# Connectivity Map for identifying drug candidates

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

# Lecture Overview

The University of Queensland
Australia

- Gene signature matching
- A database of compound gene signatures - CMap
- Generating a disease gene signature
- Querying CMap

# GWAS to medicine

Genetic variants → Disease genes → Drug candidates

- 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
  - Screening failed drugs against new indications - benefit–risk profile may vary depending on the unmet medical need
  - But…
    - Drugs with unknown mechanism of action (MoA) will be missed with this approach
    - Important disease biology may be lost under stringent p-value thresholds

The University of Queensland
Australia

# Gene signature matching

# Gene expression signature matching

**Assumption:** compounds that have the same MoA induce similar gene expression responses. Can be useful for:

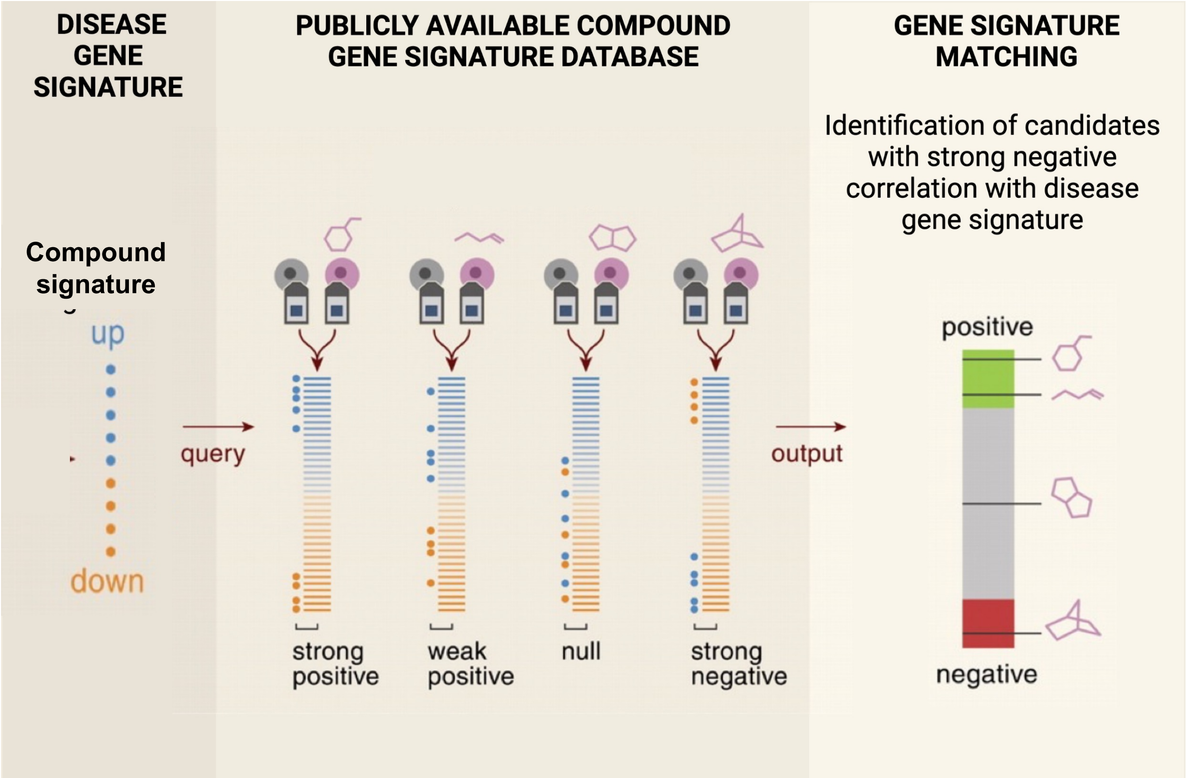
1. **Understanding MoA of a compound**

2. **Drug repurposing potential**

3. **Identifying new drug candidates**
   - Compounds that reverse gene expression changes associated with disease
   - Does not require knowledge of the drug's MoA

4. **Identifying potential drug side-effects**

Requires gene expression signatures for drugs and diseases

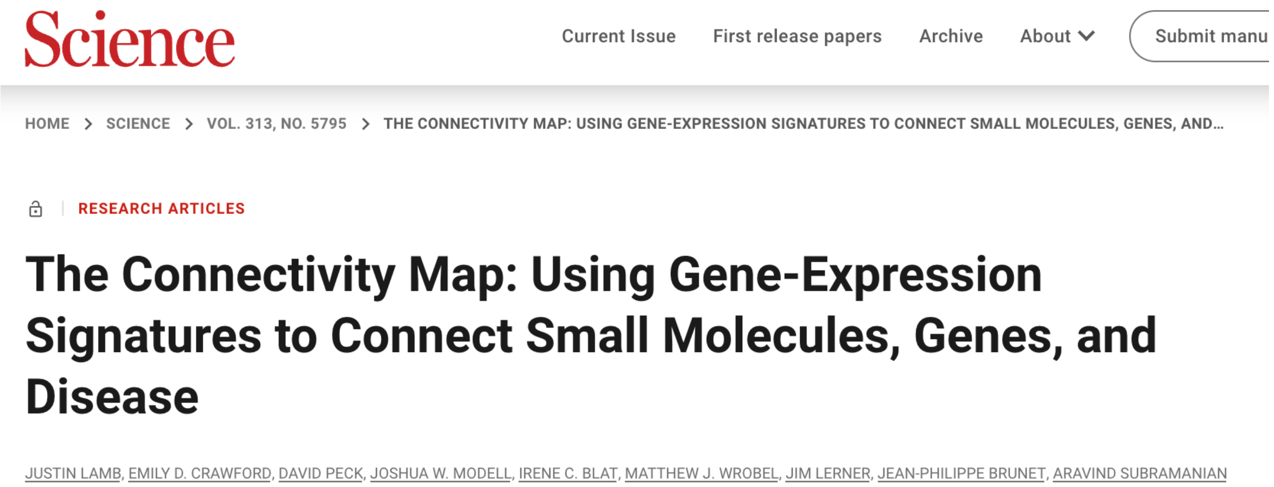


THE UNIVERSITY OF QUEENSLAND AUSTRALIA

# Connectivity Map (CMap)

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

- >1 million gene expression profiles
- ~50 different cell lines (only 4 are non-cancer cell lines)
- ~20,000 compounds (chemical perturbation)
- ~20,000 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 2013



