

Connectivity Map (Cmap) processing and querying.

Dr Solal Chauquet

**Centre for Population & Disease Genomics
Institute for Molecular Bioscience**

CMap – One dataset several names

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)
- **1,319,138 replicates measured.**

Also includes genetic perturbation.

CMap – One dataset several names

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**

Also includes genetic perturbation

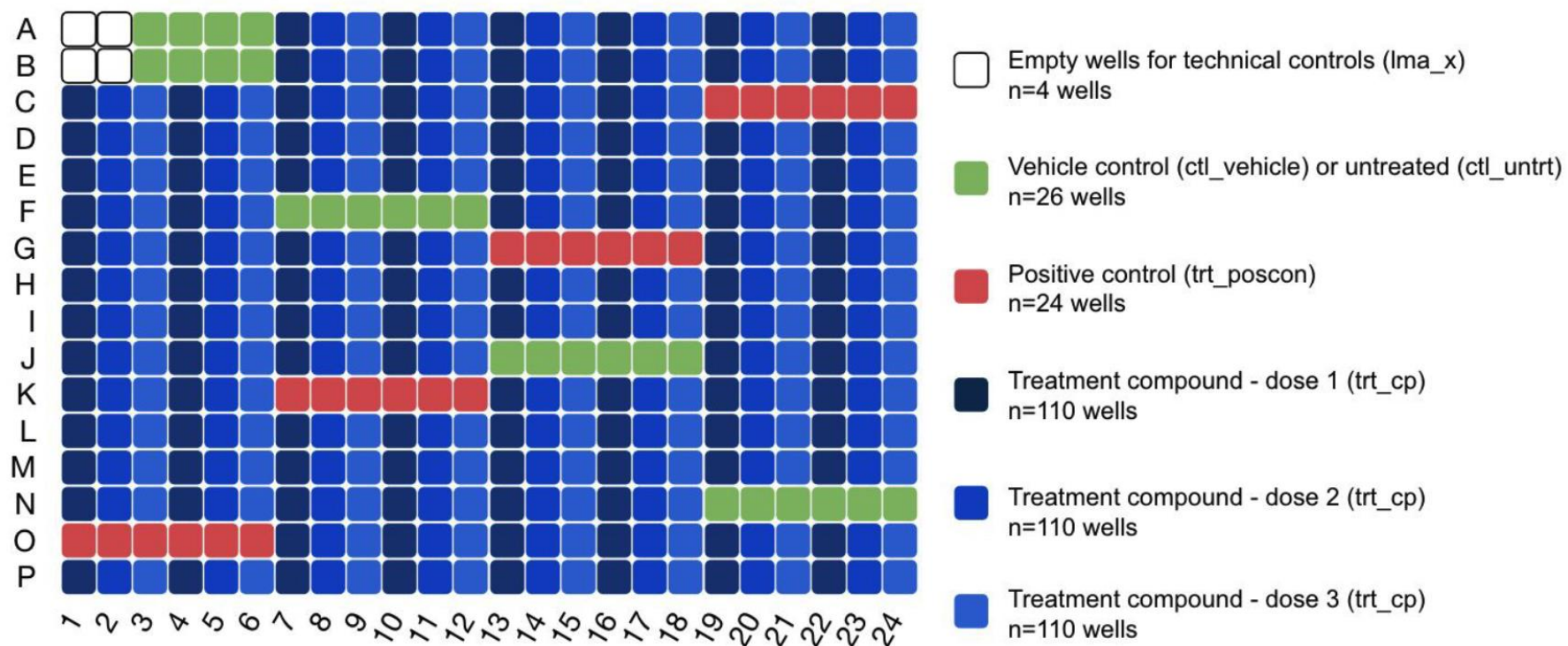
CMap – One dataset several names

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.

Measuring Gene expression:

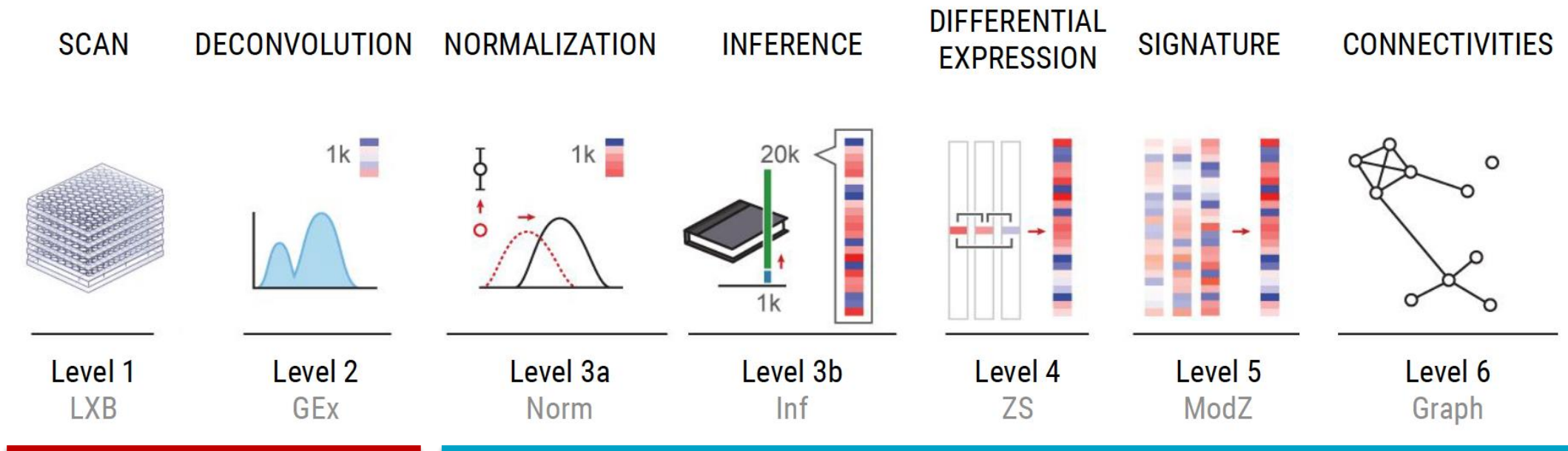
L1000 microarray:

Measures only 1,058 transcripts.

- Gene measured:
 - Landmark set:
 - 978 genes that can be used to impute the rest of the transcriptome.
 - Normalising set:
 - 80 genes being invariant within those 12,063 datasets
 - 8 genes, each at 10 levels of low to high expression

CMap – Data processing:

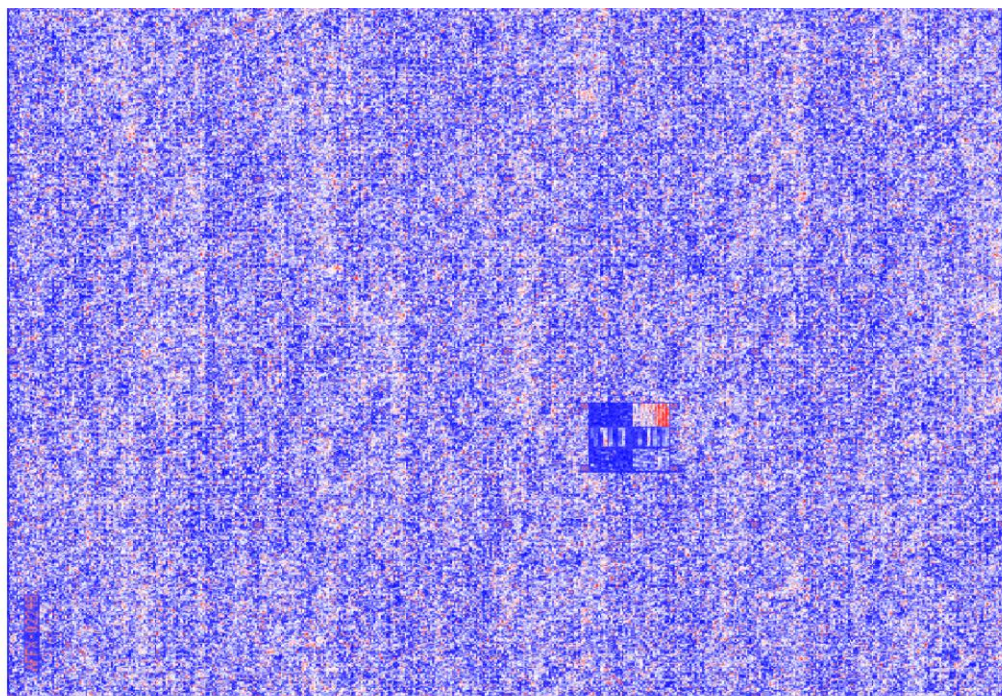
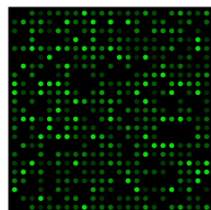
L1000 Data Processing Processing Stages Post-Detection



