SABETI, TEWHEY, & FINUCANE LABS

ENCODE Functional Characterization Plan

Focus on genetic fine-mapping and MPRA

AIM 1 (Regulatory Grammar & Variants)
Sabeti, Tewhey, Finucane Labs

- ID and predict ENCODE elements
- Incorporate regulatory variants
- Understand regulatory grammar rules
- Fine-map variants for GWAS, eQTLs, UKBB

AIM 2 (MPRA)
Tewhey Lab

Characterize variant function via high-throughput reporter assays in five ENCODE cell lines

- 100,000 GREs
- 200,000 variants

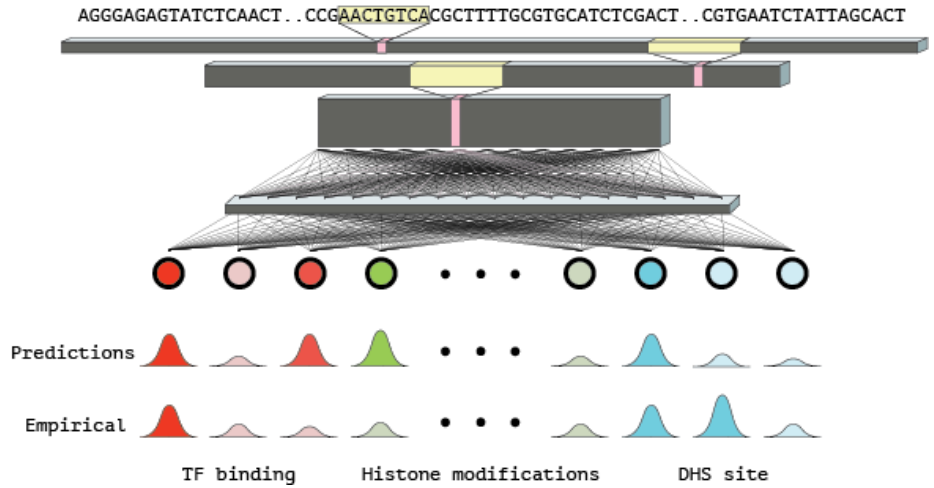
AIM 3 (CRISPR)
Sabeti Lab

Characterize complex GREs via detailed molecular dissection

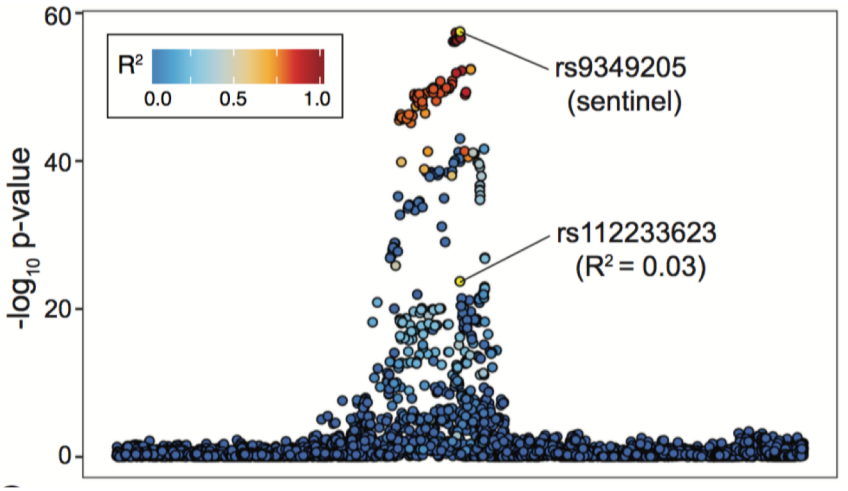
- 5 MBs of complex GREs
- CRISPRi and CRISPR bashing

AIM 4 (Disease Applications)
Sabeti, Tewhey, Finucane Labs

- Apply generalized methodologies to specific disease(s) (IBD) and biological processes
- Fine-map disease regulatory function in IBD



Computational CRE modeling



Application of human variation

AIM 1 (Regulatory Grammar & Variants)
Sabeti, Tewhey, Finucane Labs

- ID and predict ENCODE elements
- Incorporate regulatory variants
- Understand regulatory grammar rules
- Fine-map variants for GWAS, eQTLs, UKBB

AIM 2 (MPRA)
Tewhey Lab

Characterize variant function via high-throughput reporter assays in five ENCODE cell lines

- 100,000 GREs
- 200,000 variants

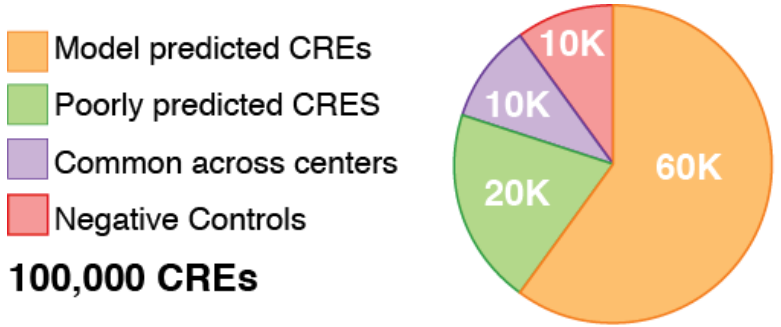
AIM 3 (CRISPR)
Sabeti Lab

Characterize complex GREs via detailed molecular dissection

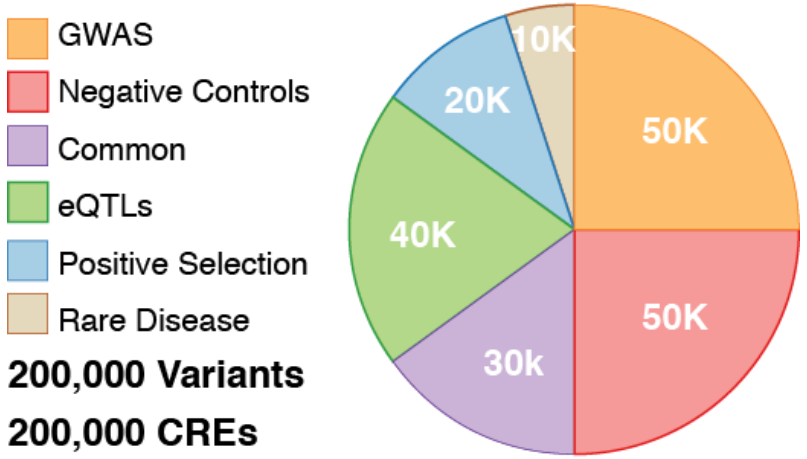
- 5 MBs of complex GREs
- CRISPRi and CRISPR bashing

AIM 4 (Disease Applications)
Sabeti, Tewhey, Finucane Labs

- Apply generalized methodologies to specific disease(s) (IBD) and biological processes
- Fine-map disease regulatory function in IBD



100,000 CREs



200,000 Variants

200,000 CREs

MPRA

5x Cell Lines: *GM12878, K562, HepG2, IMR-90, SK-N-SH*

3x Replicates

AIM 1 (Regulatory Grammar & Variants)
Sabeti, Tewhey, Finucane Labs

- ID and predict ENCODE elements
- Incorporate regulatory variants
- Understand regulatory grammar rules
- Fine-map variants for GWAS, eQTLs, UKBB

AIM 2 (MPRA)
Tewhey Lab

Characterize variant function via high-throughput reporter assays in five ENCODE cell lines