- 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

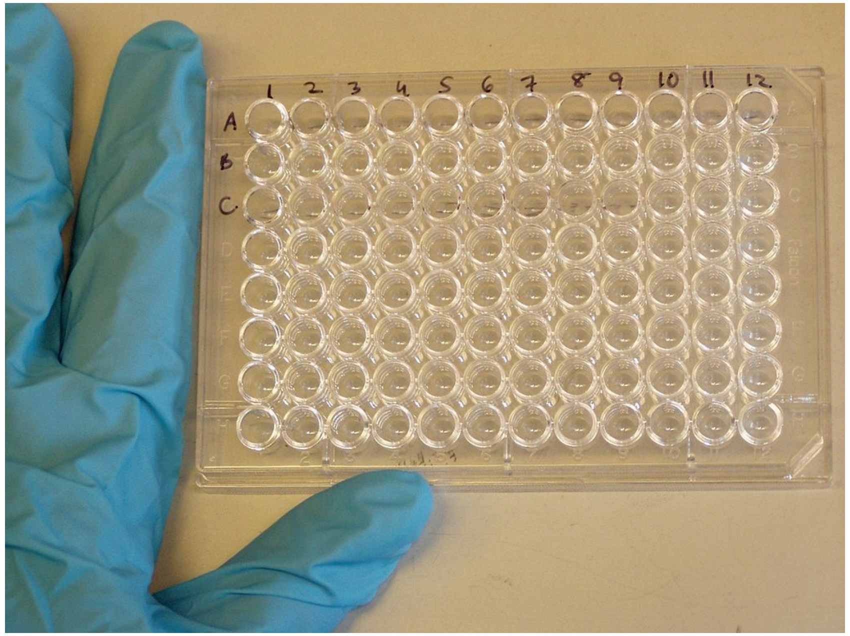
164 distinct small-molecule perturbagens, selected to represent a broad range of activities:

- 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

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

- Stably grown over long periods of time
- breast cancer epithelial cell line MCF7
  - extensively molecularly characterised,
  - used as a reference cell line
  - amenable to culture in 96-well plates
- prostate cancer epithelial cell line PC3
- nonepithelial lines HL60 (leukemia) and SKMEL5 (melanoma)
- Context-dependent gene signatures



# 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

- 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

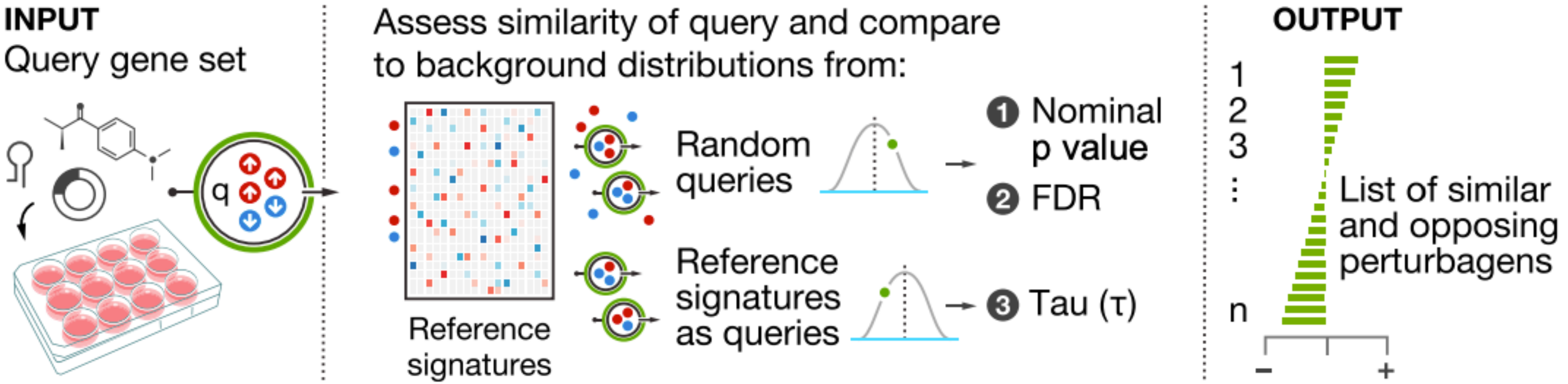
- Replicates

- 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

# Connectivity score




- Used non-parametric, rank-based pattern-matching strategy based on the Kolmogorov-Smirnov statistic (GSEA).
- Tau score - fraction of reference gene sets with a greater similarity to the perturbagen than the current query.

# Example results – HDAC inhibitors

- HDACs – remove acetyl groups on histones and regulate gene expression
- Determine if a query signature can recover compounds from the same class (same MoA).
- Query derived from 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
    - Off-target effects

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 |

# Example - Estrogens

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

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

The University Of Queensland

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

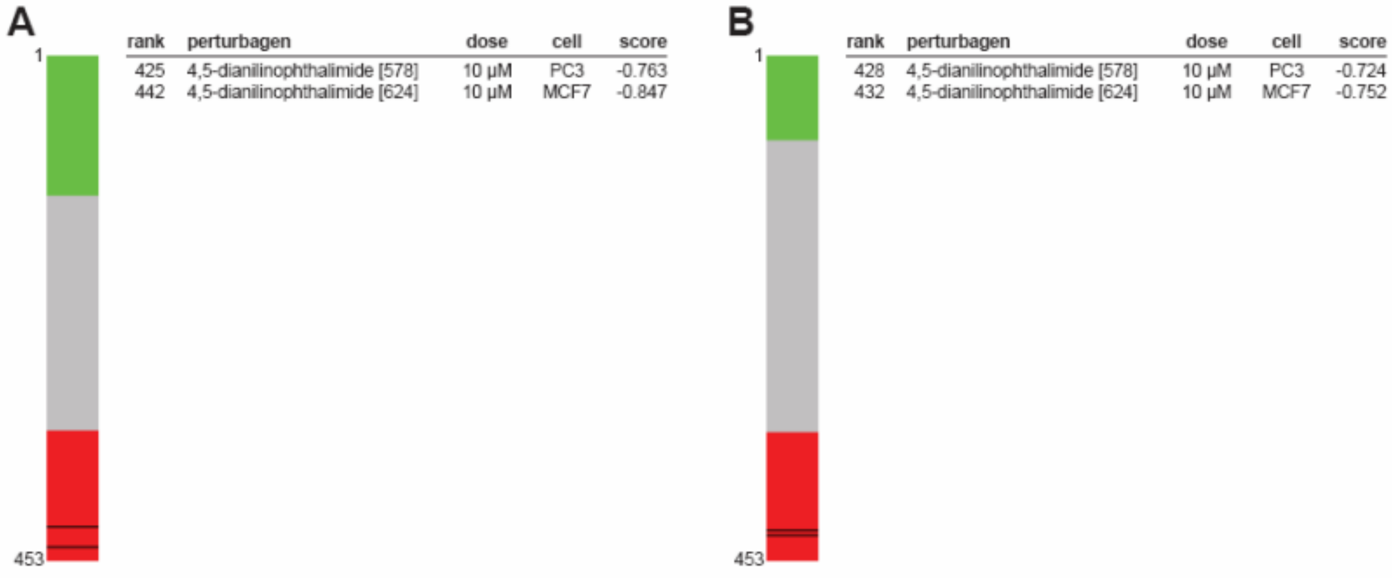

| rank | perturbagen | dose | cell | score |
|---|---|---|---|---|
| 171 | fulvestrant [704] | 1 µM | PC3 | 0 |
| 261 | fulvestrant [523] | 1 µM | ssMCF7 | 0 |
| 447 | fulvestrant [367] | 1 µM | MCF7 | -0.749 |
| 450 | fulvestrant [310] | 10 nM | MCF7 | -0.843 |
| 451 | fulvestrant [985] | 1 µM | MCF7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

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

# Connections with Disease States

- Query – DEGs from a rat model of diet-induced obesity
- Several differences in exp design: Rat vs human, exposure duration – 65 days vs 6 hrs, adipose tissue vs cell lines

**Fig. S4. PPARγ Agonists are Connected with Diet-induced Obesity in Rats.** Barview (as Fig. 2) showing all instances of troglitazone (*n*=2), rosiglitazone (*n*=1), indometacin (*n*=1) and 15-delta prostaglandin J2 (*n*=1) in PC3 cells. Unabridged results from this query are provided as Result S8.



