 3

Relations

BioProject PRJNA106517

Analyze with GEO2R

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE11393_RAW.tar	17.6 Mb	(http)(custom)	TAR (of CEL)

GEO accession GSE11393 Set

Monocyte gene expression profiling in familial combined hyperlipidemia and its modification by atorvastatin treatment

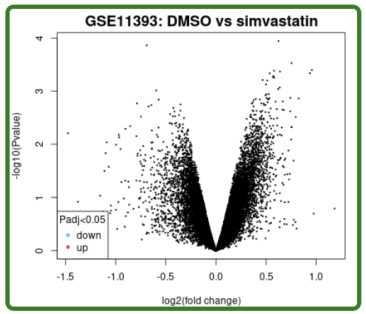
Samples Define groups Selected **6** out of **9** sam

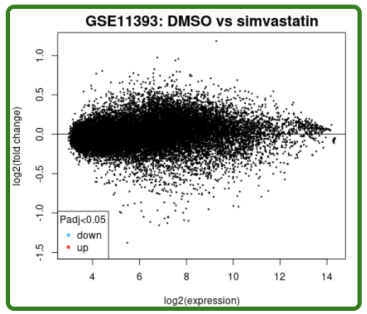
Columns

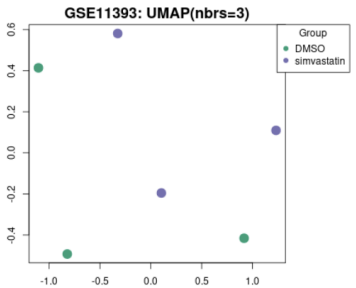
Group	Accession	Title	Source name	Characteristics
-	GSM287664	Monocyte control rep 1	Monocyte human control sample	Human healthy male s
-	GSM287665	Monocyte control rep 2	Monocyte human control sample	Human healthy male s
-	GSM287666	Monocyte control rep 3	Monocyte human control sample	Human healthy male s
DMSO	GSM287667	Monocyte FCH rep 1	Monocyte human familial combined hyperlipidemia (FCH) sample	Human male subjects
DMSO	GSM287668	Monocyte FCH rep 2	Monocyte human familial combined hyperlipidemia (FCH) sample	Human male subjects
DMSO	GSM287669	Monocyte FCH rep 3	Monocyte human familial combined hyperlipidemia (FCH) sample	Human male subjects
simvastatin	GSM287670	Monocyte ATV rep 1	Monocyte human familial combined hyperlipidemia after treatment with atorvastatin (ATV) sample	Human male subjects
simvastatin	GSM287671	Monocyte ATV rep 2	Monocyte human familial combined hyperlipidemia after treatment with atorvastatin (ATV) sample	Human male subjects
simvastatin	GSM287672	Monocyte ATV rep 3	Monocyte human familial combined hyperlipidemia after treatment with atorvastatin (ATV) sample	Human male subjects

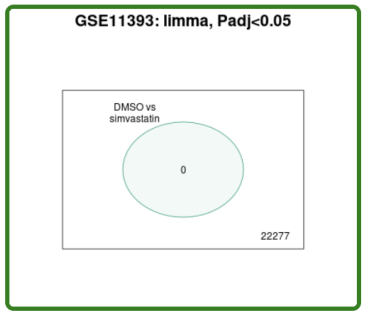
Reanalyze if you changed any options.

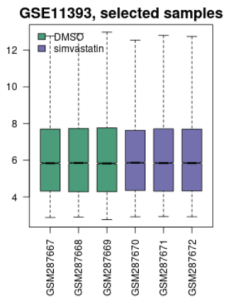
Visualization

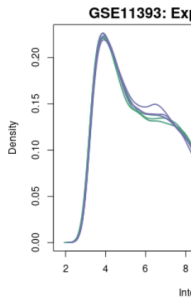












Top differentially expressed genes

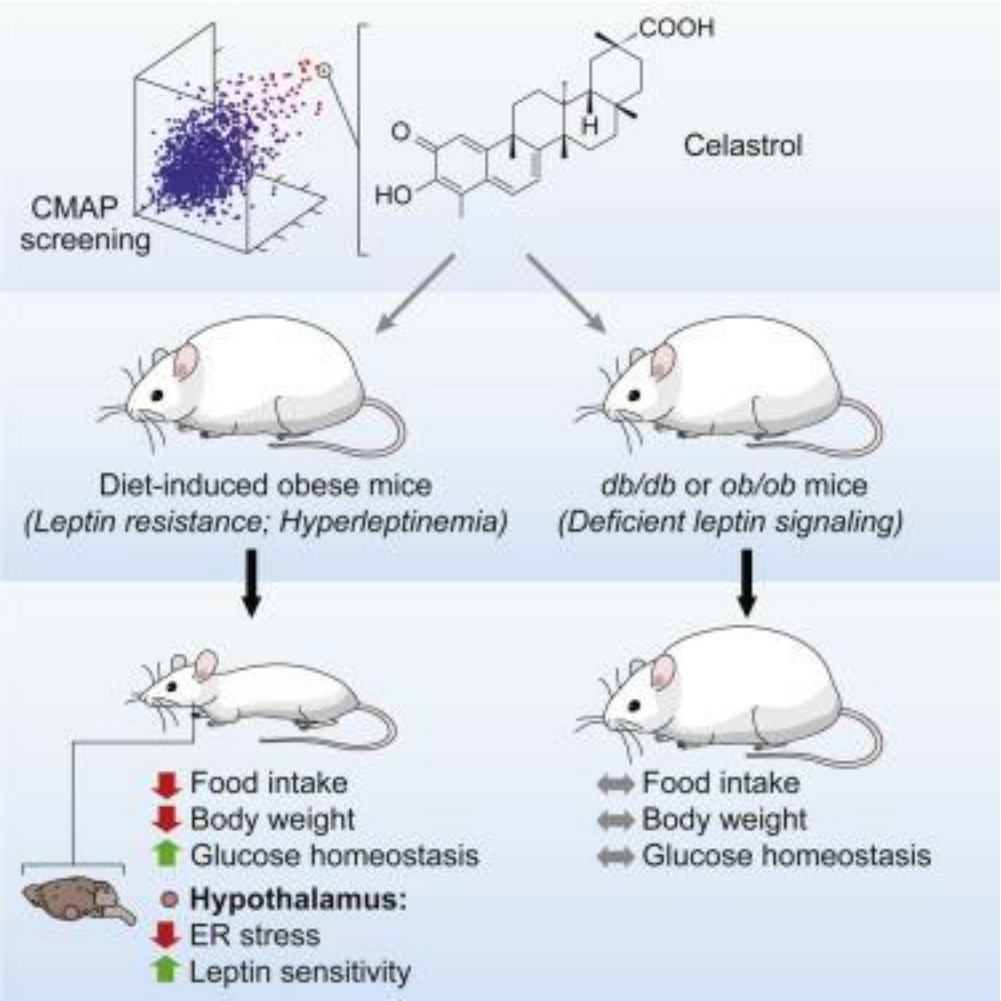
Download full table Select columns

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
214551_s_at	0.65	0.000114	7.83	-2.37	0.624	CD7	CD7 molecule
201042_at	0.65	0.000137	-7.6	-2.39	-0.69	TGM2	transglutaminase 2
204612_at	0.65	0.000297	6.71	-2.5	0.754	PKIA	protein kinase (cA..
206761_at	0.65	0.0004	6.38	-2.54	0.956	CD96	CD96 molecule
201323_at	0.65	0.000412	6.35	-2.55	0.58	MIR6733///EBNA…	microRNA 6733///…

GEO2R | Options | Profile graph | R script

```
# Version info: R 4.2.2, Biobase 2.58.0, GEOquery 2.66.0, limma 3.54.0
################################################################
#   Differential expression analysis with limma
library(GEOquery)
library(limma)
library(umap)

# load series and platform data from GEO

gset <- getGEO("GSE11393", GSEMatrix =TRUE, AnnotGPL=TRUE)
if (length(gset) > 1) idx <- grep("GPL571", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))

# group membership for all samples
gsms <- "XXX000111"
sml <- strsplit(gsms, split="")[[1]]

# filter out excluded samples (marked as "X")
sel <- which(sml != "X")
sml <- sml[sel]
gset <- gset[ ,sel]

# log2 transformation
```

2. Animal experiments



Liu et al 2015 Cell

- **Aim: Identify potential drug candidates for reducing ER stress and obesity**
- Endoplasmic reticulum (ER) stress tightly linked to obesity.

