Lecture 16: Course Project Info/Pitches

Department of Computer Science and Engineering
University of Minnesota

Reminders

- Course project proposals due in 2 weeks (Thursday, 3/30)
- Basso et al. paper discussion will be on Tuesday, 3/21
- Segal et al. paper discussion on Thursday, 3/23
- HW #2 is assigned (see Canvas); due 4/5

Modules you should watch this week

(all linked on Canvas in this week's section)

- Support Vector Machine module
- Introduction to deep learning module
- Introduction to inference of regulatory networks (background for Basso et al. discussion)

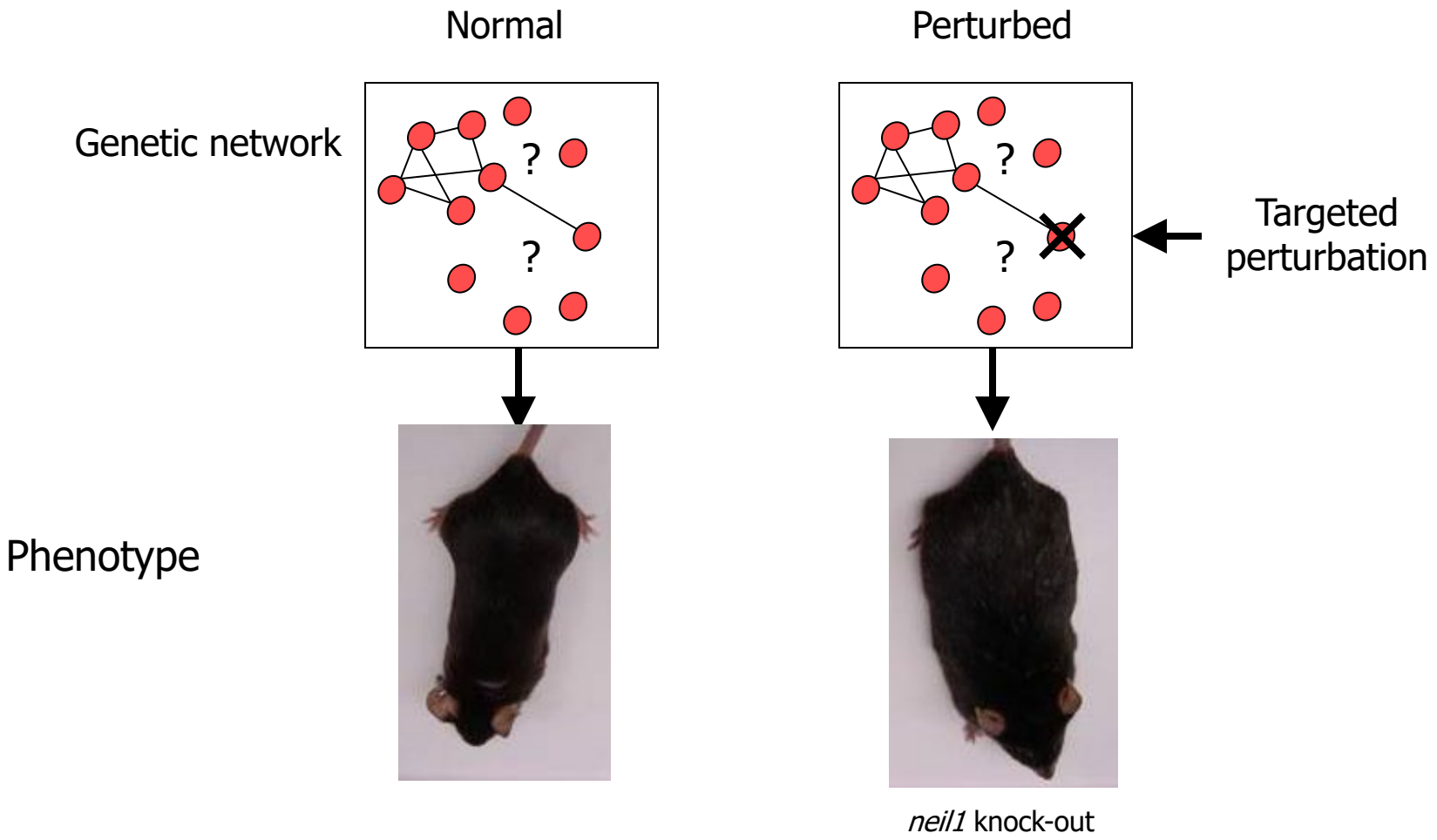
Overview of today

- I will present an overview of the Machine Learning prediction challenge (this is one possible project)
  - Q&A with Arshia
- Project Pitches from Students
  - Ibrahim/Ahmed
  - Miguel/Emma
- Remainder of lecture period: connect with other students to form project groups/discuss ideas,