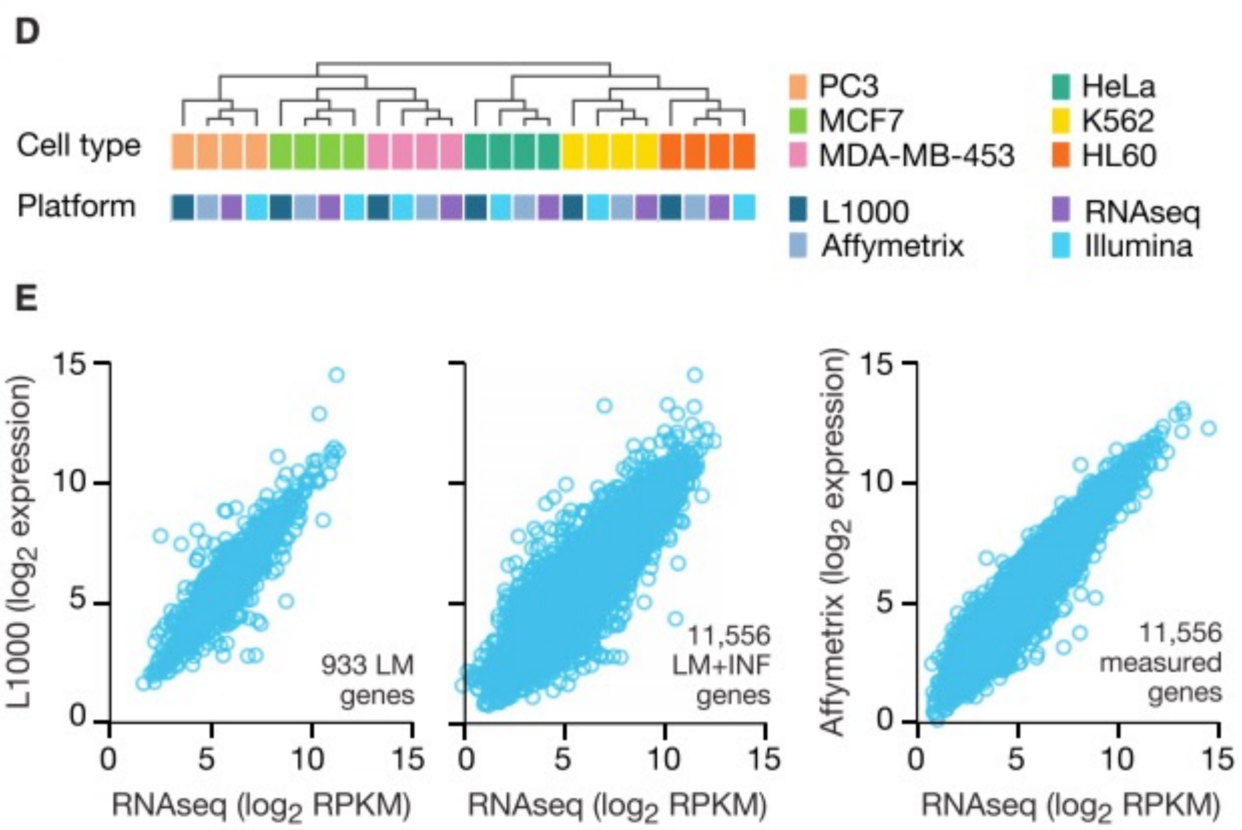
# Findings from CMap pilot study

- Genomic signatures can identify drugs with common MoA
- Discover unknown MoA e.g. HDAC activity of valproic acid (initially developed as an antiseizure drug)
- Identify potential new therapeutics using a disease-associated gene query signatures
- Signatures are often conserved across diverse cell types and settings
  - Drug target needs to be expressed in that cell line e.g estrogen receptor
- Not highly sensitive to the precise concentration of drug

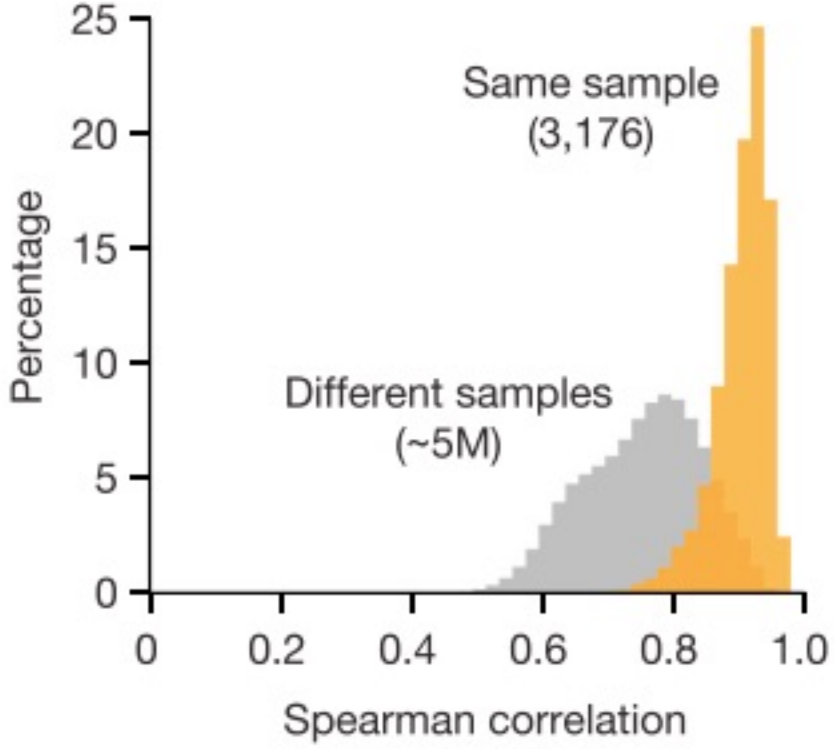
# 2nd Generation CMAP - LINCS1000

- **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.
- 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 N of informative transcripts (“landmark” transcripts)
  - Cost vs information captured
  - 1000 landmarks enough to capture 82% of full transcriptome
  - No substantial enrichment of particular protein class or developmental lineage in landmark list.