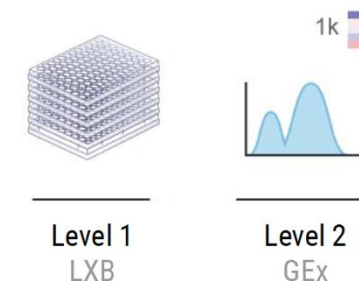
Default microarray processing

CMAP specific processing

Microarray processing:



Example Slide Affymetrix microarray (~18,000 genes)

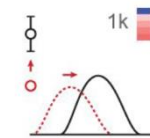


Data generation for each of the replicates:

- 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

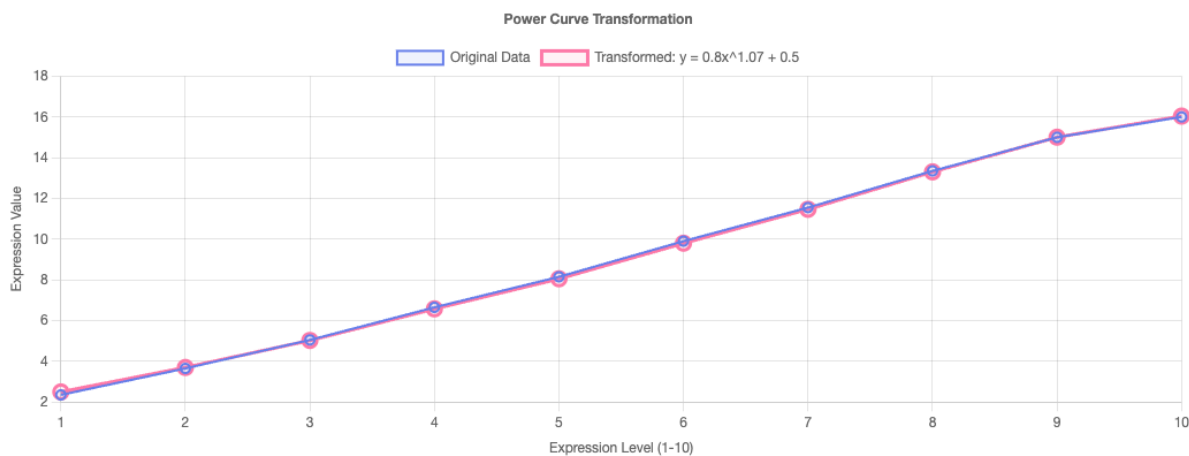


Level 3a Norm

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:

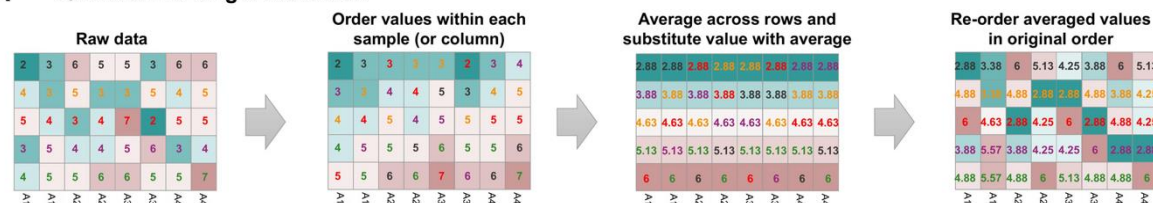
$$y_{scaled} = ay_{raw}^b + c$$

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.

A Quantile in a single class data



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:

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

Where:

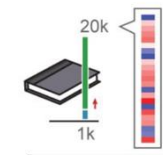
- w_i is the model weights predicted using DS_{GEO}
 - The weights are calculated using an ordinary least square approach within the DS_{GEO} dataset

This approach allows the prediction of 12,328 genes.

They benchmark this inference strategy by using the GTEx dataframe and calculate the correlation between the inferred and measure expression.

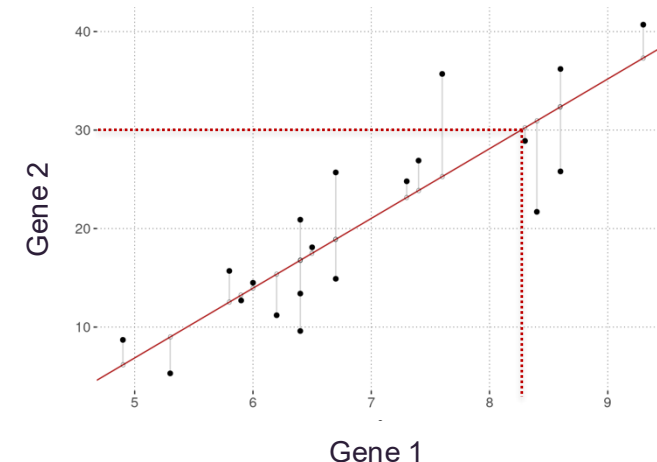
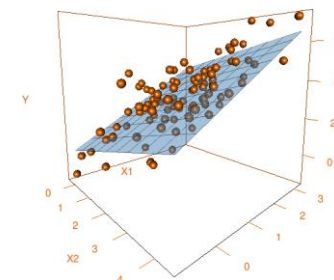
| Gene Symbol | Gene Title | Correlation | Feature set |
|-------------|--|-------------|-------------|
| ESRRA | estrogen related receptor alpha | 0.89 | BING |
| EIF3D | eukaryotic translation initiation factor 3 subunit D | 0.90 | BING |
| HAUS2 | HAUS augmin like complex subunit 2 | -0.38 | Inferred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

INFERENCE



Level 3b
Inf

2 response variables:



Example:

Gene 1: Measured value of 8.2

Gene 2: Predicted value of 30

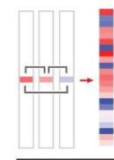
CMap processing - Level 4: Z-score scaling

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

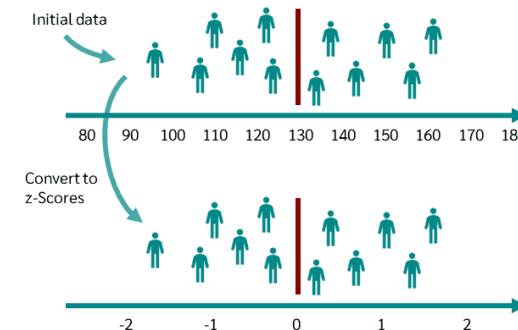
$$z_i = \frac{y_{norm} - \mu}{\sigma}$$

Where:

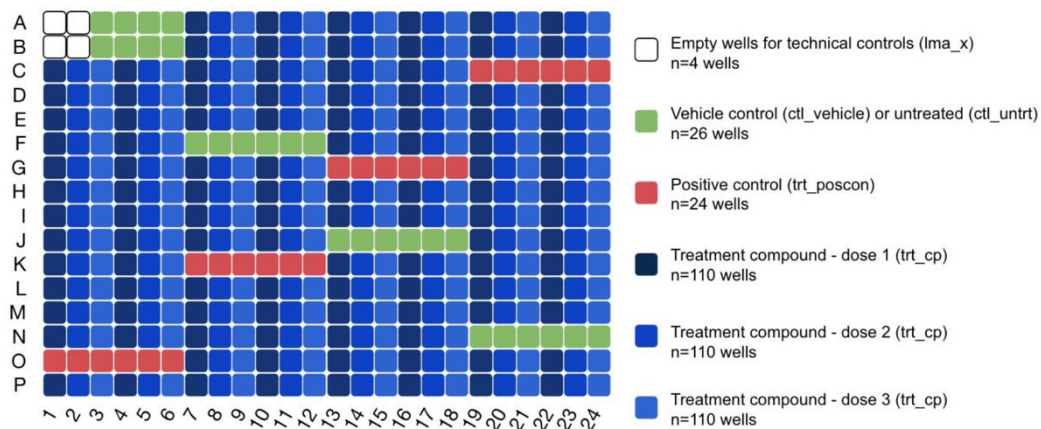
- z_i is the z-score transformed expression value.
- y_{norm} is the normalized expression (measured or inferred)
- μ is the mean normalized expression *on the plate*
- σ is the standard deviation of the normalized expression *on the plate*



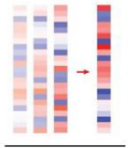
Level 4
ZS



Example plate layout: 3 dose



CMap processing – Level 5: consensus signatures



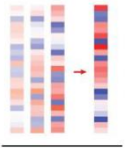
Level 5
ModZ

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.

Level 5
ModZ

CMap processing – Level 5: consensus signatures

Example 3 genes, 3 replicates:

| | Rep1 | Rep2 | Rep3 |
|--------|------|------|------|
| Gene 1 | 10 | 13 | 9 |
| Gene 2 | 8 | 6 | 0 |
| Gene 3 | 2 | 2 | 2 |

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:

$$[1.4 \quad 1.72 \quad 1.32] \quad \text{Normalizing factor: 4.44}$$

Step 4: Normalized weights:

$$[0.31 \quad 0.39 \quad 0.30]$$

Step 5: Linear combination

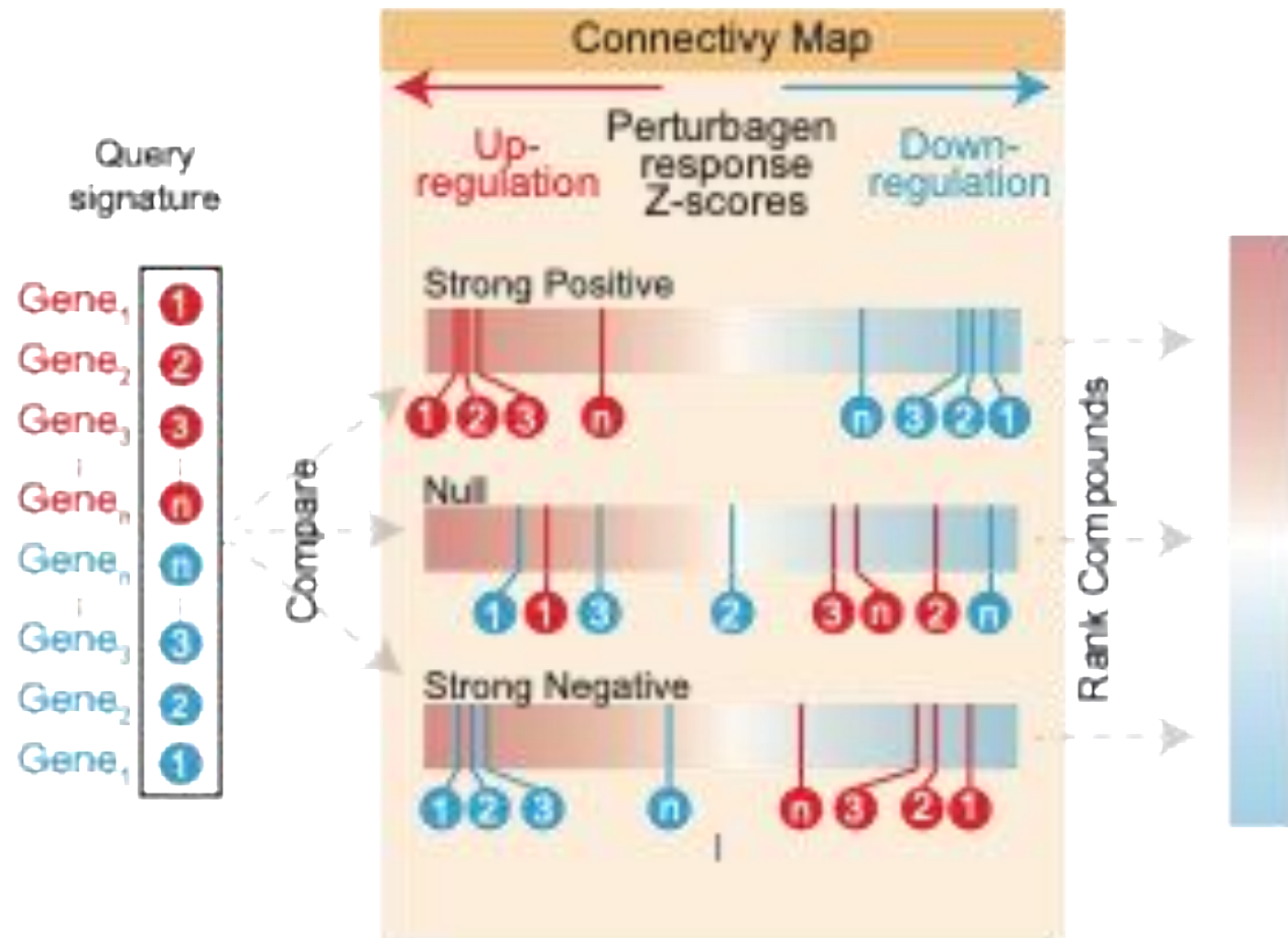
$$\begin{bmatrix} 0.31 * 10 + 0.39 * 13 + 0.30 * 9 \\ 0.31 * 8 + 0.39 * 6 + 0.30 * 0 \\ 0.31 * 2 + 0.39 * 2 + 0.30 * 2 \end{bmatrix} = \begin{bmatrix} 10.87 \\ 4.82 \\ 2 \end{bmatrix}$$