- **Gene signature** - Induced ER stress **in mice** and measured transcriptional response in the livers of these mice.
- **Query CMAP -** Celastrol (extract from Thunder god vine plant)
- **Validation in mice studies** - Celastrol increases leptin sensitivity to suppress food intake in mice.

- ERX Pharmaceuticals (founded in 2014) currently testing leptin sensitisers in Phase I clinical trials.

3. Human gene expression studies

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Gene expression differences in human cases vs controls

dataset1	Trait	g1	g2	g3
ind1	1			
ind2	0			
ind3	1			

Measure gene expression levels in cases (1) and controls (0)

		se	pval
g1			
g2			
g3			

Association between gene expression and disease status
- May be hard to get DEGs in different tissues

4a. From individual-level GWAS data using PrediXcan

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

dataset1	Trait	SNP1	SNP2	SNP3
ind1	1			
ind2	0			
ind3	1			

dataset 2
eQTL data,
training data for
prediction model

	Trait	ĝ1	ĝ2	ĝ3
ind1	1			
ind2	0			
ind3	1			

Genetically-predicted gene expression

	b	se	pval
g1			
g2			
g3			

Gene expression associated with trait

4b. From GWAS summary data using S-PrediXcan

- Requires 3 datasets
  a) GWAS data for phenotype of interest
  b) Expression QTL training set e.g. GTEx
  c) Population reference (e.g. 1000 Genomes)

3b. Gene expression signature prediction from GWAS summary data using S-PrediXcan

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

w_{gi} weight given to each SNP for predicting expression level of *g* Precomputed weights derived from a reference eQTL dataset e.g. GTEx

Association of gene *g* expression with phenotype:

+ve z-score: increased gene expression associated with increased phenotype or risk of disease

Variance of SNP_i (from reference genotype data) /

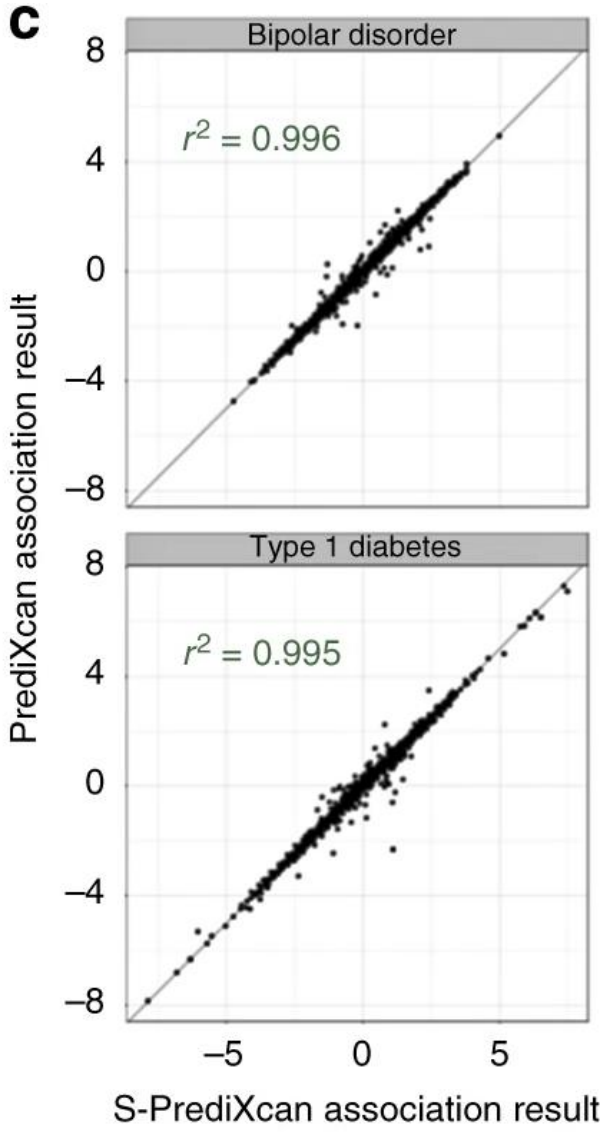
$$z_g \approx \sum_{i=1}^{k} w_{gi} \frac{\hat{\sigma}_i}{\hat{\sigma}_g} \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Effect size (and SE) of SNP_i on phenotype obtained from GWAS

Assuming set of $SNP_{1..k}$ contribute to the expression of gene *g*

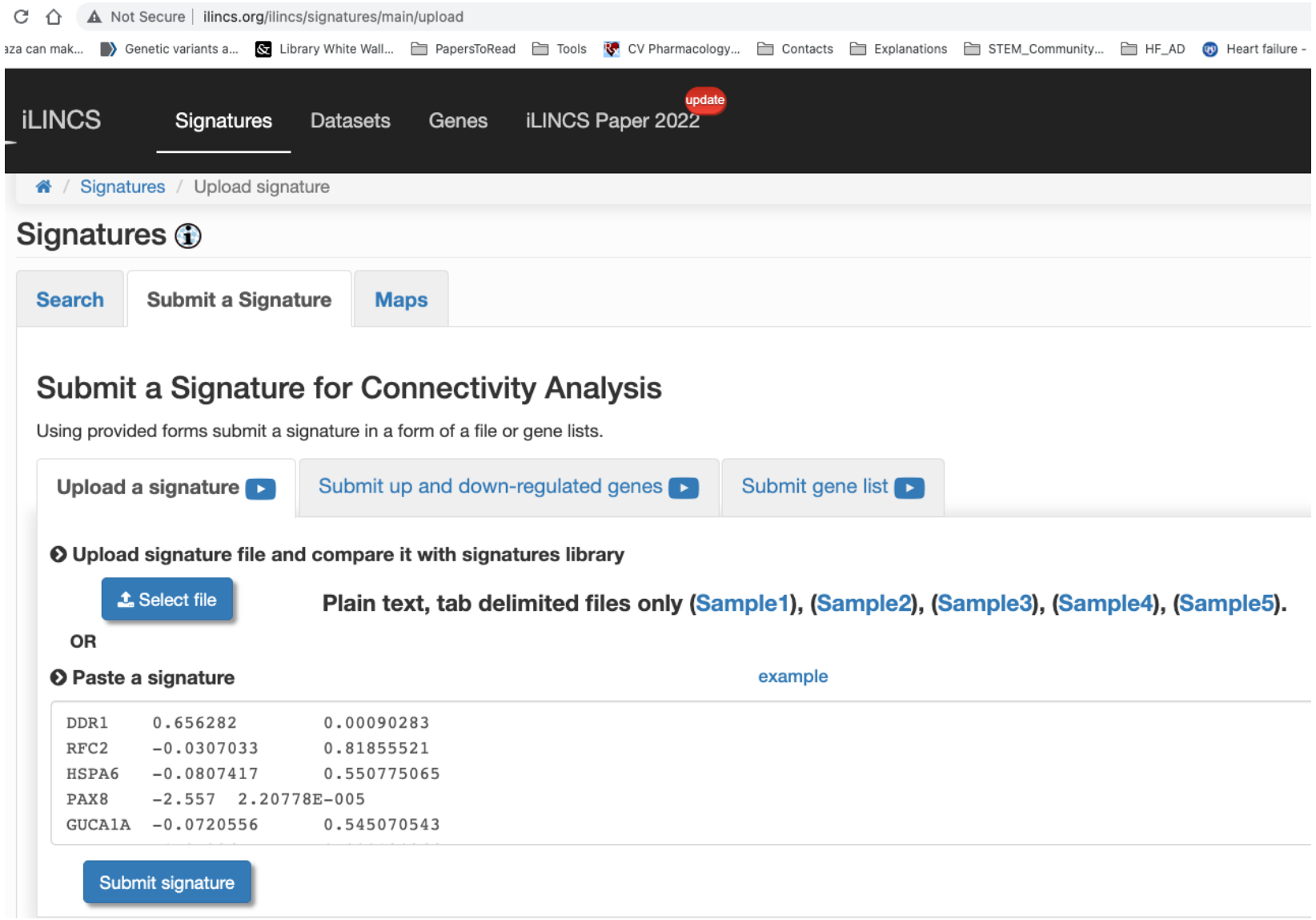
Estimated variance of predicted expression of gene *g*

