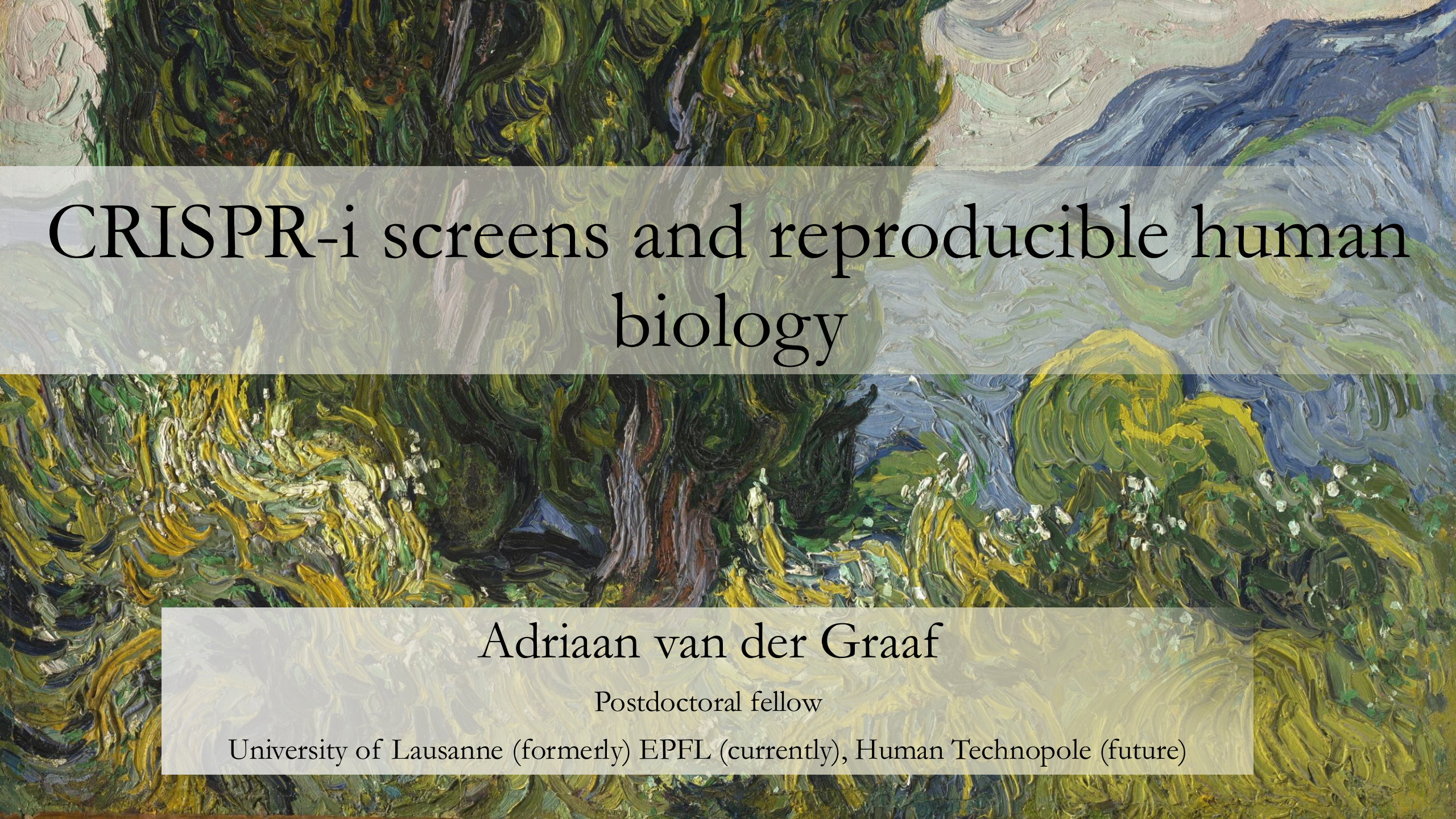


CRISPR-i screens and reproducible human biology

Adriaan van der Graaf

Postdoctoral fellow

University of Lausanne (formerly) EPFL (currently), Human Technopole (future)

About me

- PhD, Statistical Genetics
- 2017 - 2021

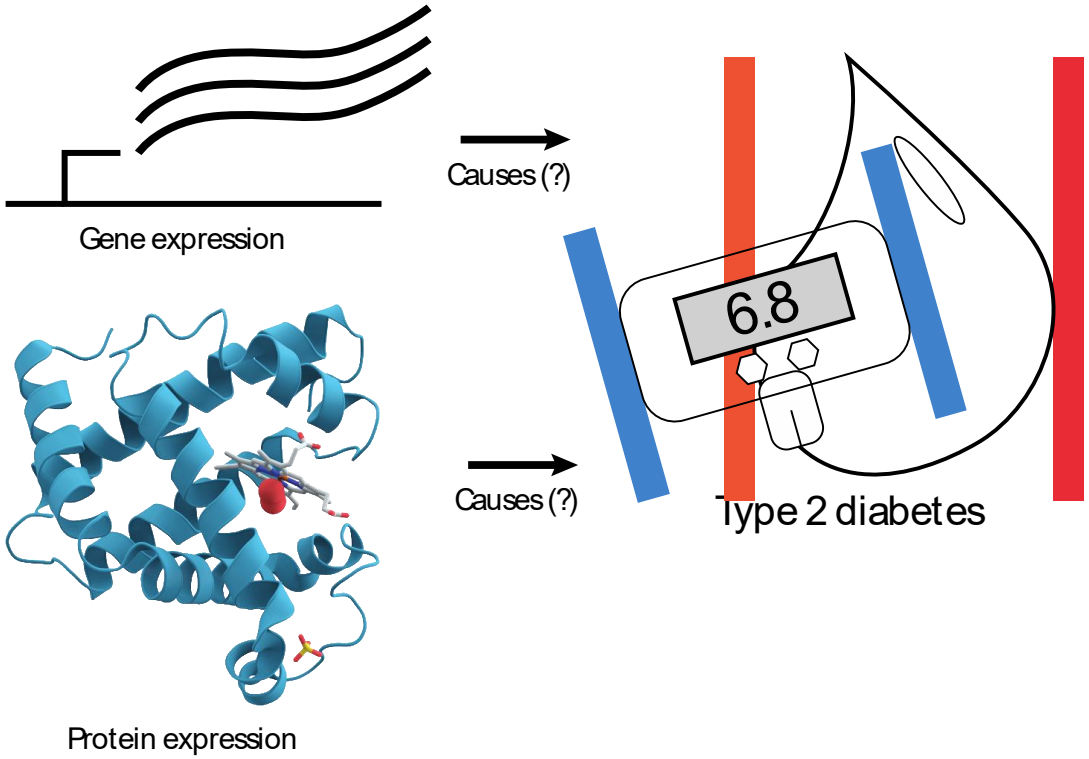
umcg:

university of groningen



- Postdoc Statistical Genetics
- 2021 – 2026 SIB Swiss Institute of Bioinformatics Unil.
- 2026 - 2027 EPFL

The questions I answer



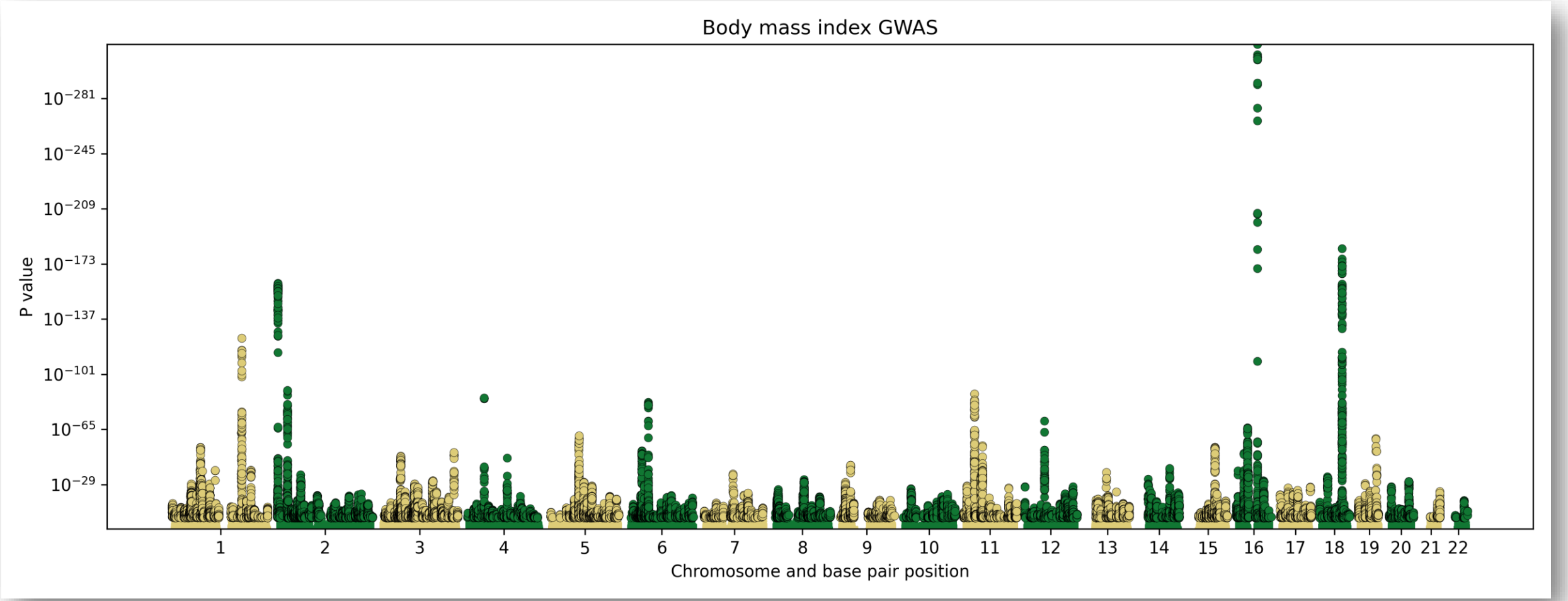
biobank uk

lifelines 20

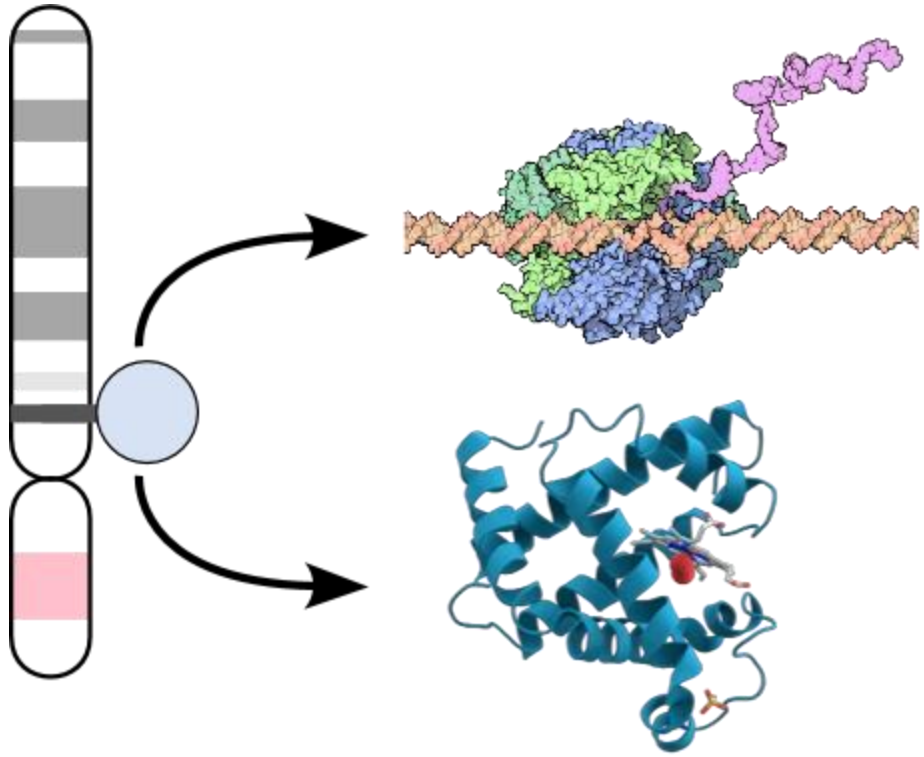
clsa élcv

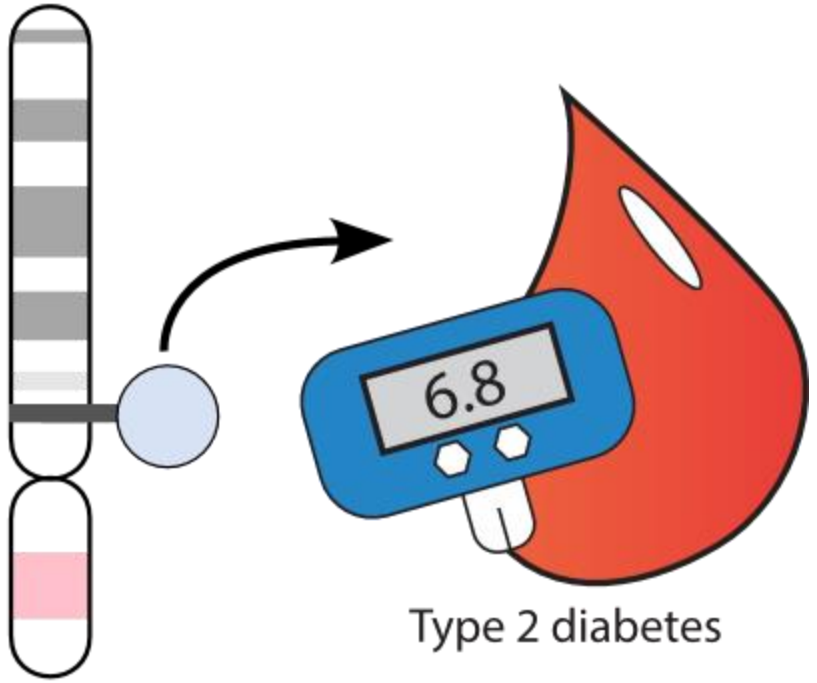
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Genome wide association studies (GWAS) map variants to traits



Matching disease loci to molecules

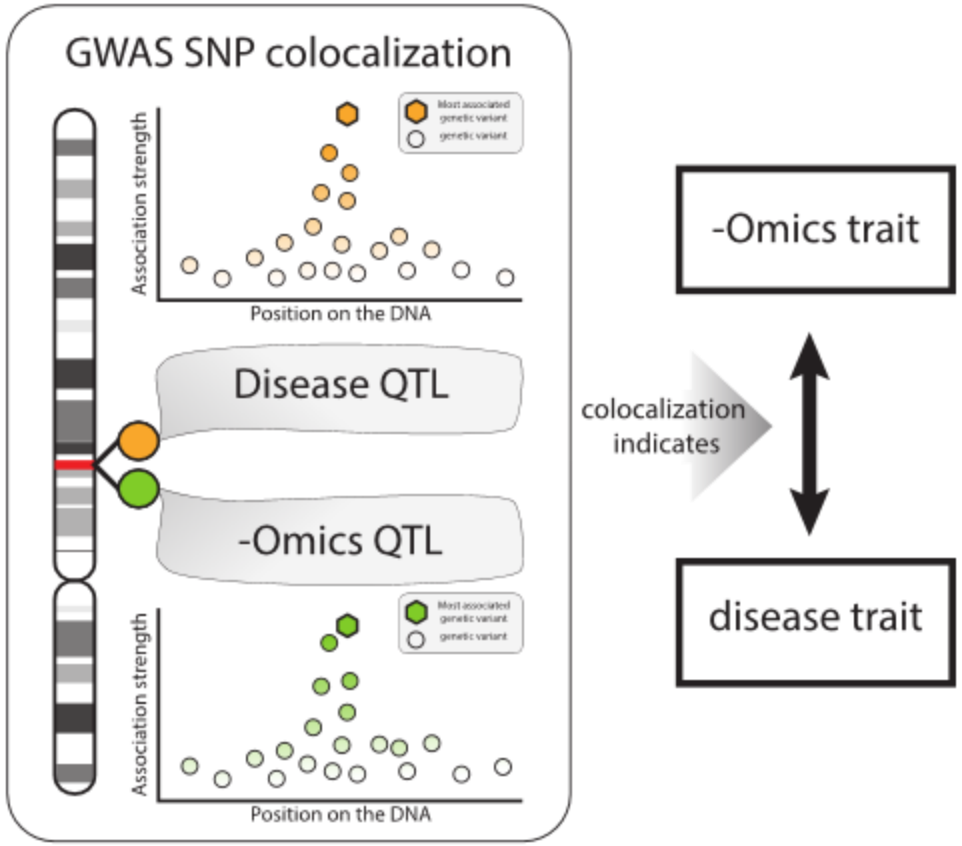




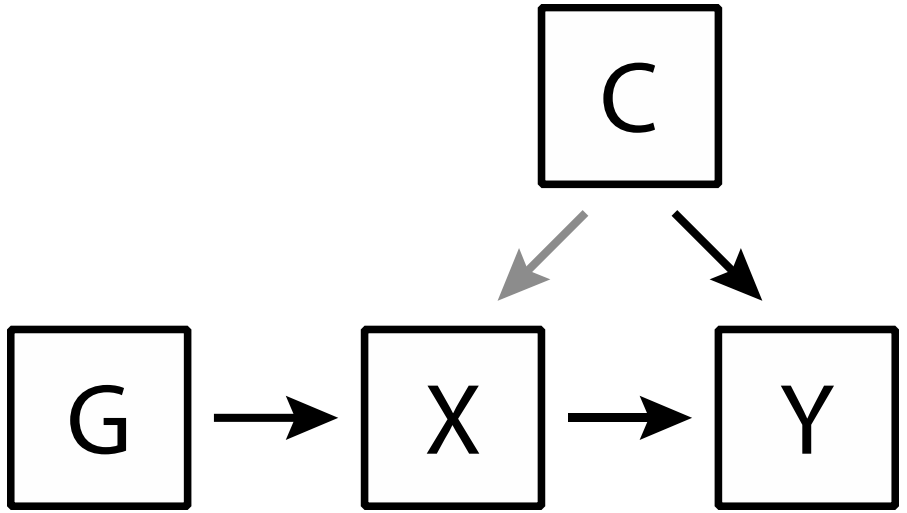
How to map disease loci to effector molecules?

Genetic methods that match loci

Colocalization



Mendelian randomization (MR)



Using genetics (G) to identify causality between X and Y, independent of confounders

GWAS loci identify important genes

Analysis

Refining the impact of genetic evidence on clinical success

https://doi.org/10.1038/s41586-024-07316-0

Received: 5 July 2023

Accepted: 14 March 2024

Published online: 17 April 2024

Open access

Check for updates

Eric Vallabh Minikel[1], Jeffery L. Painter[2,5], Coco Chengliang Dong[3] & Matthew R. Nelson[3,4]✉

The cost of drug discovery and development is driven primarily by failure[1], with only about 10% of clinical programmes eventually receiving approval[2–4]. We previously estimated that human genetic evidence doubles the success rate from clinical development to approval[5]. In this study we leverage the growth in genetic evidence over the past decade to better understand the characteristics that distinguish clinical success and failure. We estimate the probability of success for drug mechanisms with genetic support is 2.6 times greater than those without. This relative success varies among therapy areas and development phases, and improves with increasing confidence in the causal gene, but is largely unaffected by genetic effect size, minor allele frequency or year of discovery. These results indicate we are far from reaching peak genetic insights to aid the discovery of targets for more effective drugs.

