Inferring enhancer and noncoding RNA dysregulation underlying 2,419 UK Biobank Phenotypes

Alexandre Amlie-Wolf
PhD Candidate, Genomics & Computational Biology
University of Pennsylvania Perelman School of Medicine

PNGC
Penn Neurodegeneration Genomics Center

INFERNO

Perelman School of Medicine
University of Pennsylvania

Outline

- Noncoding genetics / enhancer background
- INFERNO methodology
- UK biobank data description and preprocessing
- INFERNO analysis across UK biobank phenotypes
- Signal prioritization and multiple sclerosis ICD10 results

Vast majority of GWAS signals are noncoding

- Need to characterize:
  - Affected regulatory mechanism
  - Relevant tissue context
  - Target genes
  - Downstream biological processes

Published Genome-Wide Associations as of May 2018

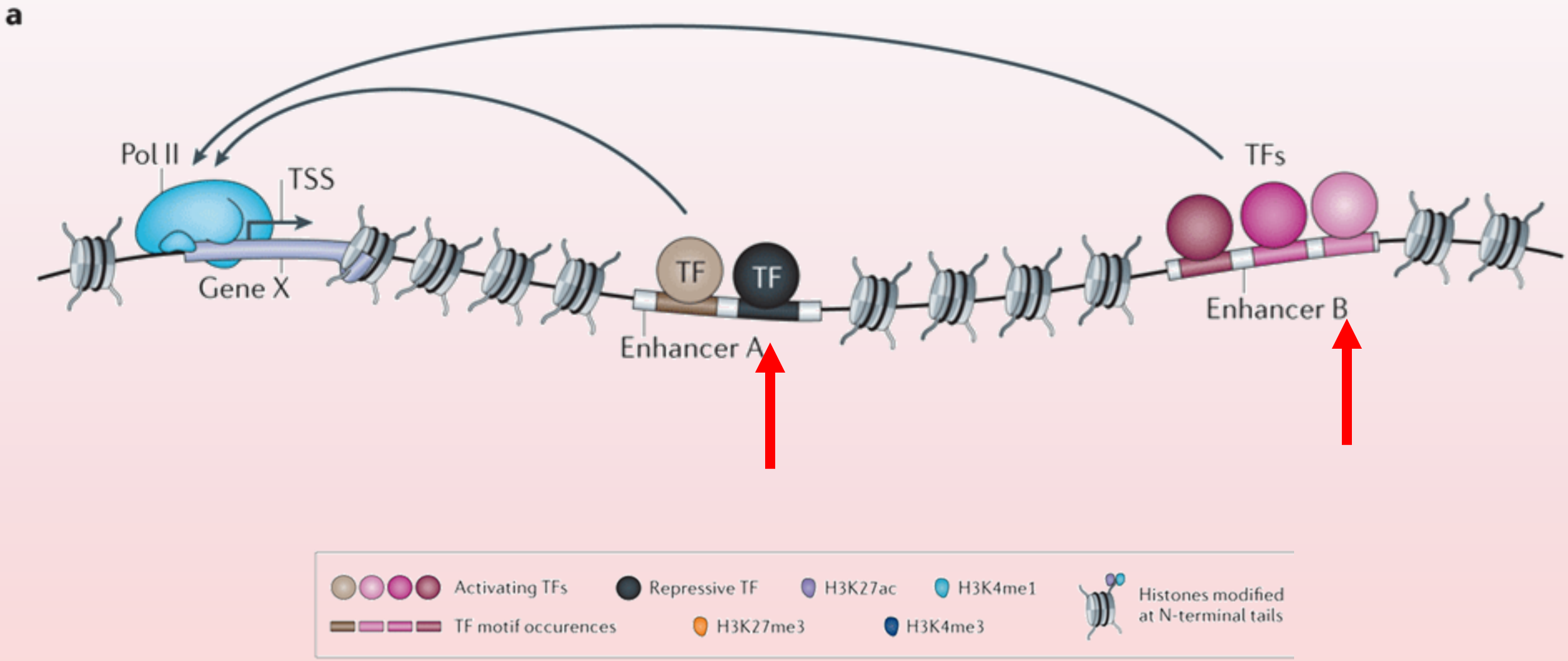
p≤5X10-8 for 17 trait categories

NIH National Human Genome Research Institute

EMBL-EBI

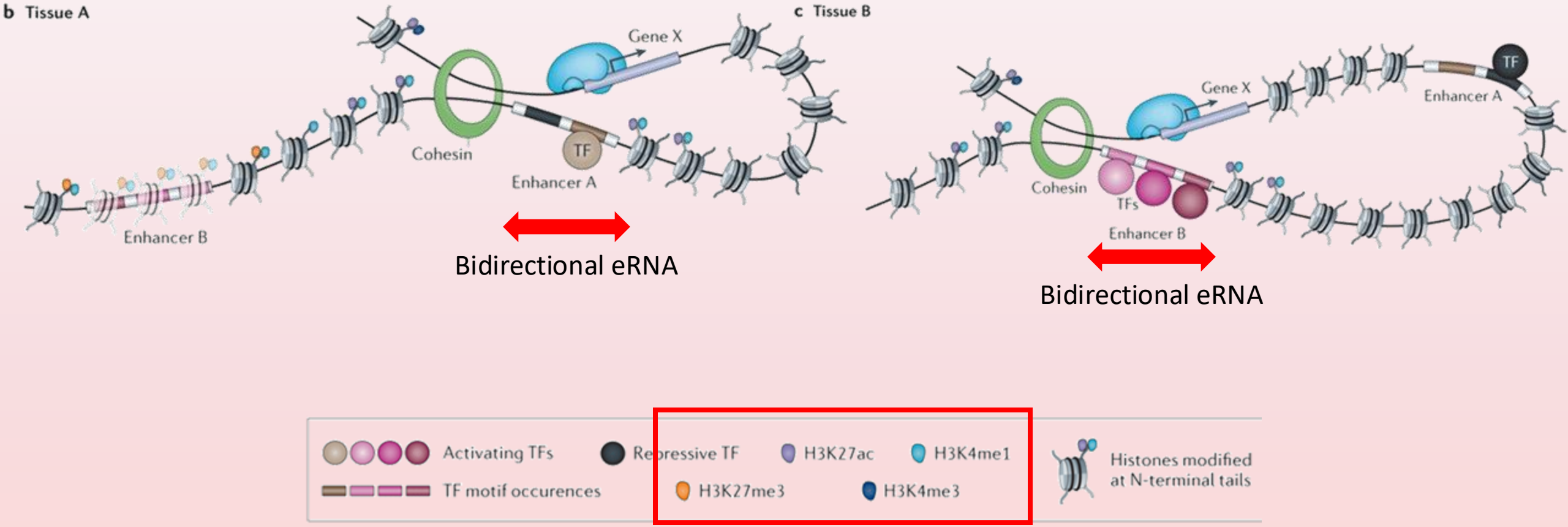
NHGRI-EBI GWAS Catalog
www.ebi.ac.uk/gwas