- 100,000 GREs
- 200,000 variants

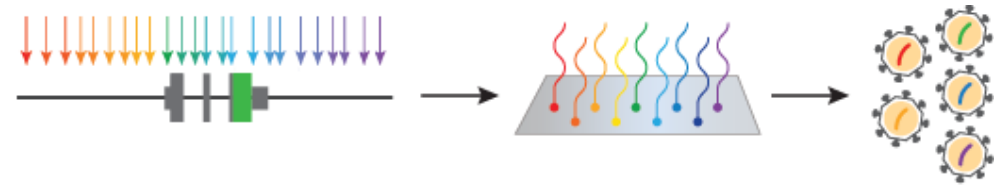
AIM 3 (CRISPR)
Sabeti Lab

Characterize complex GREs via detailed molecular dissection

- 5 MBs of complex GREs
- CRISPRi and CRISPR bashing

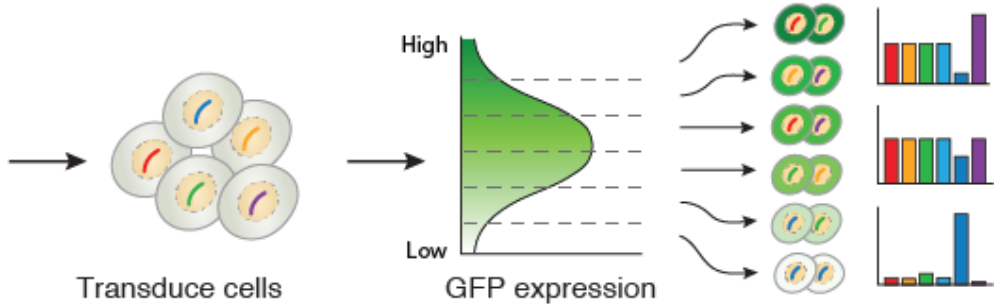
AIM 4 (Disease Applications)
Sabeti, Tewhey, Finucane Labs

- Apply generalized methodologies to specific disease(s) (IBD) and biological processes
- Fine-map disease regulatory function in IBD



Target CRIPRi/ko to genome regions adjacent to GFP tagged gene

Synthesize sgRNAs and produce lentivirus



Transduce cells

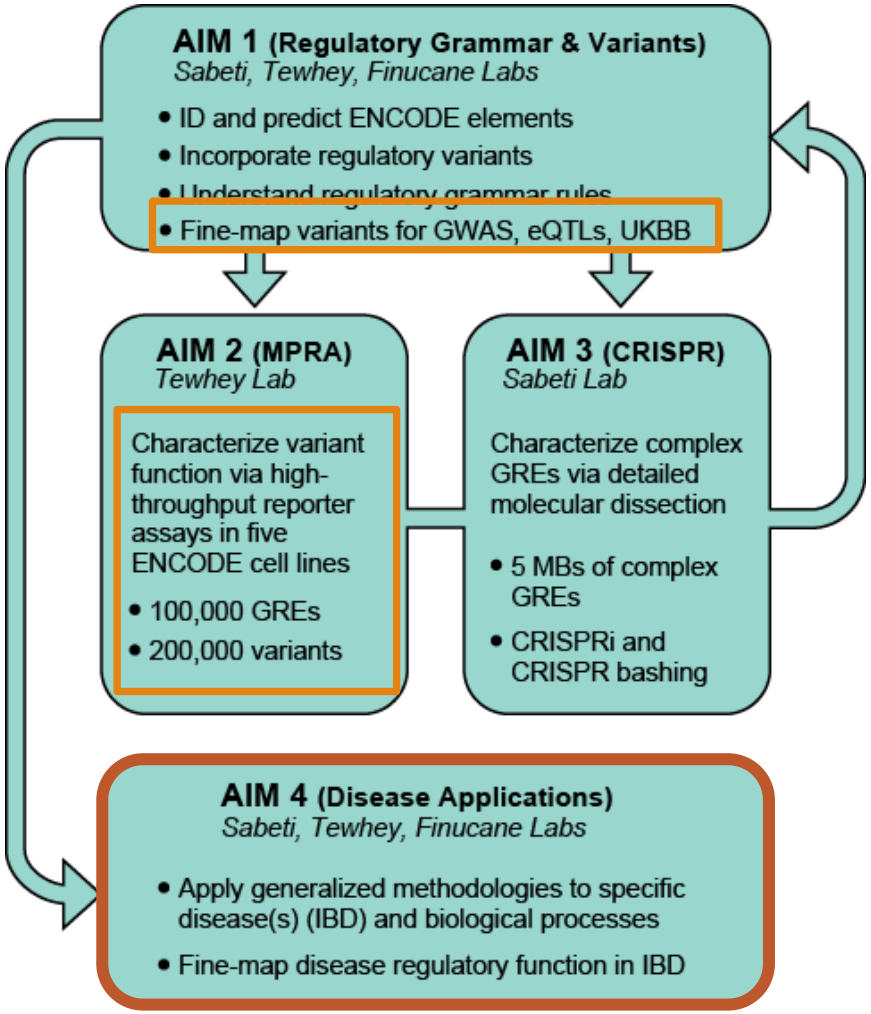
GFP expression & sequencing

5 - 1Mb Loci

~ 25 genes, 300k guides

3x Replicates

MPRA on same regions



Jacob Ulirsch
(Finucane/Sabeti Lab)

Leveraging natural variation to identify CREs most informative for learning regulatory grammar