The questions I answer come from '**hypothesis free**' data

Benefits of hypothesis free data

- Unexpected findings
- Generating of hypotheses
- Comprehensive view
- Efficient use of data

- Is **cheap** when the data is already generated

Drawbacks of hypothesis free data

- Can be hard to interpret
- Difficult to identify causality
- Can be misused through P hacking

- Is **expensive** when the data hasn't been generated yet

We need to reproduce research before we can trust the science

nature.com/articles/d41586-021-03691-0

NEWS | 09 December 2021

Half of top cancer studies fail high-profile reproducibility effort

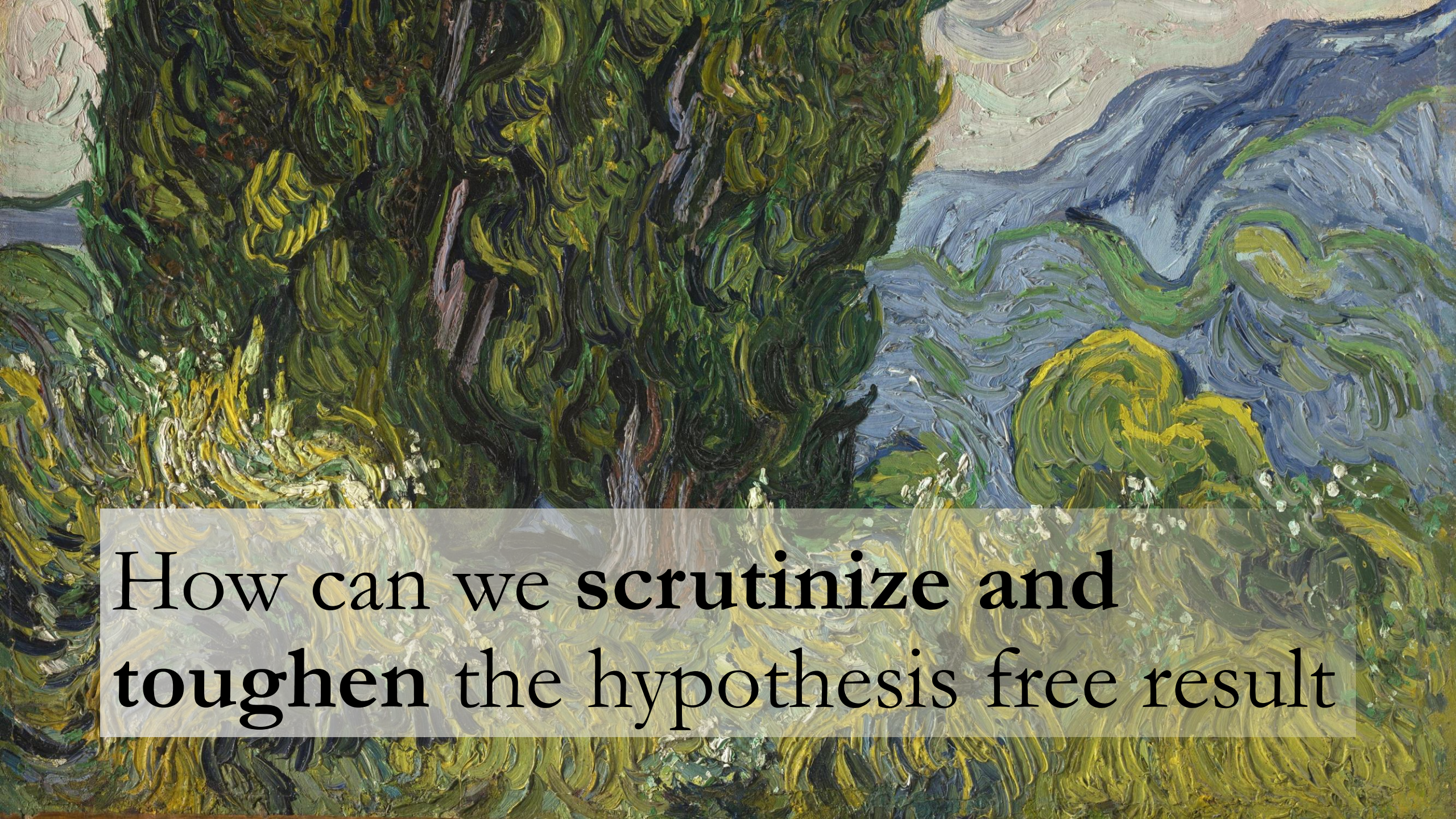
Barriers to reproducing preclinical results included unhelpful author communication, but critics argue that one-time replication attempts don't tell the whole story.

By Asher Mullard

Half of top cancer studies fail high-profile reproducibility effort

Save article





How can we **scrutinize and toughen** the hypothesis free result

What kind of strategies can we use to rethink our problem?

- Try to **break the result**
- **Test for alternatives**
- Replicate it independently

- Perform orthogonal experiments to **triangulate the evidence**



Nature **553**, 399-401 (2018)
doi: https://doi.org/10.1038/d41586-018-01023-3

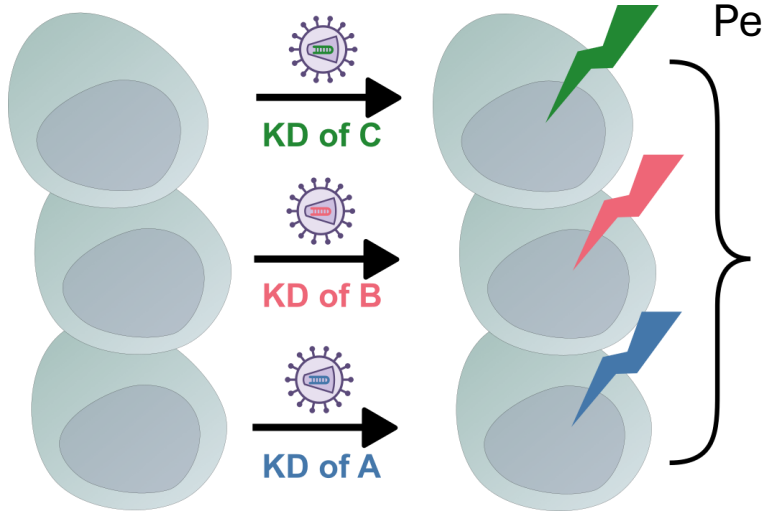
Perturb-seq



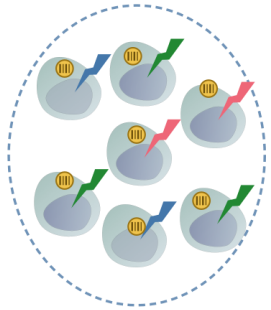
Sydney Fleming

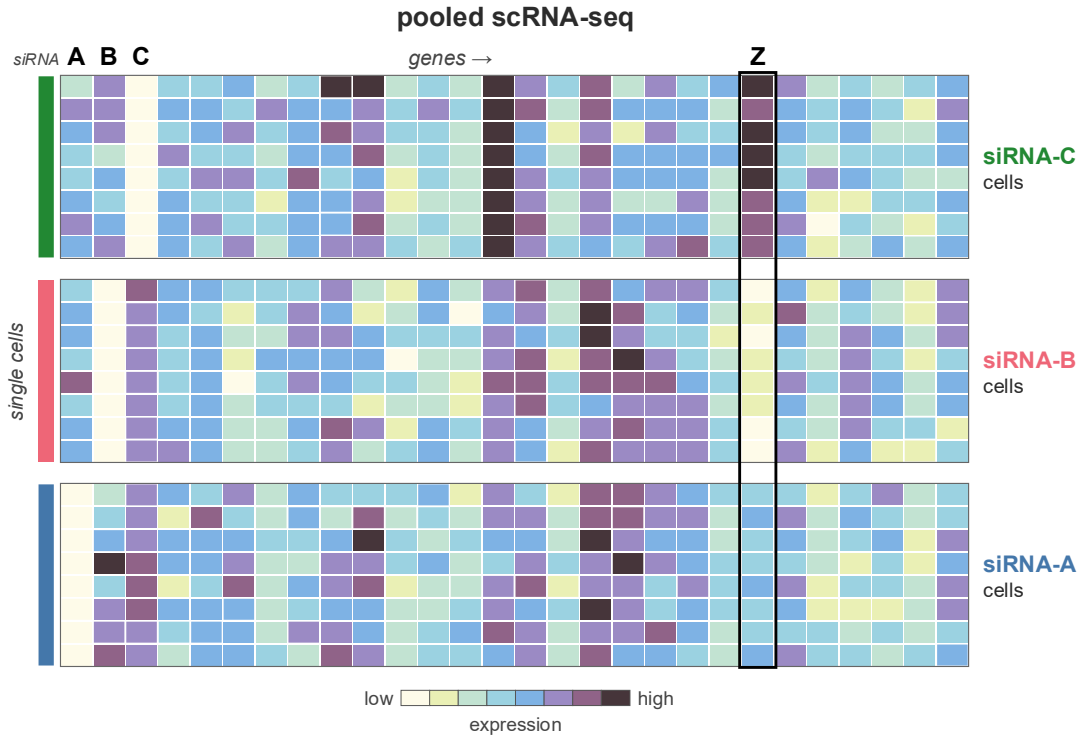
Introducing **Perturb-Seq**

Infect cells with a library of CRISPRi containing viruses



Perform single cell RNA-seq

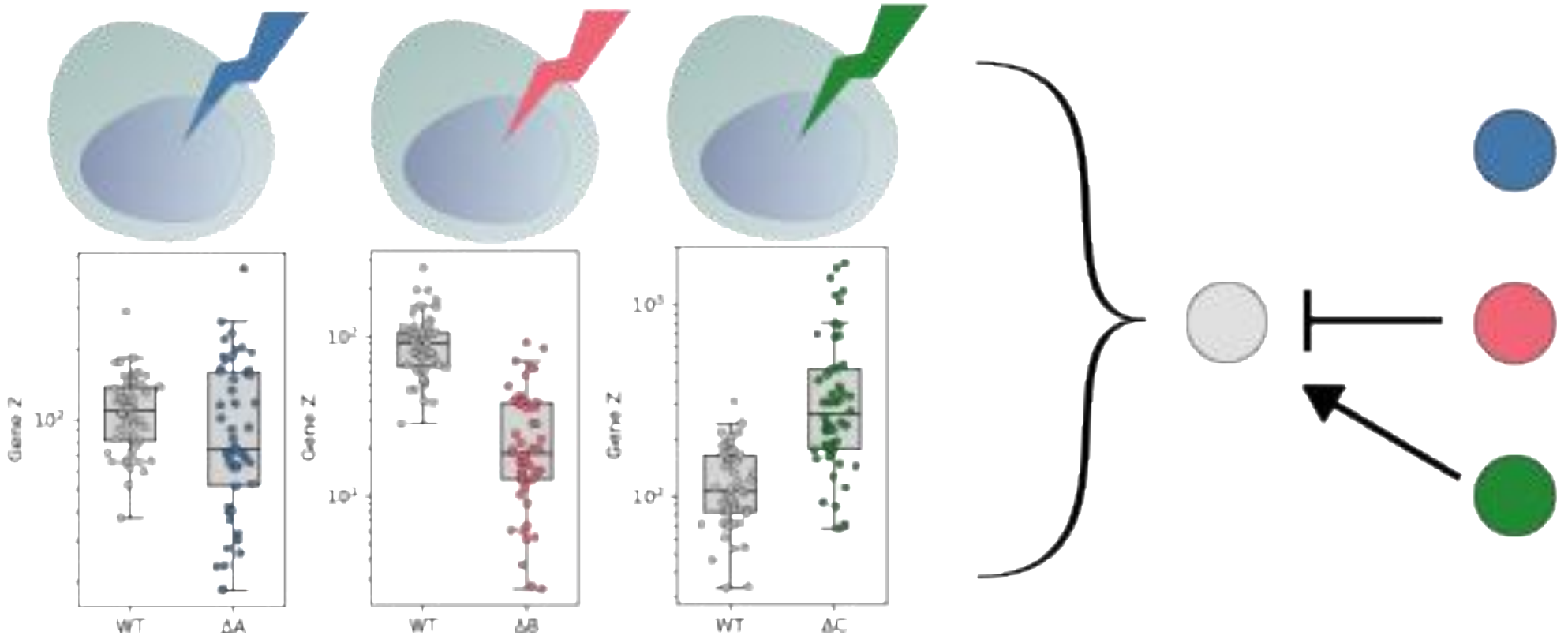




Perturb-seq is also known as crop seq

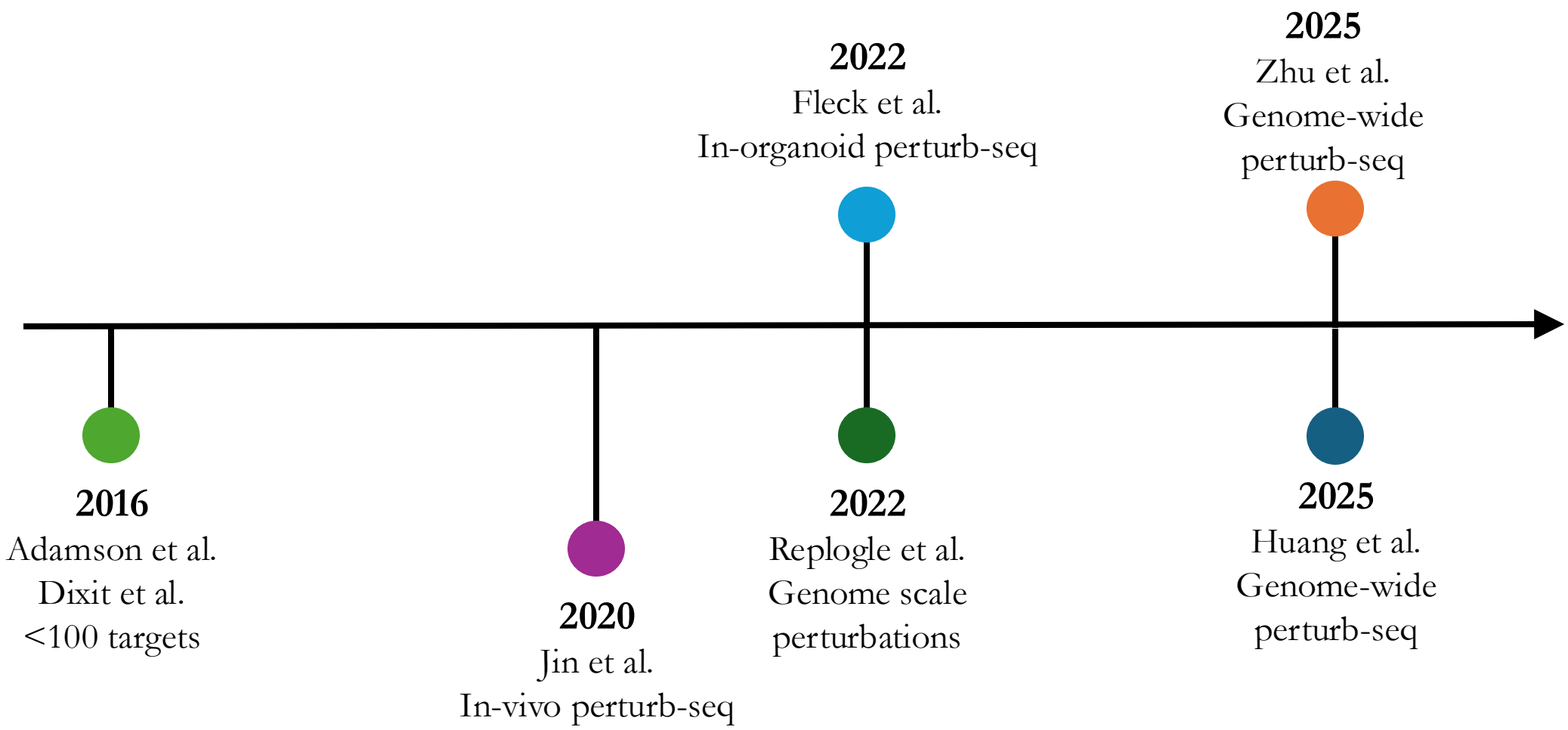
Simulated data

Perturb-seq network example



Simulated data

Perturb-seq is diverse and growing



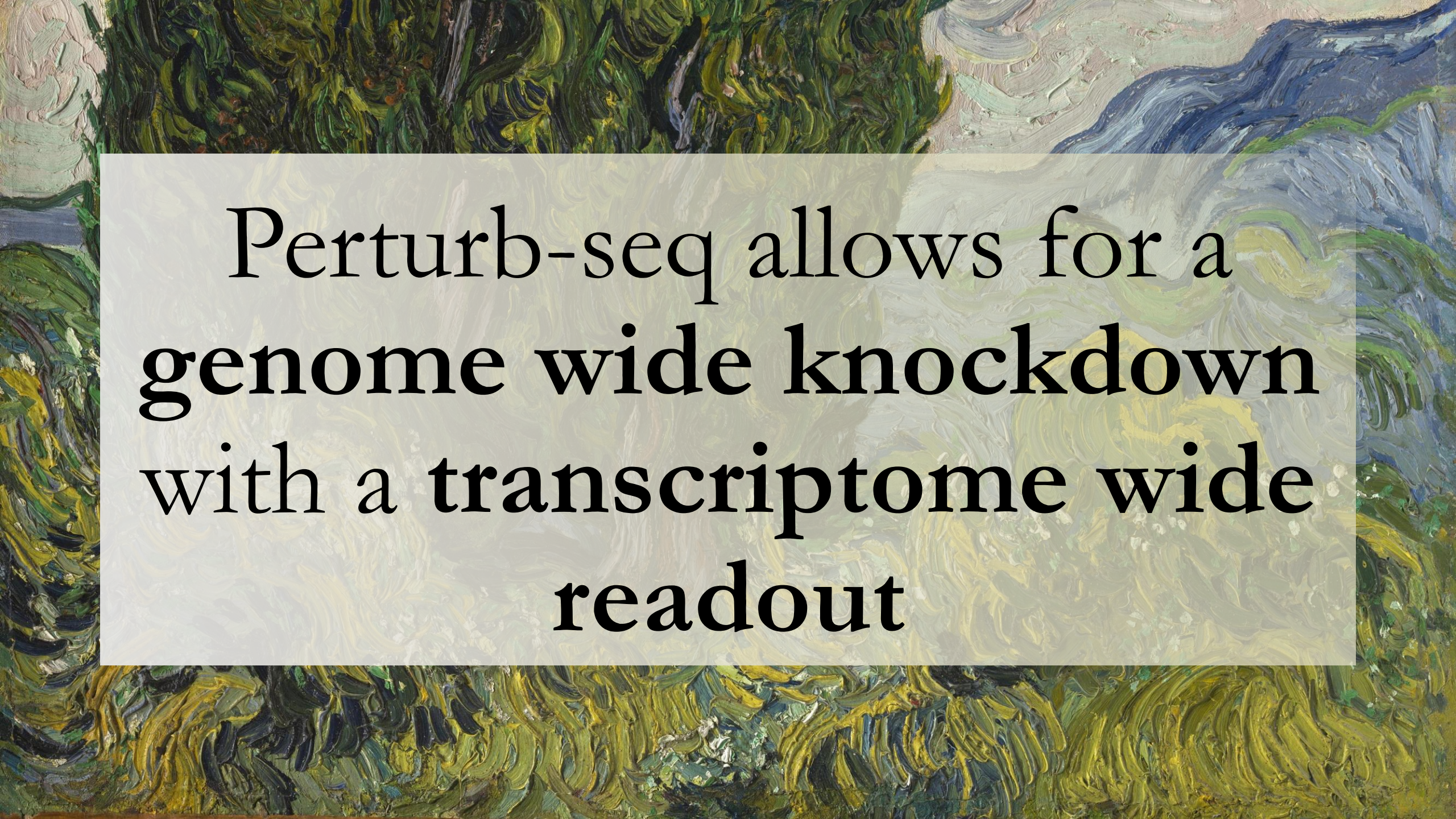
Anecdote from ASHG 2026

- Some very large institutes have started to put a lot of effort in generating perturb-seq libraries