Consensus Signature

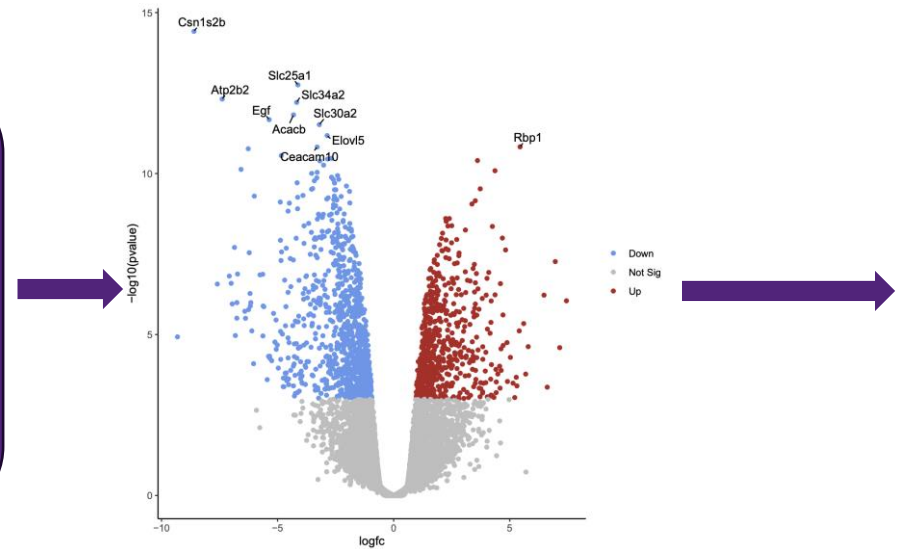
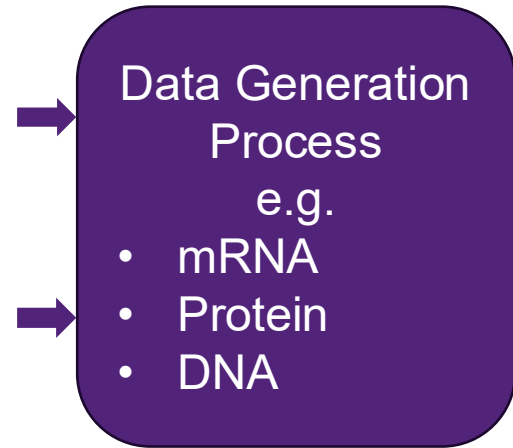
Questions?



How to query CMap:



Disease signature



| Gene | LogFC | P-value |
|---------|-------|----------|
| Csn1s2b | -8.02 | 5.29E-16 |
| atp2b2 | -6.52 | 3.52E-11 |
| slc25A | -4.53 | 2.95E-06 |
| Egf | -5.11 | 1.61E-07 |
| slc34a2 | -8.08 | 5.04E-07 |
| Slc30a2 | -3.98 | 3.45E-05 |
| Rbp1 | 5.61 | 1.01E-08 |
| Elov15 | -3.28 | 5.19E-04 |
| Ceacam | -3.68 | 1.17E-04 |

What data type should be used?

Disease signature

RNA sequencing:

Pro:

- Easy to generate
- Easy to analyze

Cons:

- **Function:** Transcriptomic changes do not necessarily have an impact on protein level
- **Causality:** RNA can be caused by the disease instead of causing the disease.
- **Sampling:** Obtaining the tissue of interest can be hard or impossible.

Protein sequencing:

Pro:

- Direct link to the biology
- (fairly) easy to analyze

Cons:

- **Causality:** Protein abundance changed can be caused by the disease instead of causing the disease.
- **Sampling:** Obtaining the tissue of interest can be hard or impossible.
- **Function:** Protein abundance do not necessarily translate to protein function

DNA sequencing:

Pro:

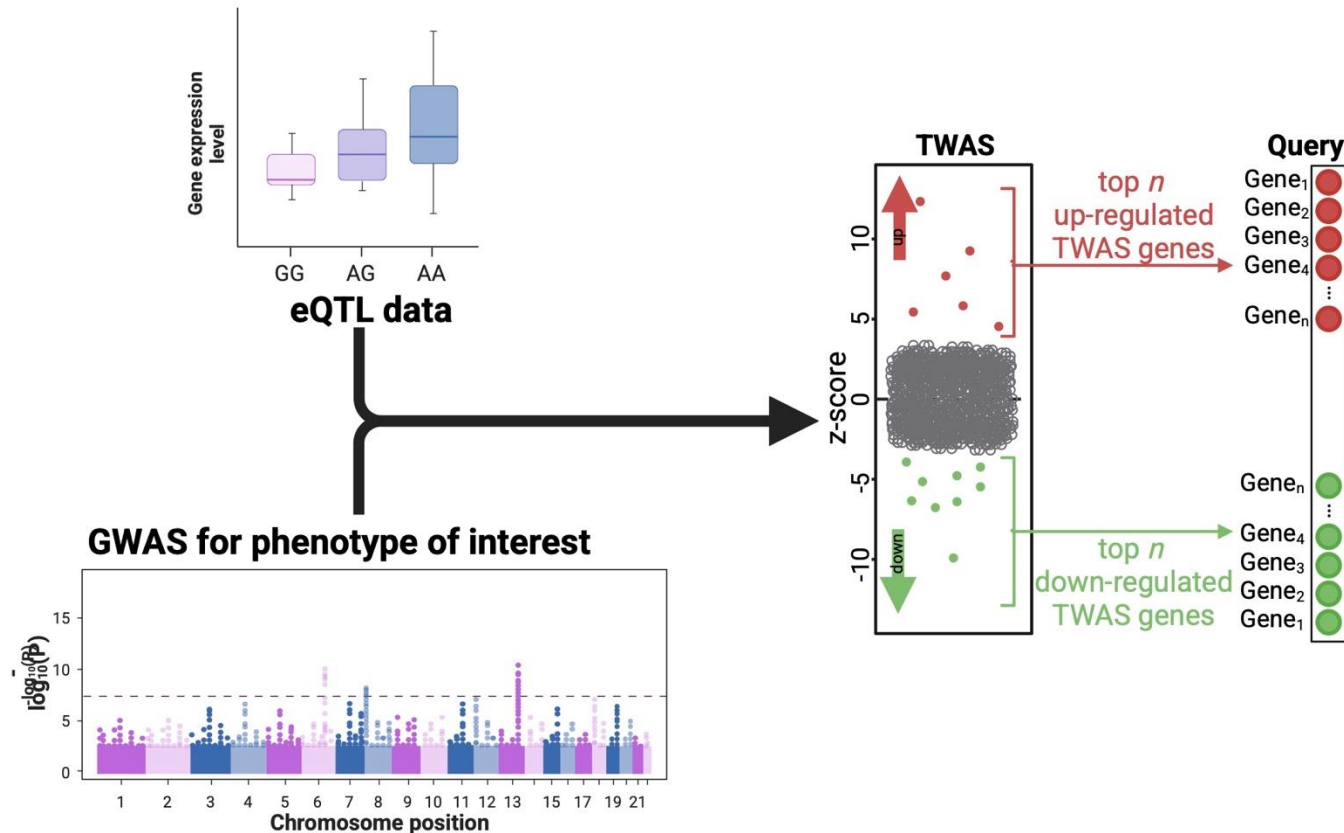
- Massive biobank available.
- **Causality:** Disease does not influence DNA.

Cons:

- **Function:** Not linked to expression changes
- **Specificity:** Not tissue specific

Using DNA for a disease signature

Transcriptome wide association study:



Intuition:

- SNP1:
 - Associated with an increase expression of gene1 (eQTL)
 - Associated with a higher prevalence of disease1

Patients with disease 1 can be predicted to have a higher expression of gene1.

- SNP2:
 - Associated with a decrease expression of gene2 (eQTL)
 - Associated with a higher prevalence of disease 1.

What can we predict about the expression of gene2?

Disease signature

How to define a disease signature?

1. What threshold do we want to use?

1. P-value or Z-score - is there a difference between the two?

2. How many genes do we want to include in the signature?

1. All significant genes?

2. Highly significant genes?

1. How do you define high here?

Don't use a single signature.