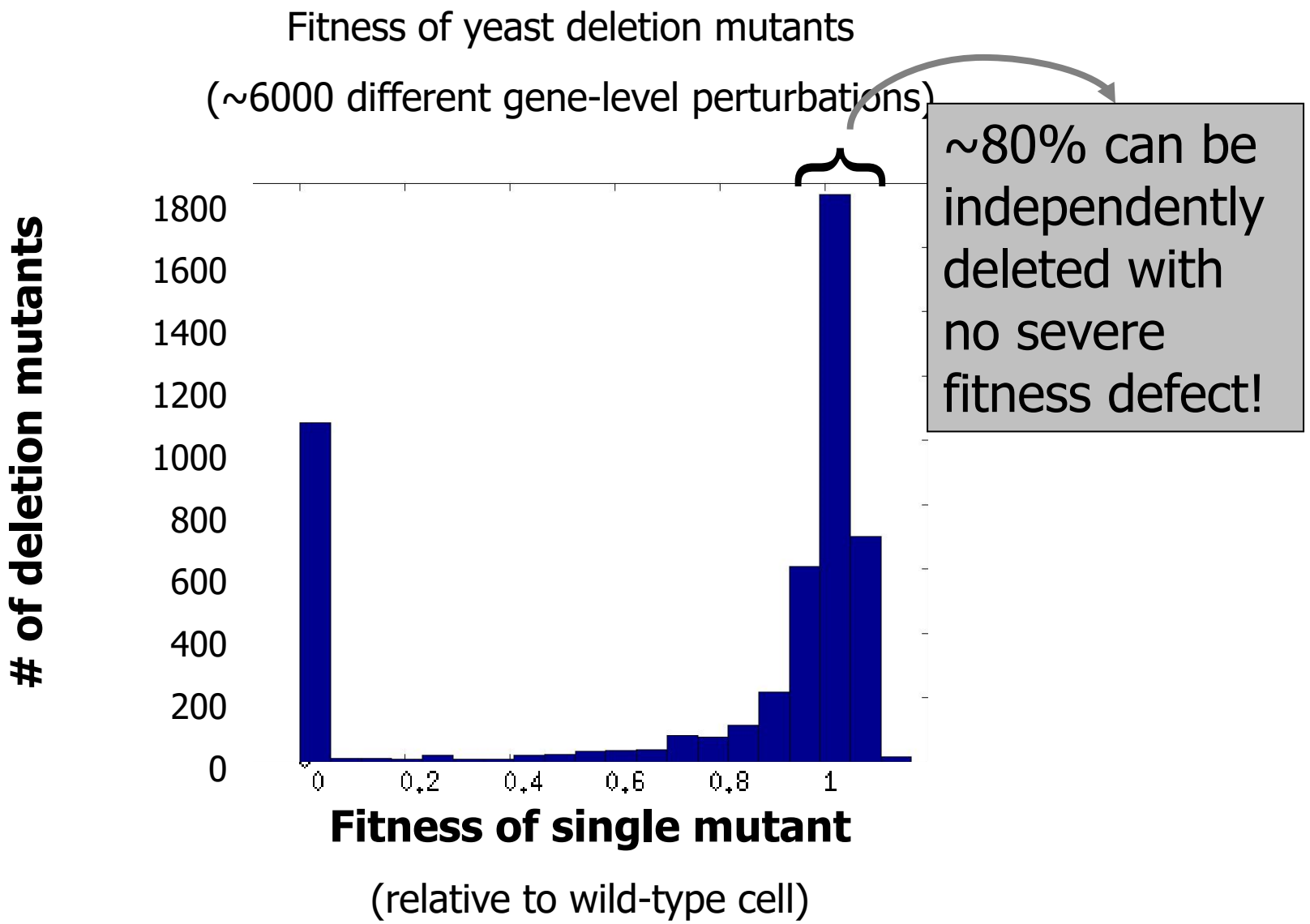
Overview of ML Prediction Challenge:

Predicting human gene function from CRISPR genetic interaction screens

Perturbation analysis: reverse engineering biology

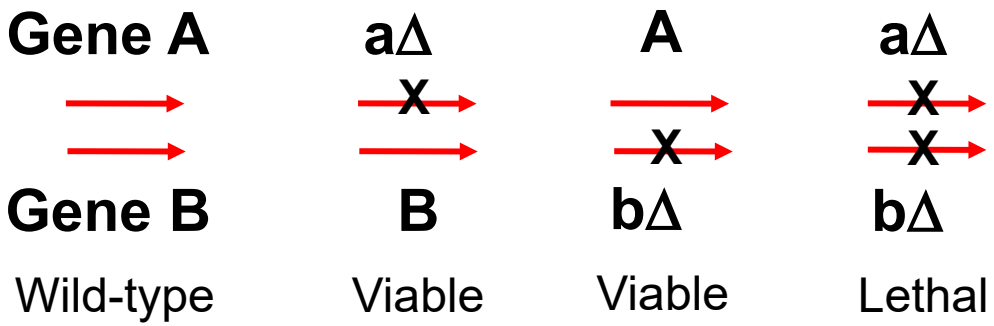


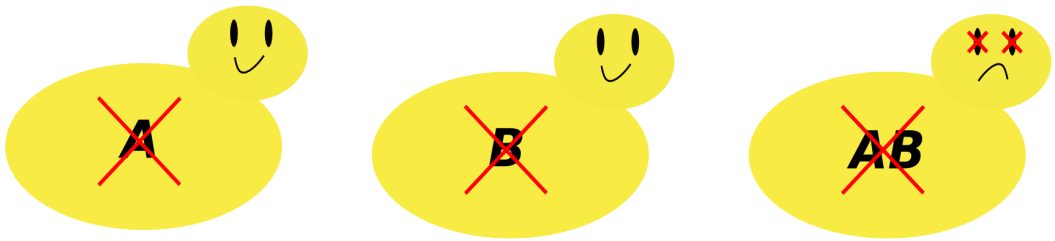
How much can we learn from single perturbations?

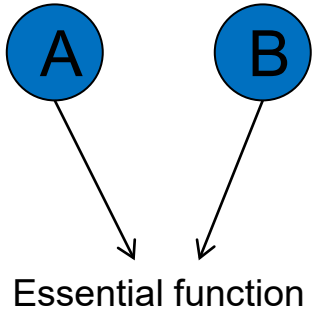


Winzeler EA *et al*: **Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis**. *Science* 1999, **285**(5429):901-906.