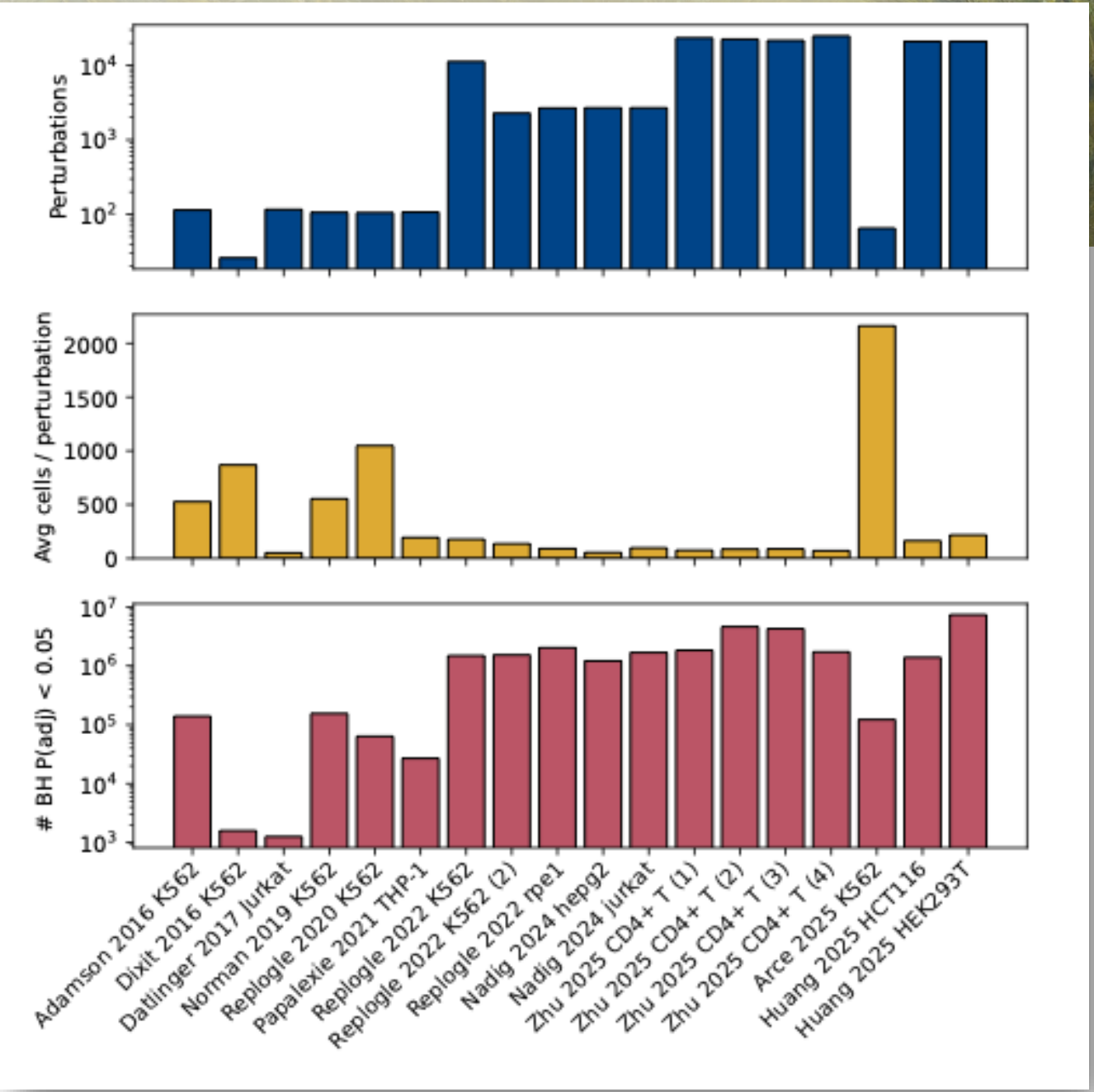


ASHG
ANNUAL MEETING
MONTRÉAL
OCTOBER 20-24, 2026


Perturb-seq allows for a
genome wide knockdown
with a transcriptome wide
readout

Perturb-seq data

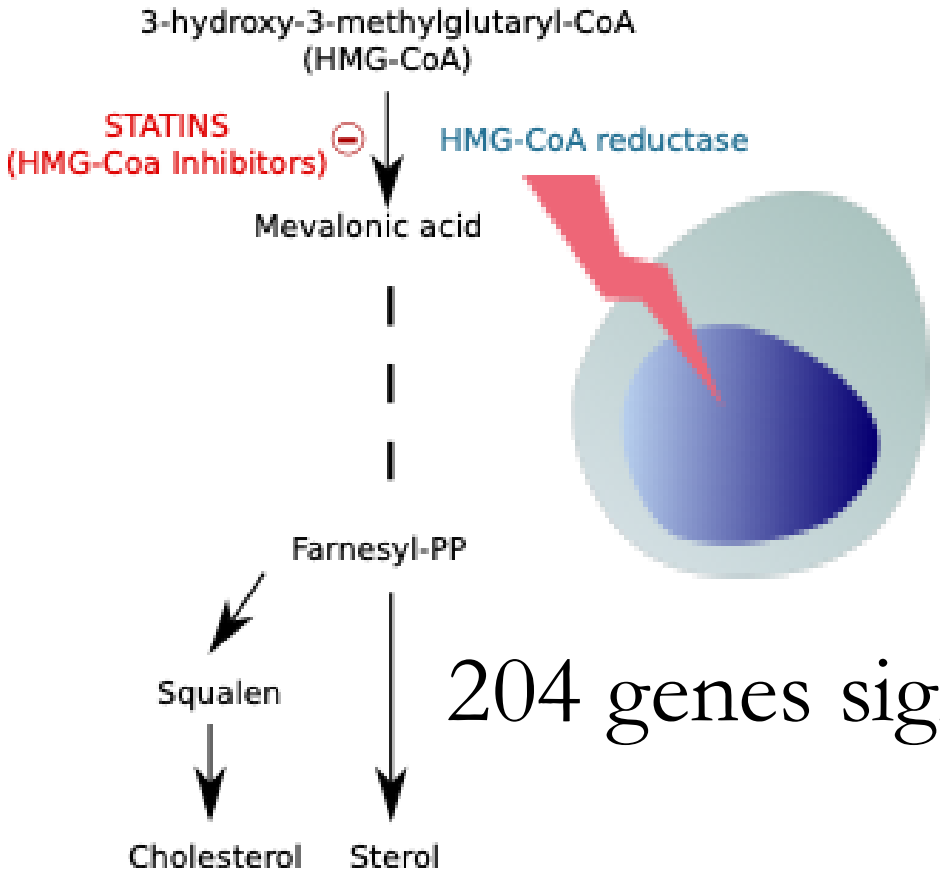
- 18 datasets from 11 studies
- 6 cell lines & CD4+ T patient derived cells
- 3.7 billion total perturbation-expression combinations
- 19,129 distinct perturbed genes