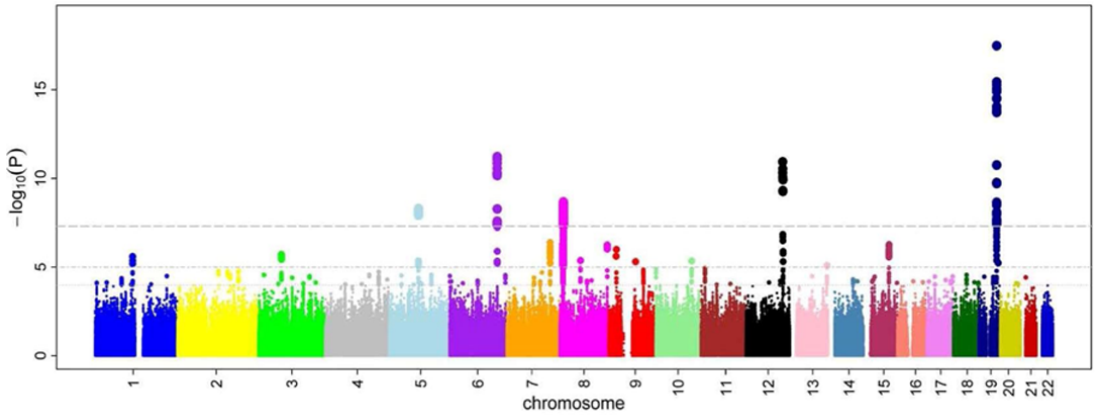
Motivation

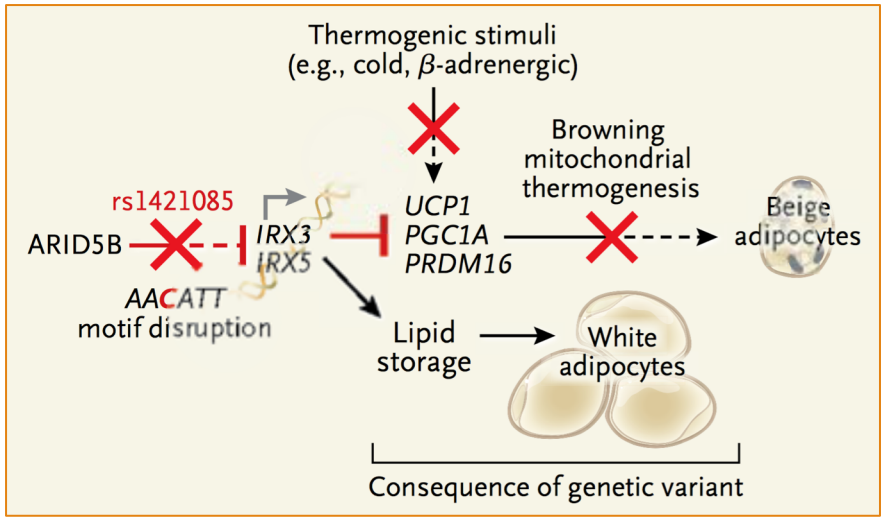
Thinking like a geneticist:

From GWAS to gene to biology

Claussnitzer et al. 2015

Complex trait associations





Causal variant → casual gene → biological insight

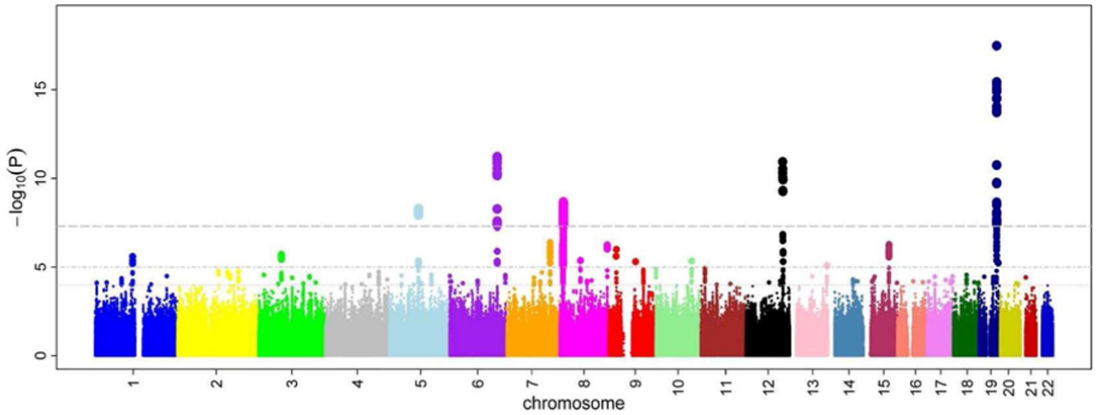
Motivation

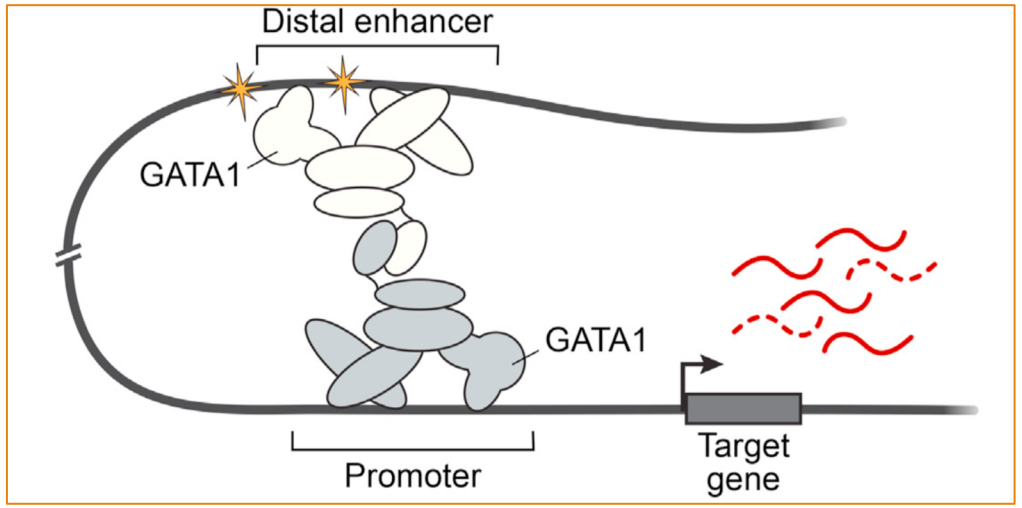
Thinking like ENCODE:

From GWAS to regulatory grammar

Claussnitzer et al. 2015
Ulirsch*, Nandakumar* et al. 2016

Complex trait associations





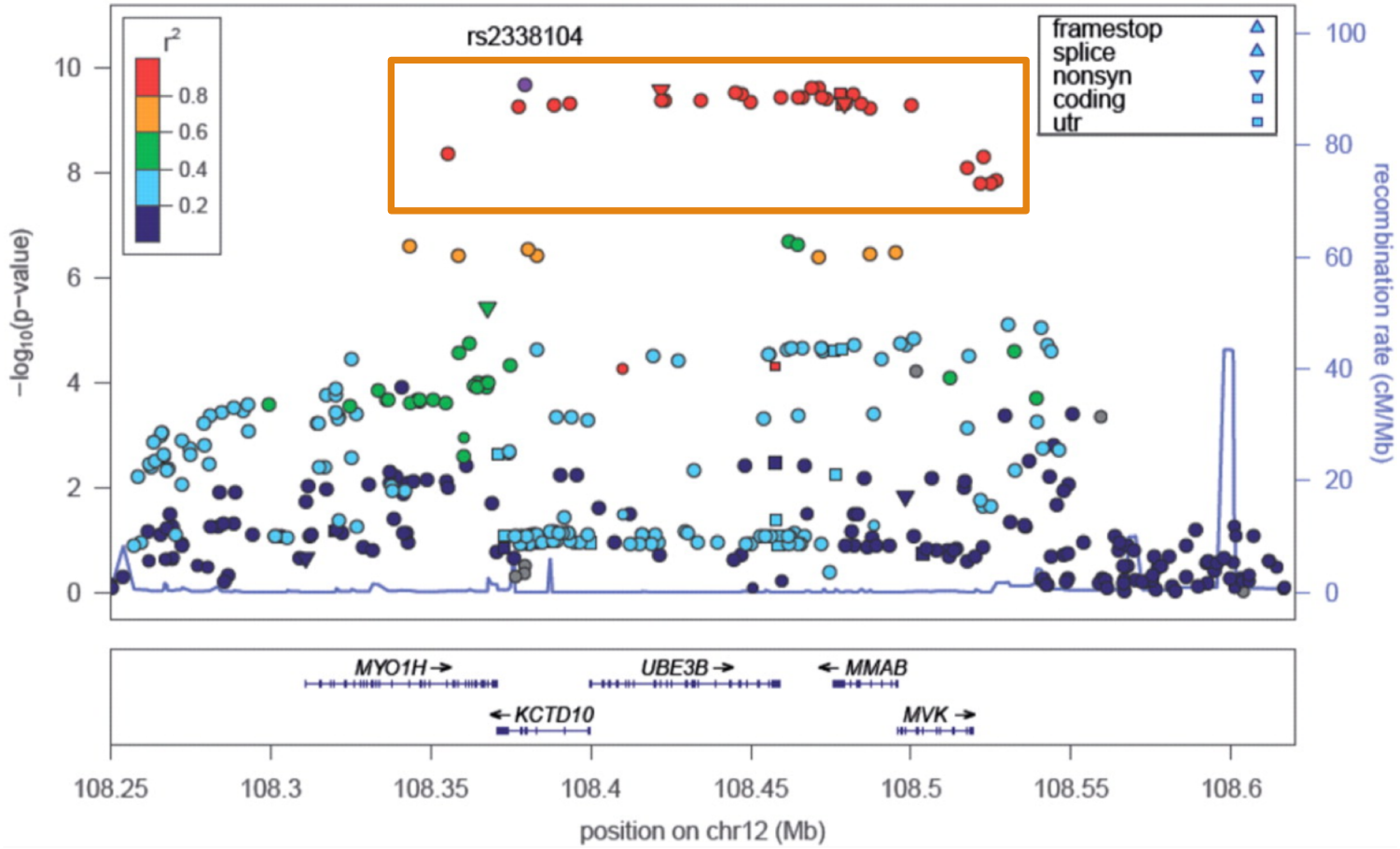
Causal variant → regulatory grammar

Motivation

Linkage disequilibrium confounds causal variant identification

Kathiresan et al. 2009

Which one(s) are causal?

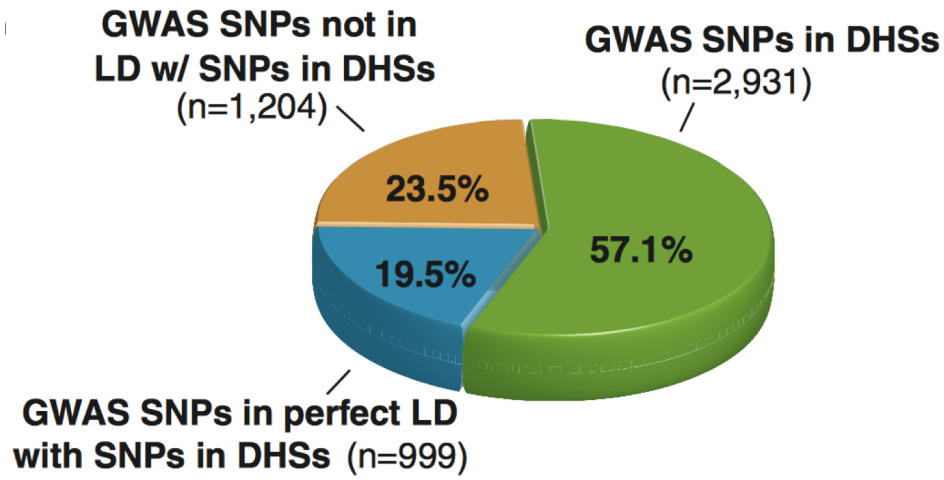


Motivation

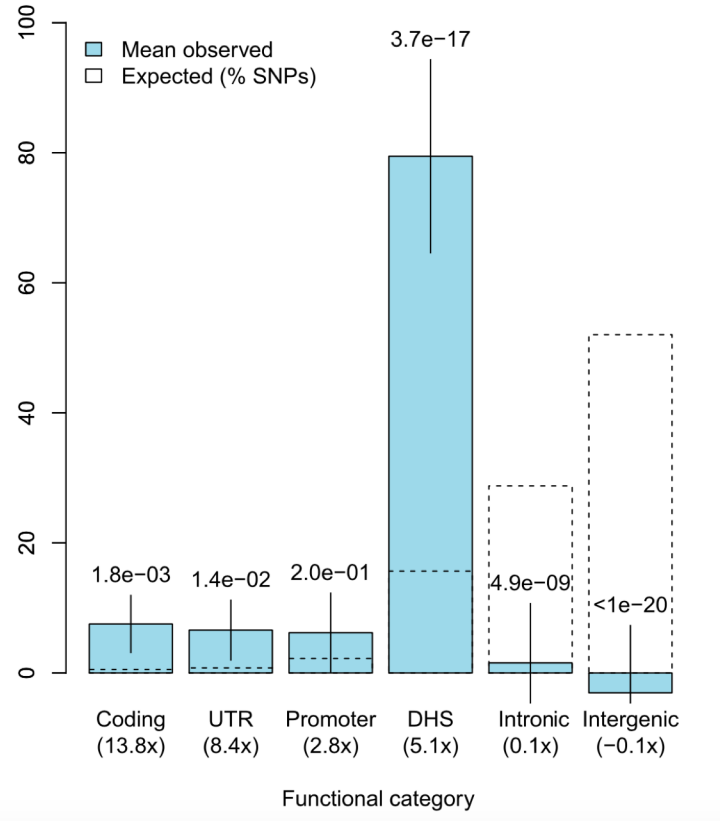
Most common variants underlying complex traits are non-coding regulatory variants