Comparison of PrediXcan and S-PrediXcan gene z-scores

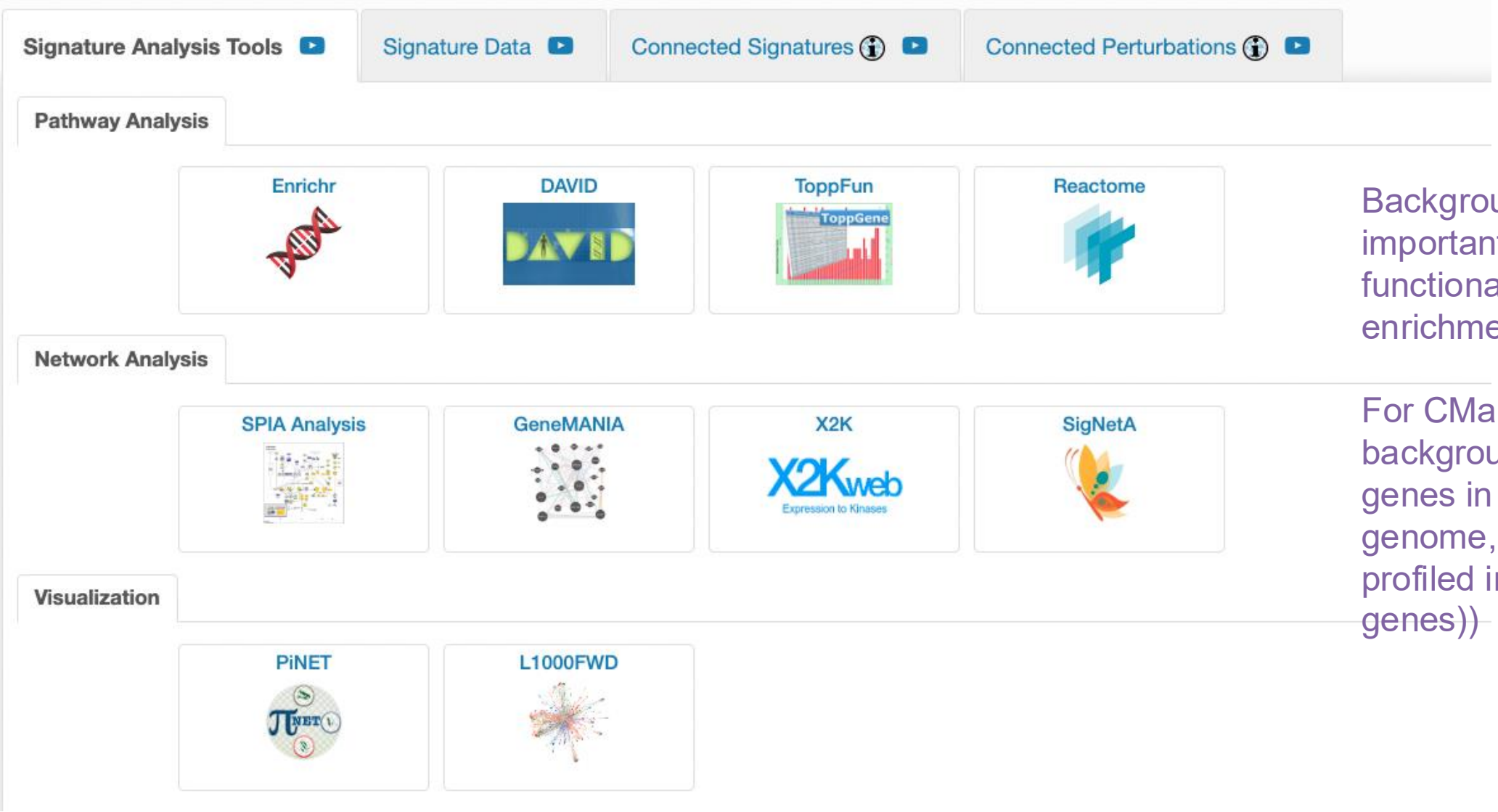


Wellcome Trust Case Control Consortium (WTCCC) data
https://www.nature.com/articles/s41467-018-03621-1

Querying CMap data with iLINCs
http://www.ilincs.org/ilincs/



https://www.ilincs.org/ilincs/signatures/main/upload



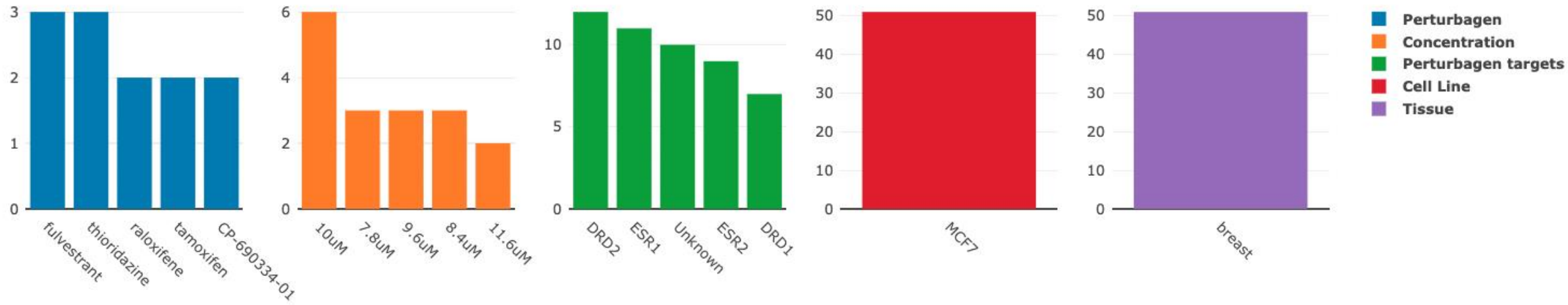
Background gene list very important when doing functional/pathway enrichment analysis.