Noncoding variants may affect transcriptional enhancers



Shlyueva et al., NRG 2014

Enhancers are tissue-specific and have stereotypical properties

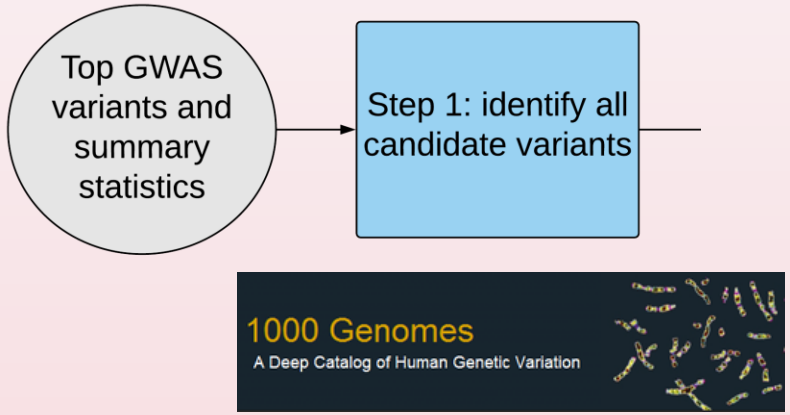


Shlyueva et al., NRG 2014

Outline

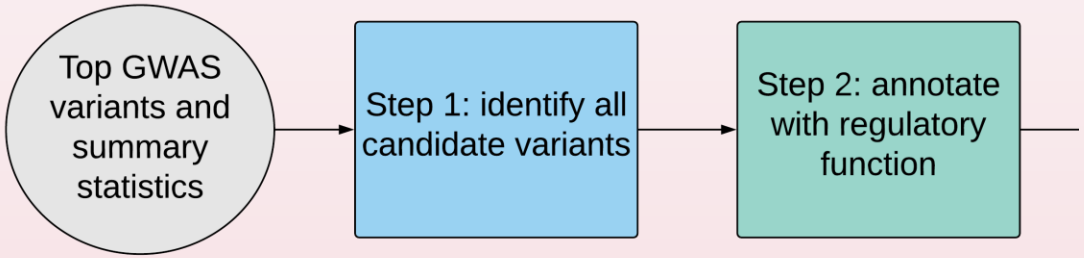
- Noncoding genetics / enhancer background
- INFERNO methodology
- UK biobank data description and preprocessing
- INFERNO analysis across UK biobank phenotypes
- Signal prioritization and multiple sclerosis ICD10 results

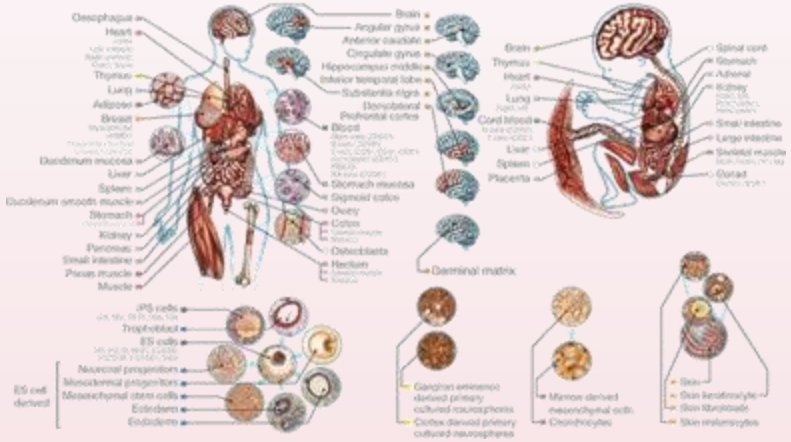
INFERNO: INFERring the molecular mechanisms of NOncoding genetic variants



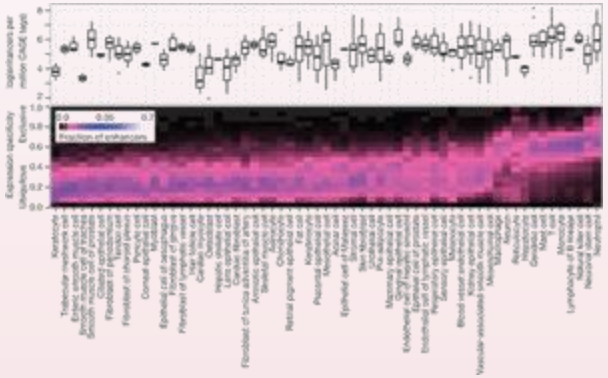
Amlie-Wolf et al., NAR 2018

INFERNO: INFERring the molecular mechanisms of NOncoding genetic variants





Roadmap ChromHMM enhancers
(127 tissues + cell types)



FANTOM5 eRNA enhancers
(112 tissues + cell types)

Leung, Y. Y., Kuksa, P. P., **Amlie-Wolf, A.**, Valladares, O., Ungar, L. H., Kannan, S., Gregory B.D., & Wang, L. S. (2016). DASHR: database of small human noncoding RNAs. *Nucleic acids research*, *44*(D1), D216-D222.

Kuksa PP, **Amlie-Wolf A**, Katanić Ž, Valladares O, Wang L-S, Leung YY. DASHR 2.0: integrated database of human small non-coding RNA genes and mature products. *Bioinformatics*. 2018.