One interesting outcome of combining gene deletions



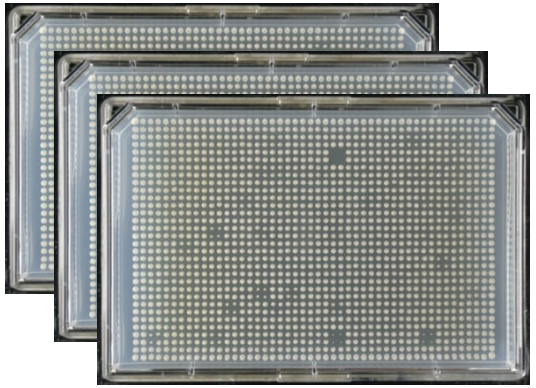




"synthetic lethality"

Or, more generally, "genetic interaction"

A near complete genetic interaction map in yeast



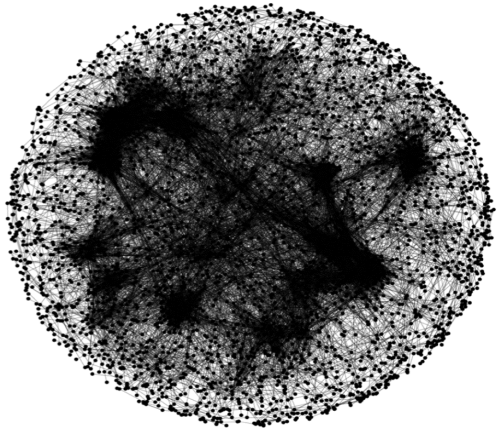
Synthetic Genetic Arrays

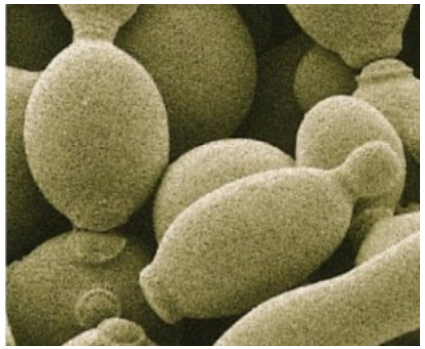
> 23 million gene pairs screened

~1 million genetic interactions discovered

- 550,000 negative interactions
- 350,000 positive interactions

Costanzo et al. 2010 *Science*
Costanzo et al. 2016 *Science*

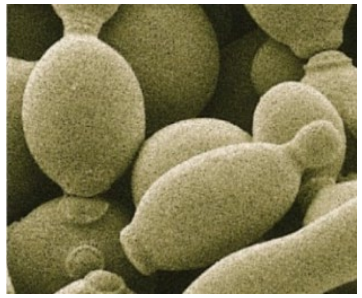


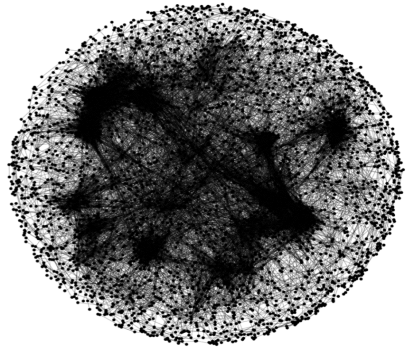


Baker's yeast
(*Saccharomyces cerevisiae*)

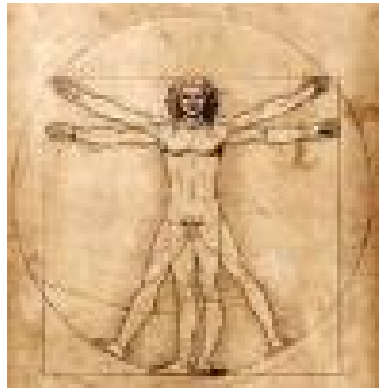
With Charlie Boone, Brenda Andrews labs (U. Toronto)

Recent focus: translating genetic interaction mapping/analysis approaches to human cells





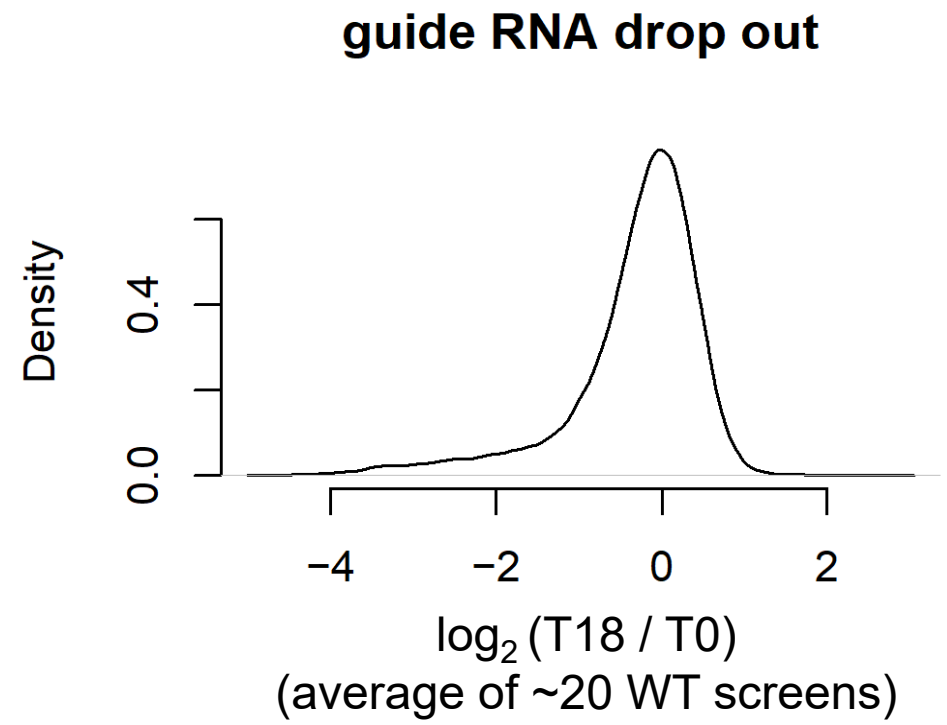
?