For CMap data, background list is not all genes in the human genome, rather all genes profiled in CMap (~12,000 genes))

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

51 of Connectivity Map signatures

Analyze | / Selection | My list | Download | Clear filters | Stats | Top 5 All Signatures



Signature Id	Perturbagen	Perturbagen targets	Concentration	Cell Line	Tissue	Concordance	pValue	nGenes
CMAP_127	raloxifene	ESR1 | ESR2	7.8uM	MCF7	breast	1.000	0	100
CMAP_128	raloxifene	ESR1 | ESR2	0.1uM,7.8uM	MCF7	breast	0.943	1.5e-48	100
CMAP_88	tamoxifen	ESR1 | ESR2	7uM	MCF7	breast	0.922	4.5e-42	100
CMAP_864	corticosterone	HSD11B1 | NR3C2	11.6uM	MCF7	breast	0.917	6.2e-41	100
CMAP_742	clomifene	ESR1	6.6uM	MCF7	breast	0.904	5.1e-38	100

5 | 25 | 50 | 100

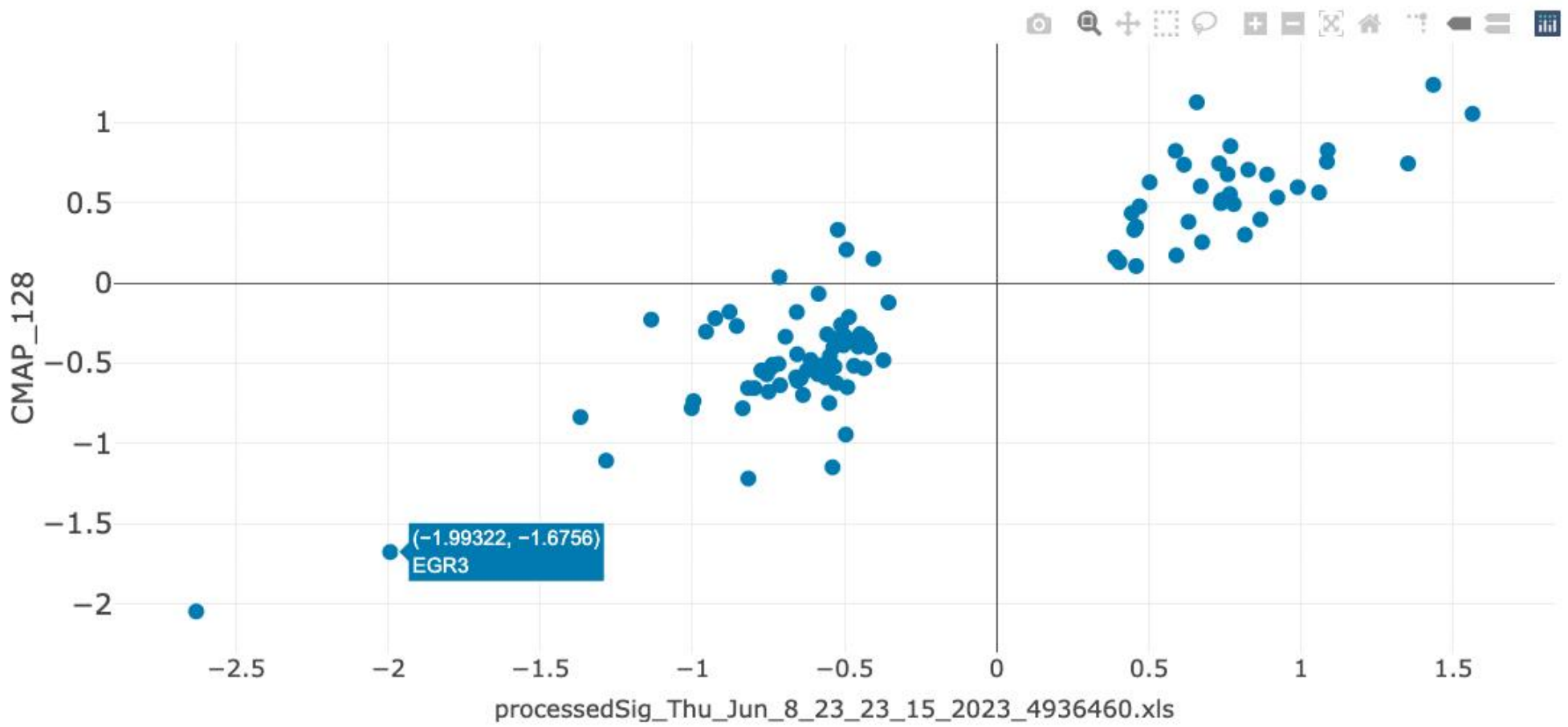
First « 1 2 3 4 5 » Last

Page 1 of 11, of 51 entries

Correlation plot

Weighted Pearson correlation: **0.943**

Pearson correlation: **0.913**



Use selected genes

Cancel

nature neuroscience

Article | Published: 14 August 2017

Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry

Hon-Cheong So, Carlos Kwan-Long Chau, Wan-To Chiu, Kin-Sang Ho, Cho-Pong Lo, Stephanie Ho-Yue Yim & Pak-Chung Sham

Nature Neuroscience **20**, 1342–1349 (2017) | Cite this article

So et al

- Brain-based eQTL models to impute disease gene expression
- Spearman, Pearson correlation, and KS-test to determine similarity between between disease and drug signature
- As there are no consensus methods to define *K*, they set different values of *K* (50, 100, 250, 500) and averaged the results for each method.
- No selection on cell type

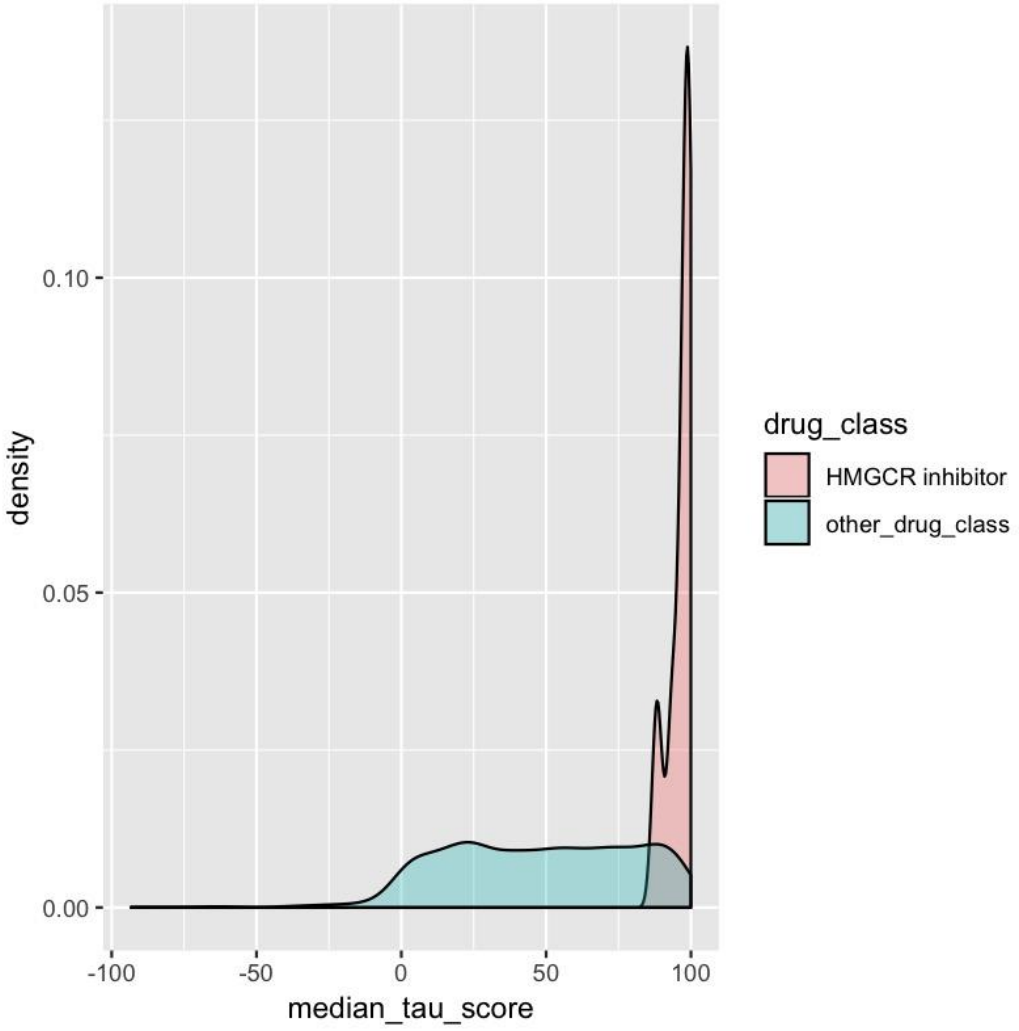
Important considerations for GWAS to drug candidate pipeline using gene expression signature mapping

- How many genes to include in your query gene list?
- what eQTL model should you use generating a disease gene signature?
  - Single-tissue model (disease tissue-specificity) vs multi-tissue model (greater power to predict gene expression)
- What cell lines should you query in iLINCs?
  - Most relevant cell type vs summary across cell lines
- How does the power of the GWAS affect identification of drug targets?