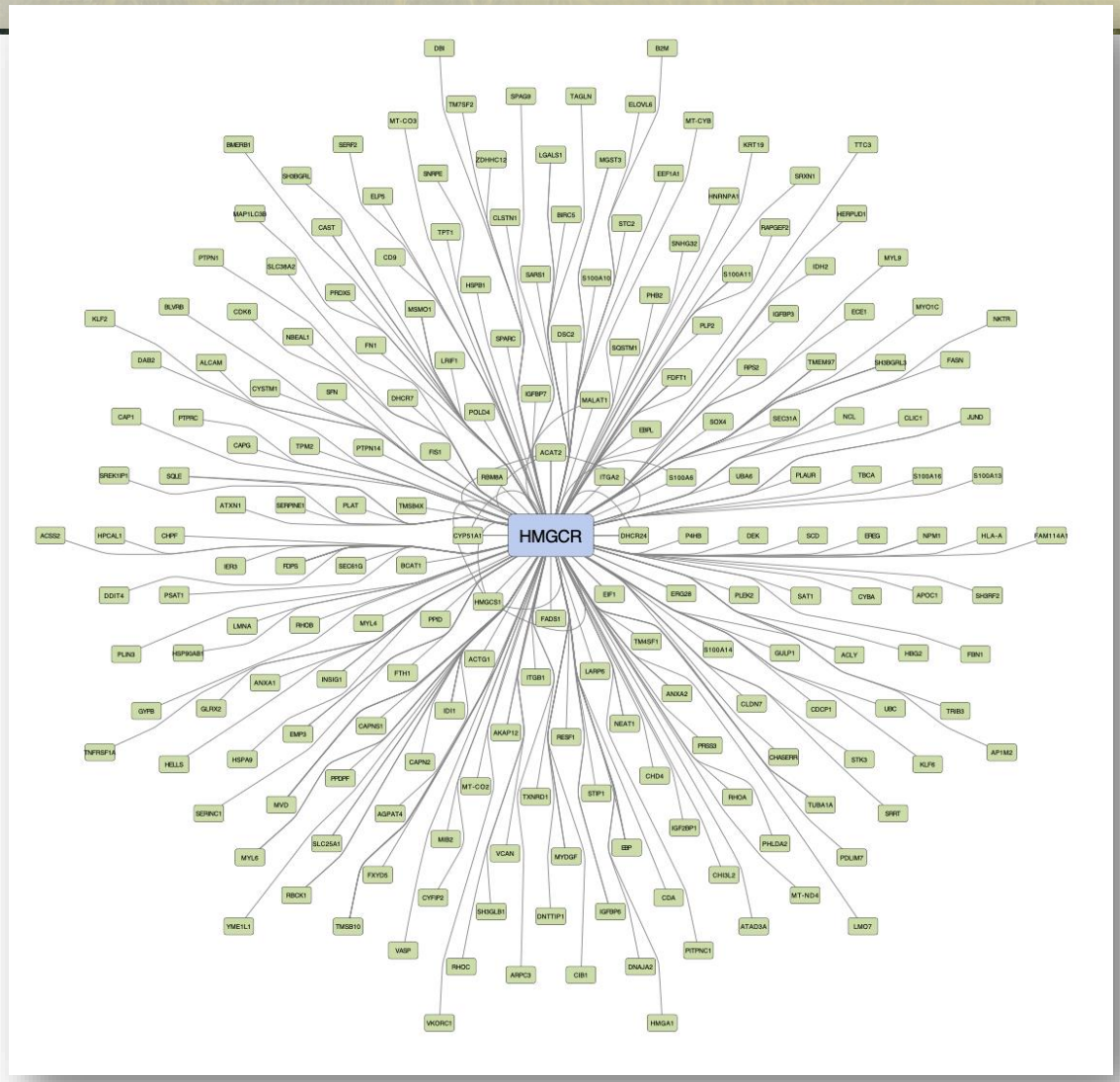


Sydney Fleming

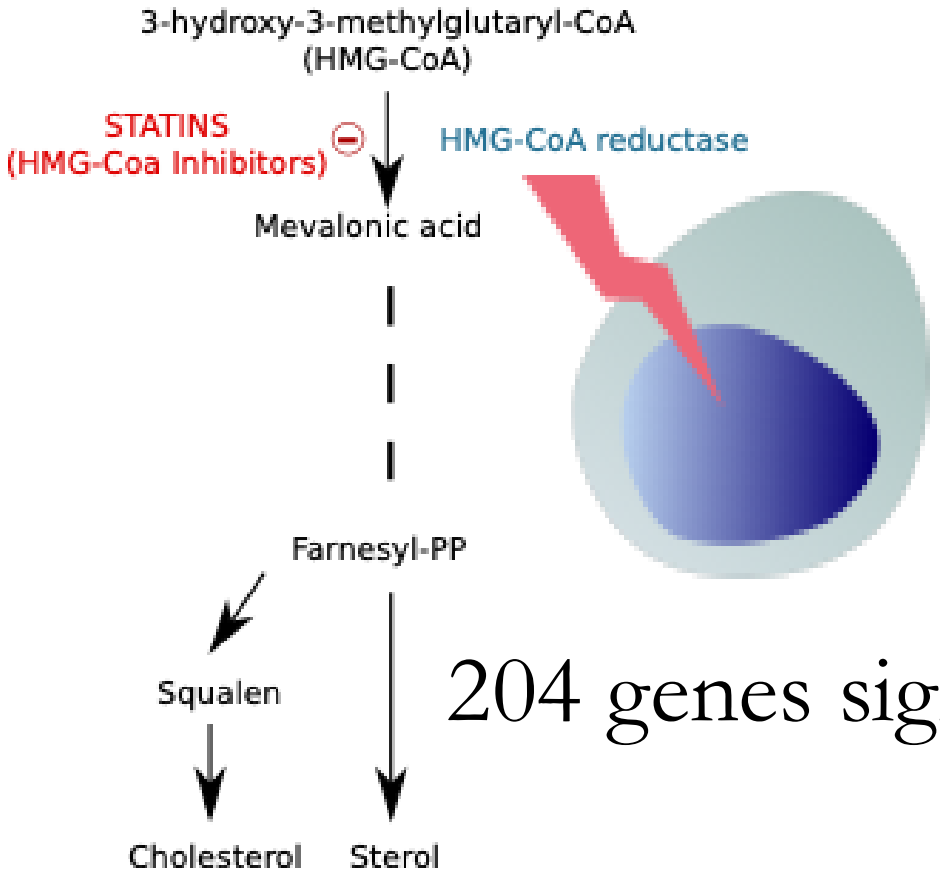
Perturb-seq results we identified



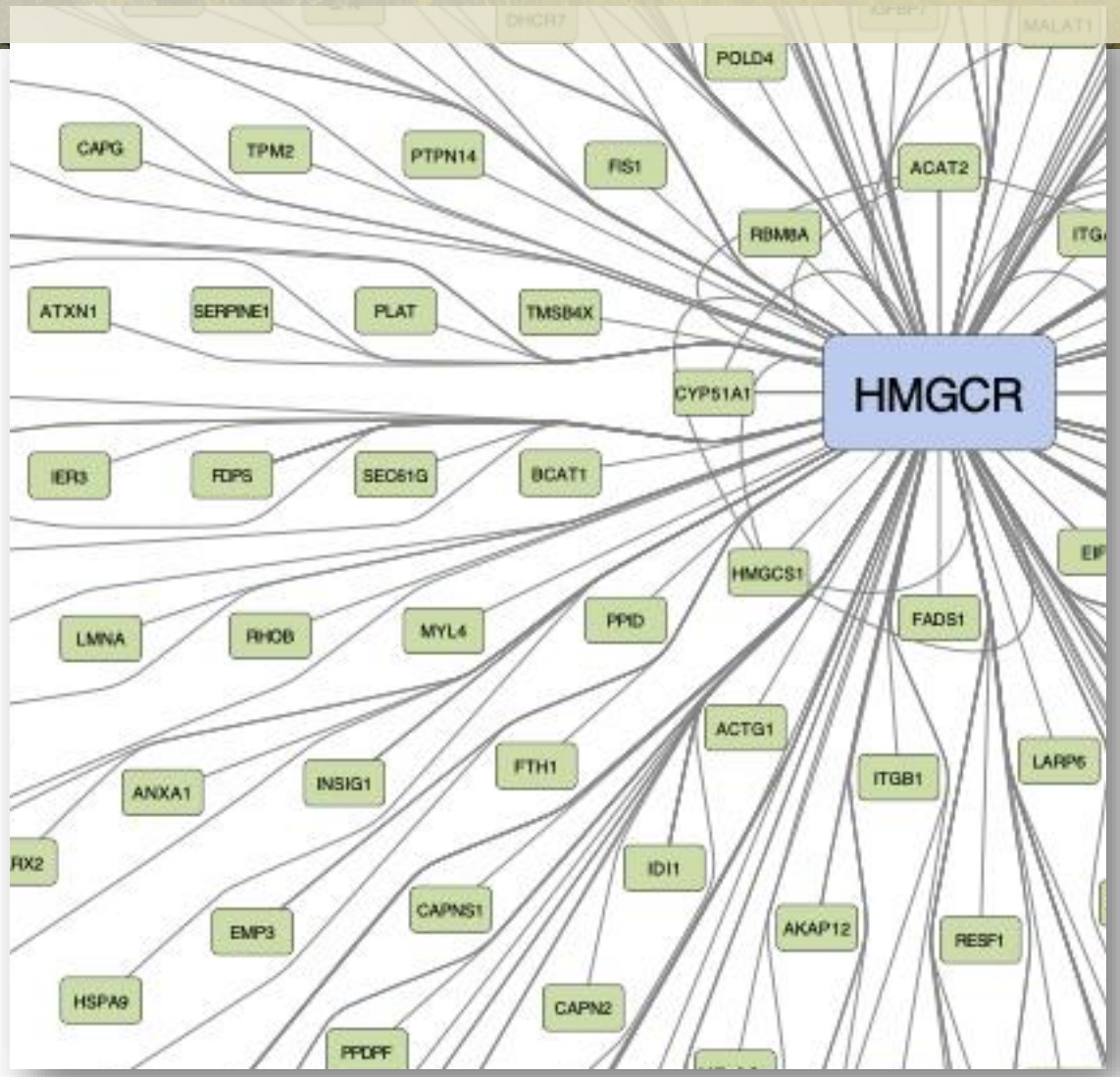
204 genes significant



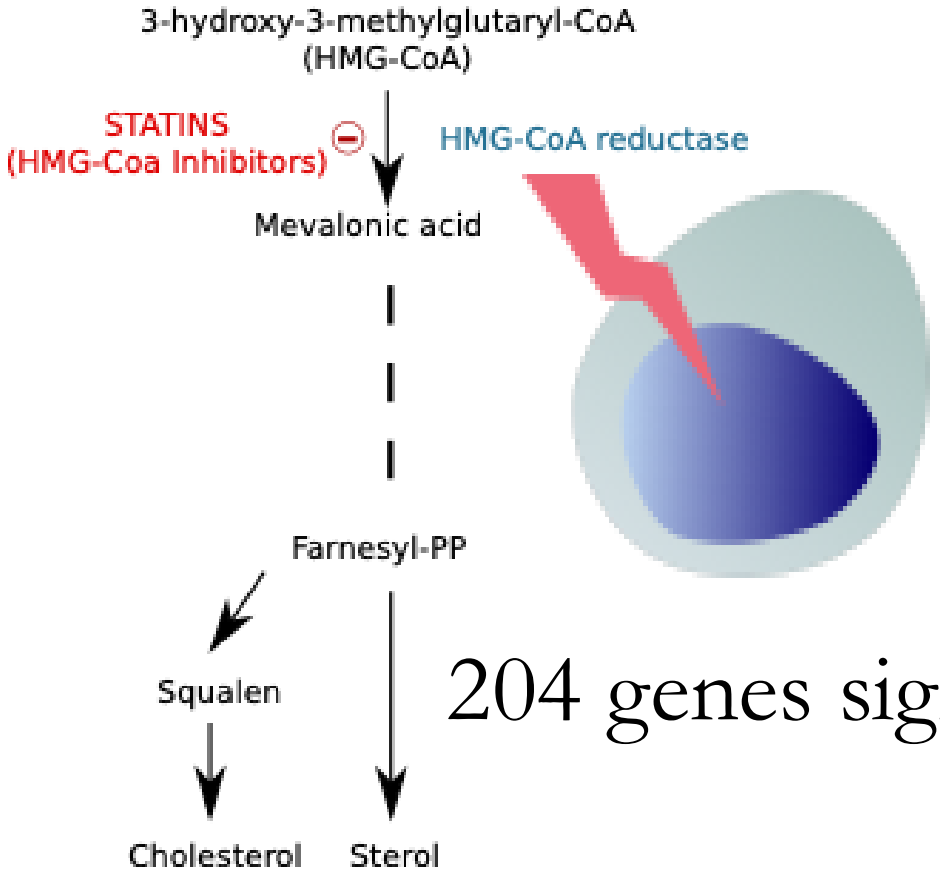
Perturb-seq results we identified



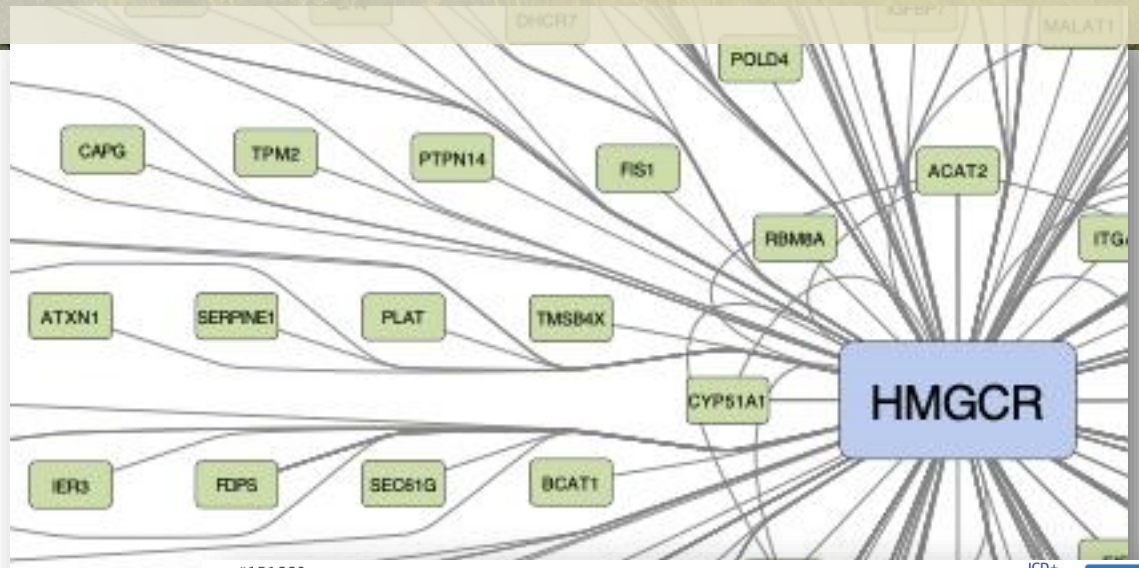
204 genes significant

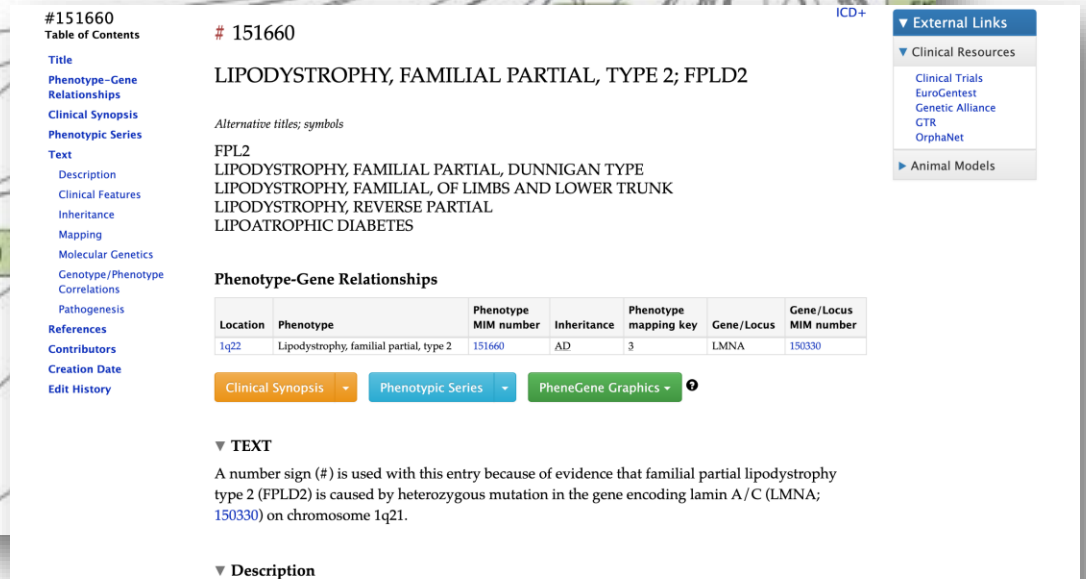


Perturb-seq results we identified

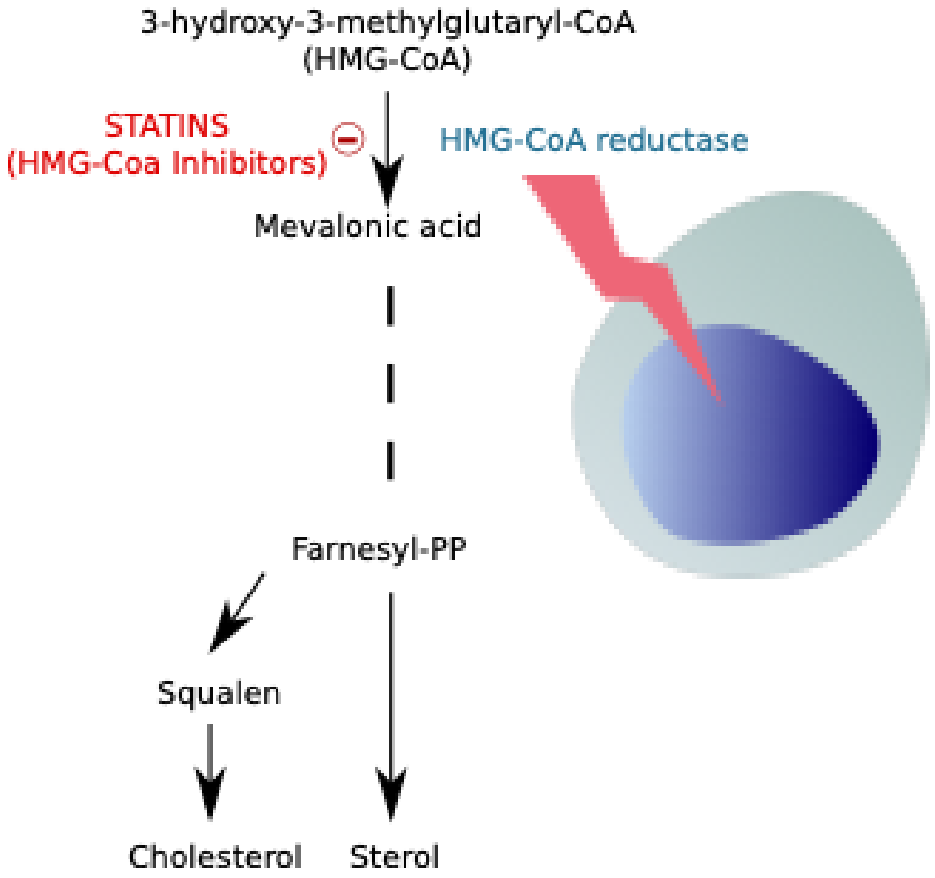


204 genes significant





Perturb-seq results we identified



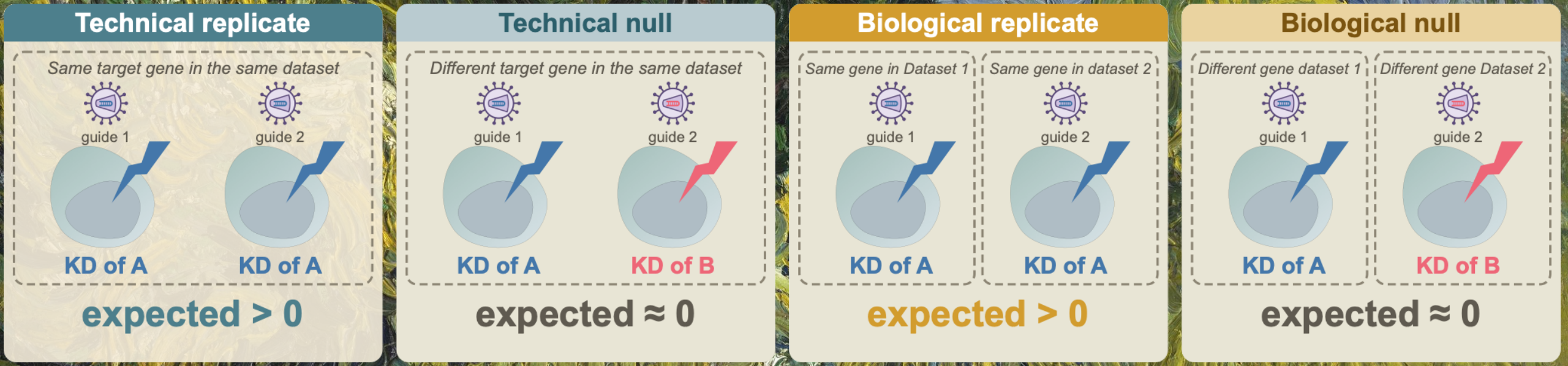
REACTOME (PWYs)

Enriched term	P (BH adj)
Cholesterol Biosynthesis	$1.54 \cdot 10^{-18}$
Activation of Gene Expression by SREBF (SREBP)	$1.15 \cdot 10^{-13}$
Regulation of Cholesterol Biosynthesis by SREBP (SREBF)	$1.15 \cdot 10^{-13}$
Metabolism of Steroids	$7.65 \cdot 10^{-12}$
Cellular Responses to Stimuli	$1.05 \cdot 10^{-7}$

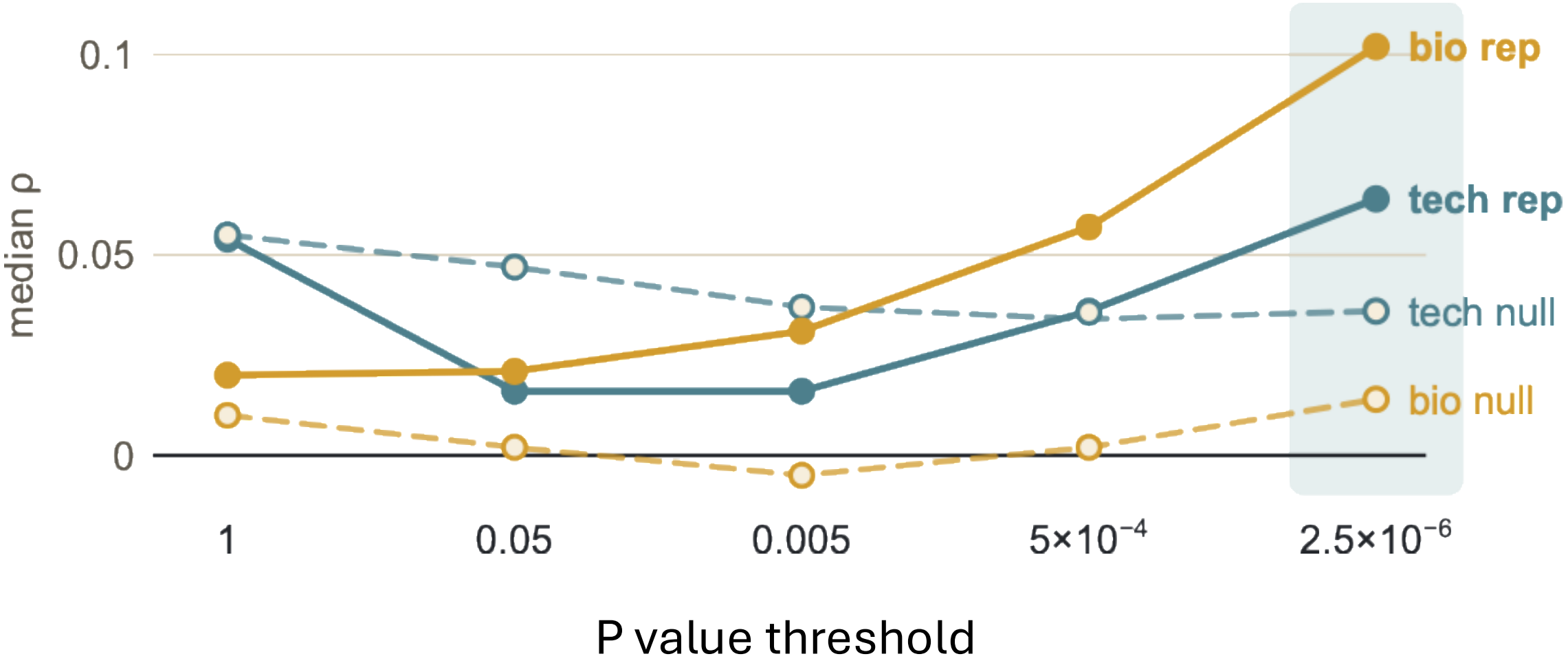
OMIM Disease

Enriched term	P (BH adj)
Leukemia	0.017
Cholesterol level	0.032

Benchmarking perturb-seq through biological and technical replicates



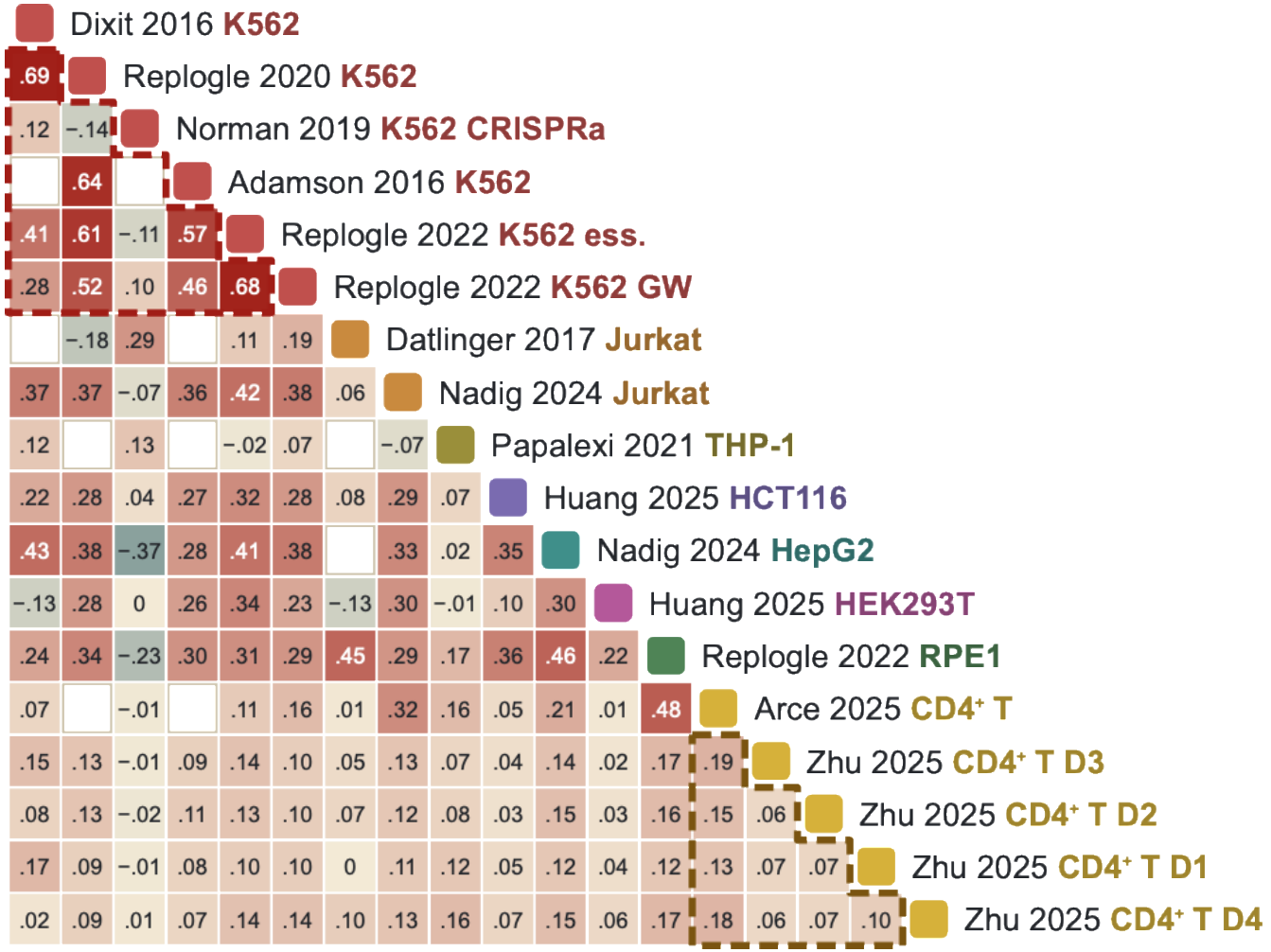
Only strict cutoffs replicate



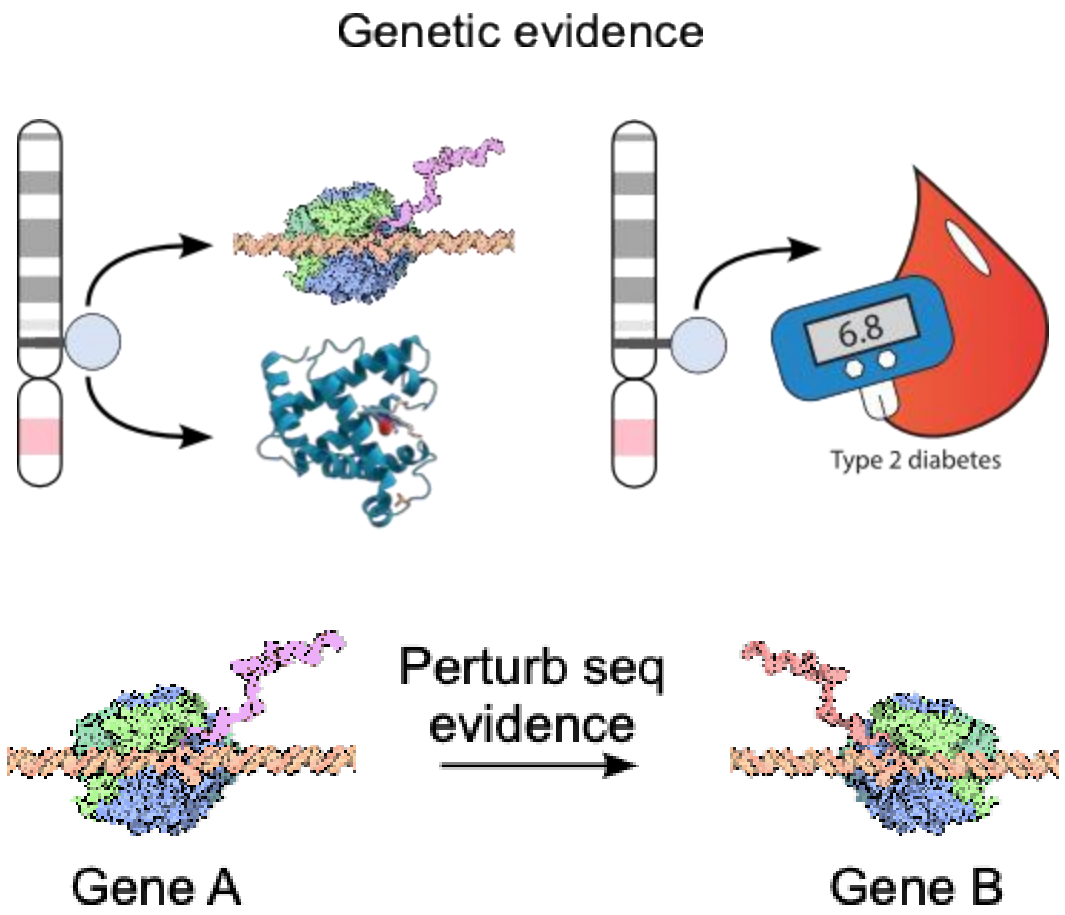
Replication between studies is dependent on context and lab