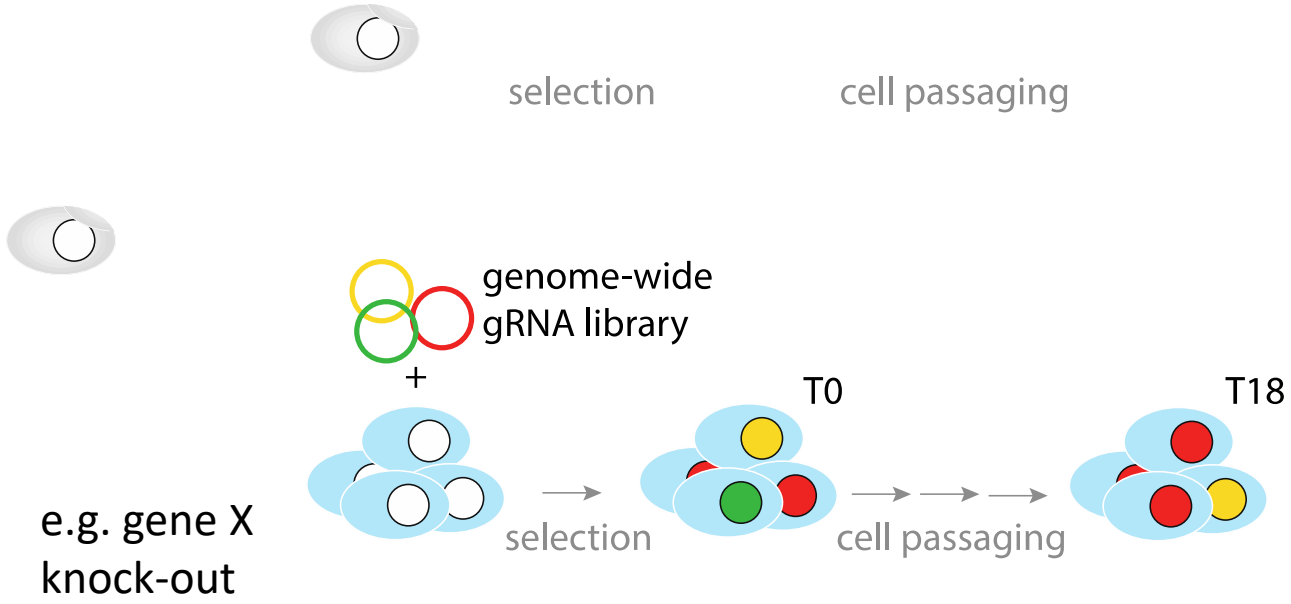
Enabling technology: CRISPR/Cas9 genome editing

CRISPR/Cas9 genome-wide screens

Fitness effect: $\log_2$ (T18 / T0)
“log fold-change"