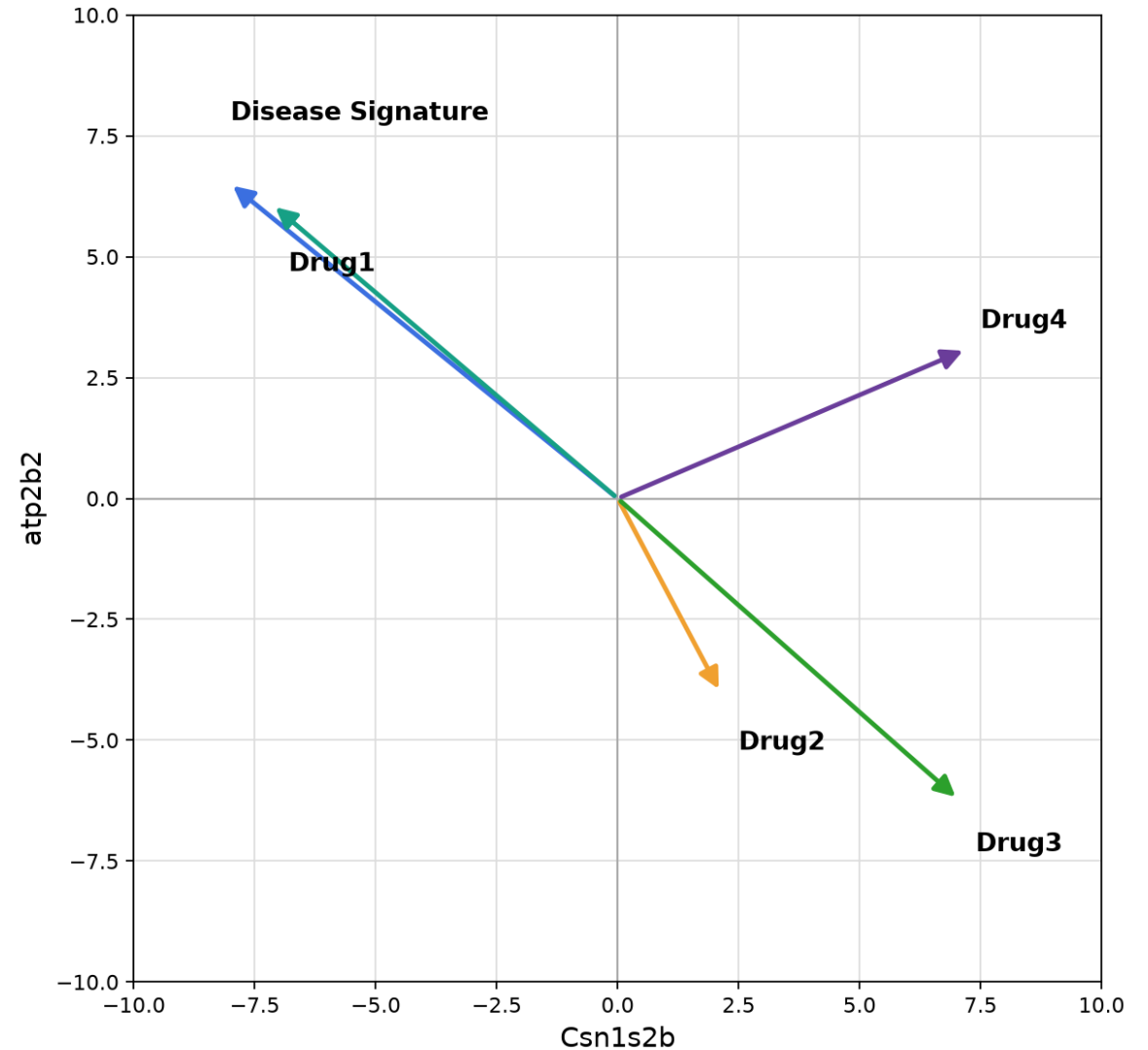
Selection bias and winner's curse can lead to inflated signal.

| Gene | Z-score | P-value |
|---------|---------|----------|
| Csn1s2b | -8.02 | 5.29E-16 |
| atp2b2 | 6.52 | 3.52E-11 |
| slc25A | -4.53 | 2.95E-06 |
| Egf | 5.11 | 1.61E-07 |
| slc34a2 | -8.08 | 5.04E-07 |
| Slc30a2 | -3.98 | 3.45E-05 |
| Rbp1 | 5.61 | 1.01E-08 |
| Elovl5 | -3.28 | 5.19E-04 |
| Ceacam | -3.68 | 1.17E-04 |

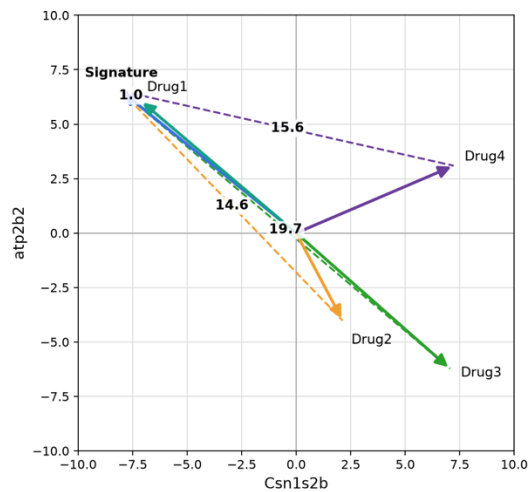
How to think about similarity?

| Gene | Disease Signature | Drug1 | Drug2 | Drug3 | Drug4 |
|----------------|-------------------|--------------|--------------|--------------|-------------|
| <u>Csn1s2b</u> | <u>-8.02</u> | <u>-7.14</u> | <u>2.13</u> | <u>7.05</u> | <u>7.22</u> |
| <u>atp2b2</u> | <u>6.52</u> | <u>6.08</u> | <u>-4.02</u> | <u>-6.23</u> | <u>3.09</u> |
| <u>slc25A</u> | -4.53 | -4.27 | 7.31 | 4.18 | -2.14 |
| <u>Egf</u> | 5.11 | 5.36 | 3.45 | -5.42 | 8.03 |
| <u>slc34a2</u> | -8.02 | -6.91 | -8.16 | 6.88 | 4.95 |

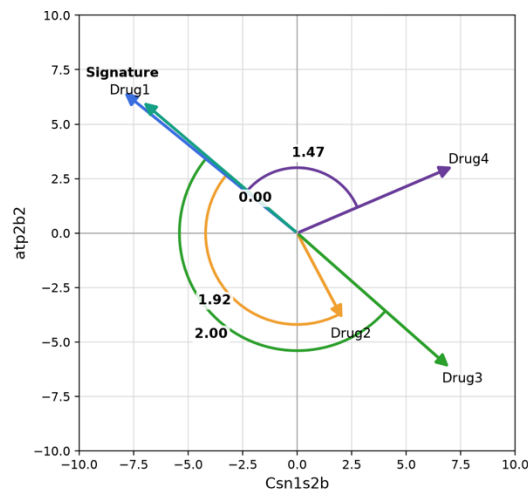
Visual representation of the perturbations in 2D