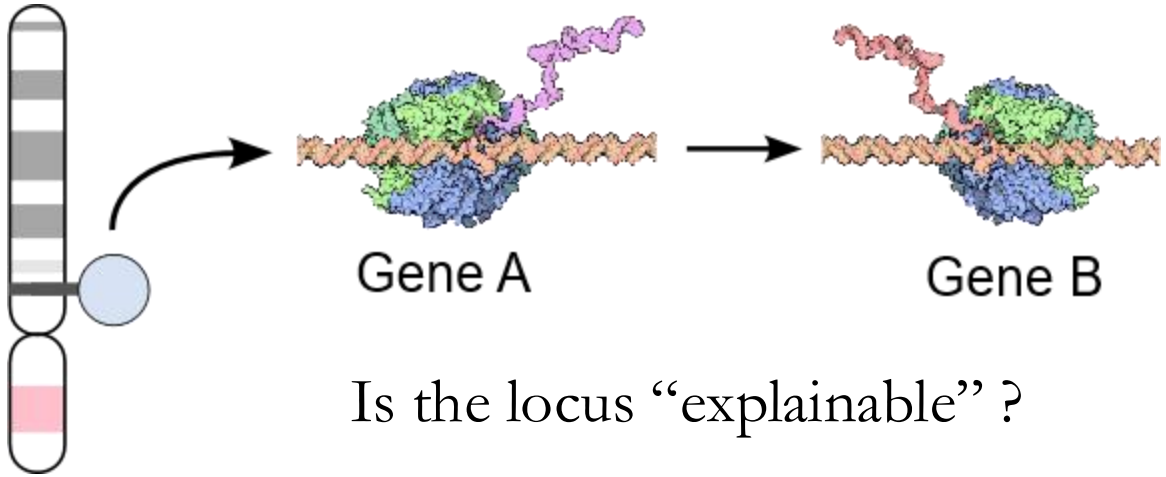
- Shown here is the correlation of the individual perturb-seq studies we analyzed

- Note K562 replication
- Note Replogle (lab)
- Note Zhu et al (2025)
- Note Norman (2019)

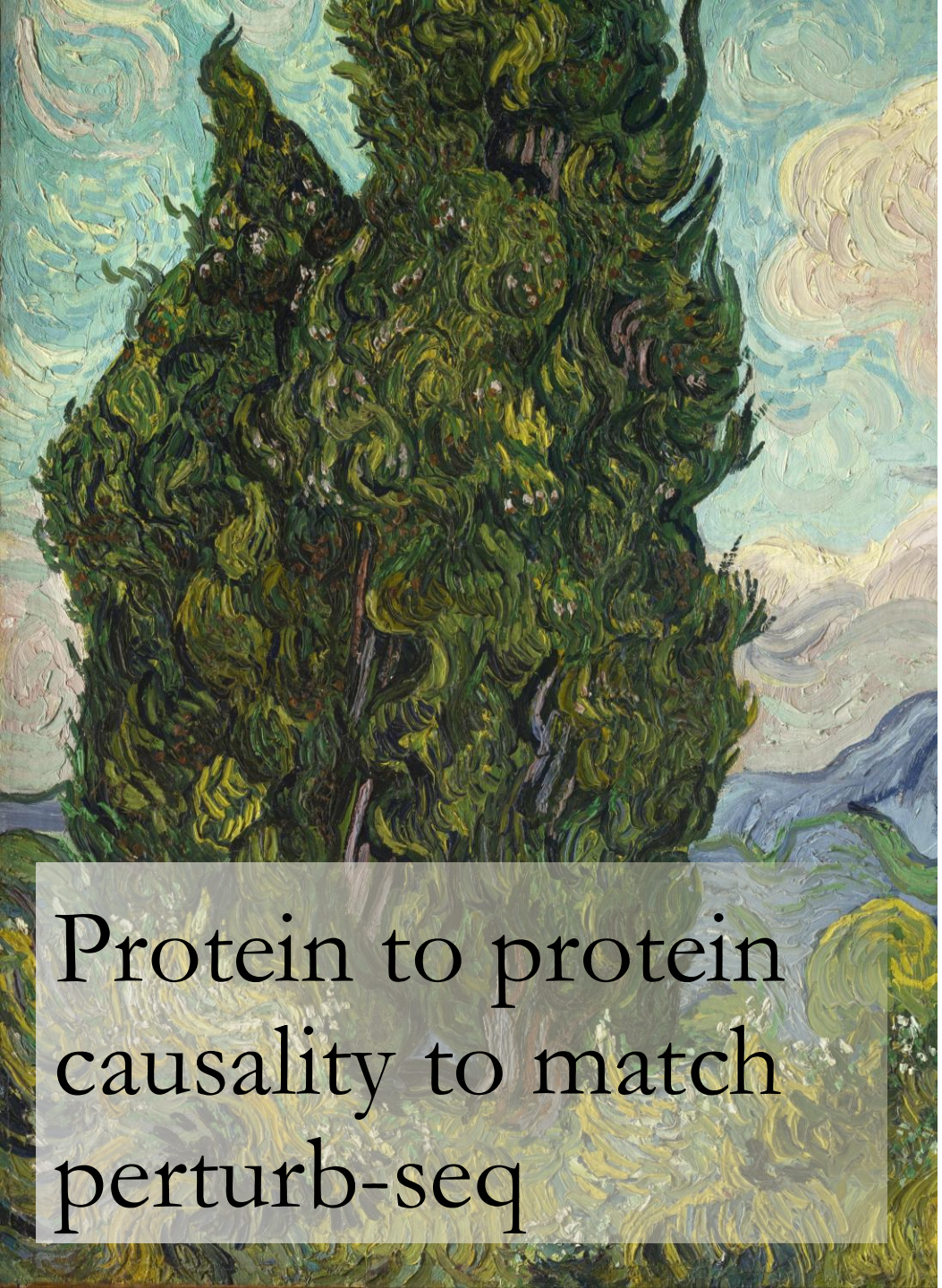


Mapping perturb-seq to disease

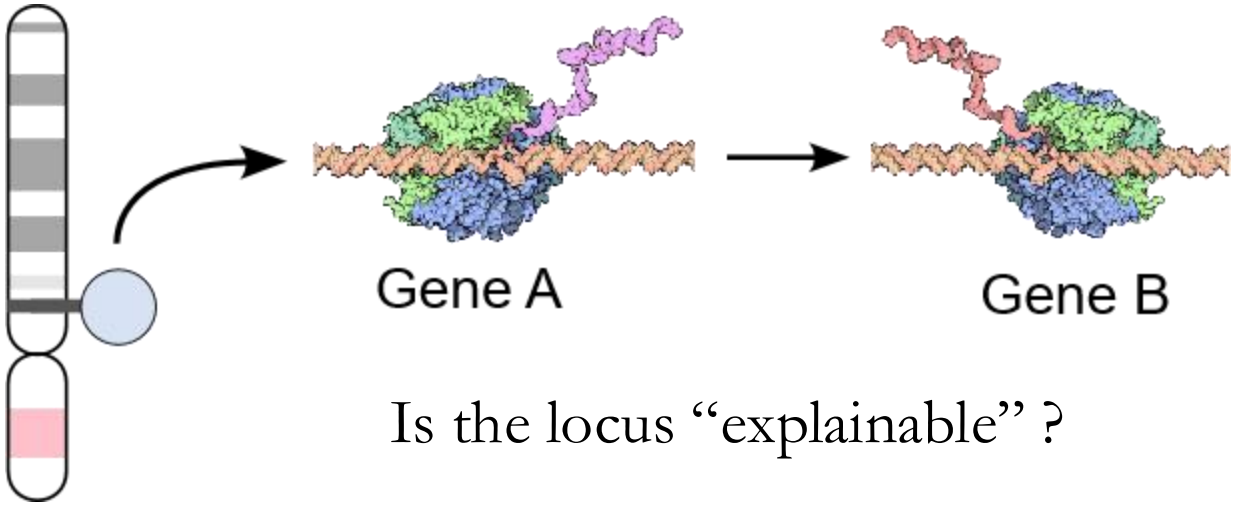




Is the locus "explainable" ?



Protein to protein causality to match perturb-seq



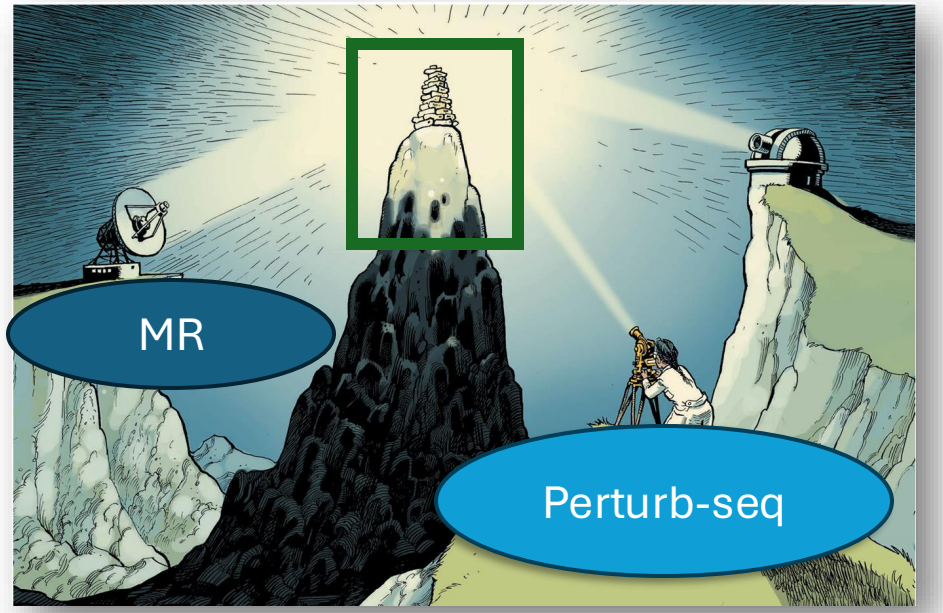
Is the locus “explainable” ?

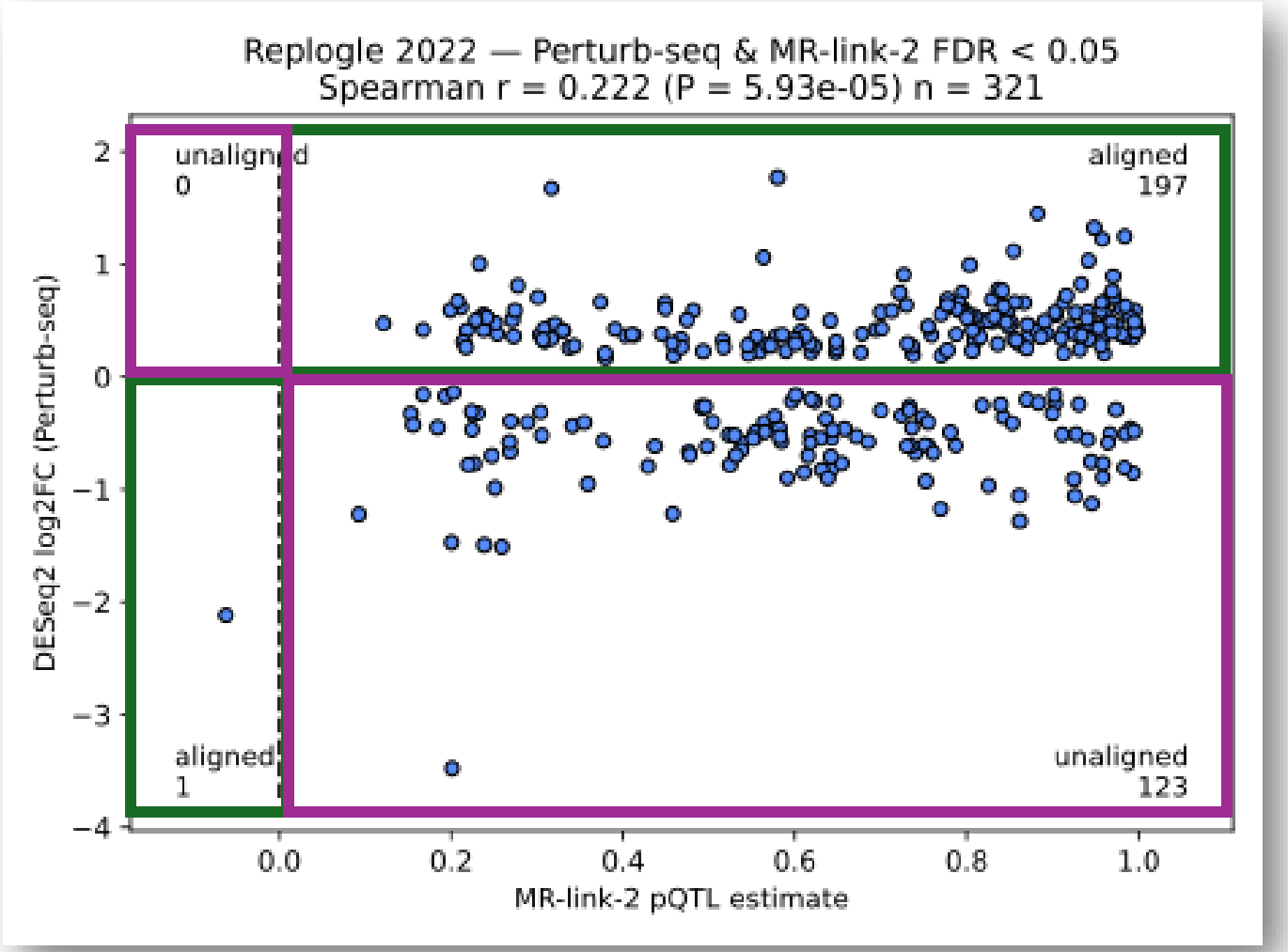
- 2,736 proteins pairwise MR
  - UKB-PPP - Olink proteomics
- 7,528,496 protein combinations
- MR-link-2

Sun, B. B. *et al*. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).