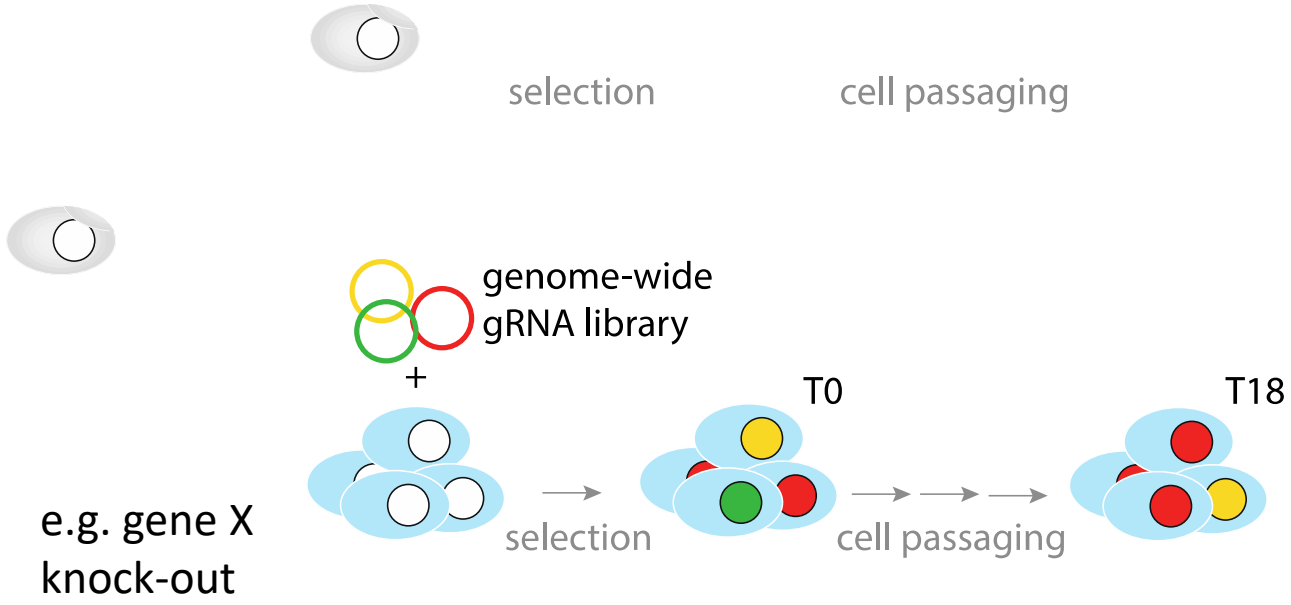


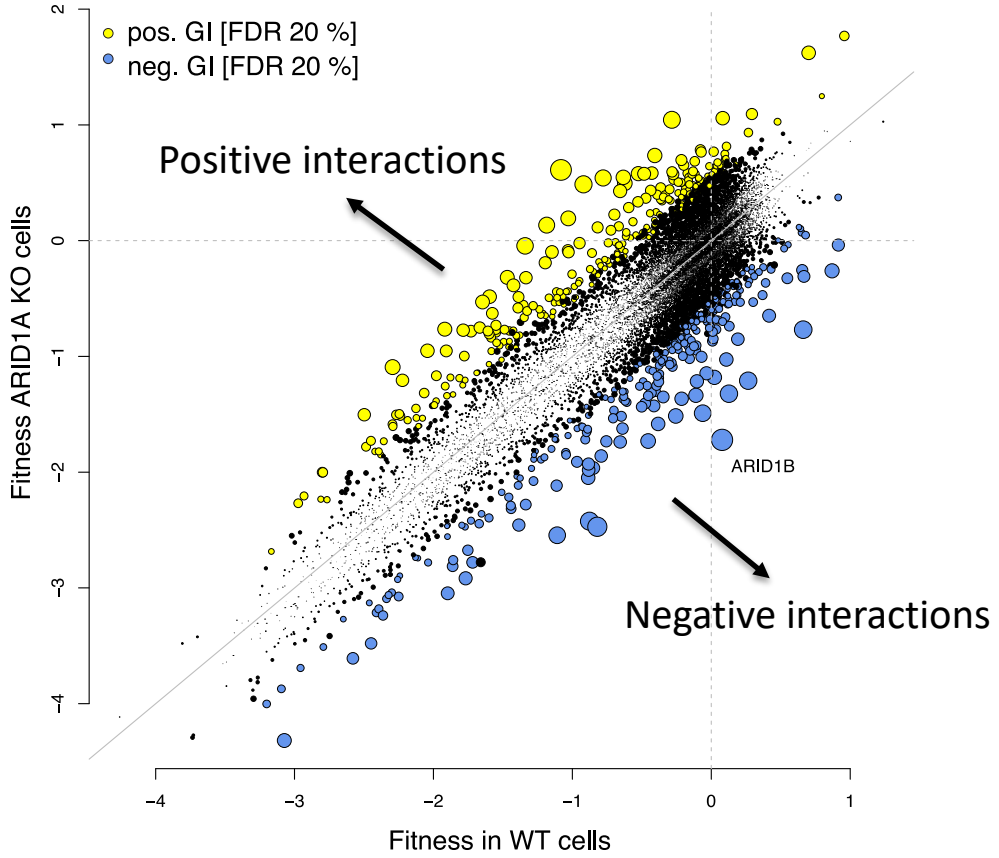
Generating genetic interactions with CRISPR/Cas9 screening



Reference: Arreger et al. Systematic mapping of genetic interactions for de novo fatty acid synthesis identifies *C12orf49* as a regulator of lipid metabolism. *Nature Metabolism* 2020.

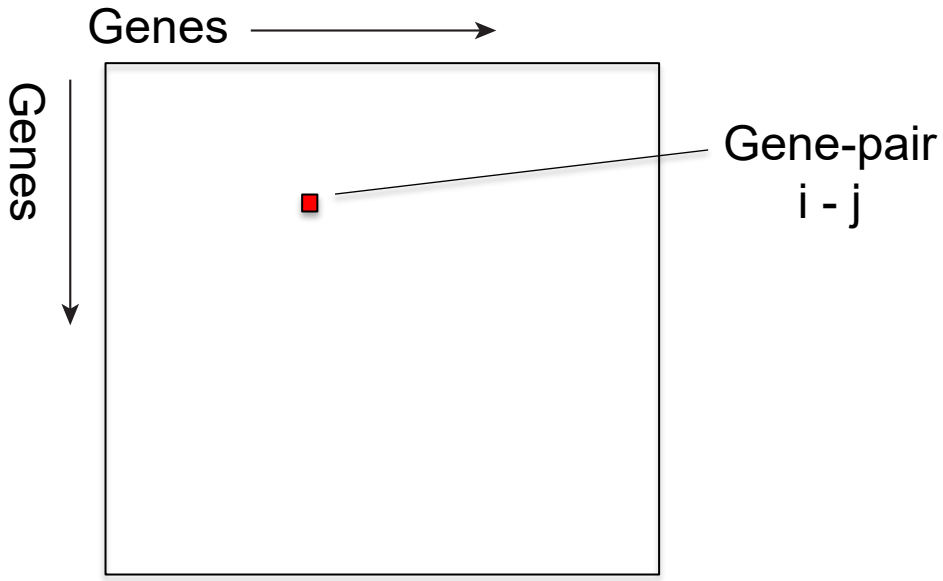
Generating genetic interactions with CRISPR/Cas9 screening

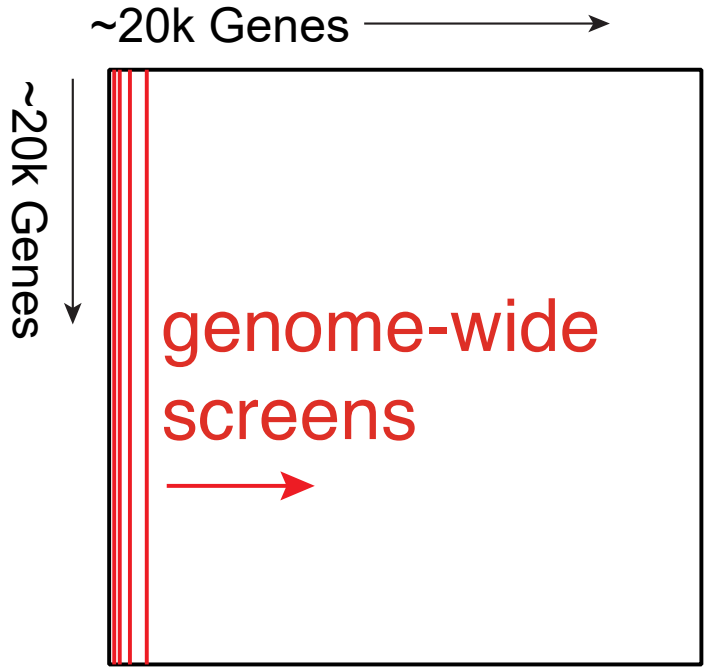




Reference: Arreger et al. Systematic mapping of genetic interactions for de novo fatty acid synthesis identifies *C12orf49* as a regulator of lipid metabolism. *Nature Metabolism* 2020.