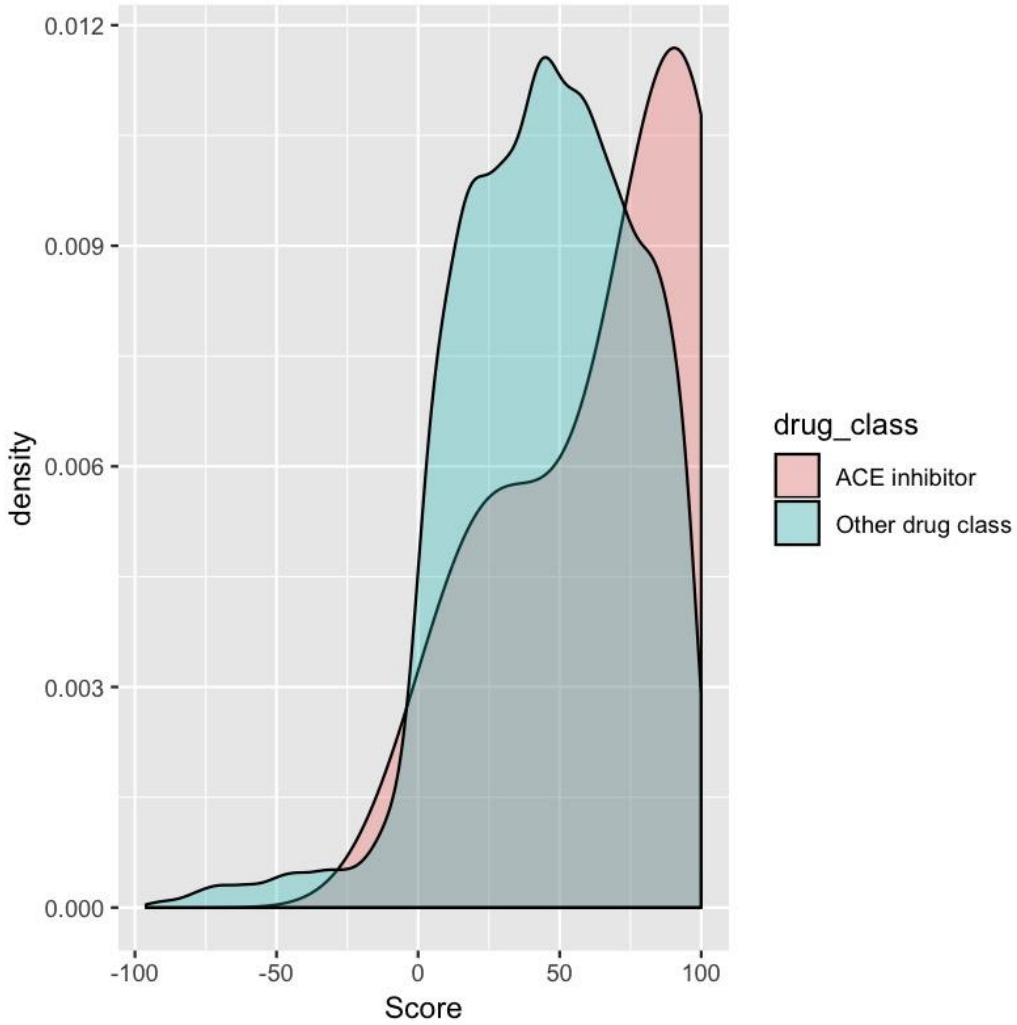
Other considerations

- iLINCS is a useful resource but requires careful manual curation
  - Check connectivity between gene knockdown/overexpression and drug
  - Check specificity of the gene signature
  - Check connectivity between compounds with same MoA
  - Compare genetic vs chemical perturbation signatures (e.g. statin vs HMGCR KO)
  - Check connectivity across cell lines
  - Drugs may not be in an active form. Need to check this from other sources e.g. DrugBank
  - Check if target is expressed in cell line before interpreting results (human protein atlas)

Connectivity of rosuvastatin with other HMGCR-inhibitors and all other compounds



Connectivity of enalapril with other ACE inhibitors and all other compounds



Statins and depression

BMJ. 1992 Feb 15; 304(6824): 431–434.
doi: 10.1136/bmj.304.6824.431

PMCID: PMC188
PMID: 153

Should there be a moratorium on the use of cholesterol lowering drugs?

G Davey Smith and J Pekkanen

DEPRESSION

A Statin Island of Woe: Are Depression and Cholesterol-lowering Connected?

Are statin drugs bringing you down?

2012 Psychology Today

Kat Lay, Health Correspondent

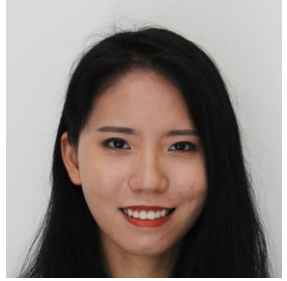
Thursday January 10 2019, 12.01am GMT, The Times

2019 The Times

Do statins have any effects on depression?