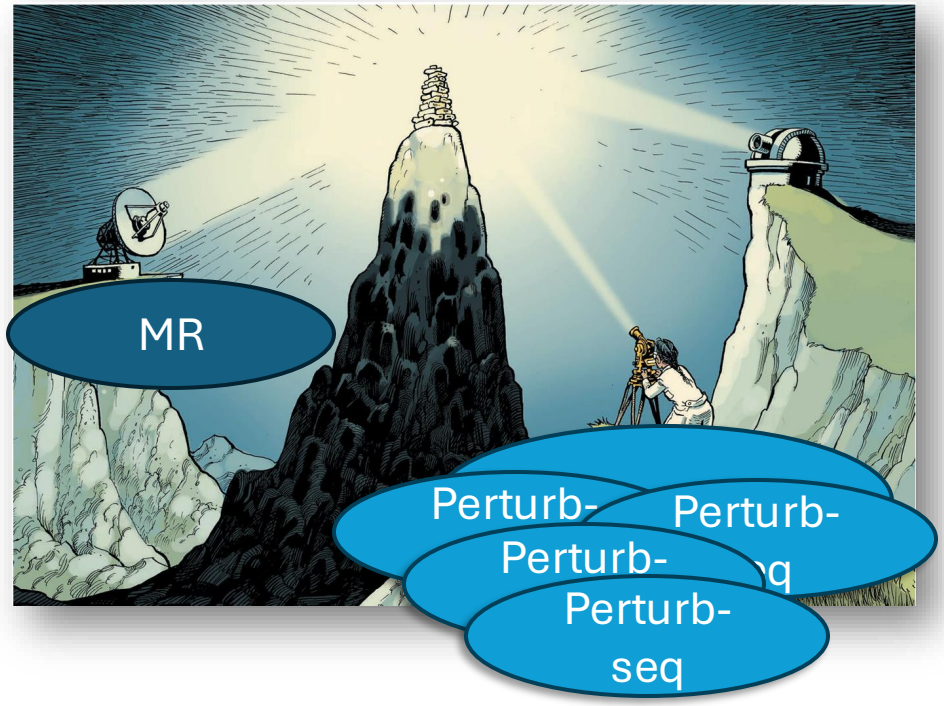
Pharma Proteomics Project

Matching protein MR to perturb-seq

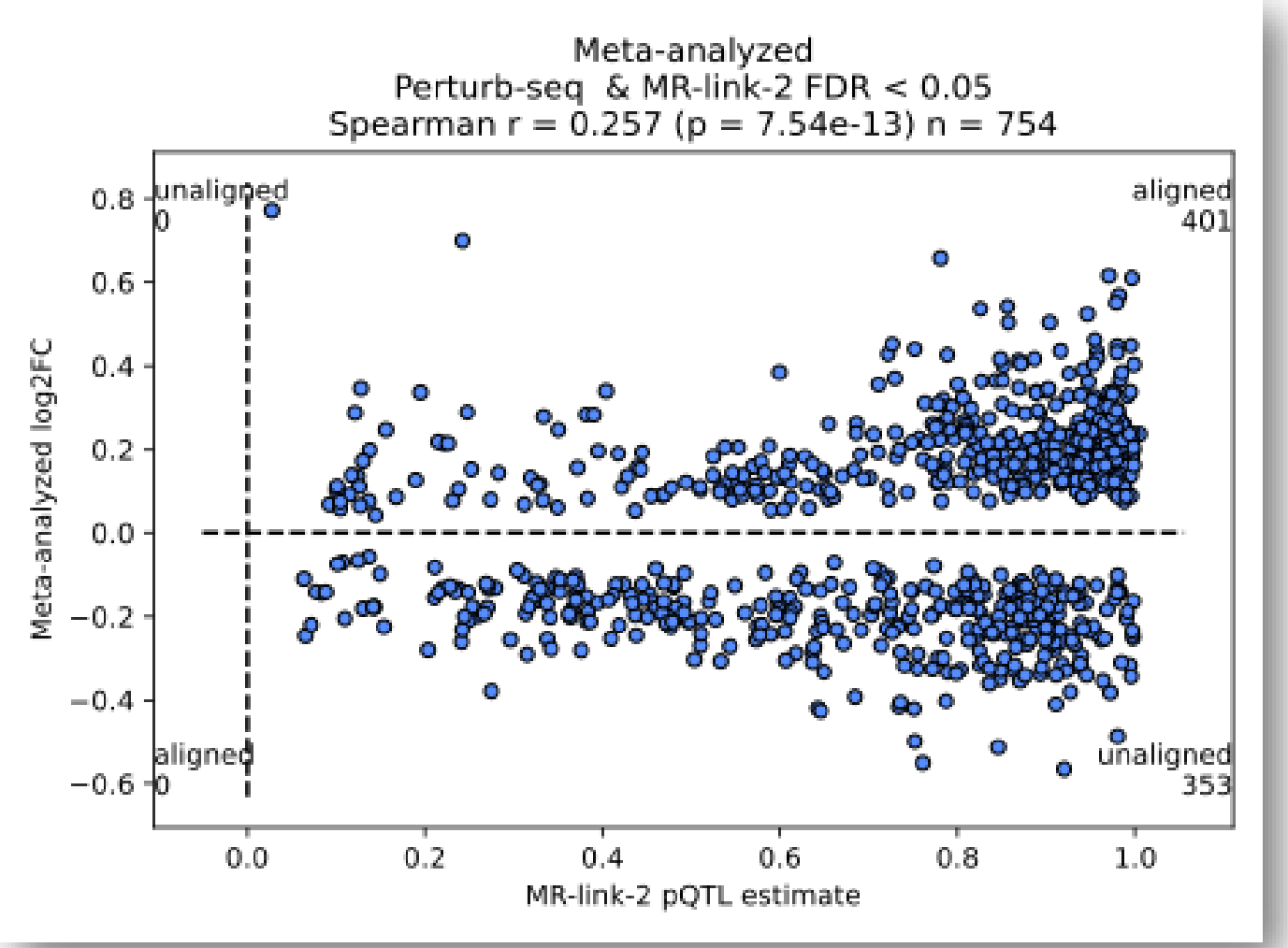




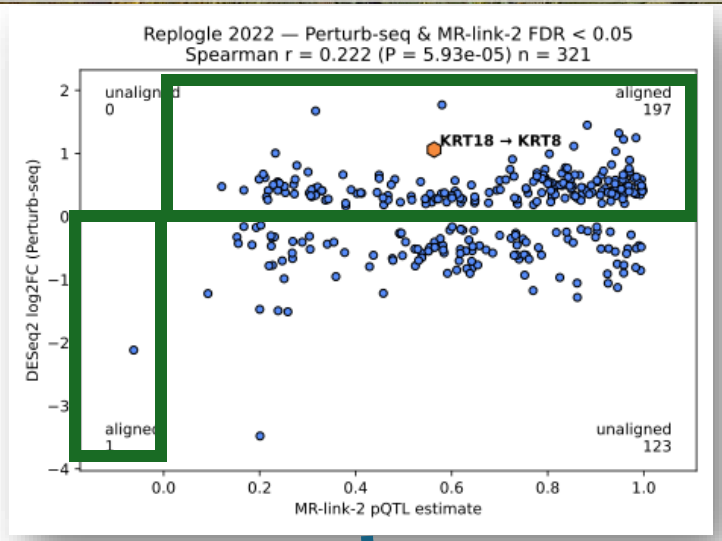
Meta-analysis improves correlation and amount of findings

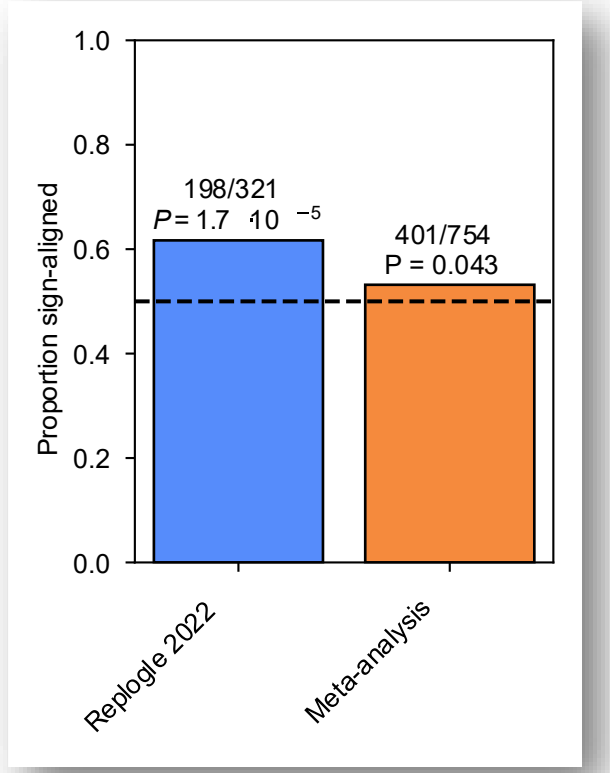


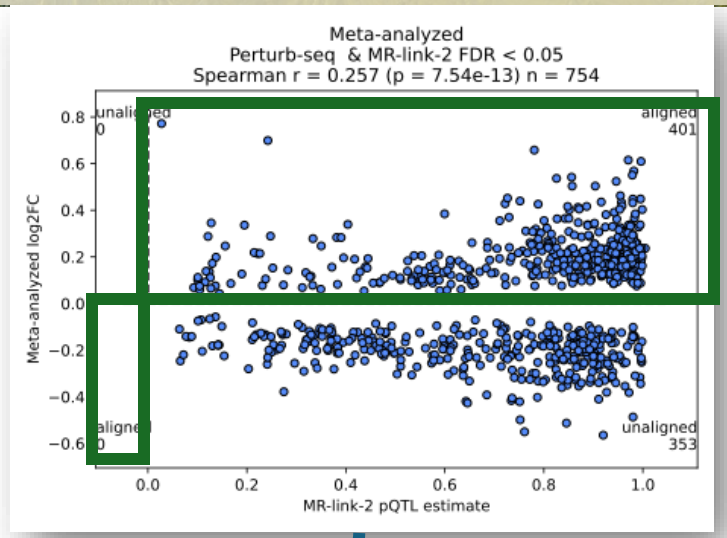
Inverse variance weighting log fold changes across studies.



Directional alignment decreases in meta-analysis

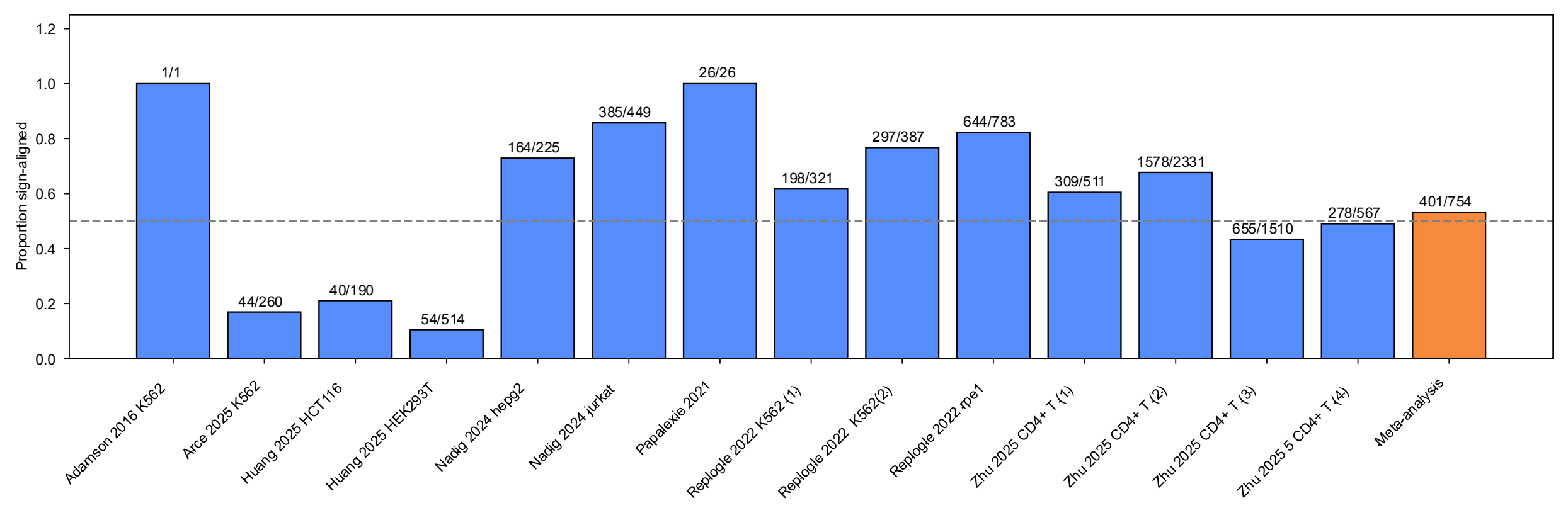






P from one sided binomial test

Alignment differs between experiments

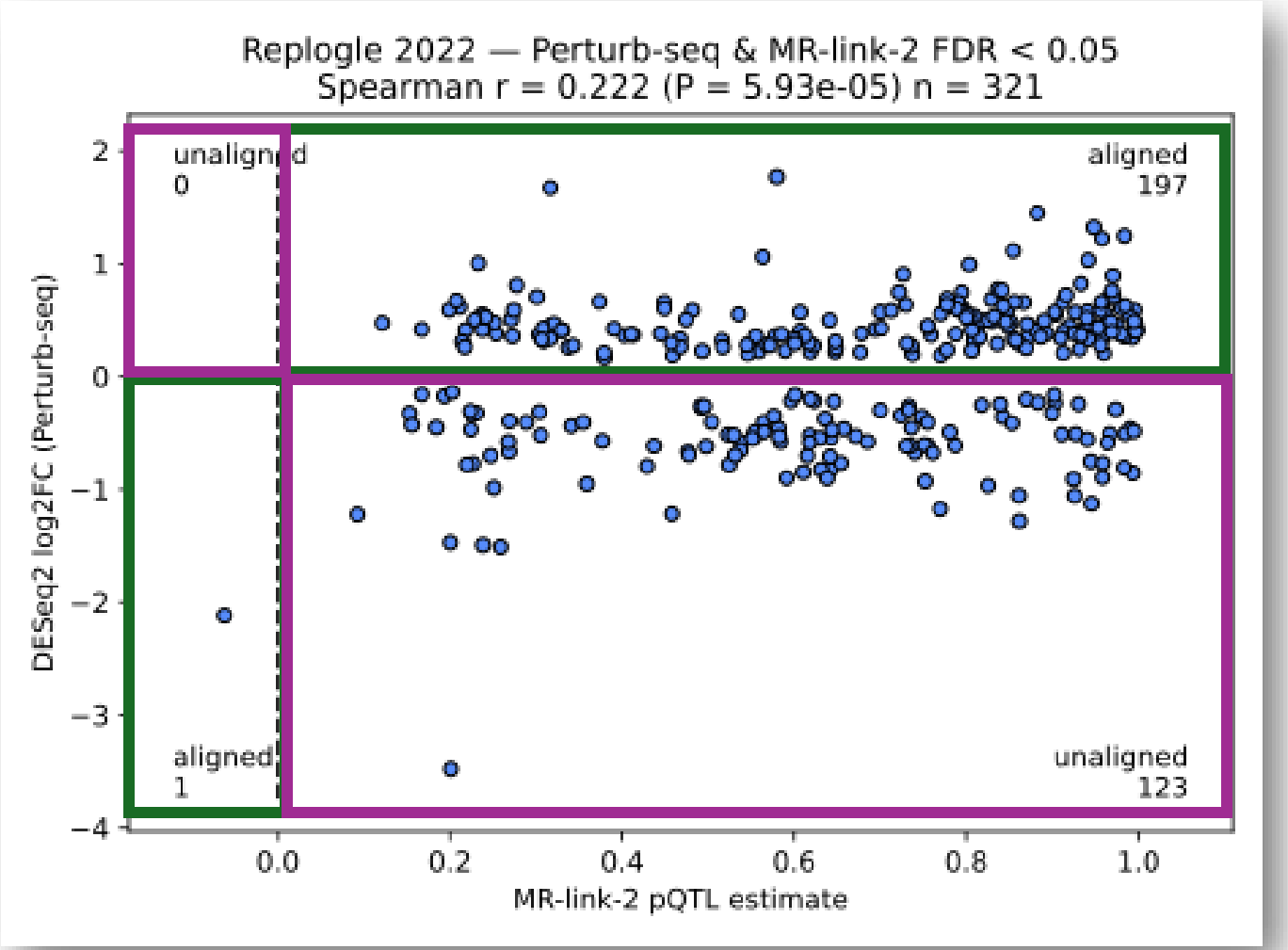


P from one sided binomial test

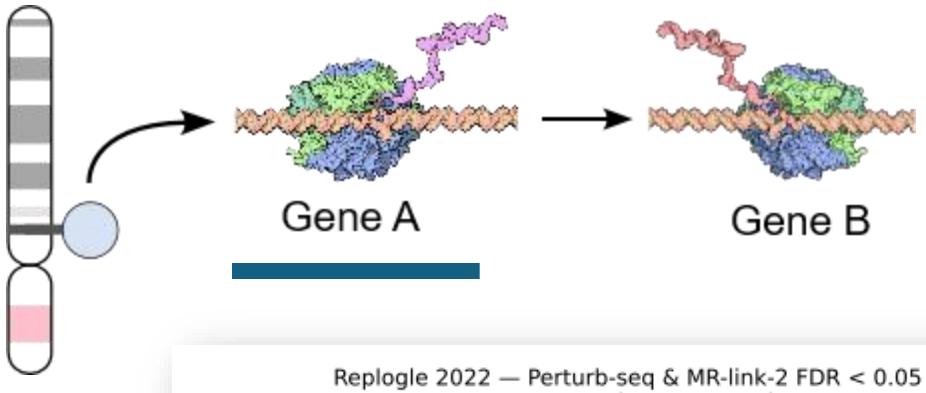
Why alignment can be low

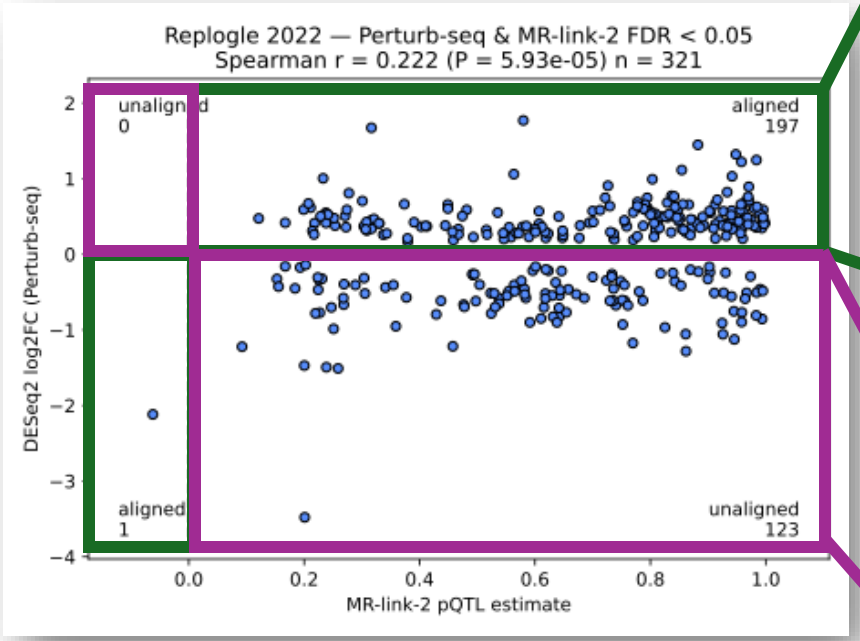
Reasons for discrepancy

- mRNA to protein boundary
- Cell type and context
- Experimental bias
- Cell line vs. in-vivo
- Perturbation differences
- Time differing biology



Aligned loci are enriched for biology





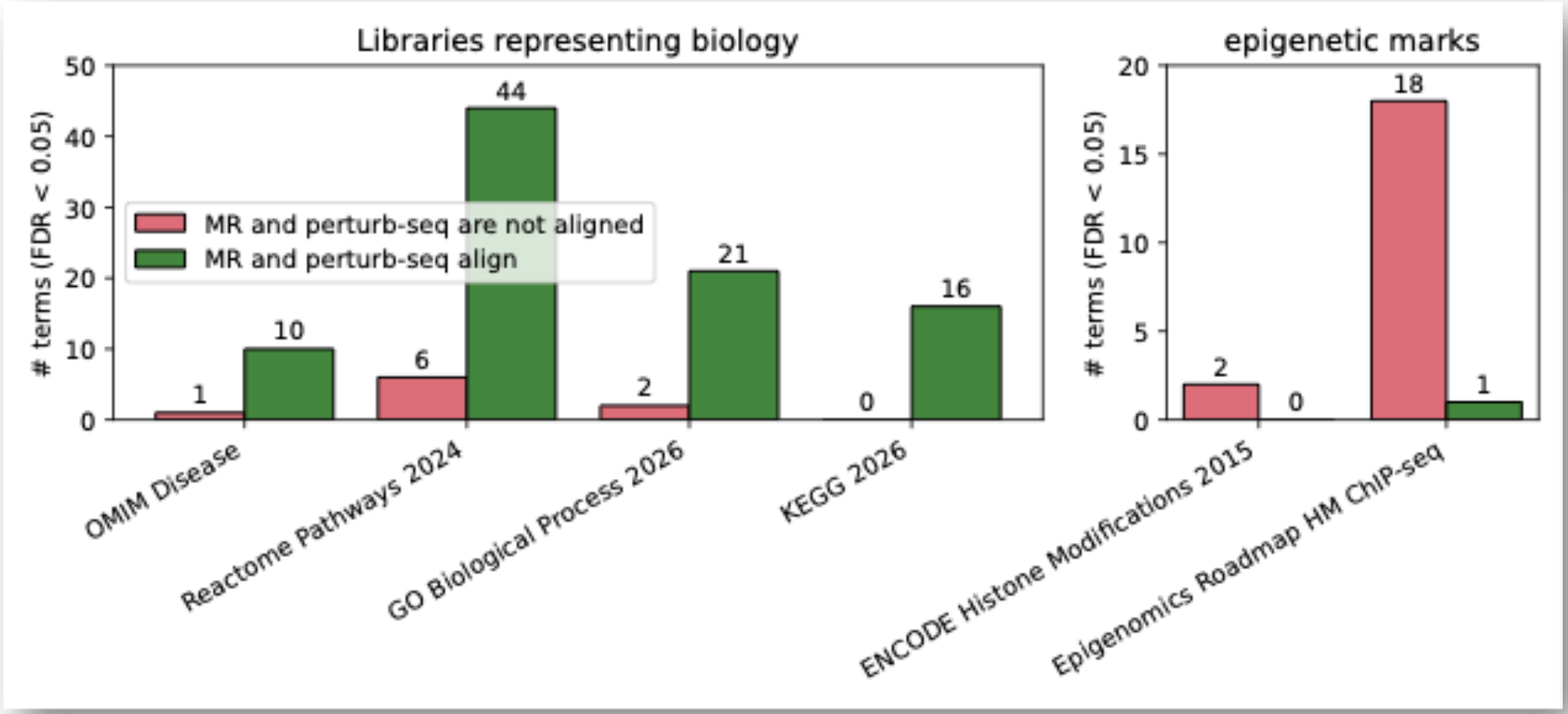
GO BIOLOGICAL PROCESS ('26)