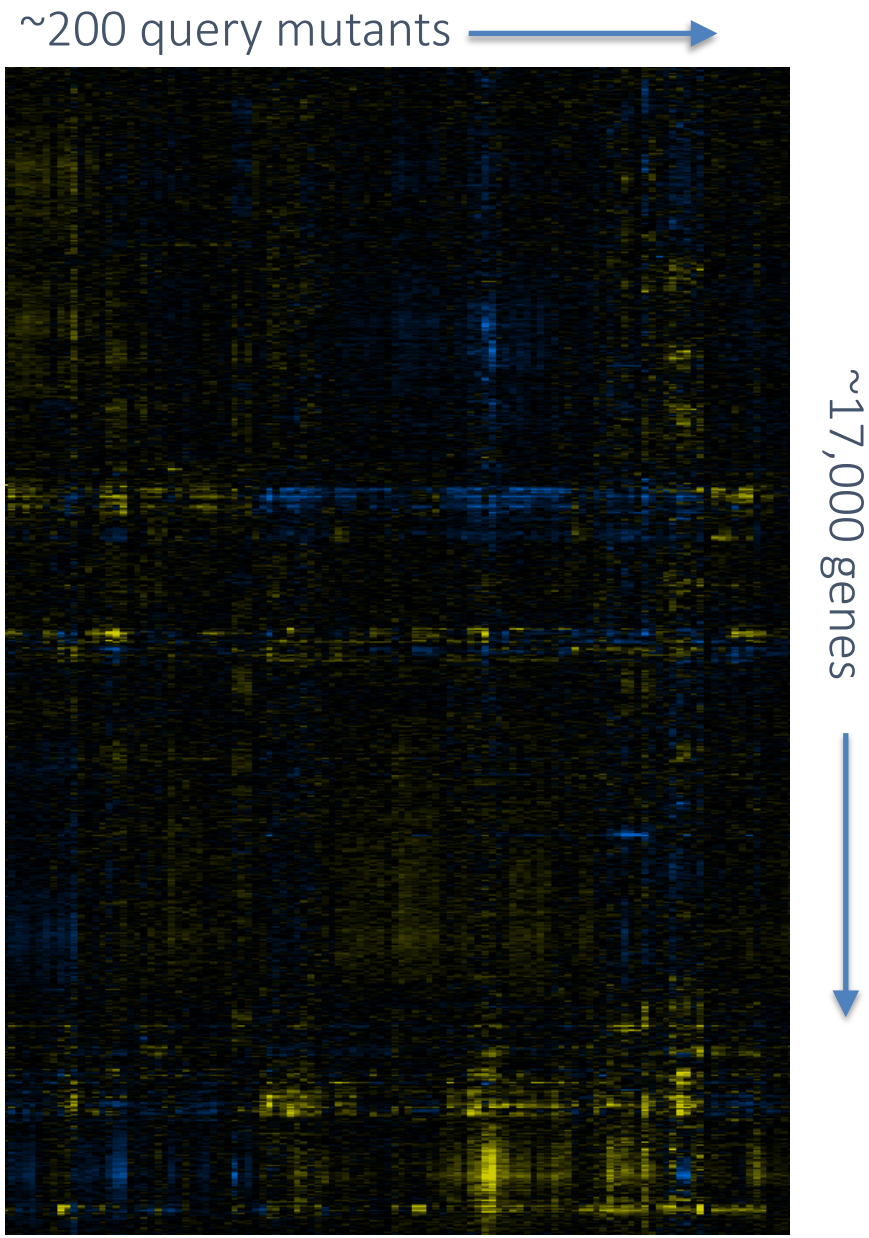
Toward a global human genetic interaction network





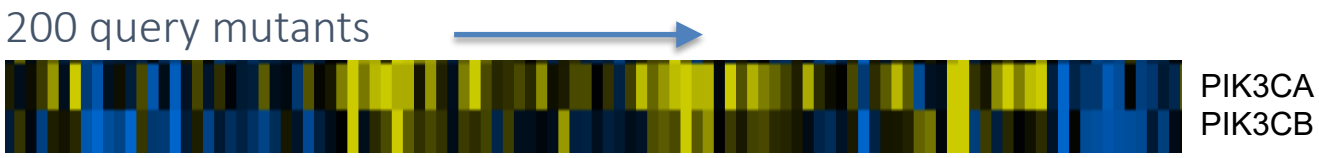
Element (i, j) – genetic interaction score for double mutant in genes i and j

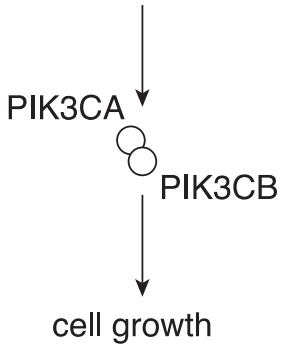
Progress toward a global human genetic interaction network

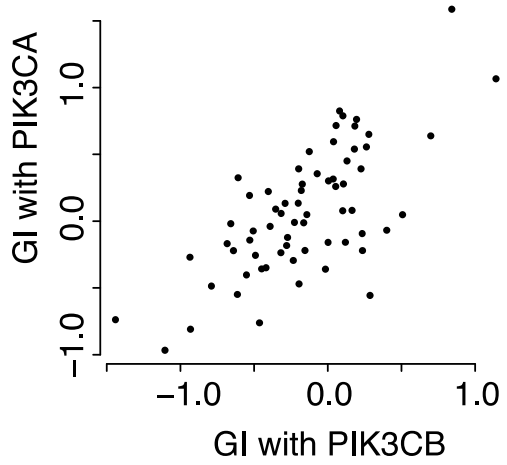


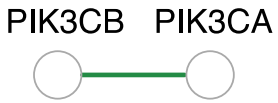
~200 query mutants x ~17,000 genes
~3.5 million gene pairs tested