Euclidean distance



Cosine distance



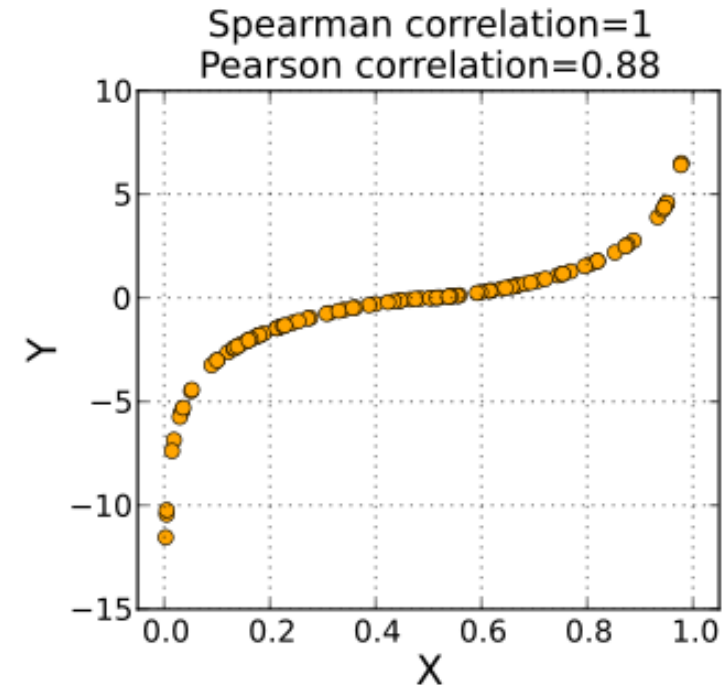
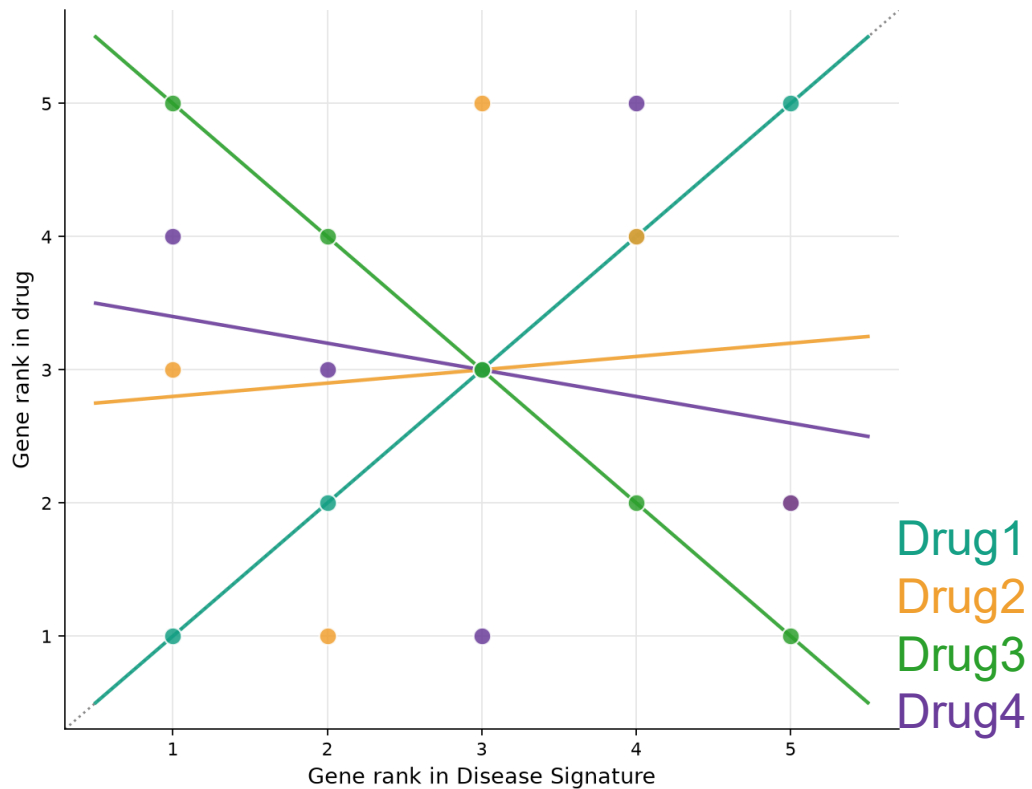
How to think about similarity?

There are no obviously 'better' metric than another.

But... some metrics have a natural advantage:

- Spearman correlation is resistant to scale effect.

Similarity based on rank



Drug recommendations:

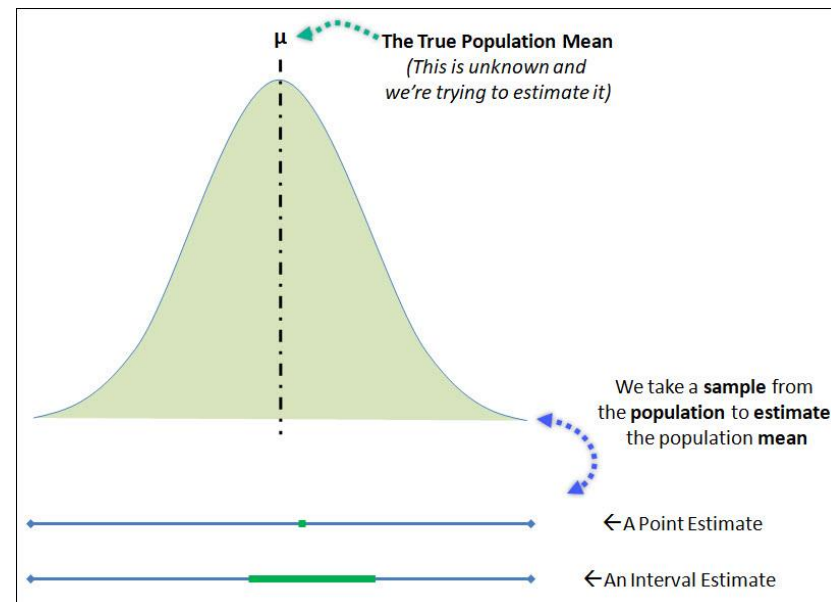
We rank drugs based on their similarity:

1. Drug1
2. Drug2 *Based on our hypothesis Drug3 is therefore the most likely to be a candidate to help patients with the disease*
3. Drug4 *patients with the disease*
4. Drug3

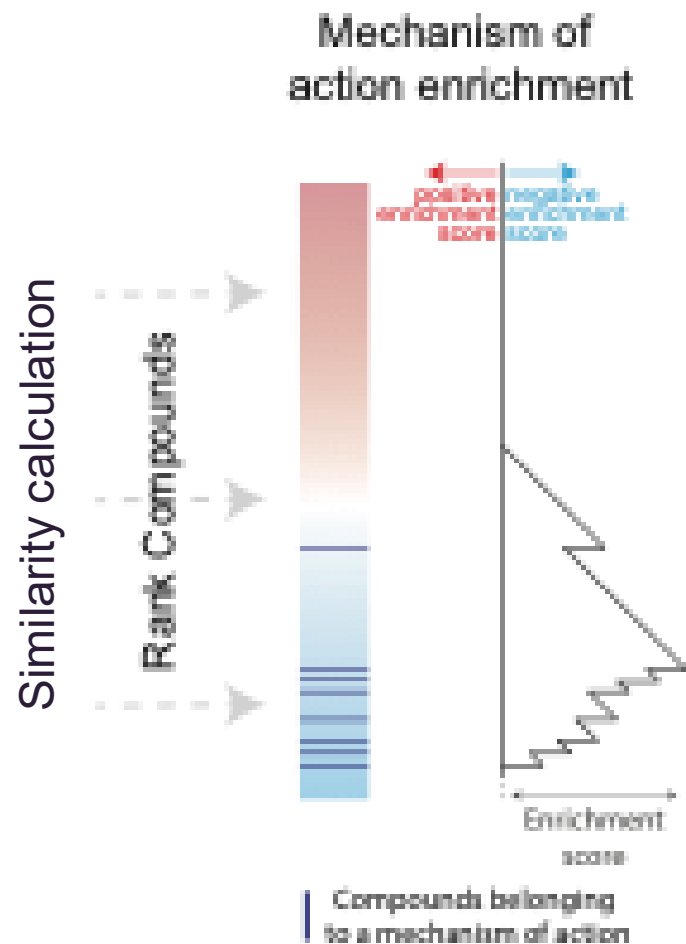
Two possible ways of improving our drug prediction:

1. Increase the sample size of our perturbation dataset (more accurate measurement of every gene)
2. Group similar compounds together.

How likely is it to be true?