Sonia Shah

Chenwen Hu



Jiayue Clara Jiang

1. Connectivity map (CMap) analysis

2. Mendelian randomisation analysis

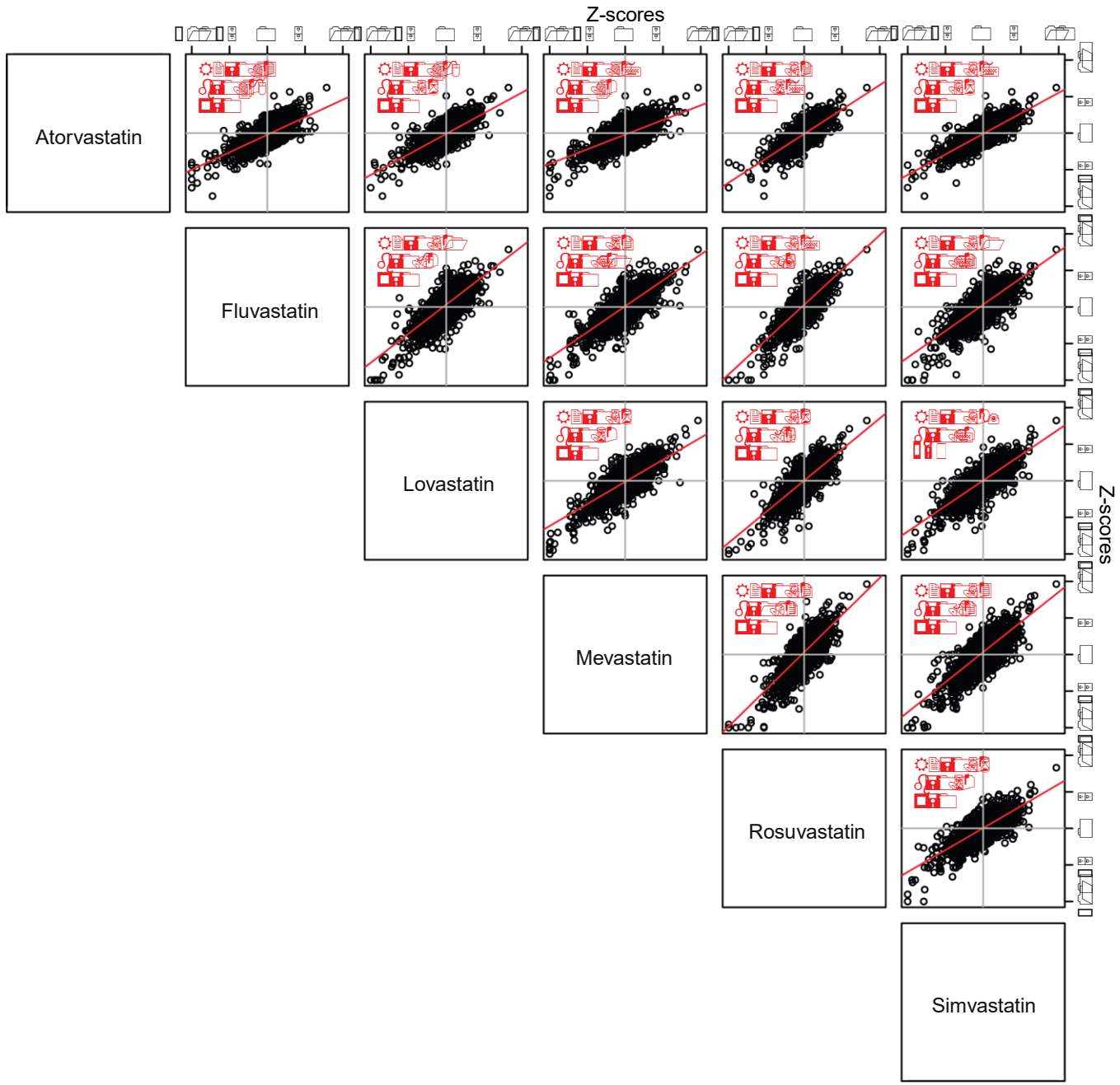
Translational Psychiatry

Article | Open access | Published: 04 April 2023

Investigating the potential anti-depressive mechanisms of statins: a transcriptomic and Mendelian randomization analysis

Jiayue-Clara Jiang, Chenwen Hu, Andrew M. McIntosh & Sonia Shah

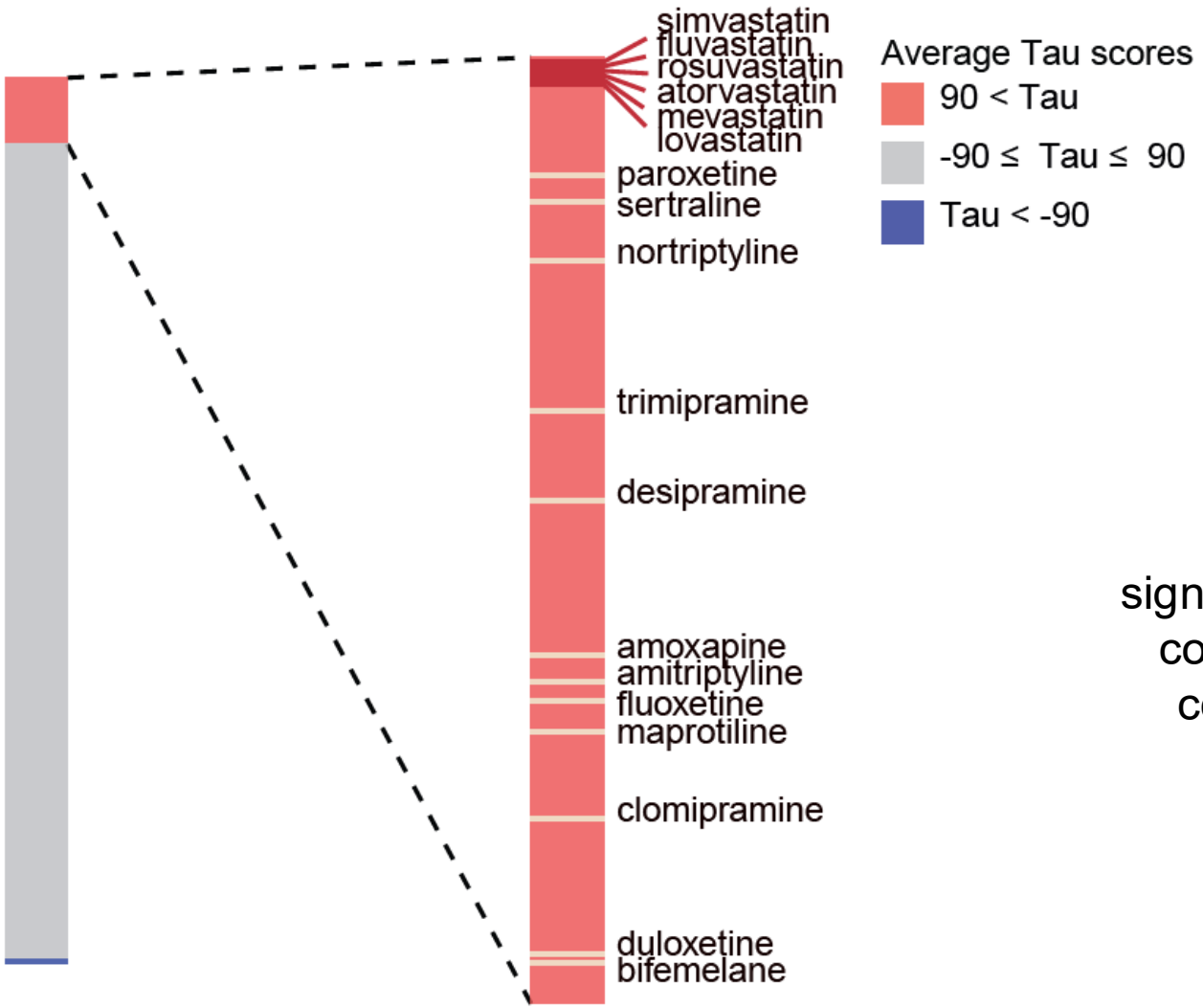
Translational Psychiatry **13**, Article number: 110 (2023) | Cite this article