Genetic interaction profile similarity identifies functionally related genes

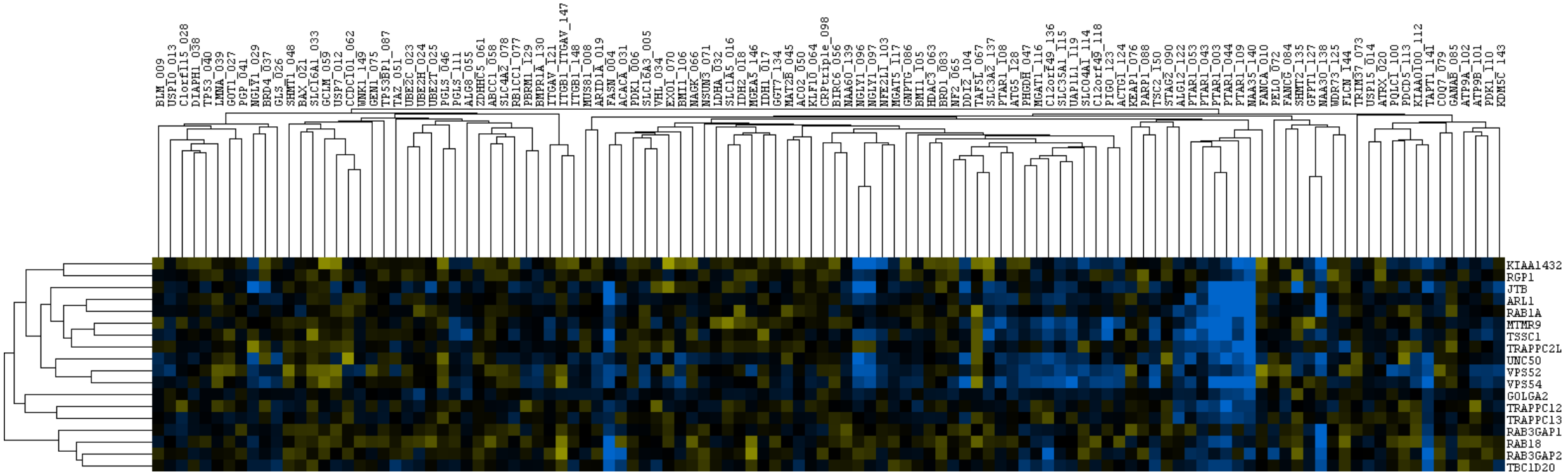






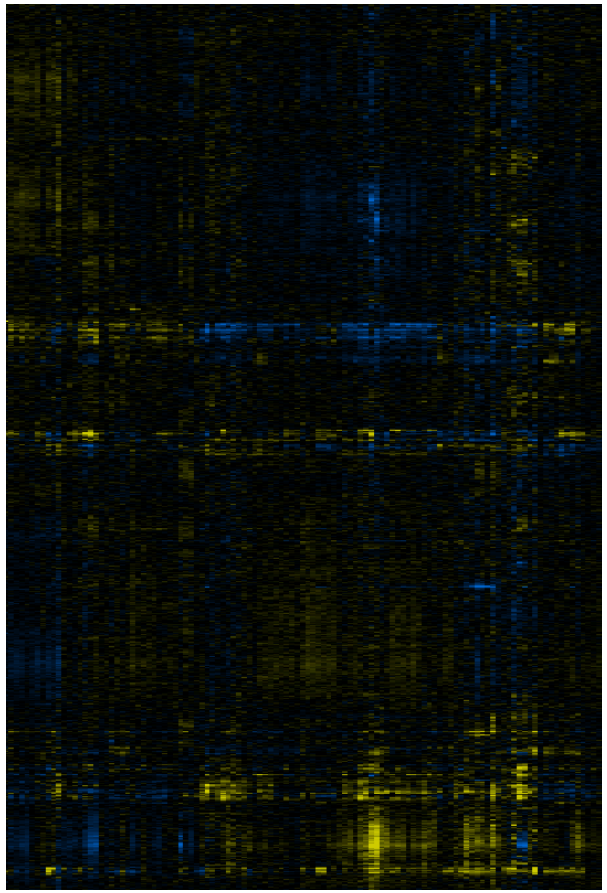


Example cluster: TRAPP complex/ER-Golgi transport



Prediction Challenge A

~200 query mutants →



~17,000 genes →

GO term annotation matrix:

~1,000 GO terms

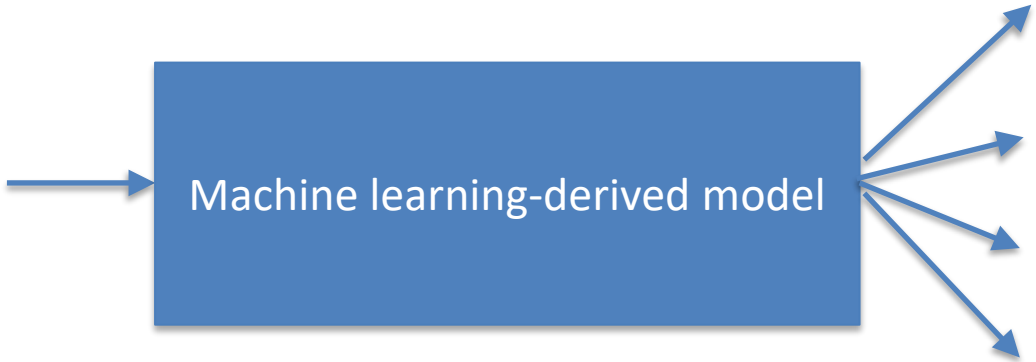
~17k genes

Binary GO annotations

Challenge: develop machine learning model to predict genes' functions for the 17k genes from the 1x200 interaction profile
Input: interaction matrix + GO term labels for a subset of genes
We'll withhold a subset of known gene→ GO term labels for validation