dashr

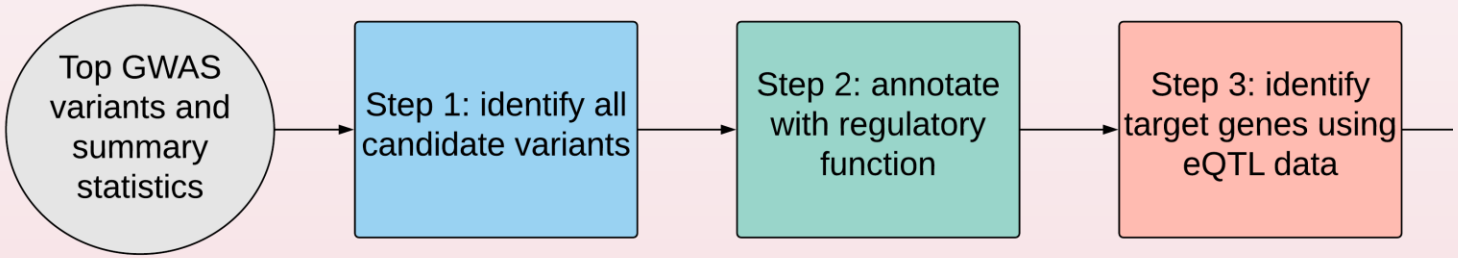
Database of Small Human Noncoding RNAs

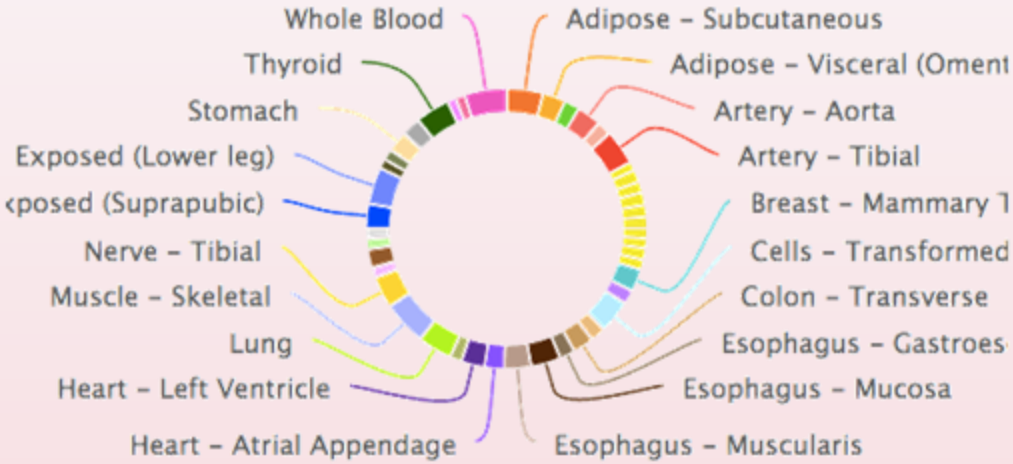
185 tissues + cell type

HOMER
TFBSs

Amlie-Wolf et al., NAR 2018

INFERNO: INFERring the molecular mechanisms of NOncoding genetic variants

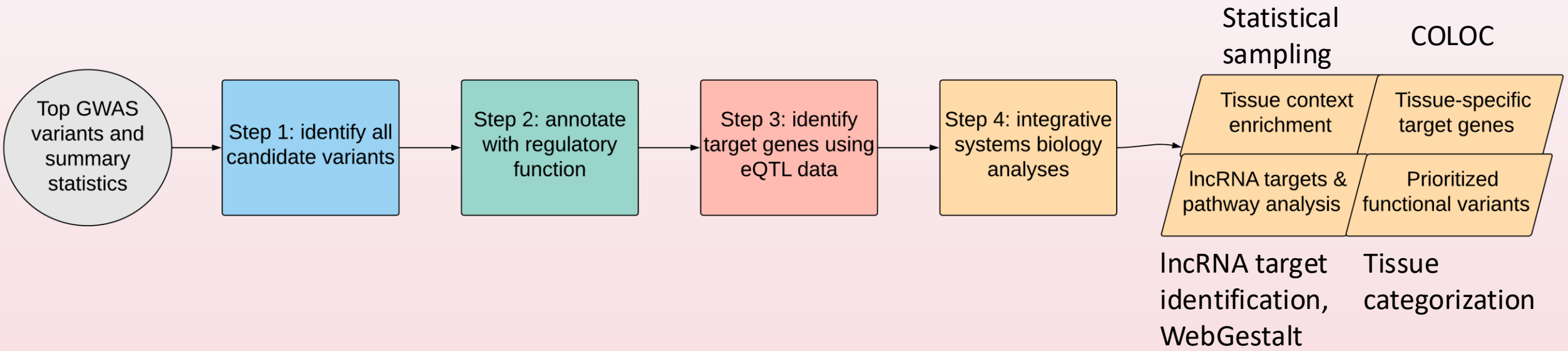




GTEx eQTLs (44 tissues + cell types)

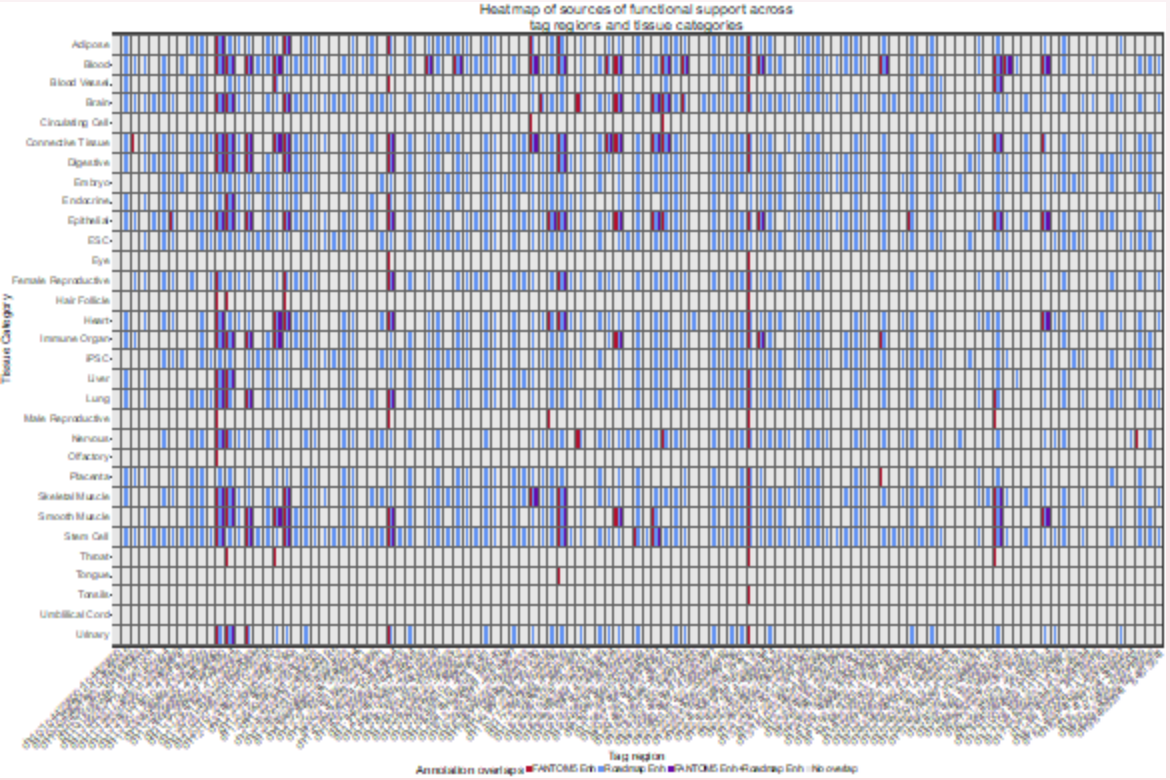
Amlie-Wolf et al., NAR 2018

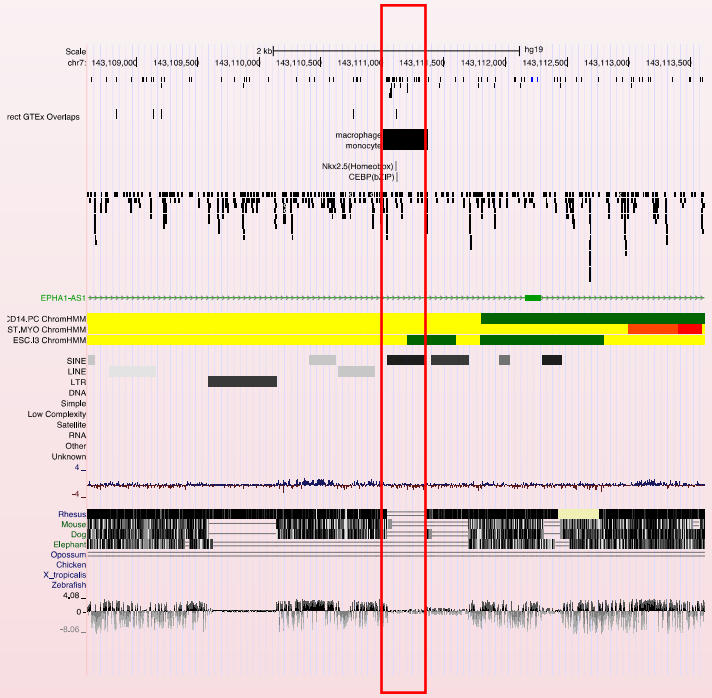
INFERNO: INFERring the molecular mechanisms of NOncoding genetic variants