Maurano et al. 2012
Gusev et al. 2014

GWAS loci-based



Heritability-based

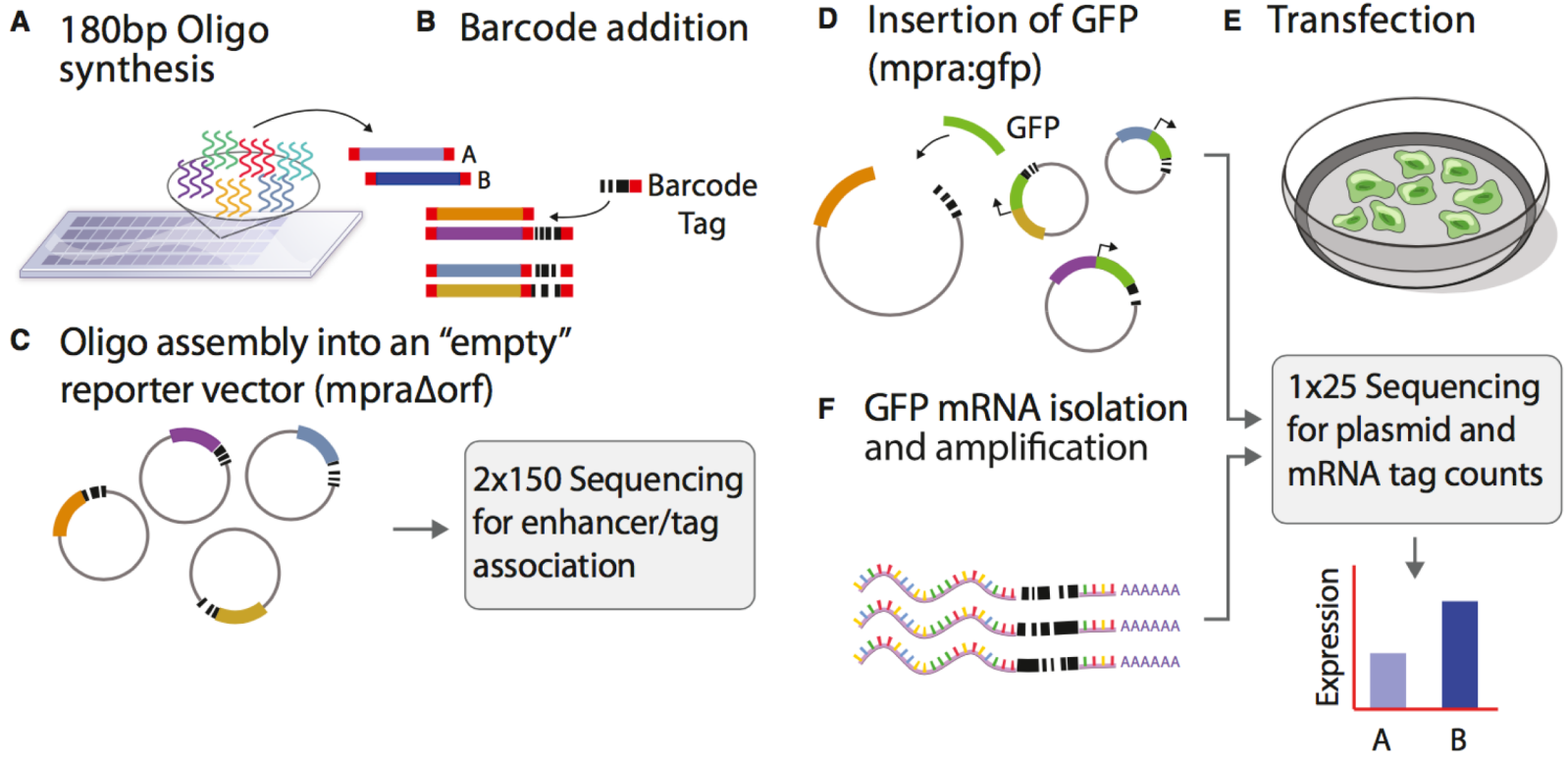


Motivation

Can we learn regulatory grammar by functionally characterizing variants related to human health and disease?

Aim 2

Direct identification of CREs and noncoding regulatory variants via high-throughput reporter assays



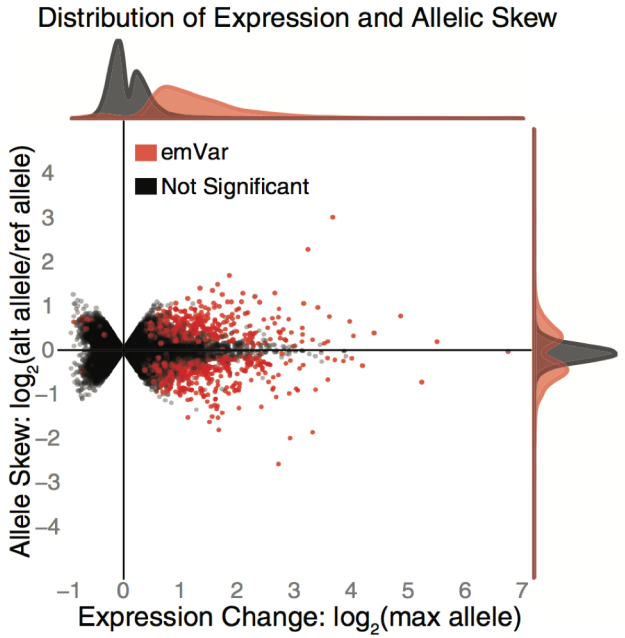
Melnikov et al. 2012
Patwardhan et al. 2012
Ulirsch et al. 2016
Tewhey et al. 2016

Aim 2

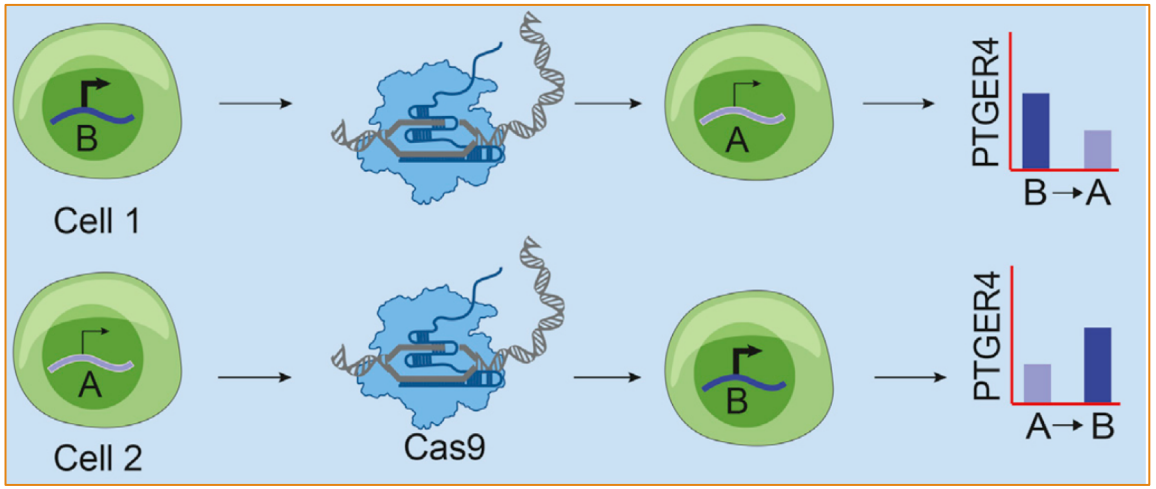
Direct identification of CREs and noncoding regulatory variants via high-throughput reporter assays