# Comparison of L1000 with RNAseq

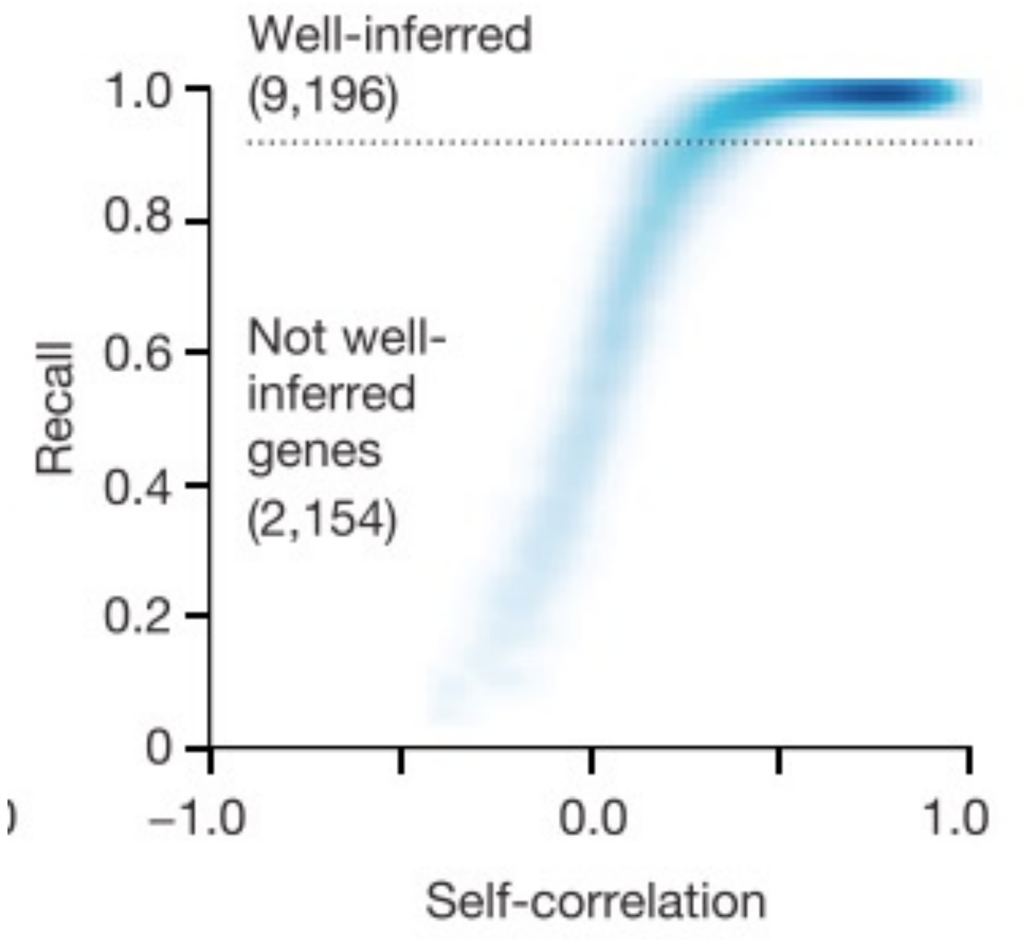



RNAseq vs L1000 in ~3K samples




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

# Imputation of GTEx data




~1000 landmark genes
~9200 well-inferred genes
~2000 (not well-)inferred genes

Only landmark and well-inferred genes used in analyses.

# CMap-L1000v1



- 19,811 compounds profiled in triplicate (at 6 and/or 24 hrs)
- Genetic perturbation (KD or overexpression) of 5075 genes measured after 96 hrs (triplicates)
- 77 cell lines
- 470K gene signatures from ~42K perturbagens – 1000-fold increase of CMap pilot dataset.
- All data (at multiple processing levels) available in GEO (GSE92742)
- Web-based tool to query database https://clue.io

# Generating disease gene expression signatures for querying CMap

# 1. Your own experiments

- Gene expression differences in cases vs controls

# 2. 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

# 3a. Gene expression signature prediction from individual-level GWAS data using PrediXcan

The University of Queensland
Australia

- A gene-level association approach that tests the mediating effects of gene expression levels on phenotypes.
- 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)

| | Trait | g1 | g2 | g3 |
|---|---|---|---|---|
| ind1 | | | | |
| ind2 | | | | |
| ind3 | | | | |

# 3a. Gene expression signature prediction from individual-level GWAS data using PrediXcan

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

| dataset1 | Trait | g1 | g2 | g3 |
|---|---|---|---|---|
| ind1 | | | | |
| ind2 | | | | |
| ind3 | | | | |

| | b | se | pval |
|---|---|---|---|
| g1 | | | |
| g2 | | | |
| g3 | | | |

Gene expression associated with trait

dataset 2
eQTL data,
training data for
prediction model

| dataset1 | Trait | SNP1 | SNP2 | SNP3 |
|---|---|---|---|---|
| ind1 | | | | |
| ind2 | | | | |
| ind3 | | | | |

| | Trait | ĝ1 | ĝ2 | ĝ3 |
|---|---|---|---|---|
| ind1 | | | | |
| ind2 | | | | |
| ind3 | | | | |

Genetically-predicted gene expression

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

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

Gene expression change associated with phenotype: z-score for gene *g*

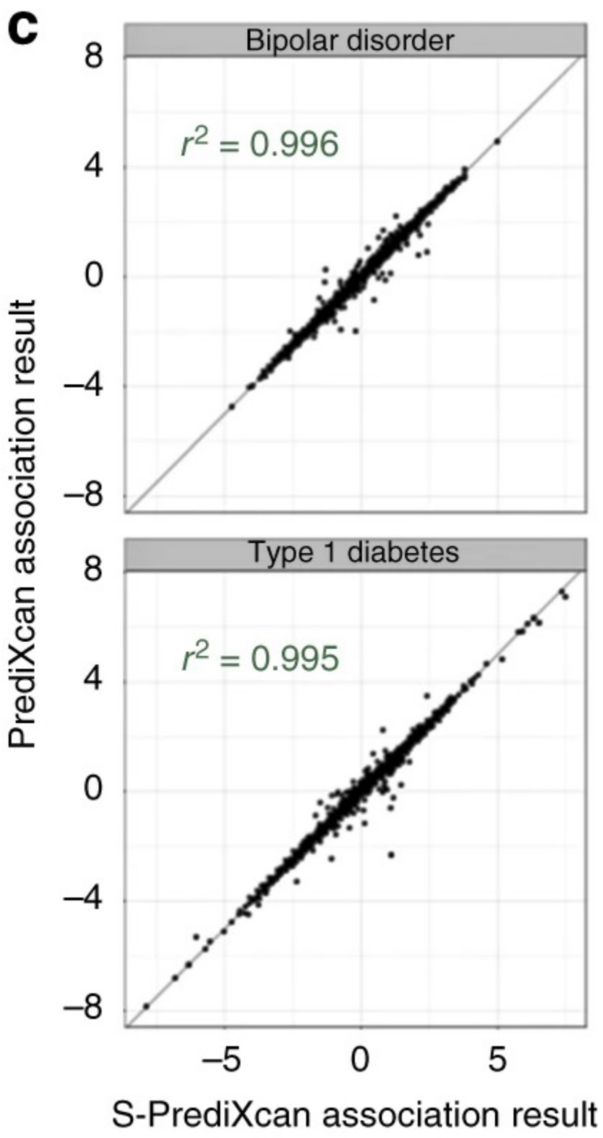
$$z_g \simeq \sum_{i=1}^{k} w_{gi} \frac{\hat{\sigma}_i}{\hat{\sigma}_g} \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Summary z-statistic of $\text{SNP}_i$ for the disease trait obtained from GWAS