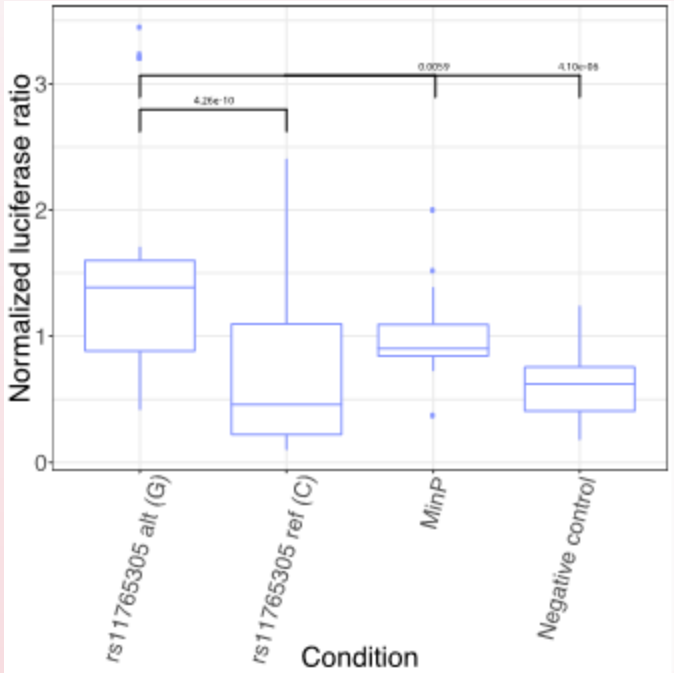


- Open source pipeline implemented in R, Python, and bash
- **Amlie-Wolf A**, Tang M, Mlynarski EE, Kuksa PP, Valladares O, Katanic Z, Tsuang D, Brown CD, Schellenberg GD, Wang LS. INFERNO: inferring the molecular mechanisms of noncoding genetic variants. Nucleic Acids Research 2018:211599. doi:10.1093/nar/gky686.

INFERNO disease applications







- In schizophrenia, recapitulated known disease genes including *CACNA1C* (Amlie-Wolf et al., NAR 2018)

- In Alzheimer's disease, recapitulated known signals and identified novel lncRNA mechanisms (Amlie-Wolf et al., bioRxiv 2018)
- https://bit.ly/2No5MIn

Outline

- Noncoding genetics / enhancer background
- INFERNO methodology
- UK biobank data description and preprocessing
- INFERNO analysis across UK biobank phenotypes
- Signal prioritization and multiple sclerosis ICD10 results

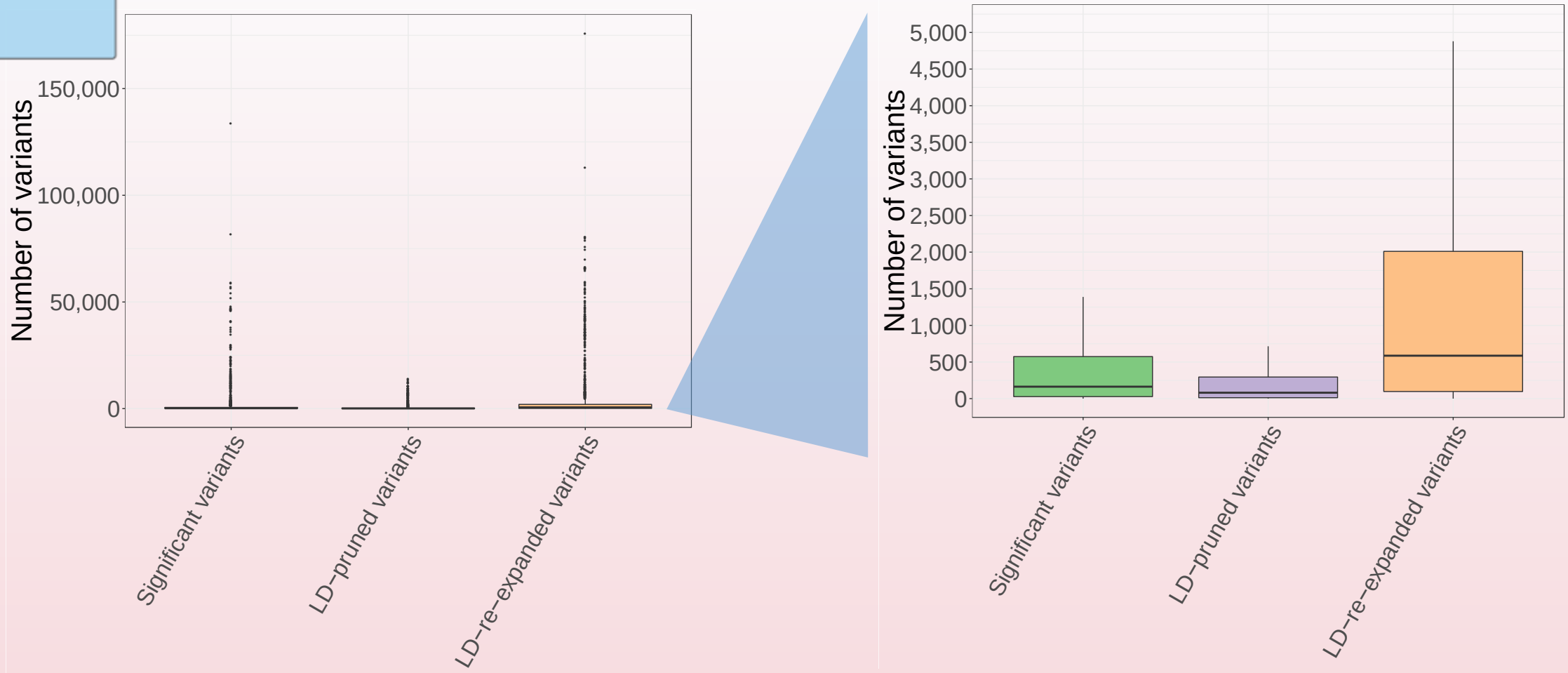
biobank uk
Improving the health of future generations

- Using Ben Neale's Round 1 Hail GWAS results on ~337,000 individuals
- 2,419 phenotypes: 191 quantitative and 2,228 case/control

Parent Category	# Case/Control	# Quantitative	Child Categories
Anthropometry	0	36	Body size measures; Impedance measures
Cognitive function	0	4	Fluid intelligence test; Pairs matching test; Prospective memory test; Reaction time test
Health and medical history	139	8	Artery disease; Breathing; Cancer screening; Chest pain; Eyesight; General health; Hearing; Medical conditions; Medication; Mouth; Operations; Pain
Health-related outcomes	783	2	Cancer register; Death register; ICD10 diagnosis
Lifestyle and environment	44	57	Alcohol; Physical activity; Sleep; Smoking
Physical activity measurement	0	1	Wear time duration
Physical measures	4	29	Arterial stiffness; Blood pressure; Bone-densitometry of heel; ECG during exercise; Hand grip strength; Spirometry; Urine metabolites
Population characteristics	0	1	Baseline characteristics
Psychosocial factors	35	15	Mental health
Recruitment	10	0	Reception
Sex-specific factors	11	16	Female-specific factors; Male-specific factors
Sociodemographics	20	8	Education; Employment
Touchscreen	49	7	Early life factors; Family history
Verbal interview	1118	6	Early life factors; Employment; Medical conditions; Medication; Operations
Work environment	15	1	Medical information