Enrichment:



Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian¹, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, Jill P Mesirov

Affiliations + expand

PMID: 16199517 PMCID: [PMC1239896](https://pubmed.ncbi.nlm.nih.gov/16199517/) DOI: [10.1073/pnas.0506580102](https://doi.org/10.1073/pnas.0506580102)

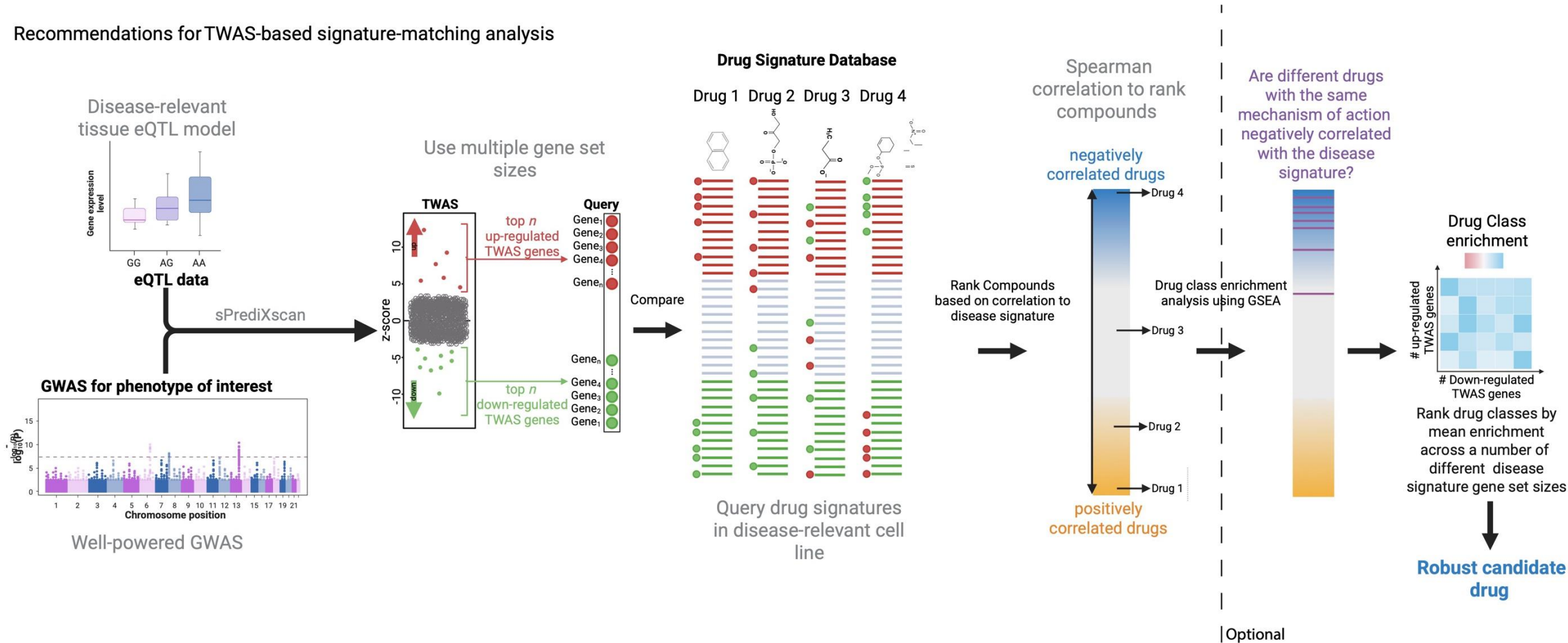
Grouping compounds based:

- Known mechanism of action
- Indications
- ...
- Hypothesis based group
- Chemical structure

This allows us to increase the amount of evidence underlying the proposed drugs.

Putting it all together and more questions:

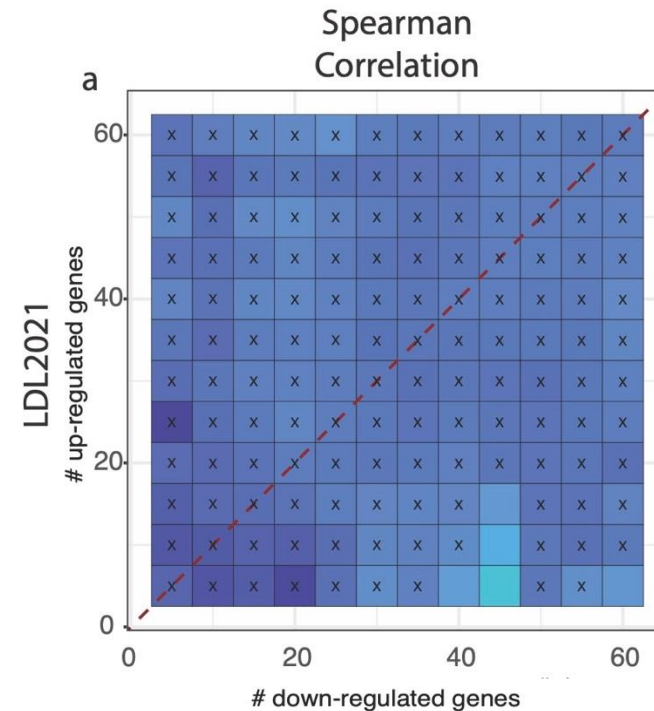
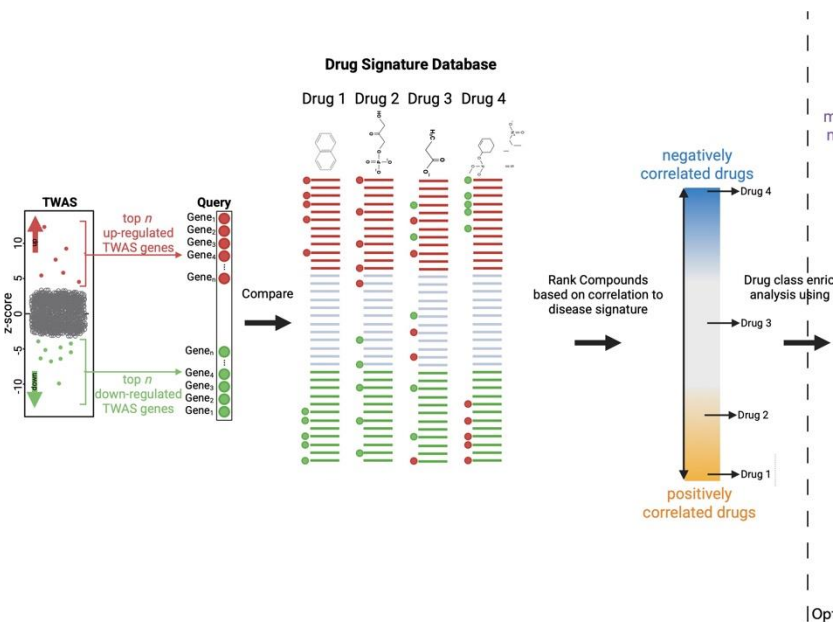
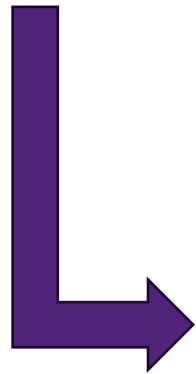
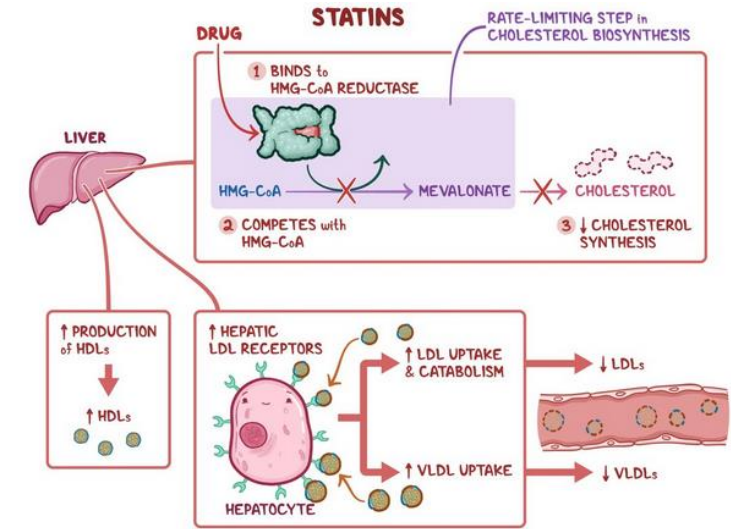
Recommendations for TWAS-based signature-matching analysis