Assuming set of $\text{SNP}_{1..k}$ contribute to the expression of gene *g*

Variance of $\text{SNP}_i$ and gene *g* estimated from reference genotype

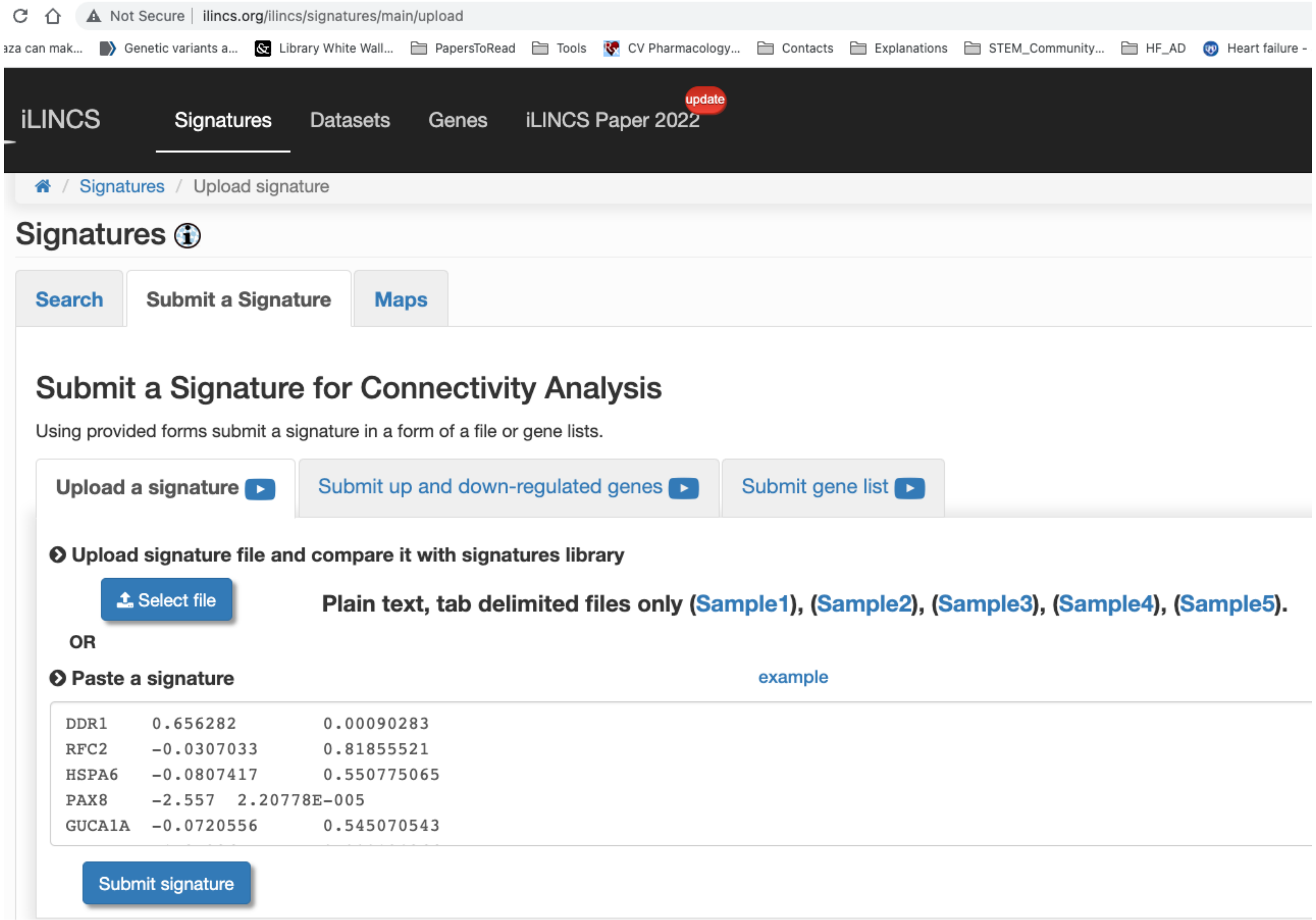
# Comparison of PrediXcan and S-PrediXcan gene z-scores