Within-statin gene expression signature correlation

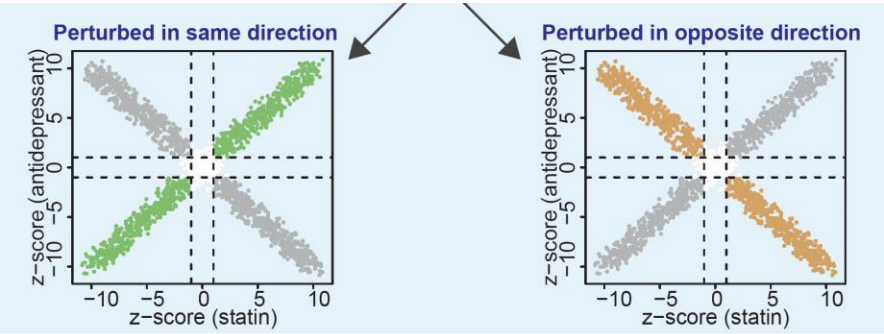
statin-statin tau scores > 95

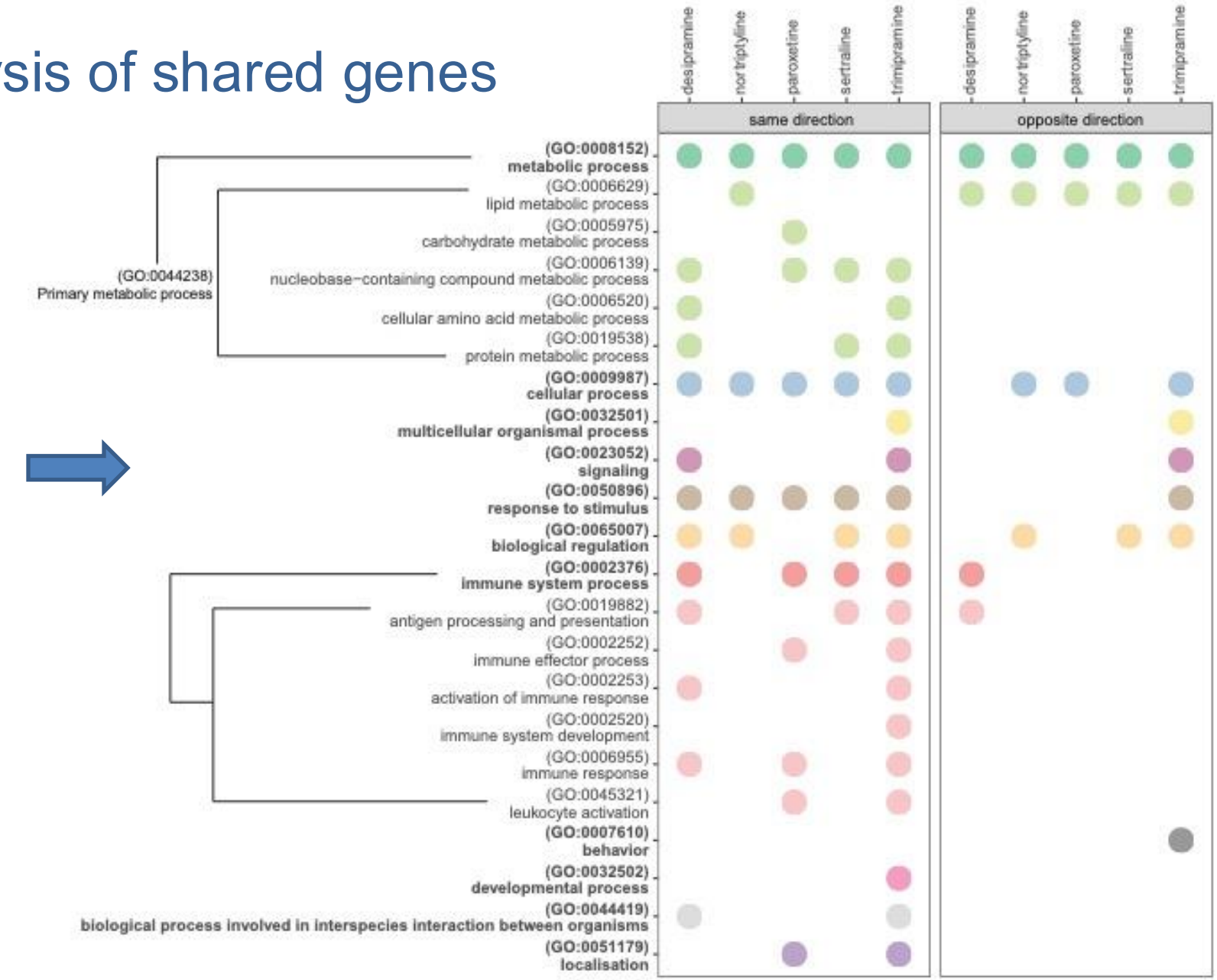
Enrichment of antidepressants amongst statin-connected compounds



Antidepressants are significantly enriched amongst compounds that have high connectivity (tau > 90) to statins

Gene-set enrichment analysis of shared genes

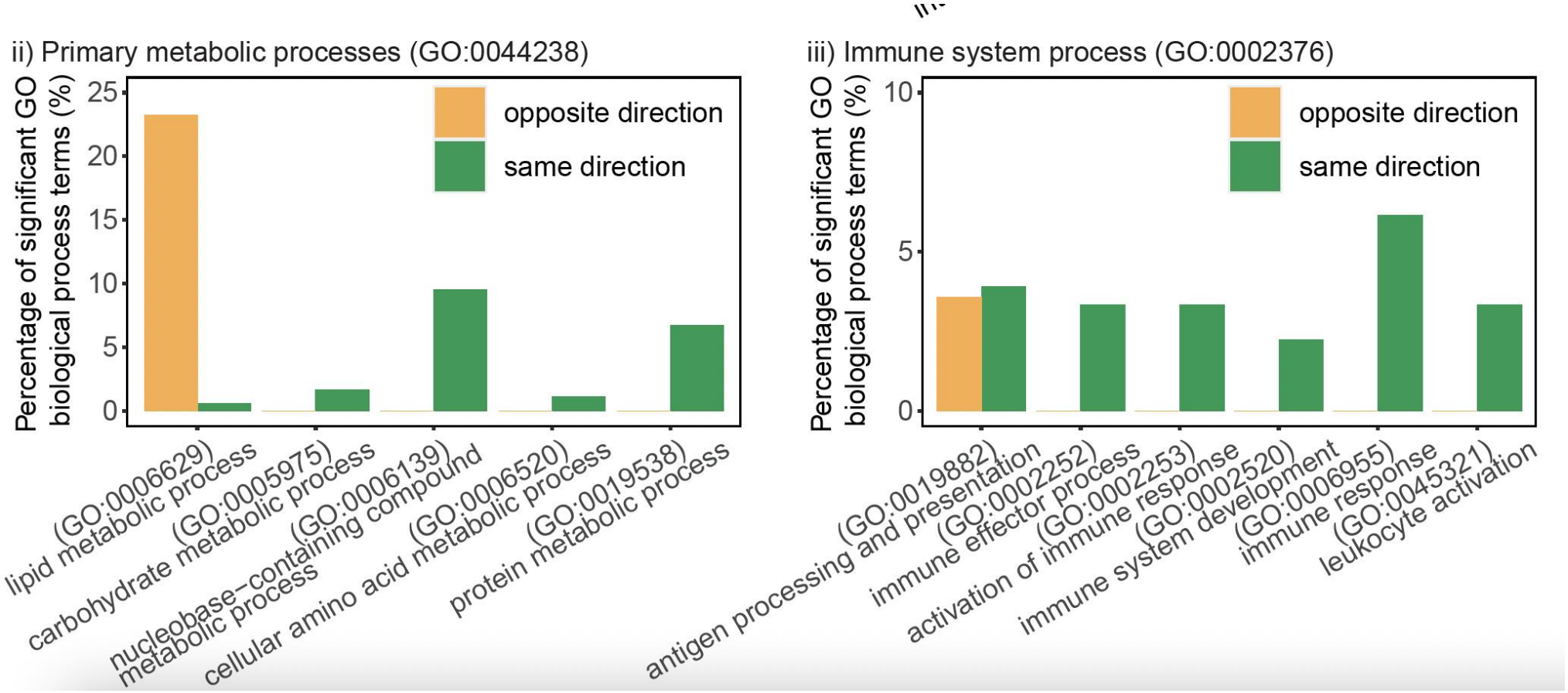




Functional terms enriched amongst genes perturbed in the same and opposite direction by both statins and antidepressants