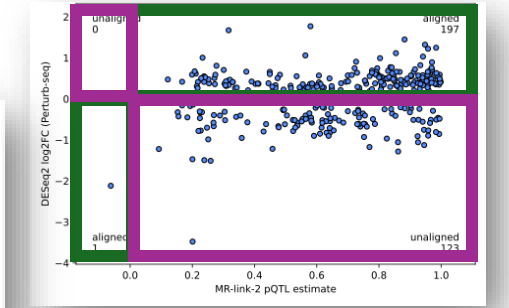
Experiment	Enriched term	P (BH adj)
Zhu 2025 CD4+ T (3)	Positive Regulation of Coagulation (GO:0050820)	0.011919
Huang 2025 HCT116	Nuclear Membrane Organization (GO:0071763)	0.04
Huang 2025 HCT116	Nuclear Membrane Reassembly (GO:0031468)	0.04
Huang 2025 HCT116	Regulation of Signal Transduction by P53 Class Mediator (GO:1901796)	0.04

17 more rows

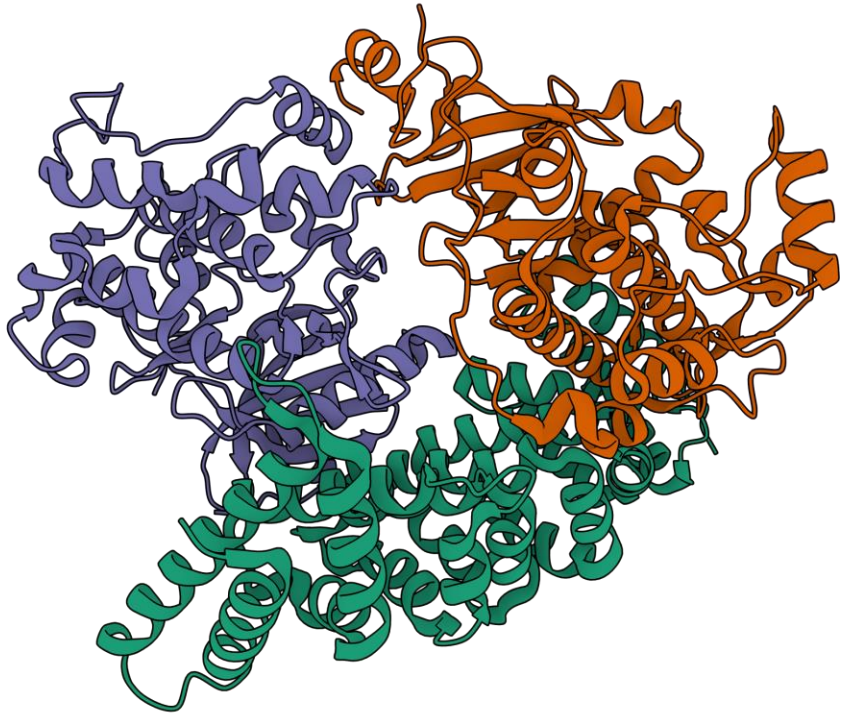
Experiment	Enriched term	P (BH adj)
Replogle 2022 rpe1	Negative Regulation of Transcription by RNA Polymerase II (GO:0000122)	0.041661
Zhu 2025 CD4+ T (4)	Regulation of Interferon-Alpha Production (GO:0032647)	0.044085

Aligned loci are enriched for biology

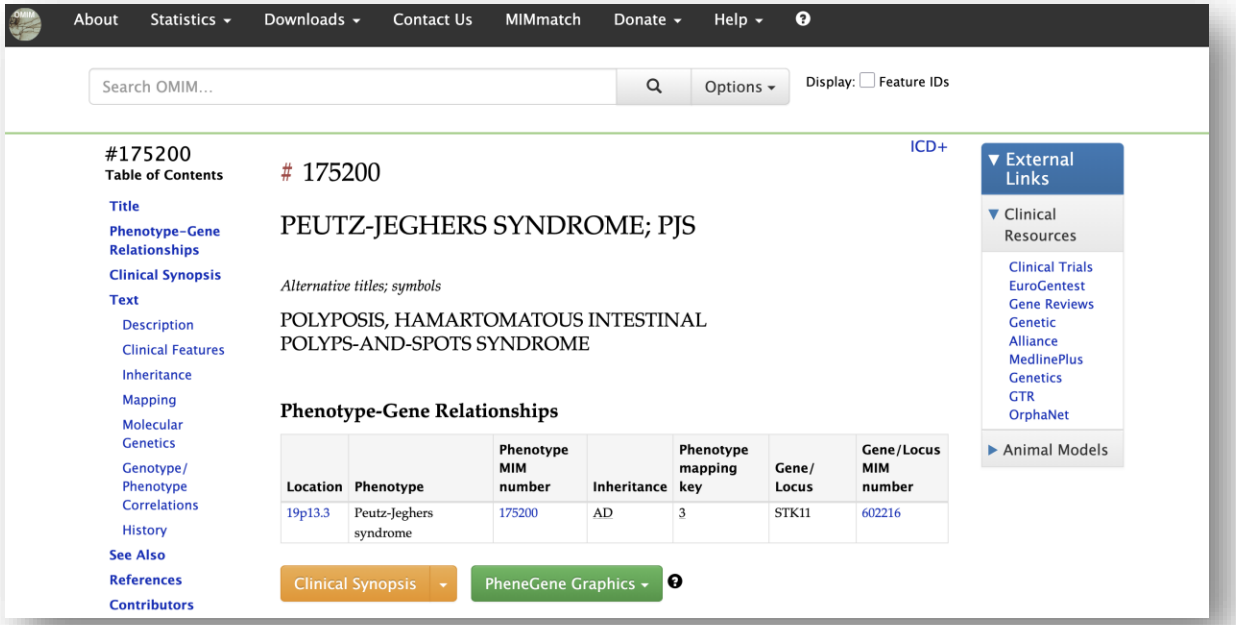




Serine Threonine Kinase 11 (*STK11*)



8VSU



Disease trait	Study Count
Type 2 diabetes	2
Chronic inflammatory diseases (Crohn's disease, etc.)	1

GWAS Catalog
The NHGRI-EBI Catalog of published genome-wide association studies

STK11 is aligned **and** unaligned

Replogle et al. K562
Blood derived





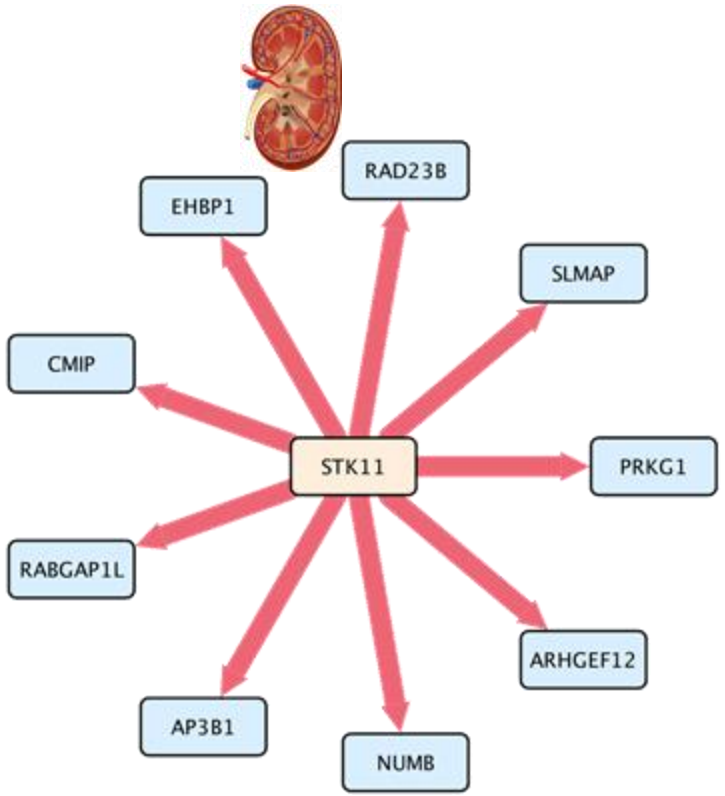
Huang et al. HCT116
Colon derived





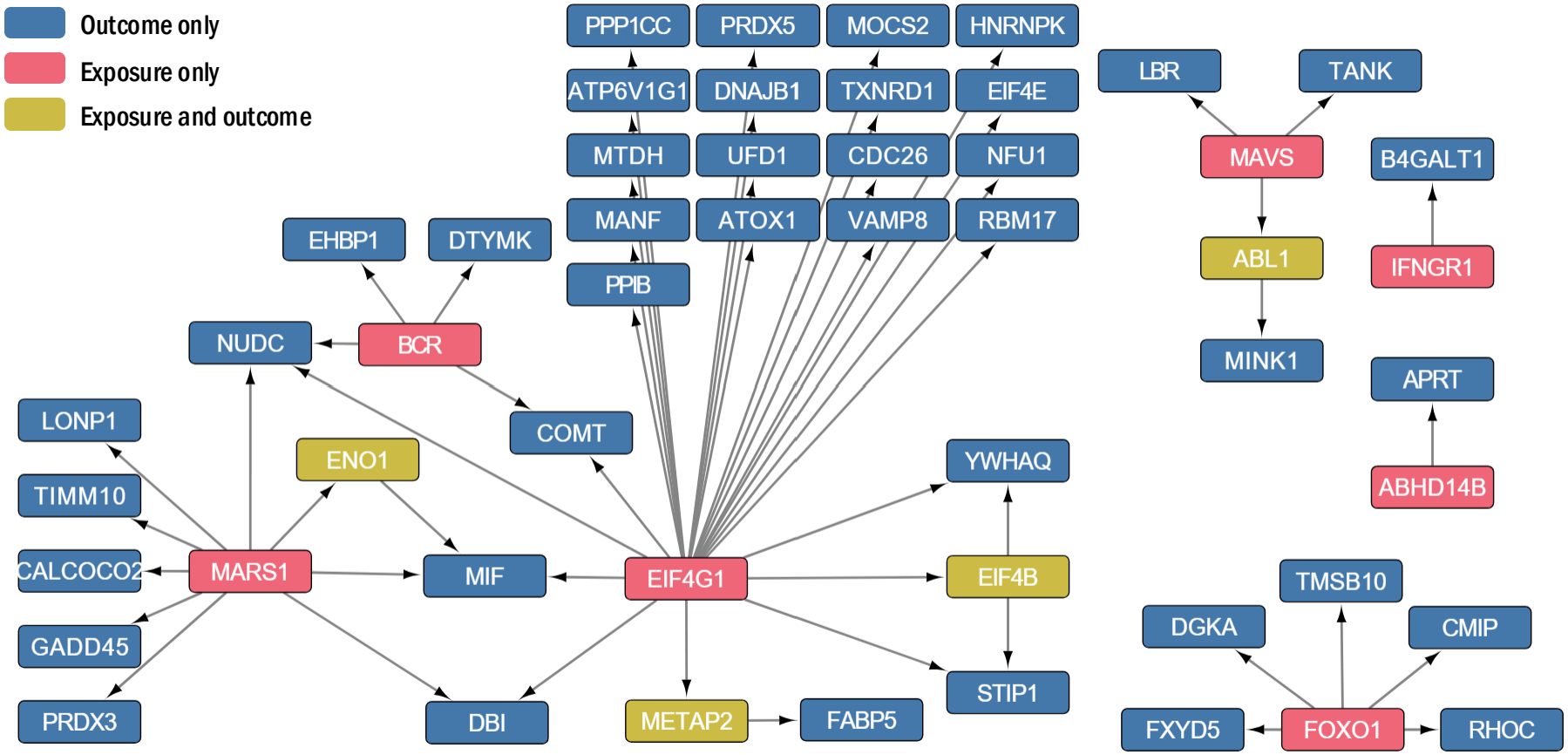
Alignment is seen of the same gene across 2 experiments

Huang et al. HEK293T
Kidney derived

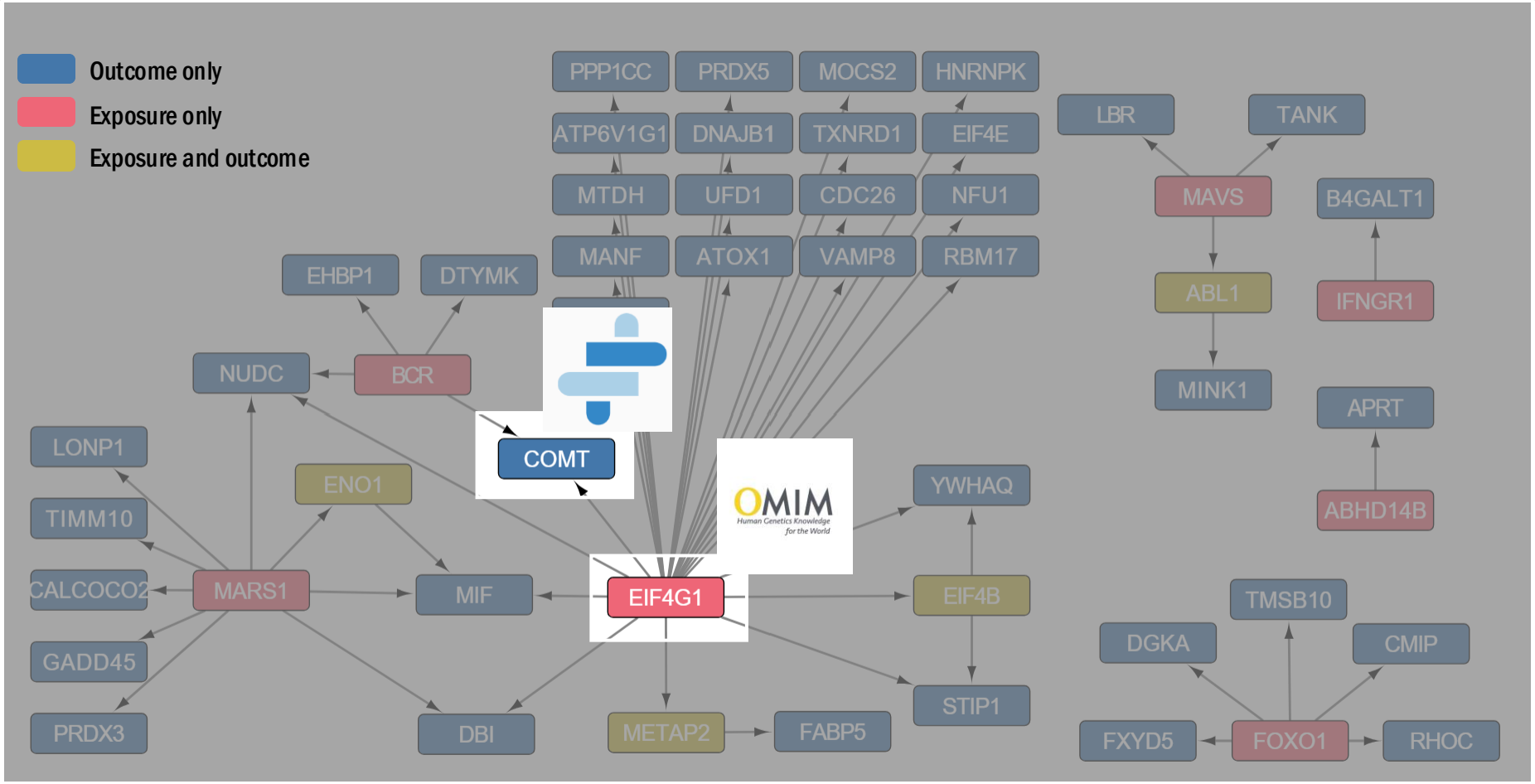


No alignment

Nonetheless, significant results show interesting biology

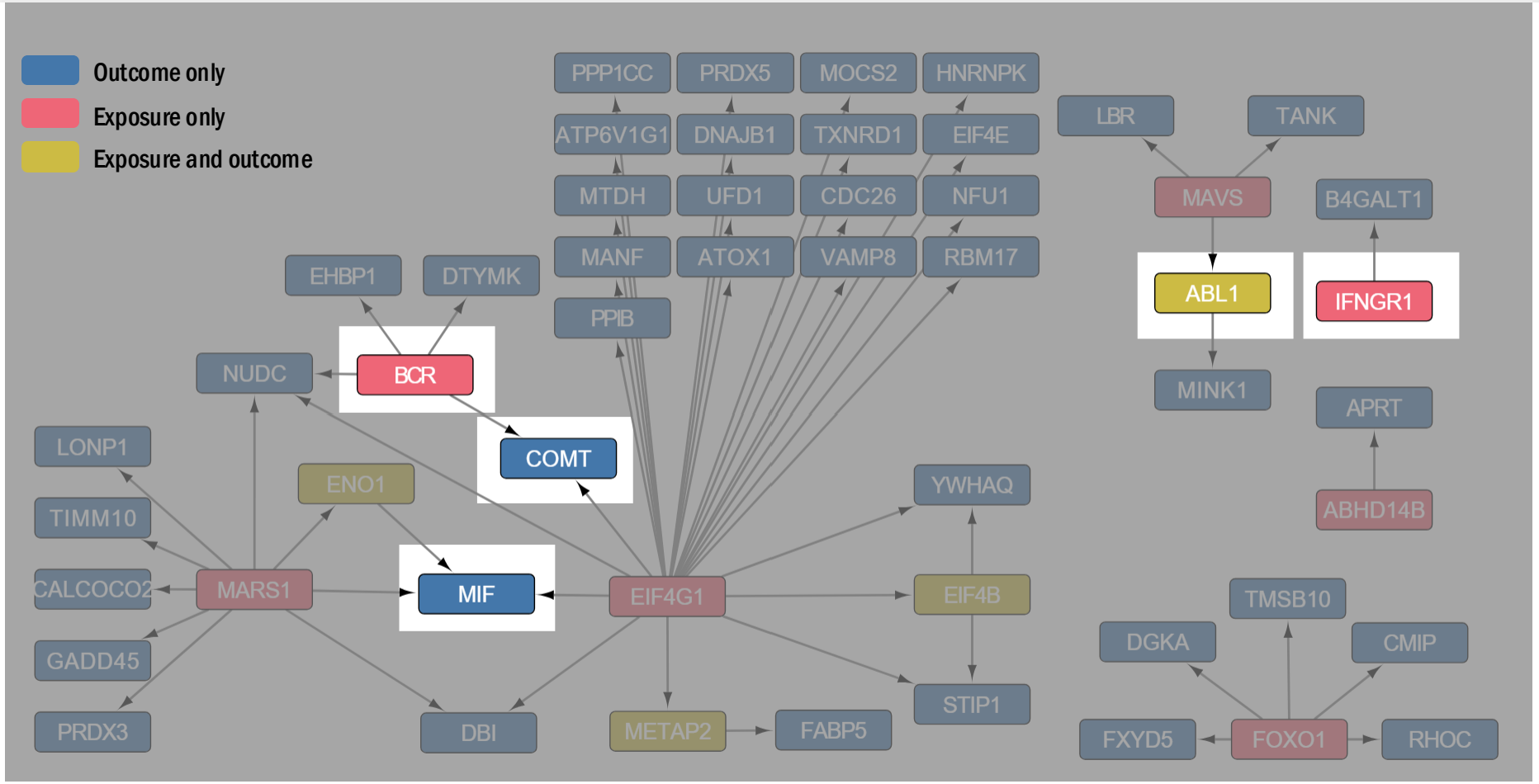


Nonetheless, significant results show interesting biology



Parkinson's disease associated genes

Nonetheless, significant results show interesting biology



Drug target genes

Conclusions on our work

- Perturb-seq identifies genome wide knock down experiments
  - Through a transcriptome wide read-out

- Matching perturb-seq to genetic inference identifies:
  - Shared regulation across the transcriptome → proteome barrier
  - The genetic loci that are experimentally translatable
  - Unaligned loci may be context specific

- **Understanding when and how perturb-seq is appropriate is still a challenge**


What could you write a review on?