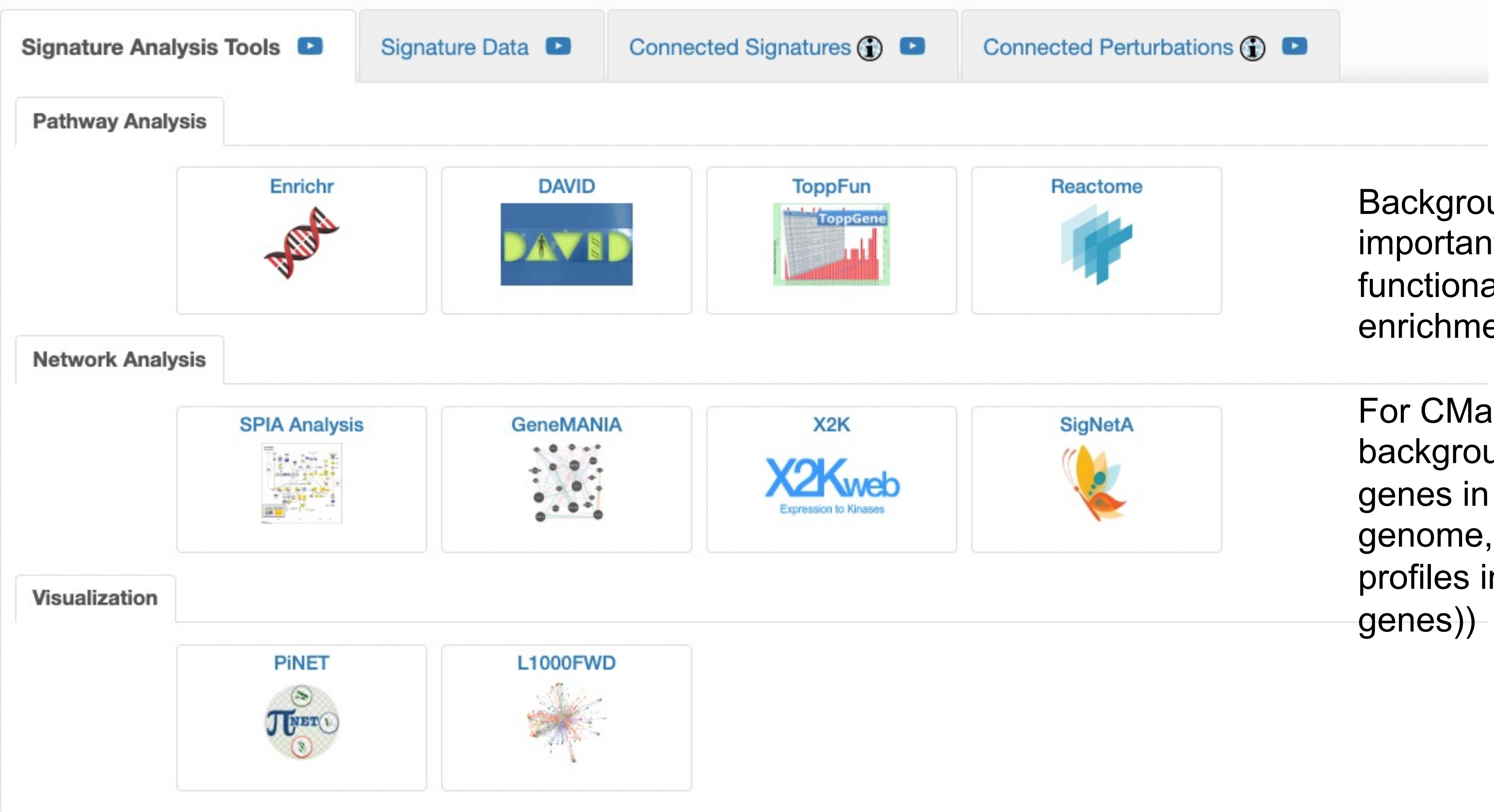



Wellcome Trust Case Control Consortium (WTCCC) data
https://www.nature.com/articles/s41467-018-03621-1

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

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA





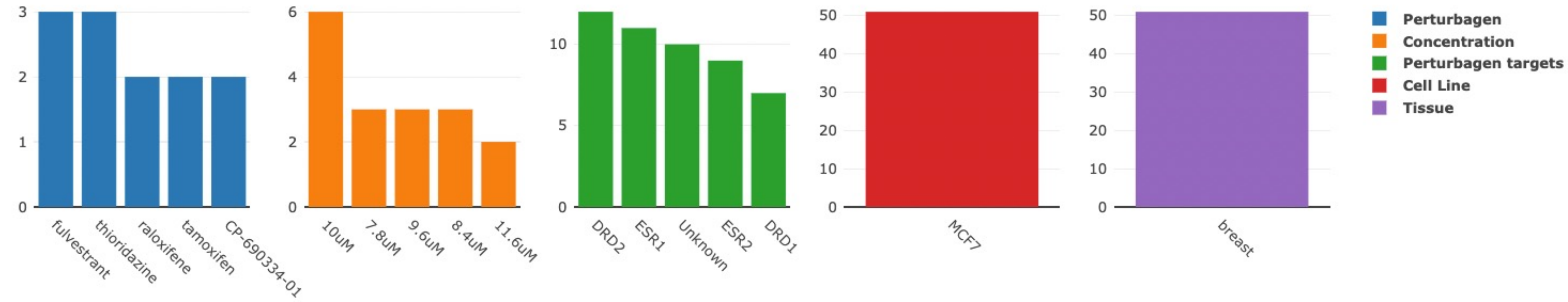

Background gene list very important when doing functional/pathway enrichment analysis.

For CMap data, background list is not all genes in the human genome, rather all genes profiles 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