Step 1: identify all candidate variants

Identifying significant UK biobank signals



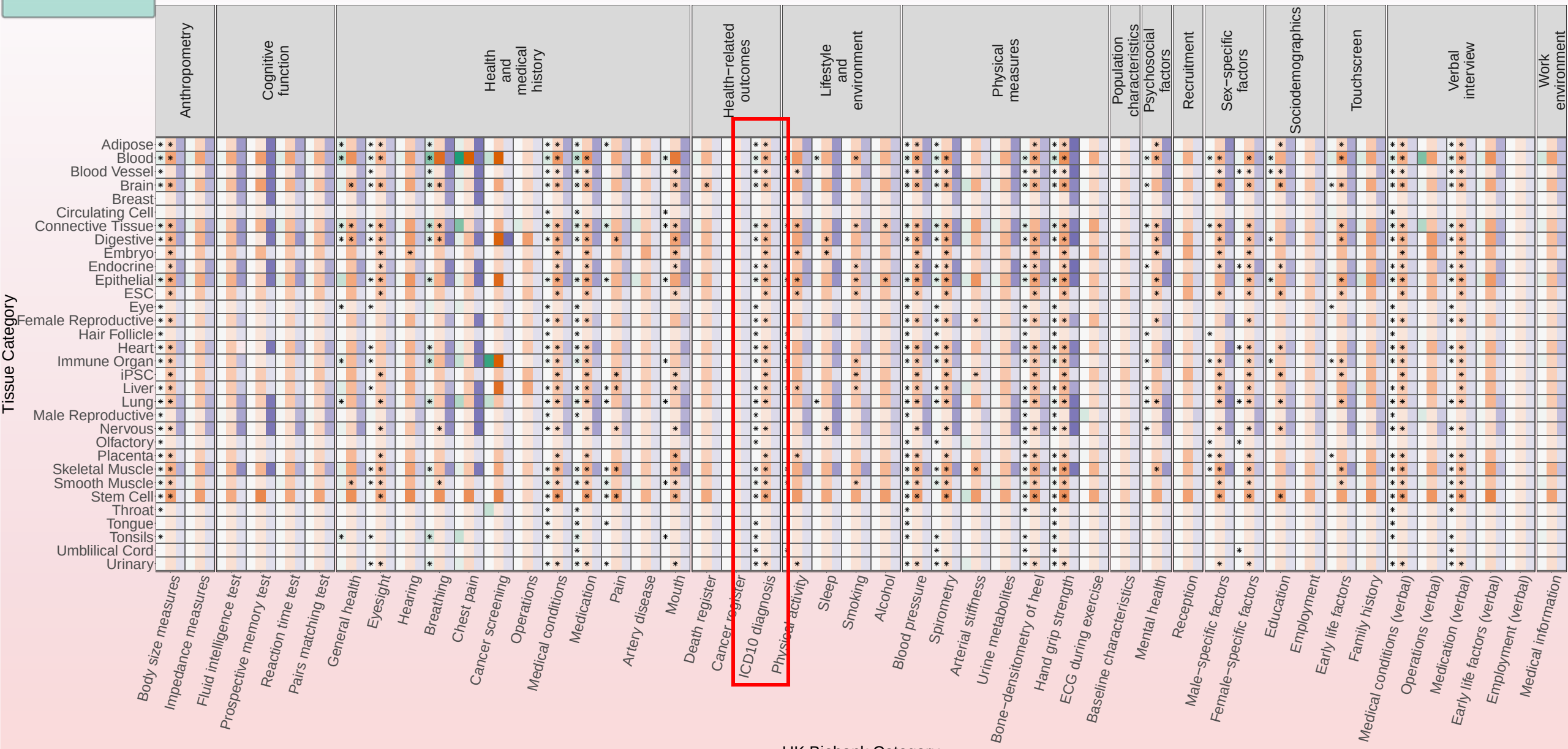
1. Identify 1,389,198 significant variants ($p \leq 5 \times 10^{-8}$) in 2,298 phenotypes
2. LD prune significant variants ($r^2 \geq 0.7$) to yield "tag" variants (INFERNO input)
3. Re-expand by LD to identify all putative causal variants

Outline

- Noncoding genetics / enhancer background
- INFERNO methodology
- UK biobank data description and preprocessing
- INFERNO analysis across UK biobank phenotypes
- Signal prioritization and multiple sclerosis ICD10 results

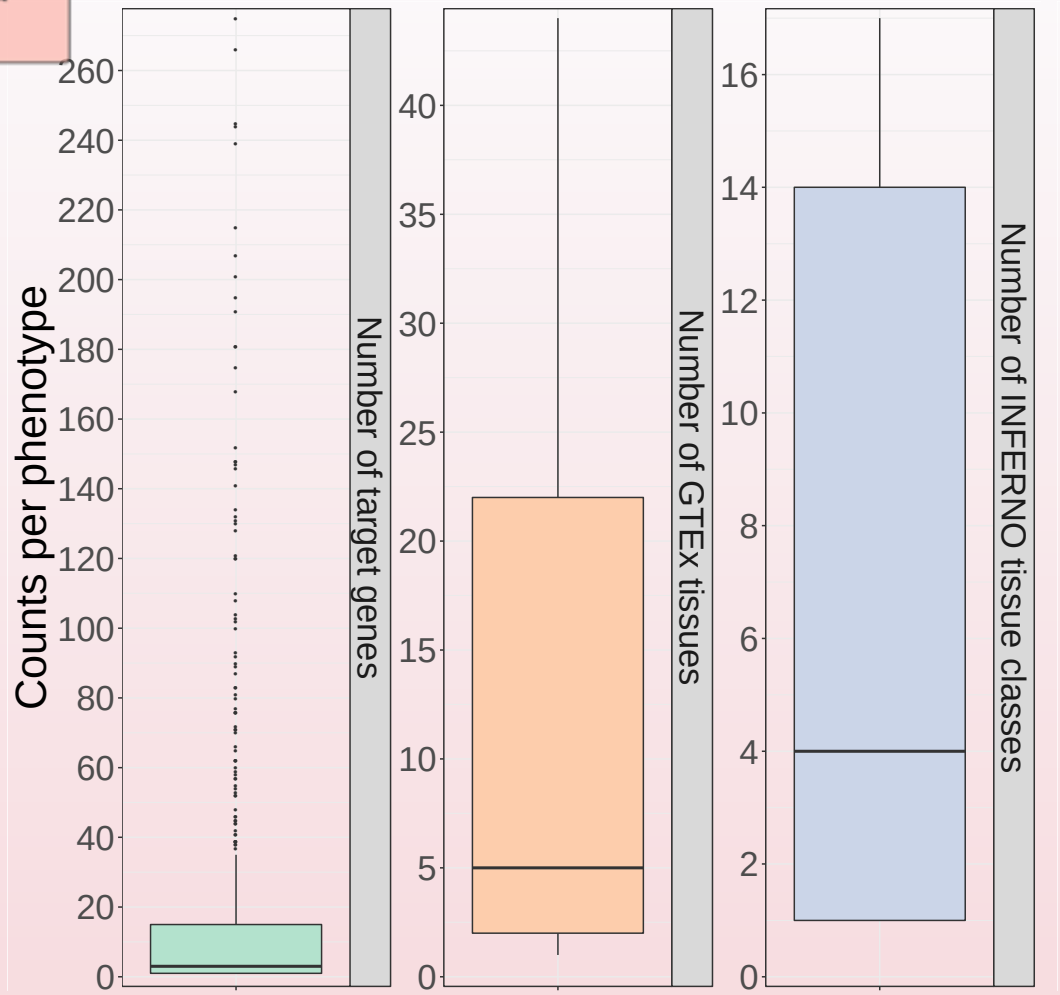
Step 2: annotate with regulatory function