MPRAs can identify functional variants at scale

Tewhey et al. 2016



From 29,173 candidates to 842 regulatory variants

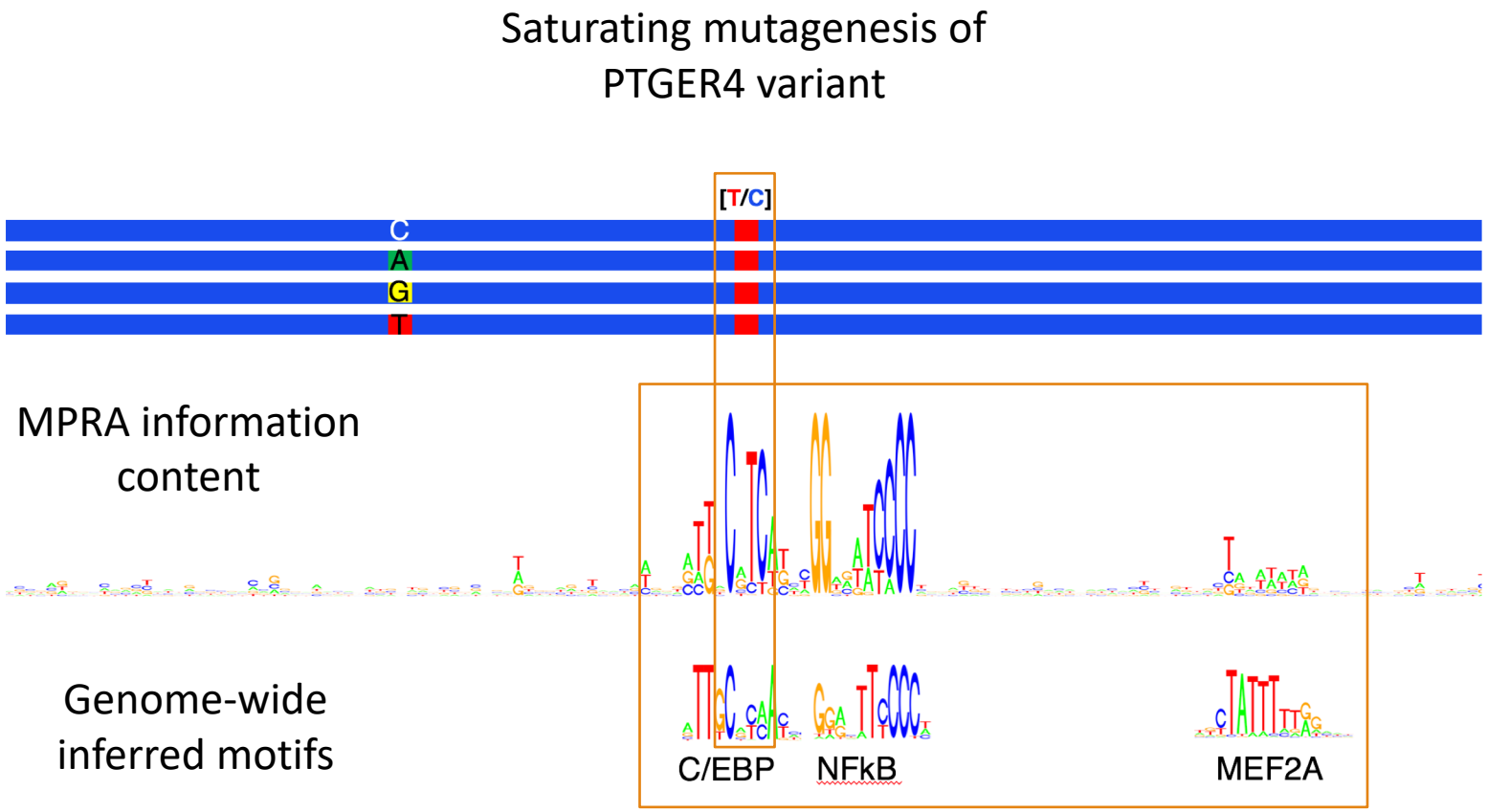


Aim 2

Direct identification of CREs and noncoding regulatory variants via high-throughput reporter assays

From variant to regulatory grammar

Tewhey lab, unpublished



De novo reconstruction of regulatory grammar at **individual** elements

Aim 1

Genetic fine-mapping of complex traits and eQTLs

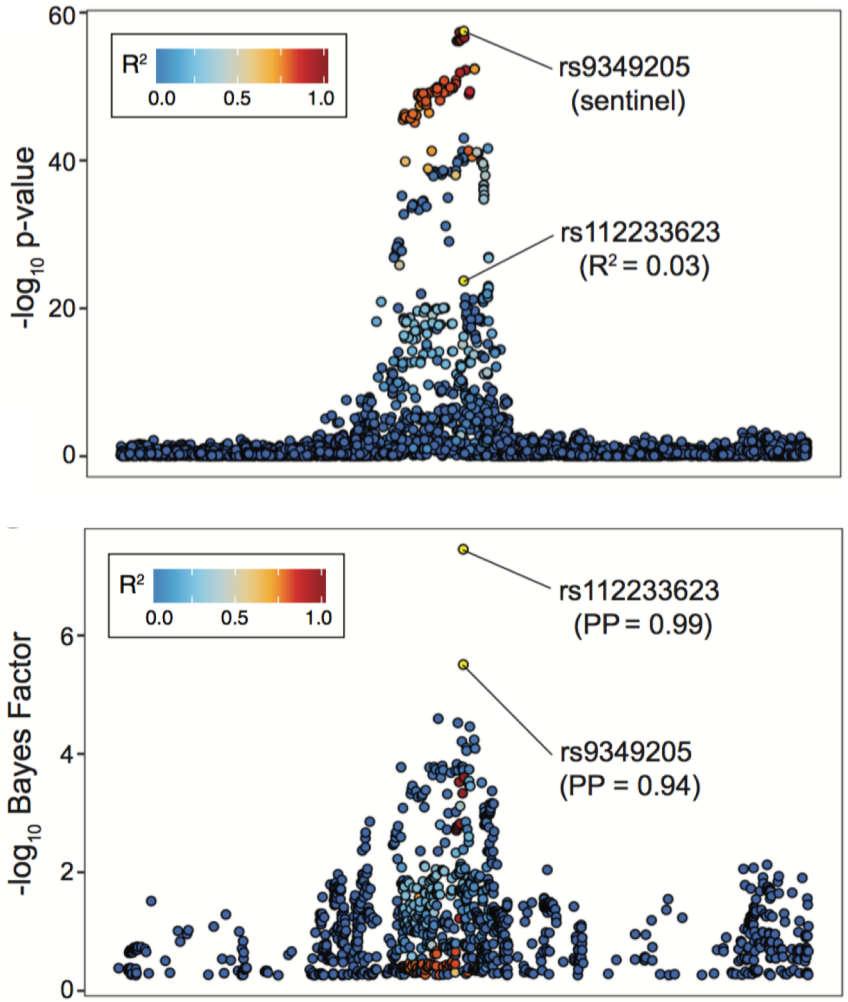
How can we be thoughtful about what variants to include?
(massive ≠ infinite)

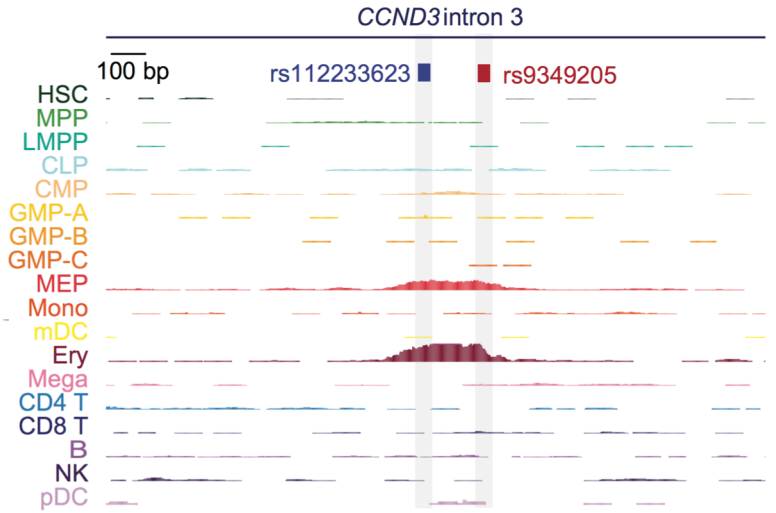
Genetic fine-mapping!