Sonia Shah

Gene-set enrichment analysis of shared genes

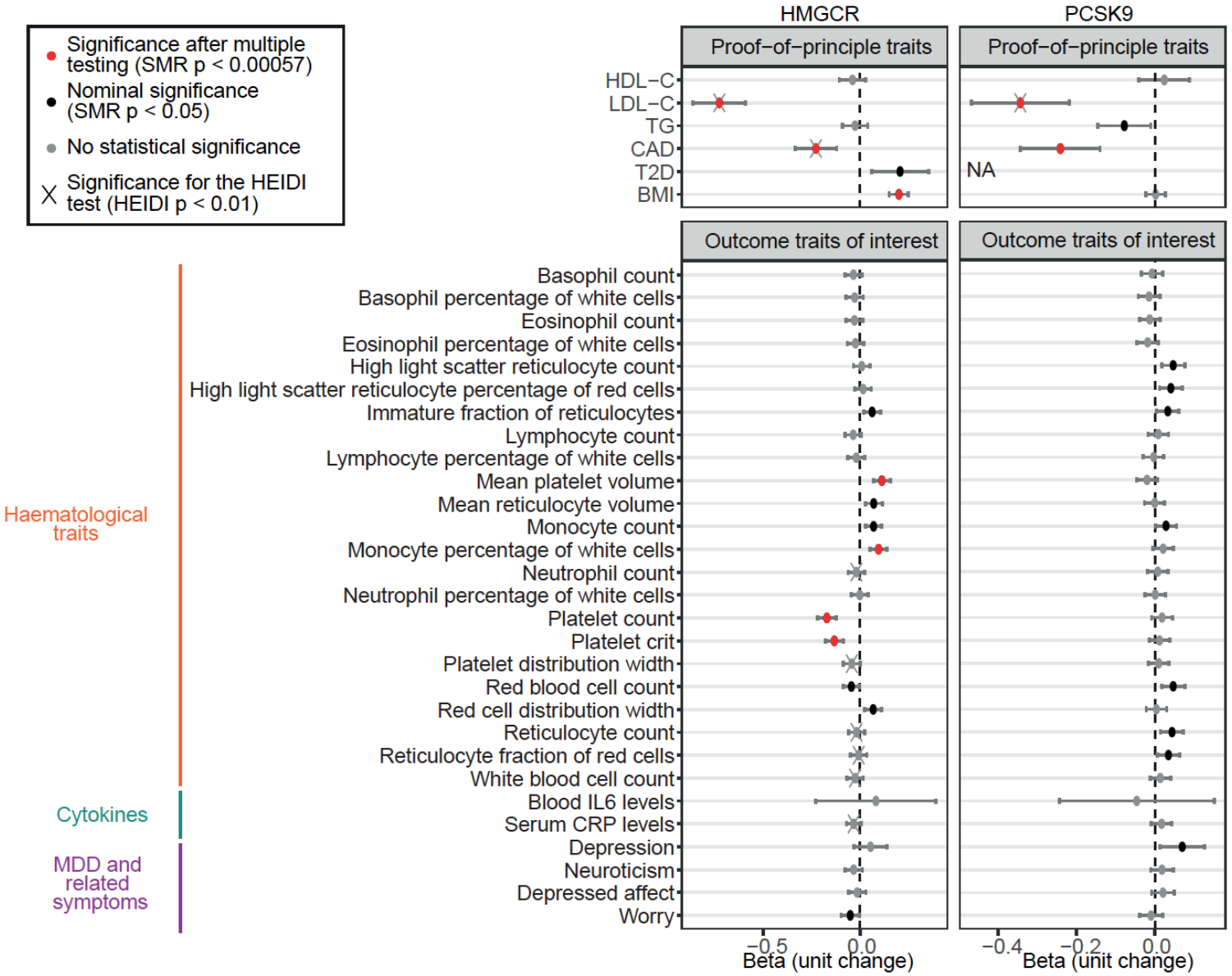


MR analysis HMGCR/ITGAL/HDAC2 gene expression

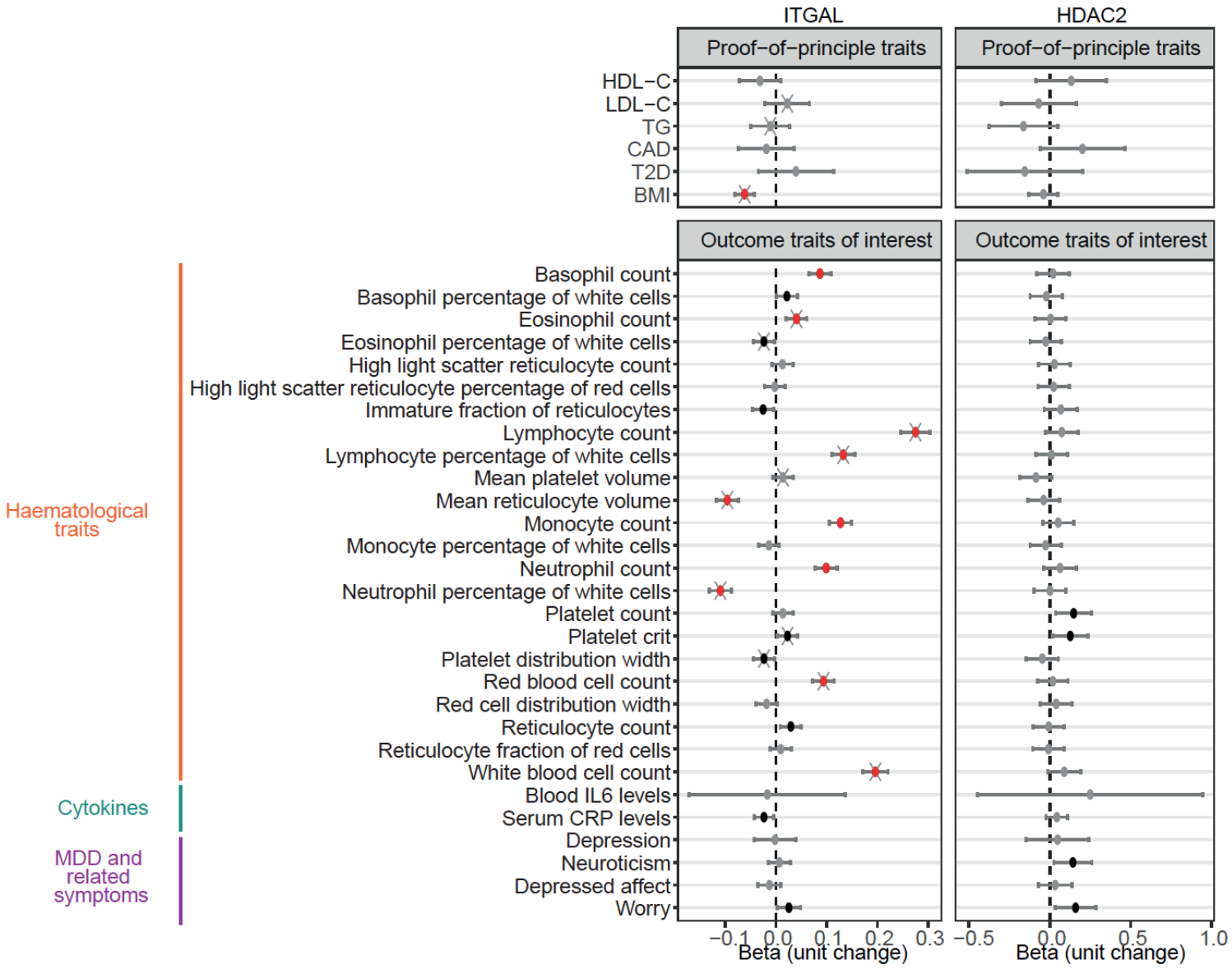
Statin	Gene targets
simvastatin	HMGCR, ITGAL, HDAC2
atorvastatin	HMGCR, DPP4, AHR, HDAC2, NR1I3
rosuvastatin	HMGCR, ITGAL
lovastatin	HMGCR, ITGAL, HDAC2
fluvastatin	HMGCR, HDAC2
mevastatin	HMGCR
cerivastatin	HMGCR
pitavastatin	HMGCR, ITGAL
pravastatin	HMGCR, HDAC2

Drug bank database

Lower HMGCR expression associated with platelet measures



Lower ITGAL expression associated with white blood cell counts



Your feedback would be greatly appreciated so we can improve on our content next year

- Things you enjoyed and why
- Things you didn’t enjoy and why
- Suggestions on how we could improve
- Anything you were hoping we would cover but we didn’t?