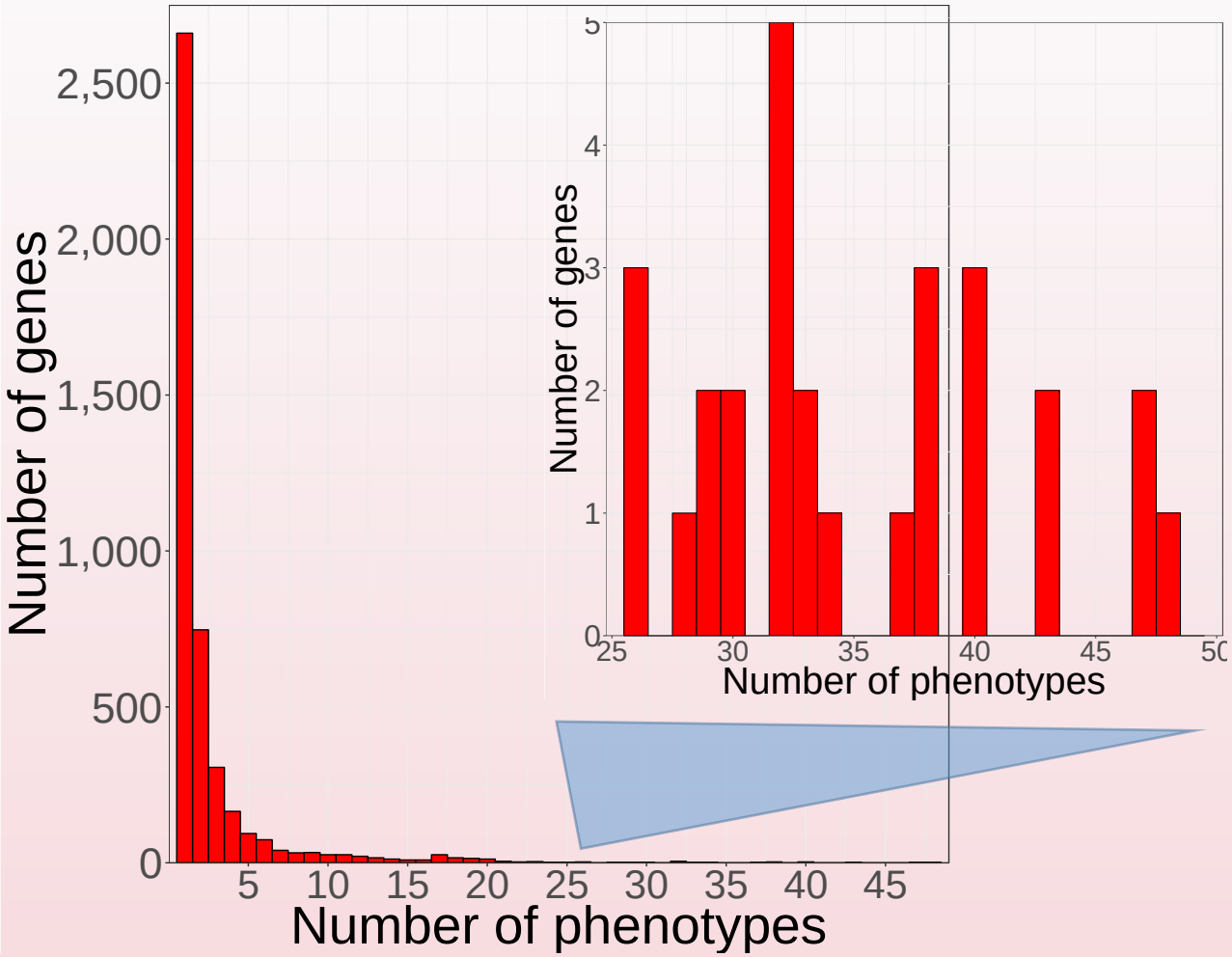
Widespread tissue-specific regulatory enrichment



Step 3: identify target genes using eQTL data

Co-localized eQTL / GWAS signals in a subset of phenotypes





- eQTL targets included 522 lncRNAs regulating ~16k genes
- Leukocyte antigen MHC gene ***LY6G5B*** co-localized with 48 phenotypes spanning mental health, medical conditions, general health, and specific medications