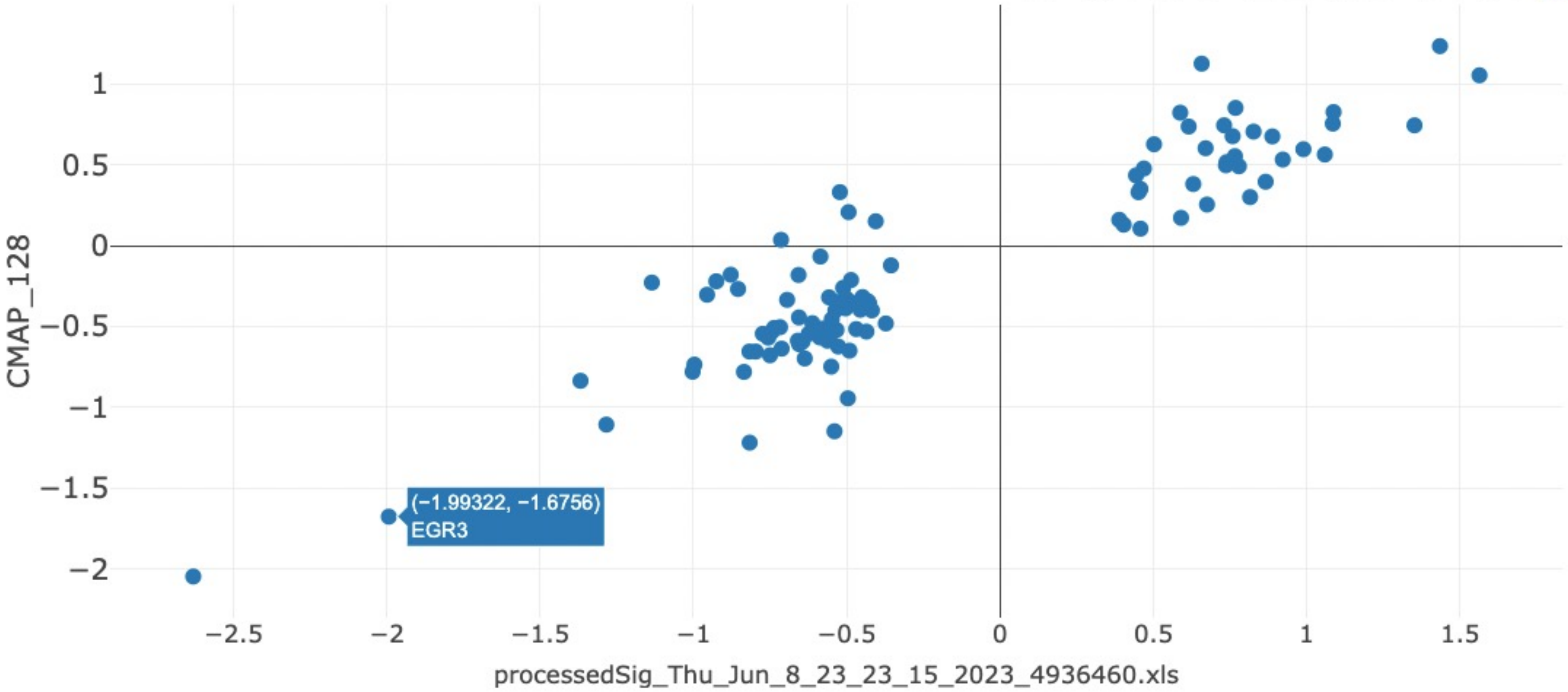
First « 1 2 3 4 5 » Last

Page 1 of 11, of 51 entries

# Correlation plot

Weighted Pearson correlation: **0.943**

Pearson correlation: **0.913**

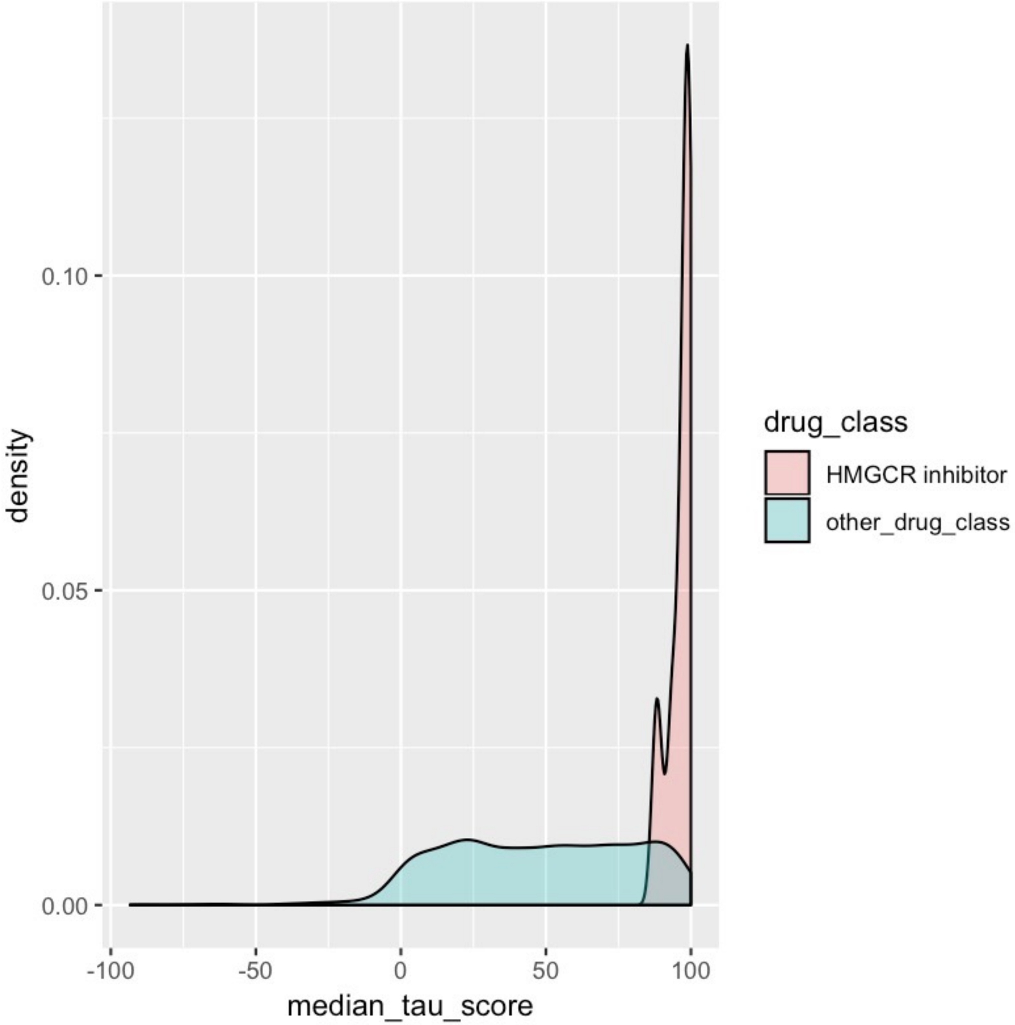



Use selected genes▾ Cancel

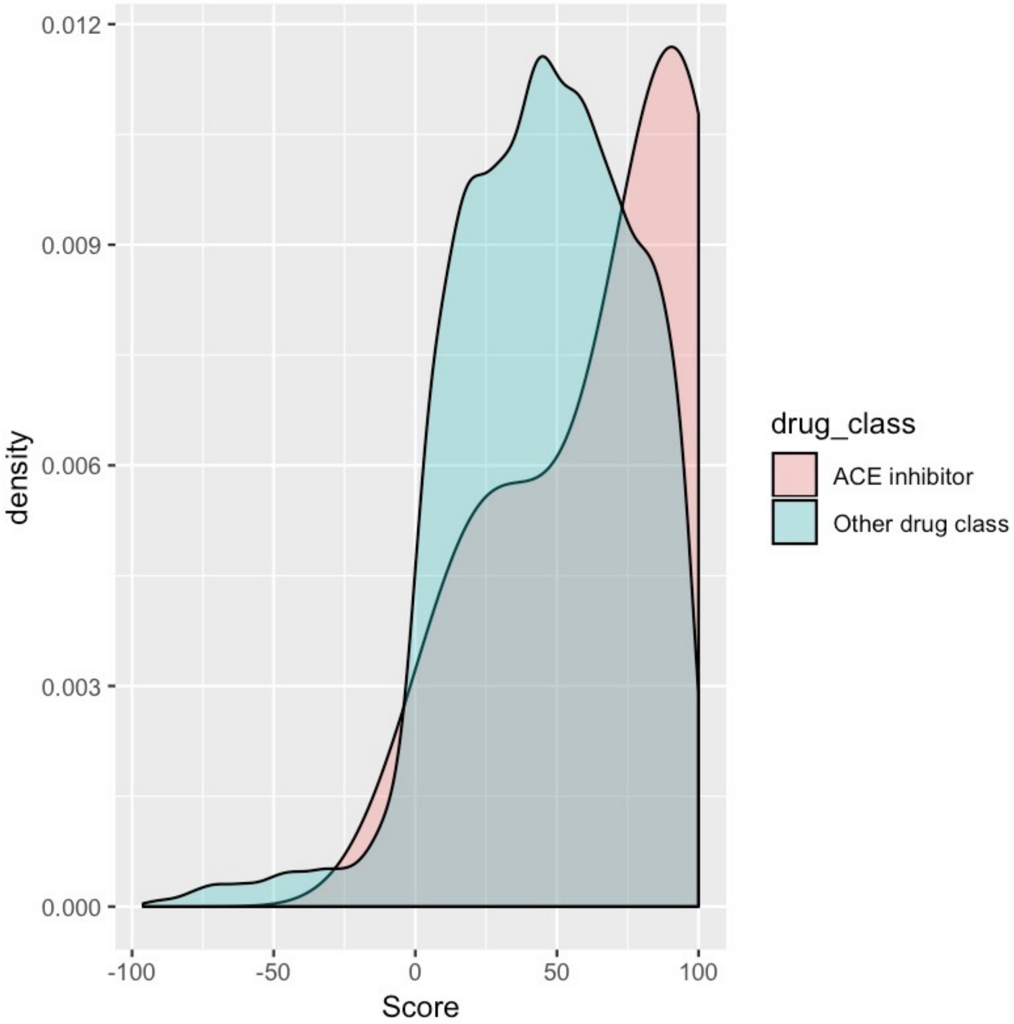
# Take home messages

- 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

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



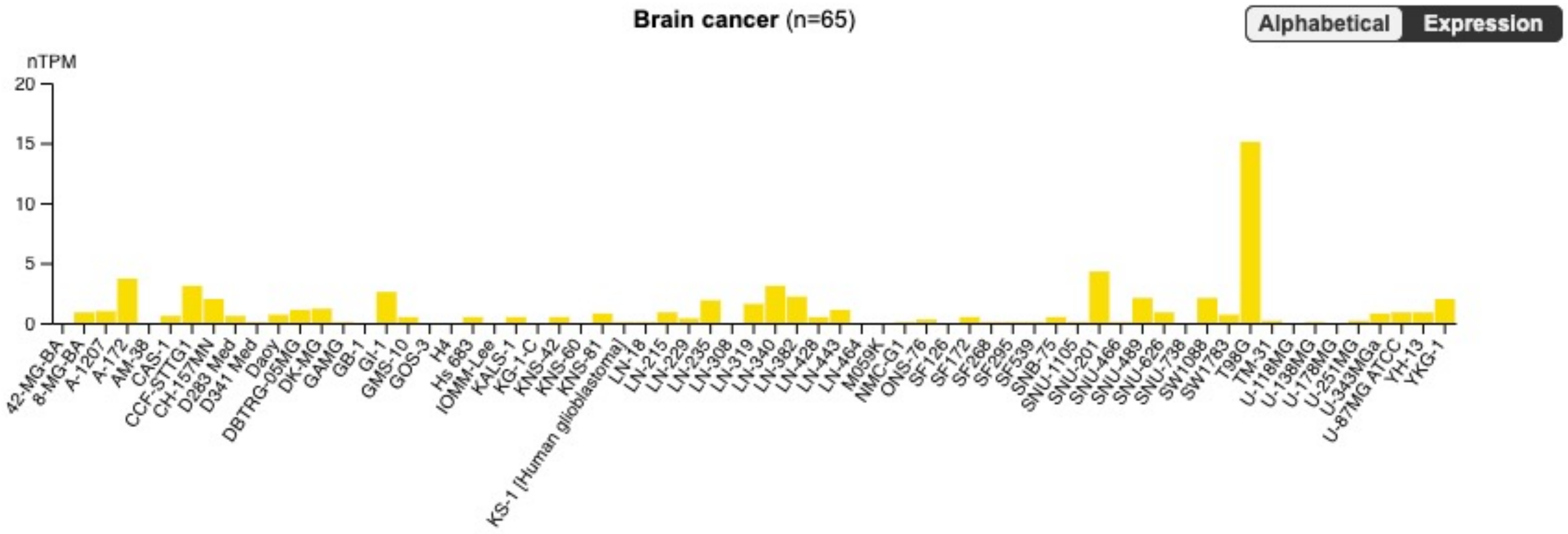

Connectivity of enalapril with other ACE inhibitors and all other compounds



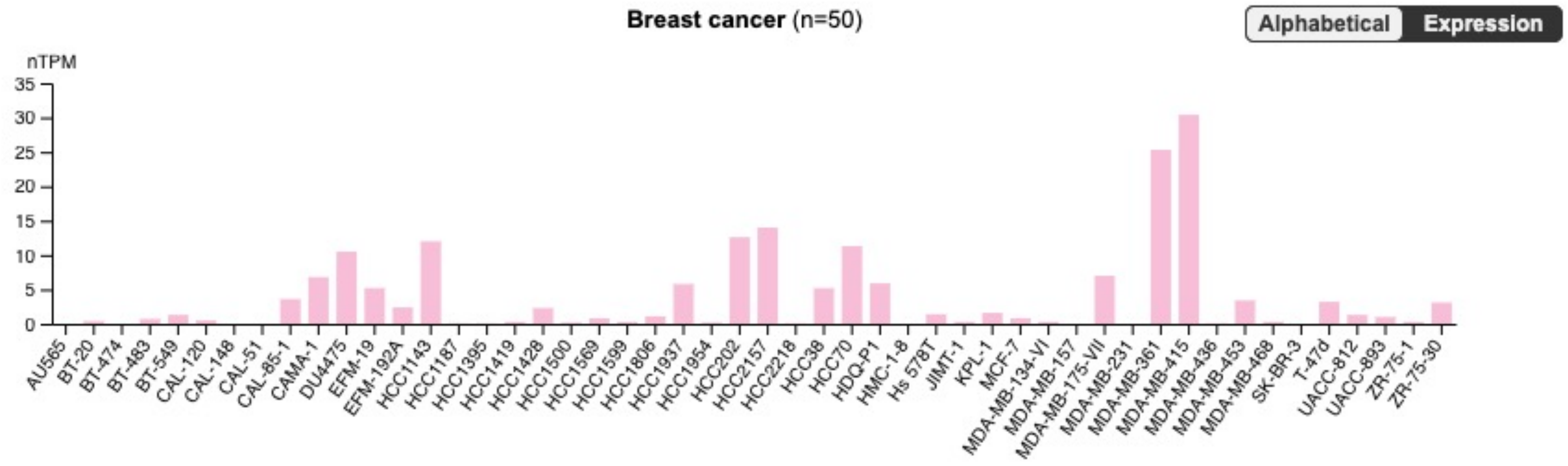
# Take home messages

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

- 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
  - 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)



Brain cancer (n=65)
Alphabetical
Expression
nTPM
20
15
10
5
0
42-MG-BA
8-MG-BA
A-1207
A-172
AM-38
CAS-1
CCF-STTG1