Real example:

TWAS based repurposing:

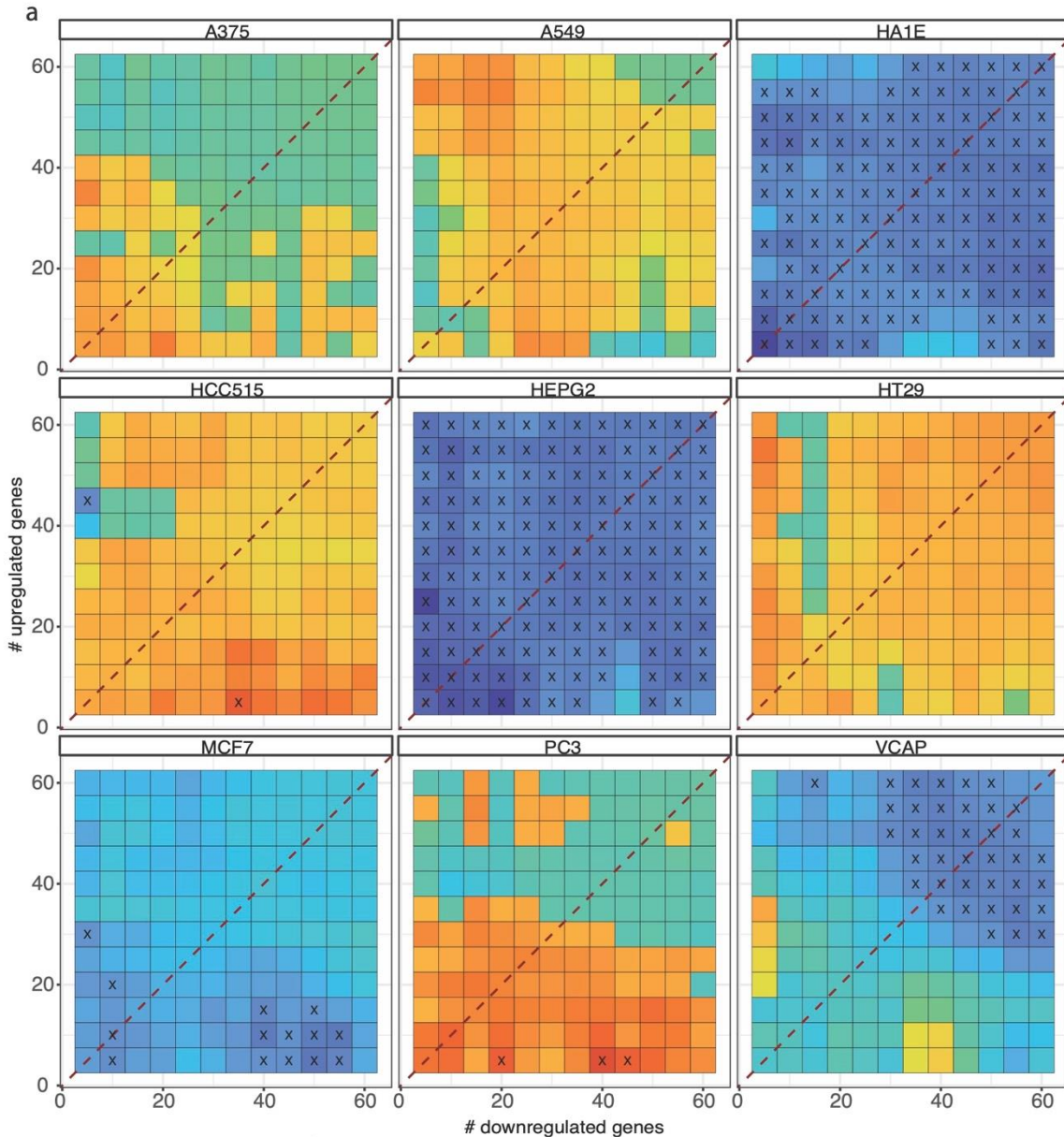
| | | |
|--------------|--------------|-------------|
| GWAS: | LDL2021 | N=1,320,016 |
| eQTL: | Liver Tissue | N=262 |



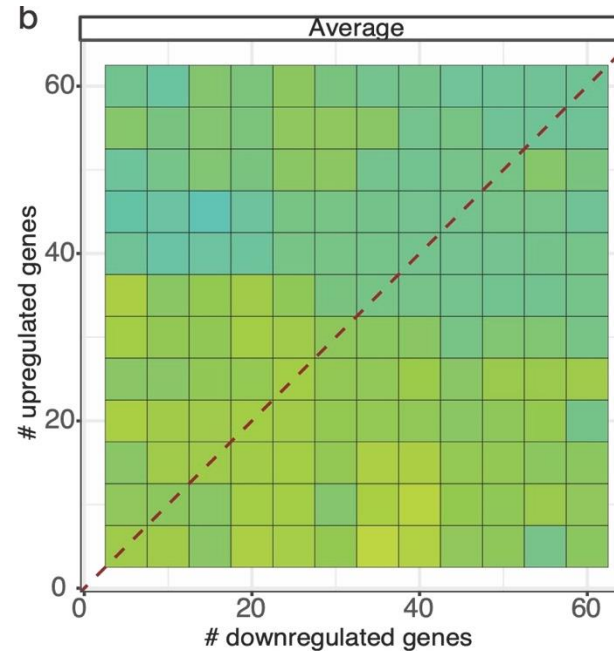
Enrichment of HMGCR inhibitors for LDL-C:

- Spearman correlation:
- Enrichment: -1.76 ± 0.15 , $p = 0.014$

Analytic considerations:



HMGCR inhibitors
Normalised enrichment Score:
-2 -1 0 1 2



Profiling drugs in different cell line completely change the enrichment

From GWAS to drug: A framework for drug candidate prioritisation using a gene expression signature matching approach

medRxiv

THE PREPRINT SERVER FOR HEALTH SCIENCES

Other perturbagen dataset:

PerturBase: a comprehensive database for single-cell perturbation data analysis and visualization



[Zhiting Wei](#), [Duanmiao Si](#), [Bin Duan](#), [Yicheng Gao](#), [Qian Yu](#), [Zhenbo Zhang](#) ✉, [Ling Guo](#) ✉, [Qi Liu](#) ✉ [Author Notes](#)

Tahoe-100M: A Giga-Scale Single-Cell Perturbation Atlas for Context-Dependent Gene Function and Cellular Modeling

THE PREPRINT SERVER FOR BIOLOGY

Article

Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq

[Joseph M. Replogle](#)^{1 2 3 4 5 14}, [Reuben A. Saunders](#)^{2 3 4 5 14}, [Angela N. Pogson](#)^{3 4 5}, [Jeffrey A. Hussmann](#)^{3 4 5}, [Alexander Lenail](#)^{4 5}, [Alina Guna](#)⁵, [Lauren Mascibroda](#)⁶, [Eric J. Wagner](#)^{6 7}, [Karen Adelman](#)⁸, [Gila Lithwick-Yanai](#)⁹, [Nika Iremadze](#)⁹, [Florian Oberstrass](#)⁹, [Doron Lipson](#)⁹, [Jessica L. Bonnar](#)^{3 4 5}, [Marco Jost](#)^{3 10}, [Thomas M. Norman](#)¹¹  ✉, [Jonathan S. Weissman](#)^{3 4 5 12 13 15}  ✉