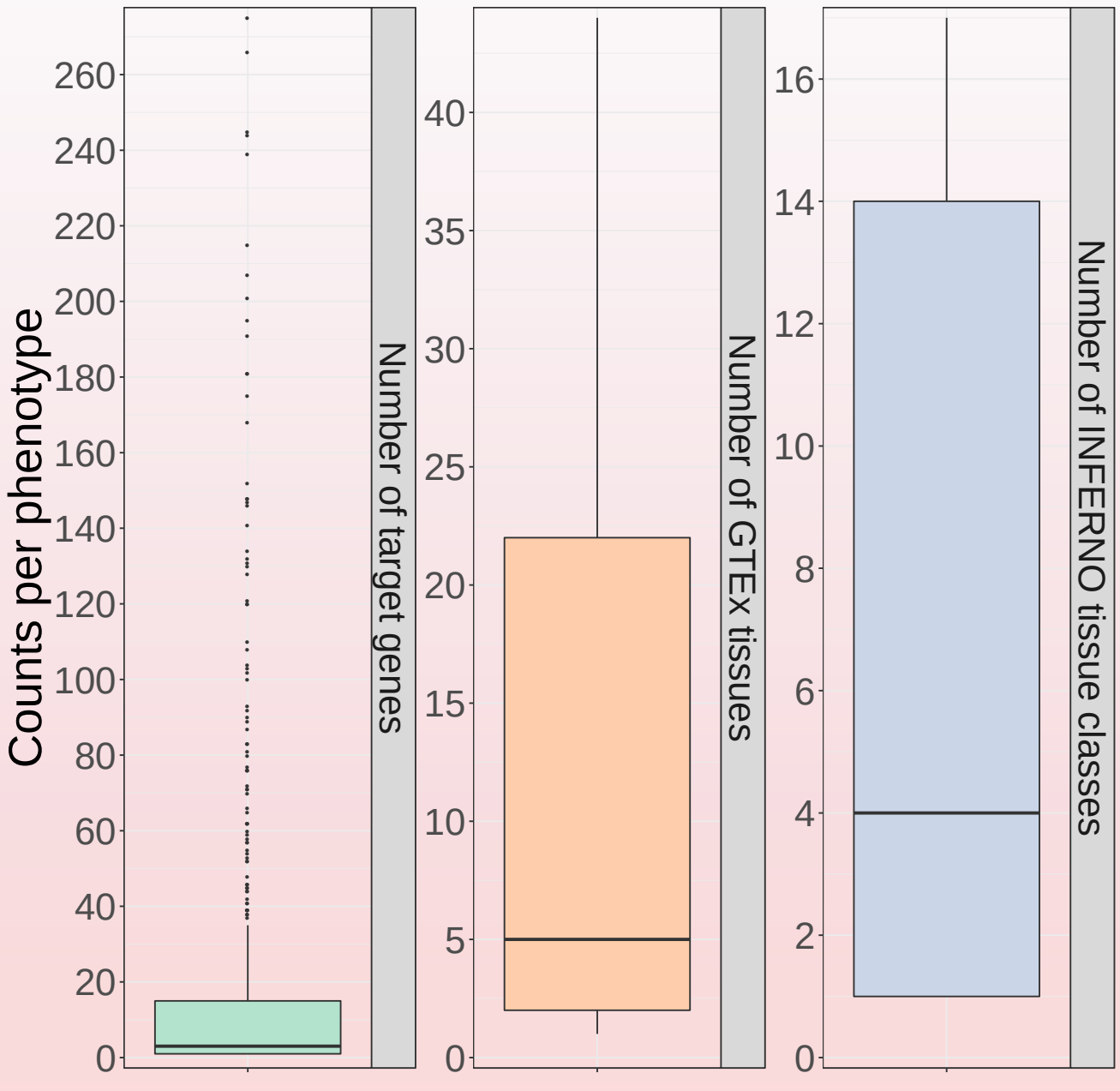
Outline

- Noncoding genetics / enhancer background
- INFERNO methodology
- UK biobank data description and preprocessing
- INFERNO analysis across UK biobank phenotypes
- Signal prioritization and multiple sclerosis ICD10 results

Step 4: integrative systems biology analyses

Integrative signal prioritization

Co-localization signals
4,381 genes
17 INFERNO categories
(all 44 GTEx tissues)
616 phenotypes

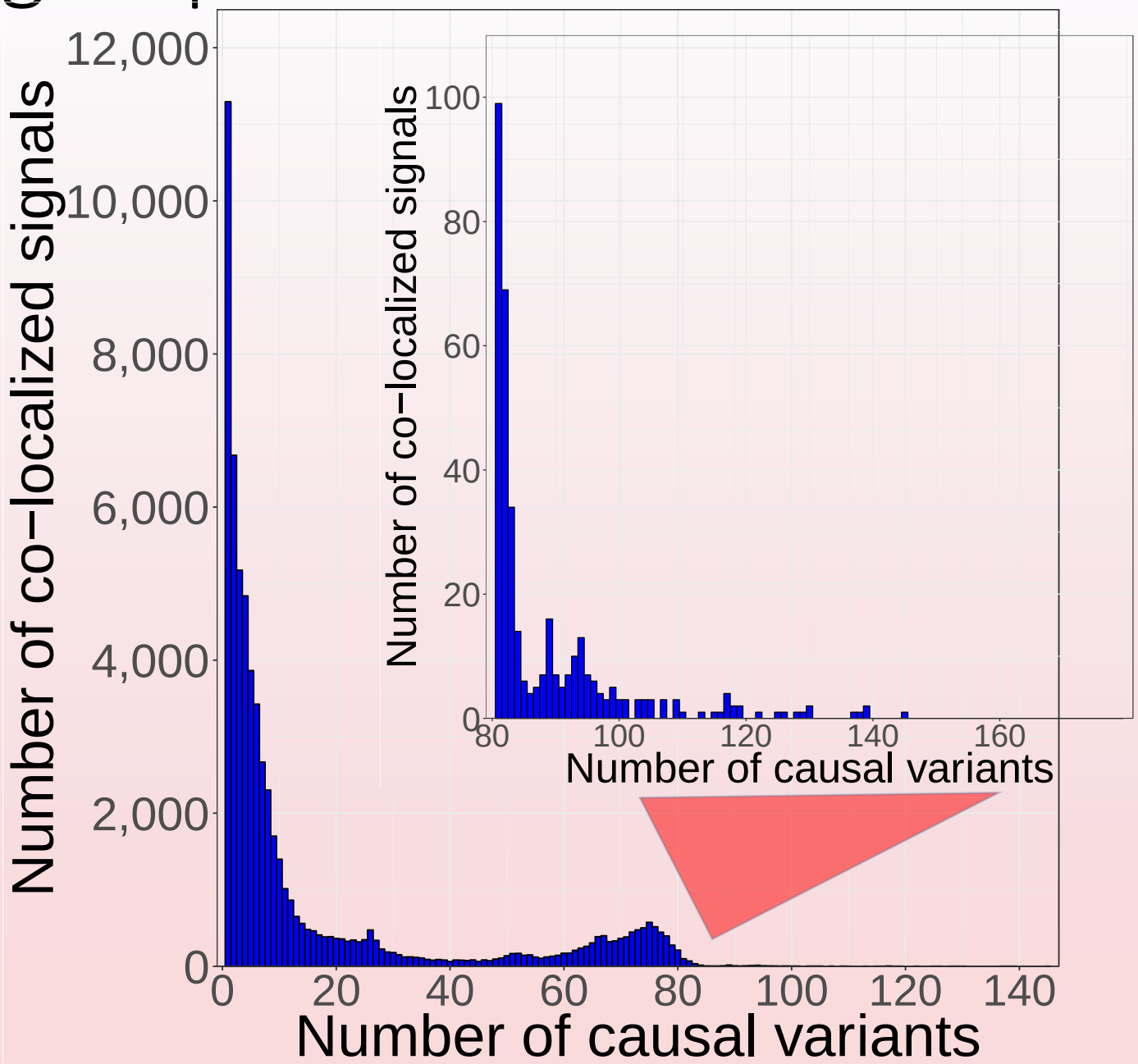


Step 4: integrative systems biology analyses

Integrative signal prioritization

Co-localization signals
4,150 genes
17 INFERNO categories
(all 44 GTEx tissues)
562 phenotypes

Causal variant sets
Median of 5 variants per tissue-specific co-localized signal



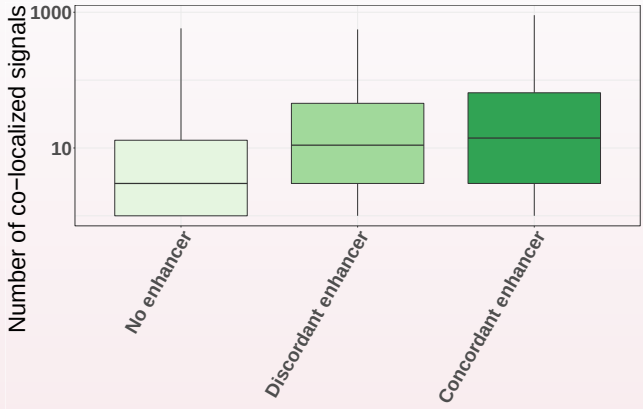
Integrative signal prioritization

Step 4: integrative systems biology analyses

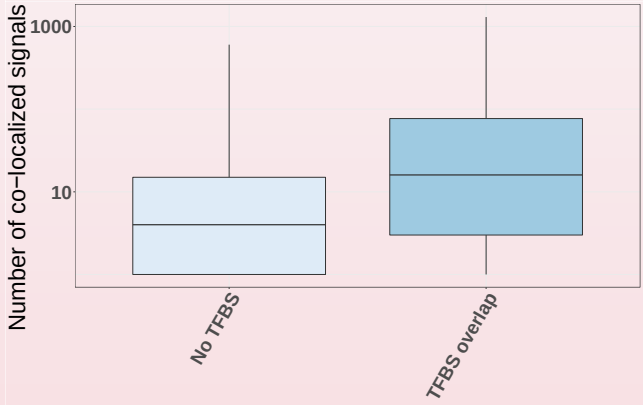
Co-localization signals
4,150 genes
17 INFERNO categories
(all 44 GTEx tissues)
562 phenotypes