CH-157MN
D283 Med
D341 Med
Daoy
DBTRG-05MG
DK-MG
GAMG
GB-1
GI-1
GMS-10
GOS-3
H4
Hs 683
IOMM-Lee
KALS-1
KG-1-C
KNS-42
KNS-60
KNS-81
KS-1 [Human glioblastoma]
LN-18
LN-215
LN-229
LN-235
LN-308
LN-319
LN-340
LN-382
LN-428
LN-443
LN-464
M059K
NMC-G1
ONS-76
SF126
SF172
SF268
SF295
SF539
SNB-75
SNU-1105
SNU-201
SNU-466
SNU-489
SNU-626
SNU-738
SW1088
SW1783
T98G
TM-31
U-118MG
U-138MG
U-178MG
U-251MG
U-343MGa
U-87MG ATCC
YH-13
YKG-1

Breast cancer (n=50)
Alphabetical
Expression
nTPM
35
30
25
20
15
10
5
0
AU565
BT-20
BT-474
BT-483
BT-549
CAL-120
CAL-148
CAL-51
CAL-85-1
CAMA-1
DU4475
EFM-19
EFM-192A
HCC1143
HCC1187
HCC1395
HCC1419
HCC1428
HCC1500
HCC1569
HCC1599
HCC1806
HCC1937
HCC1954
HCC202
HCC2157
HCC2218
HCC38
HCC70
HDQ-P1
HMC-1-8
Hs 578T
JIMT-1
KPL-1
MCF-7
MDA-MB-134-VI
MDA-MB-157
MDA-MB-175-VII
MDA-MB-231
MDA-MB-361
MDA-MB-415
MDA-MB-436
MDA-MB-453
MDA-MB-468
SK-BR-3
T-47d
UACC-812
UACC-893
ZR-75-1
ZR-75-30