Prediction Challenge A

1 gene's 1x200 interaction profile



Output prediction score for each of the ~1000 GO term classes

Prediction Challenge B

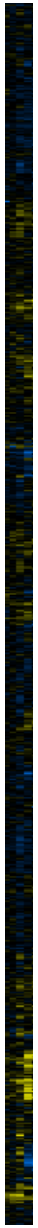
Gene X

1 x 200 library genetic interaction profile

Machine learning model

17k x 1 query genetic interaction profile

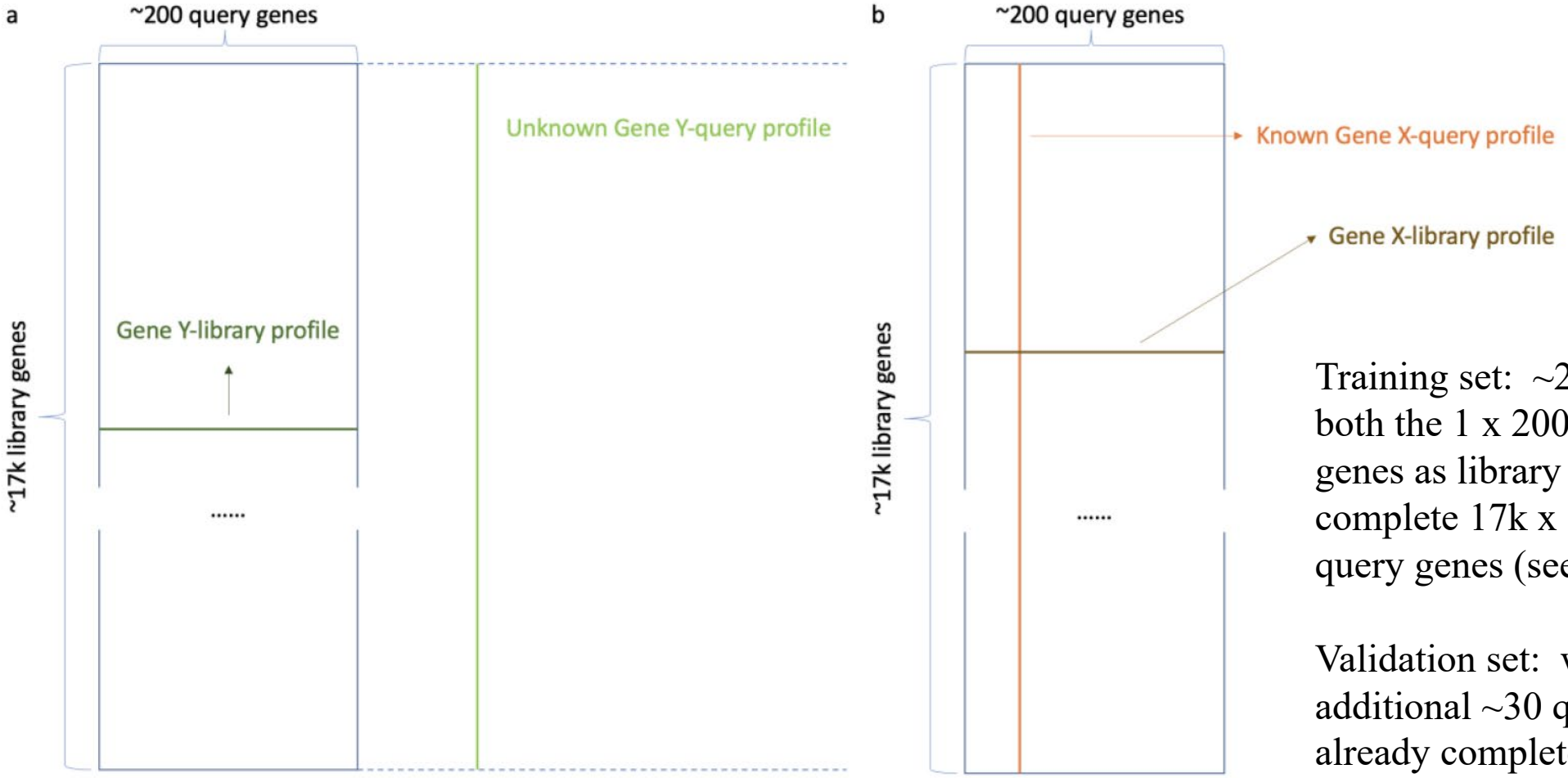
Gene X



Prediction Challenge B

(a) Gene Y only has its library profile available in our dataset and has not been screened as a query yet. The goal is to predict its query genetic interaction profile (17k x 1) based on its known library genetic interaction profile (1 x 200). (b) Gene X is present as both a query gene and library gene in our dataset.

Prediction Challenge B



Training set: ~200 genes which we have both the 1 x 200 interaction profile for those genes as library genes as well as their complete 17k x 1 interaction profile as query genes (see Gene X in Fig. 2b)

Validation set: we will withhold an additional ~30 query genes for which we've already completed genome-wide screens

Instructions for participating in these challenges

- If you choose to work on this project, you may work on one or both of these challenges (A, or B, or both)

- Email chadm@umn.edu and csci5461-help@umn.edu if your team would like to work on one or both challenges– we will provide you with the corresponding datasets

- Submission deadline for predictions: 5/1 at midnight (we will independently evaluate all groups' submissions on withheld data)

References

- Wang et al. Identification and characterization of essential genes in the human genome. Science. 2015 Nov 27;350(6264):1096-101. doi: 10.1126/science.aac7041.
- Meyers et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nature Genetics 2017 October 49:1779–1784. doi:10.1038/ng.3984.
- Costanzo et al. A global genetic interaction network maps a wiring diagram of cellular function. Science. 2016 Sep 23;353(6306). pii: aaf1420.
- Arreger et al. Systematic mapping of genetic interactions for de novo fatty acid synthesis identifies *C12orf49* as a regulator of lipid metabolism. *Nature Metabolism* 2020.