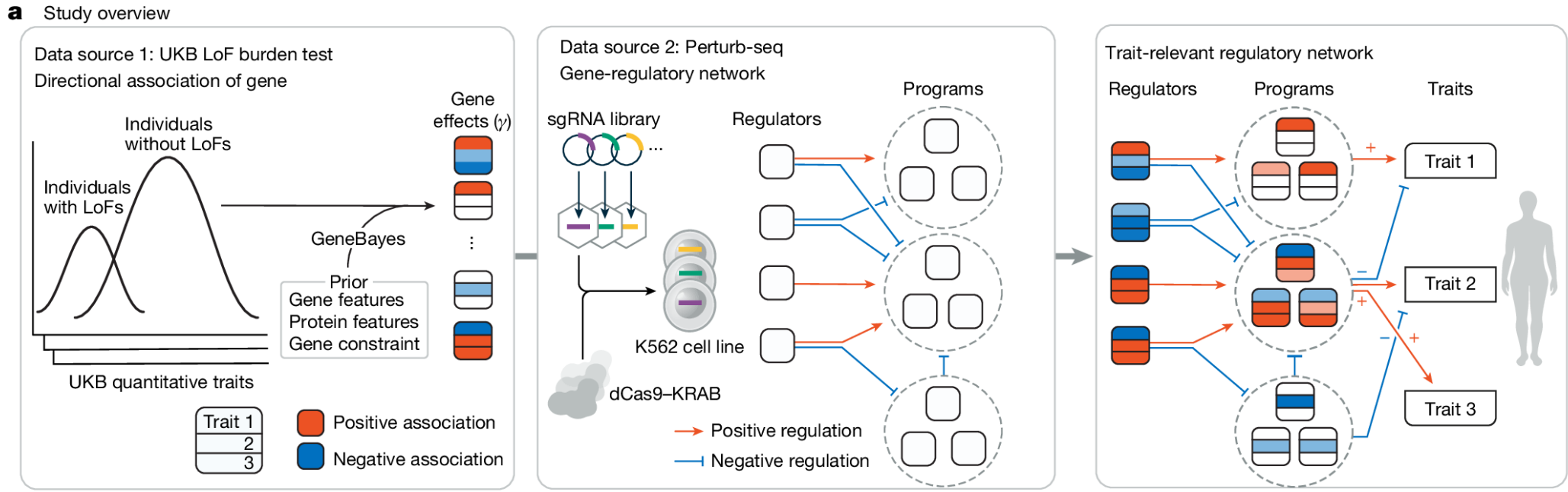
Suggestions for points to address

Guide to newcomers

- What is the state of the field?
- How do people analyze this data?
- What is the reproducibility of this data?
- What should you do once you encounter a perturb-seq dataset?

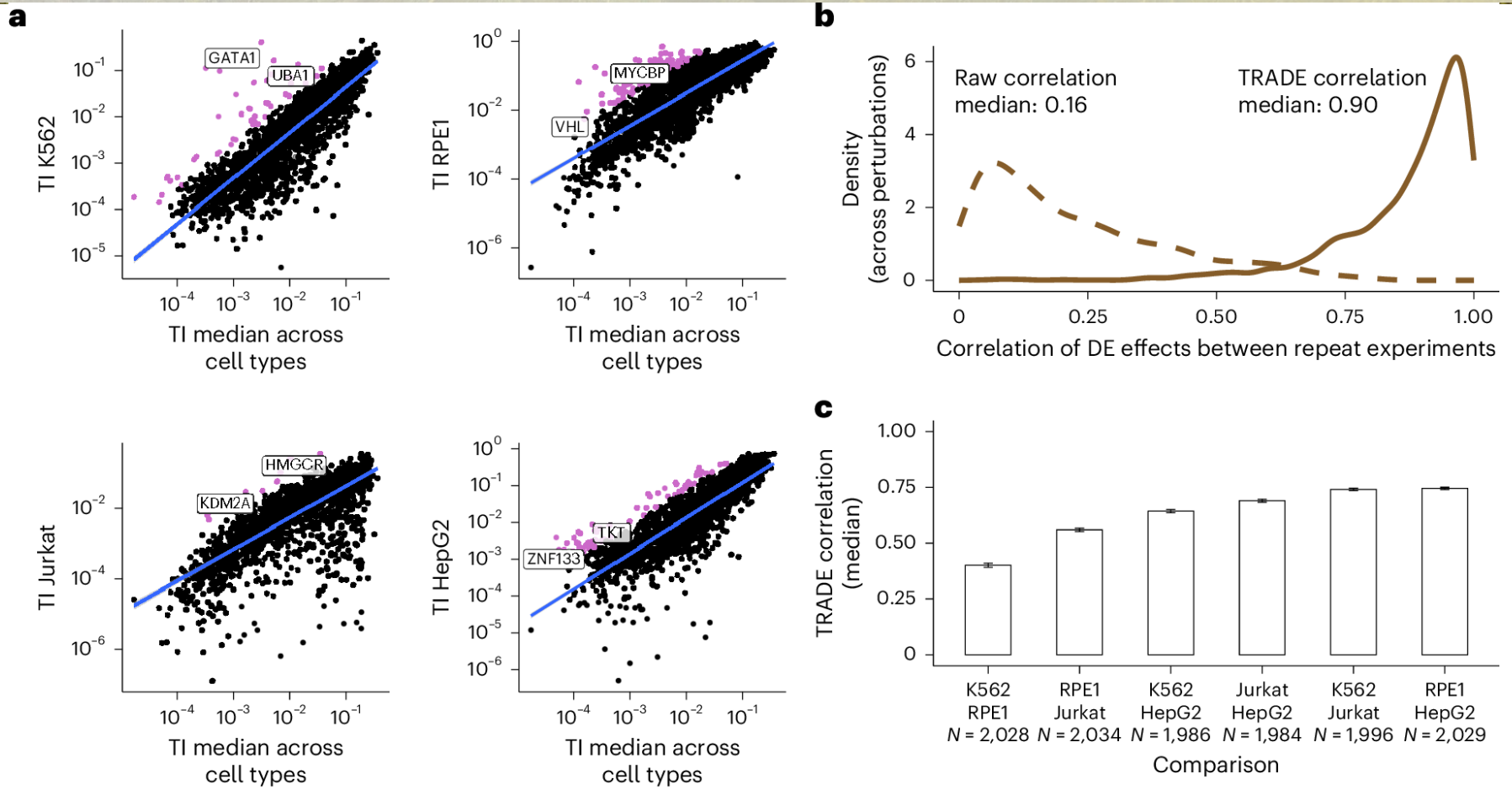
Things that are important to address: **gene programs**

- Perturb-seq is relatively sparse data, and so investigators refer to **gene programs**



Ota, M., Spence, J.P., Zeng, T. *et al.* Causal modelling of gene effects from regulators to programs to traits. *Nature* **650**, 399–408 (2026). https://doi.org/10.1038/s41586-025-09866-3

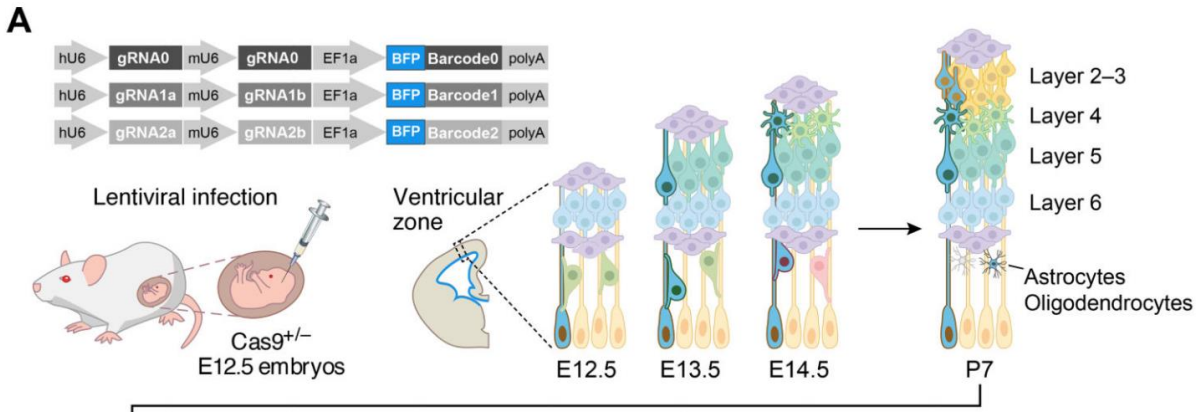
Things that are important to address: **Reproducibility**



Nadig, A., Replogle, J.M., Pogson, A.N. *et al.* Transcriptome-wide analysis of differential expression in perturbation atlases. *Nat Genet* **57**, 1228–1237 (2025). https://doi.org/10.1038/s41588-025-02169-3

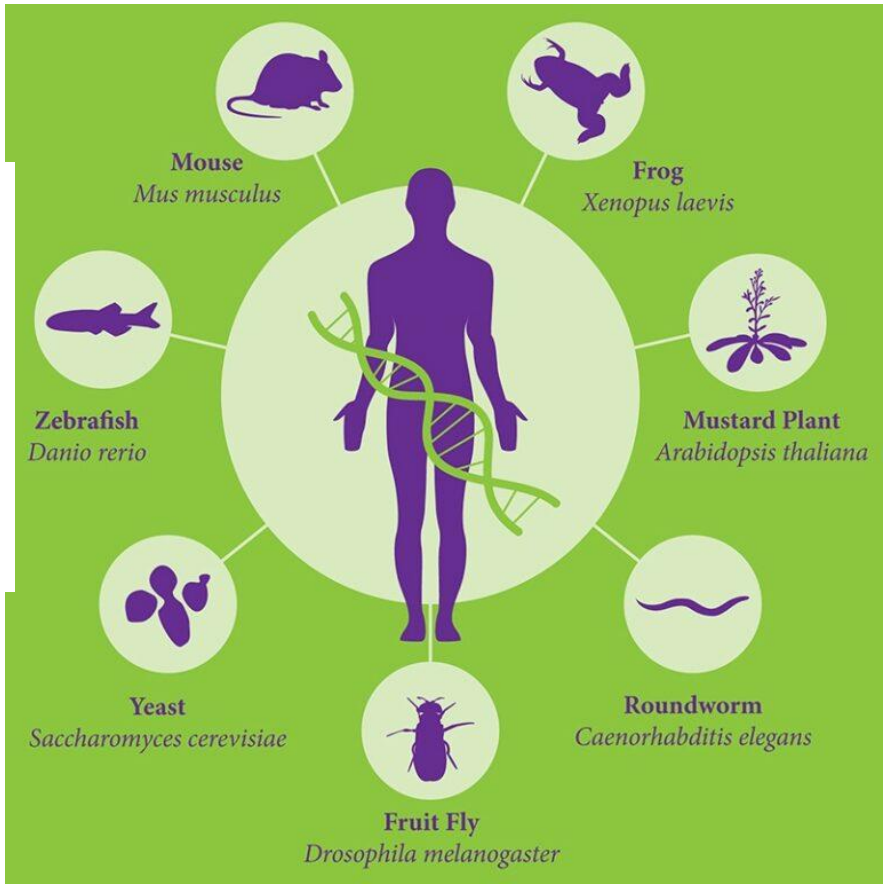
Things that are important to address: **Multiple organisms**

- My work has been on *human* cellular datasets



1.

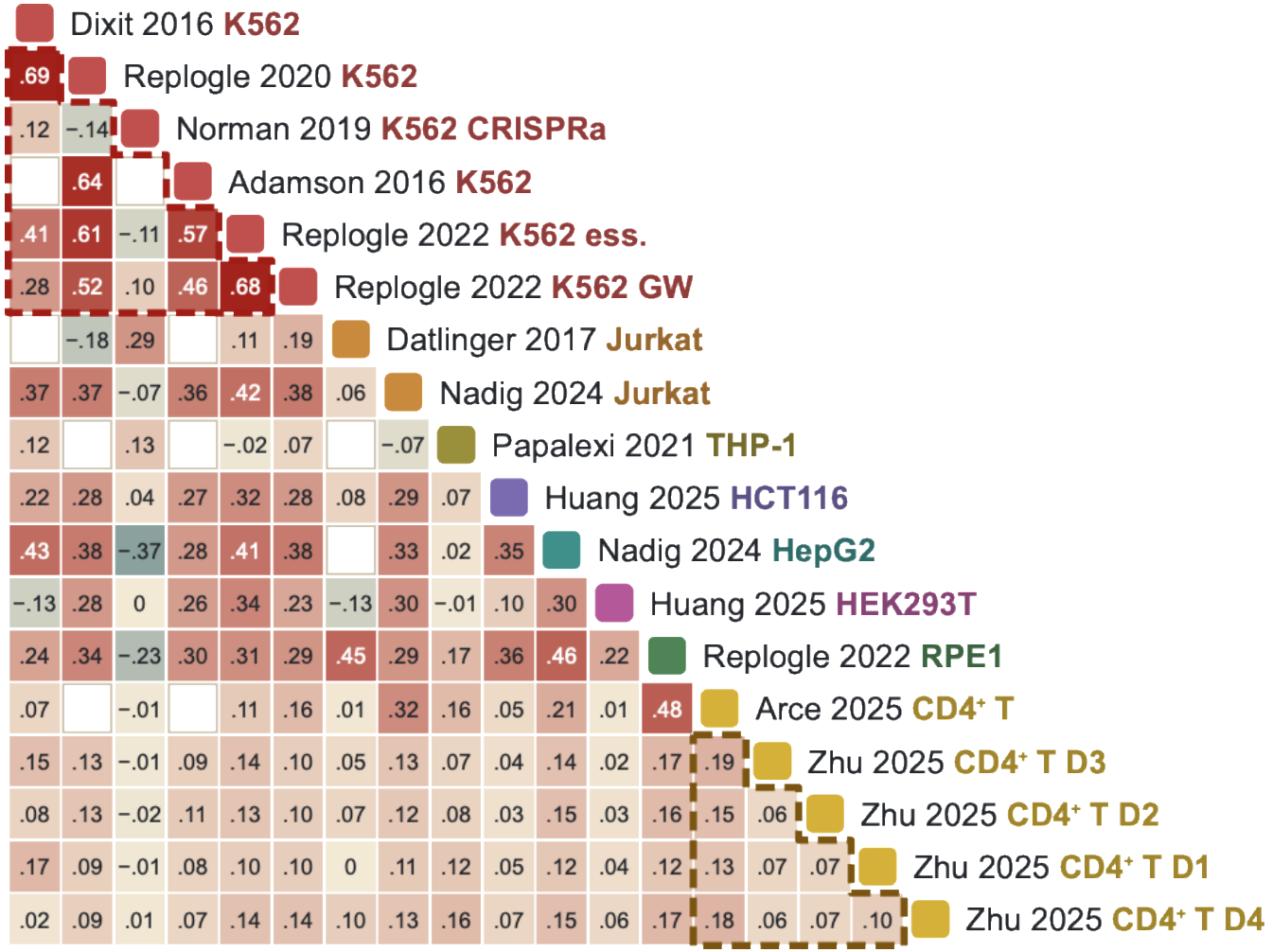
Jin, X. *et al.* In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science* **370**, eaaz6063 (2020).



https://bionomous.ch/small-model-organisms/

Things that are important to address: **CONTEXT**

- The context matters!



Suggestions for points to address

What is unexplored?

- What is being perturbed?
  - Inhibition, activation
- What is being measured?
  - Currently: scRNA-seq

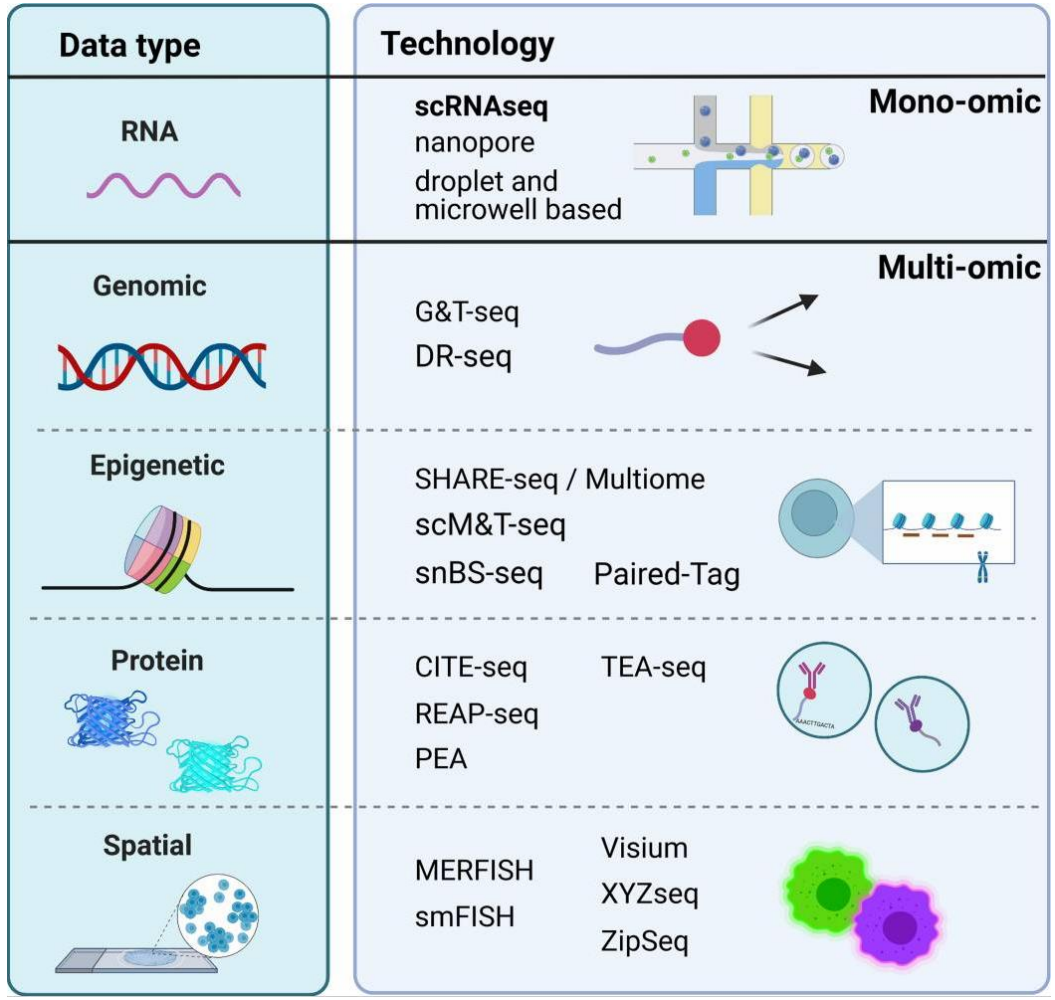
What is being perturbed?

- Currently, what is being perturbed is parallel CRISPR-I and CRISPR-A
- These techniques perturb the mRNA expression of a gene

- What about perturbing the DNA?
- What about perturbing proteins somehow?
- What about perturbing splicing specifically?

Things that are can be addressed: **Multiomics**

- Single cell screens can be extended
- How do you measure proteins?
  - What is affordable?
- How could you measure spatial transcriptomics?
  - Difficult combined with crispr-i



Flynn et al. Annu Rev Biomed Data Sci. 2023 May 9;6:313–337. doi: 10.1146/annurev-biodatasci-020422-050645

Suggestions for points to address

What is new?

- What papers are going around on (bio)Rxiv

- How can perturbation screens be used to answer questions in your field?
  - Existing data
  - Newly designed experiment

Thank you!