Aim 1

Genetic fine-mapping of complex traits and eQTLs

Building fine-mapping intuition through an example



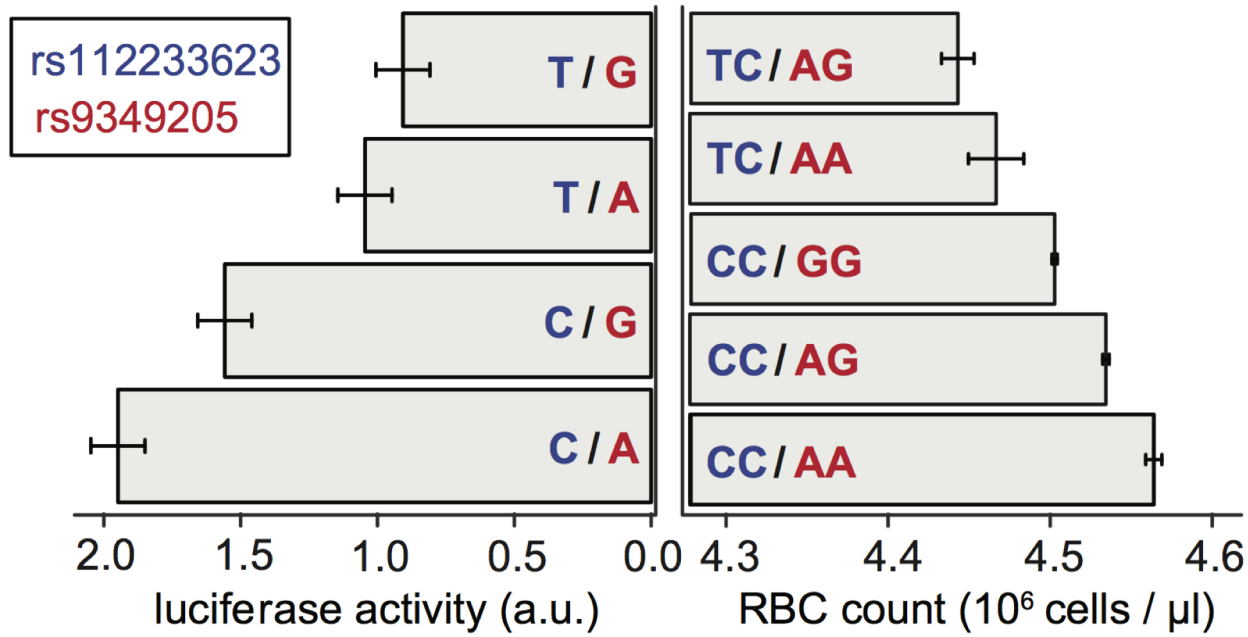


Lareau*, Ulirsch*, Bao* et al. bioRxiv

Aim 1

Genetic fine-mapping of complex traits and eQTLs

Building fine-mapping intuition through an example



Big picture – how does fine-mapping work?

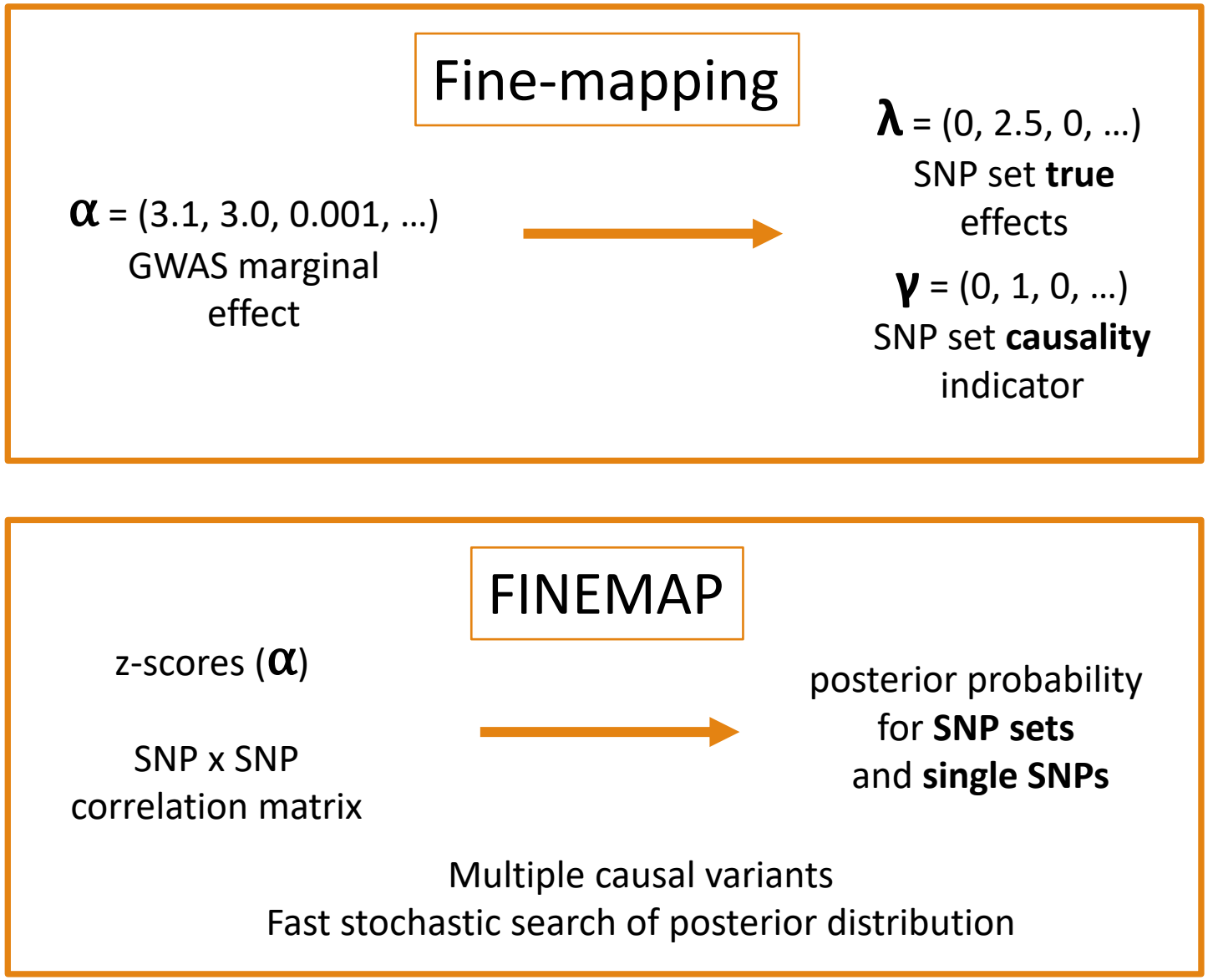
Lareau*, Ulirsch*, Bao* et al. bioRxiv

Aim 1

Genetic fine-mapping of complex traits and eQTLs

Our method (FINEMAP)

Benner et al. 2016
Benner et al. bioRxiv
Lareau*, Ulirsch*, Bao* et al. bioRxiv



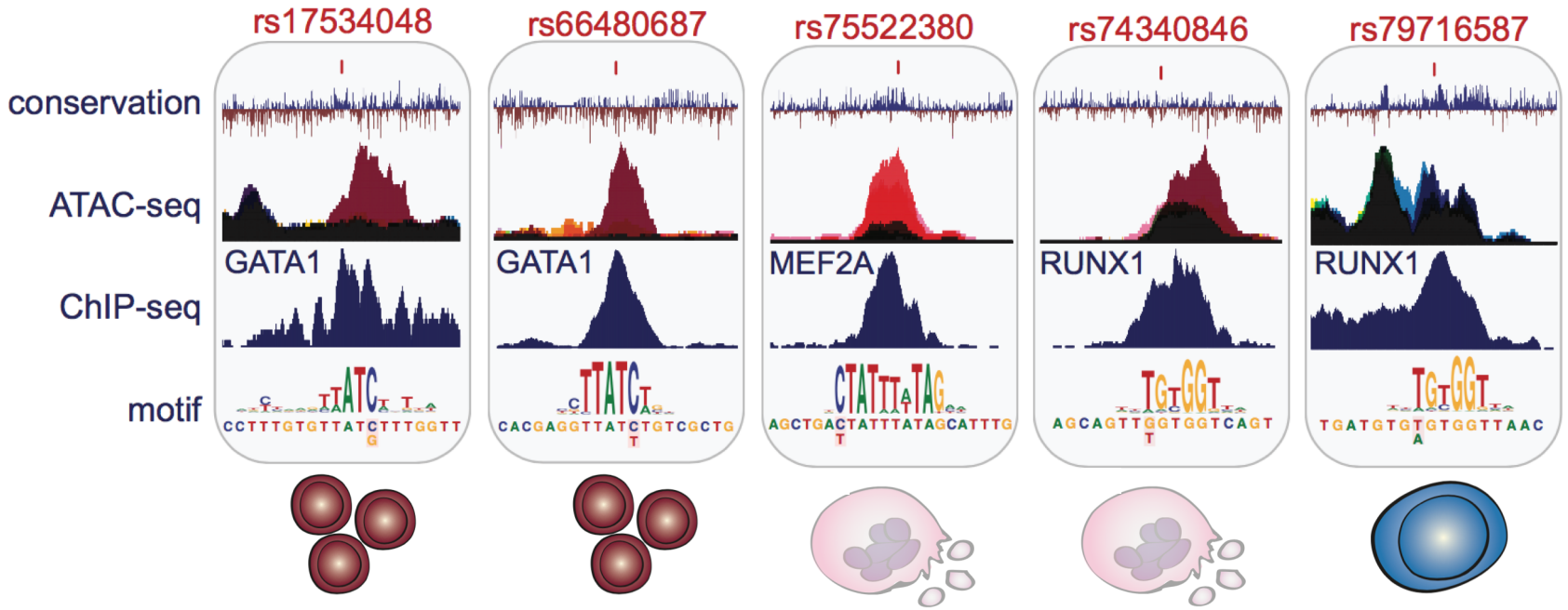
Aim 1

Genetic fine-mapping of complex traits and eQTLs

Overlap of fine-mapped blood cell trait variants with ChIP-seq and known motifs

Lareau*, Ulirsch*, Bao* et al. bioRxiv

Examples from fine-mapping 16 blood cell traits



Functions unclear for most fine-mapped variants

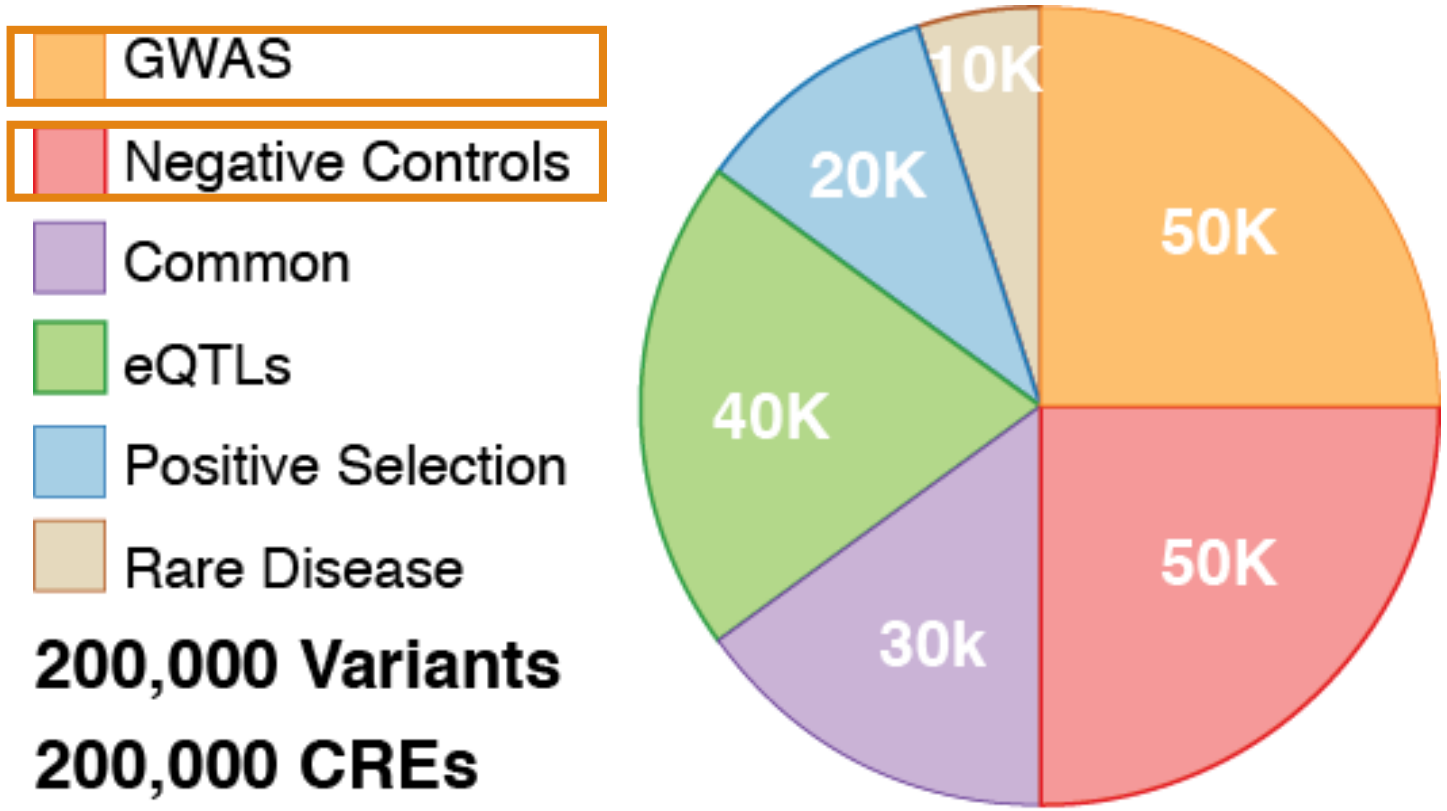
Fine-mapping cannot resolve high LD regions

→ MPRA

Aim 2

Direct identification of CREs and noncoding regulatory variants via high-throughput reporter assays

Overall experimental design



MPRA

5x Cell Lines: *GM12878, K562, HepG2, IMR-90, SK-N-SH*

3x Replicates

Aim 2

Direct identification of CREs and noncoding regulatory variants via high-throughput reporter assays

UK Biobank – many phenotypes

Public resource!
~500,000 genotyped individuals
~11,000,000 high quality variants
~2,000 phenotypes

Aim 2

Direct identification of CREs and noncoding regulatory variants via high-throughput reporter assays

GWAS test sets

~20 heritable phenotypes

- Diabetes, white blood cell count, BMI, education, CVD, etc.

Tier 1) Fine-mapped variants

- All variants > 10% posterior probability

Tier 2) LD blocks for top associations

- All variants with $R^2 > 0.8$ for top 20 GWAS "hits" for each trait

Tier 3) Annotation nominated variants

- All variants > 1% posterior probability in ATAC-seq peaks

Tier 4) Sub-projects

- Haplotypes, regions with > 3 signals, pleiotropic regions, saturation mutagenesis, etc.

Aim 2

Direct identification of CREs and noncoding regulatory variants via high-throughput reporter assays

Control sets

Tier 1) Controls for fine-mapped variants

- Position matched to fine-mapped variants (< 2kb)
- Low LD to fine-mapped variant
- High p-value, low posterior probability

Tier 2) Distribution matched controls

- MAF
- Imputation quality
- ENCODE annotation matched
- LDscore

Tier 3) Random negative controls

- Not in LD with GWAS loci
- Not strong GTEx eQTL