Causal variant sets
Median of 5 variants per tissue-specific co-localized signal

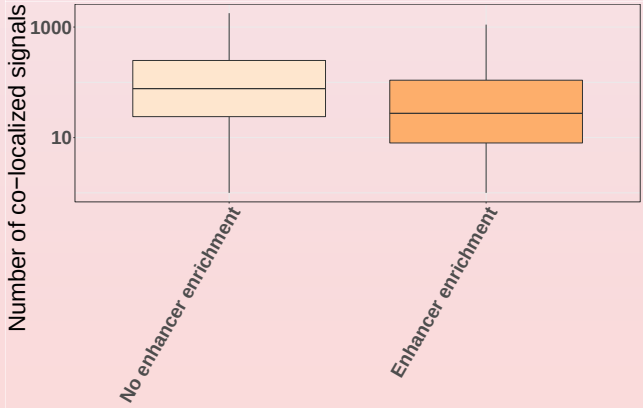
Concordant enhancer overlaps
2,242 genes
15 tissue categories
279 phenotypes



TFBS overlaps
3,347 genes
17 tissue categories
344 phenotypes

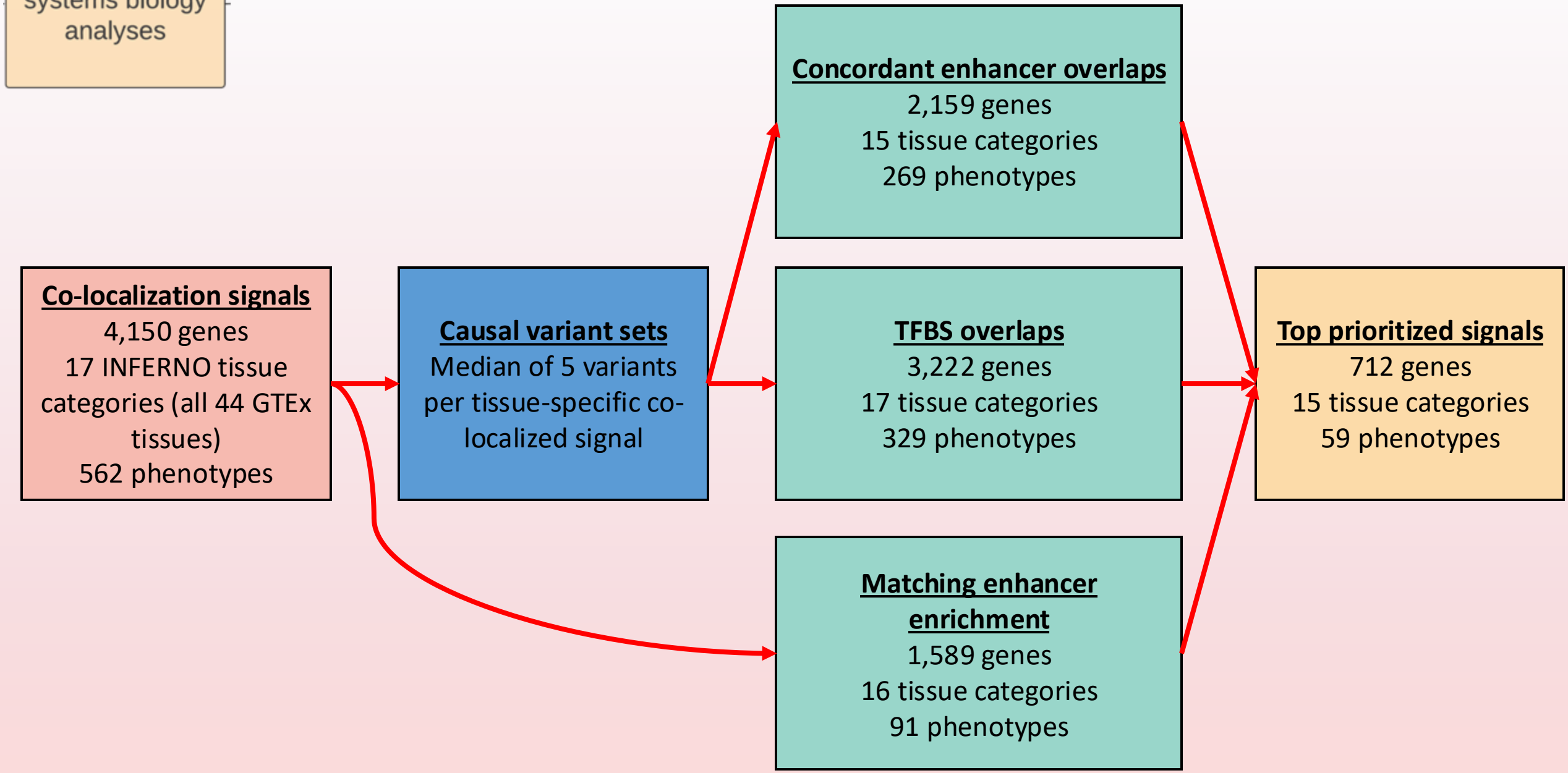


Matching enhancer enrichment
1,589 genes
16 INFERNO categories
92 phenotypes



Step 4: integrative systems biology analyses

Integrative signal prioritization



INFERNO+UK Biobank discovers MS ICD10 signals

INFERNO Tissue Class	Field	GTEx Tissues	Target Genes	Matching Enhancer Enrichments
Blood	Diagnoses - main ICD10: G35 Multiple sclerosis	Cells EBV-transformed lymphocytes; Whole Blood	*HLA-DQB1; LY6G5B; XXbac-BPG254F23.6; HLA-DRB1; HLA-DQB1-AS1; HLA-DRB5*	FANTOM5
Brain	Diagnoses - main ICD10: G35 Multiple sclerosis	Brain Cerebellum; Brain Putamen basal ganglia; Brain Hippocampus; Brain Hypothalamus; Brain Cerebellar Hemisphere; Brain Nucleus accumbens basal ganglia; Brain Frontal Cortex BA9	*HLA-DRB5*	Roadmap
Digestive	Diagnoses - main ICD10: G35 Multiple sclerosis	Pancreas; Stomach; Esophagus Mucosa; Small Intestine Terminal Ileum; Esophagus Gastroesophageal Junction; Esophagus Muscularis; Colon Transverse; Colon Sigmoid	*VWA7; HLA-DQB1; HLA-DRB6; XXbac-BPG254F23.6; PRRT1; HLA-DQB1-AS1; LY6G5B; HLA-DQA2; TAP2; PPP1R2P1; HLA-DRB5; HLA-DRB1*	FANTOM5, Roadmap, and Both

- Multiple sclerosis is a demyelinating CNS disease (brain) mediated by immune system dysfunction (blood category and HLA genes)
- MS can also lead to significant gastrointestinal problems (digestive)

Conclusions

- INFERNO provides a useful tool for integrating functional genomics data to generate post-GWAS hypotheses
- UK Biobank provides a rich resource for exploring genetic associations with a range of traits
- INFERNO identified enhancer dysregulation and affected target genes in a variety of phenotypes including multiple sclerosis

http://inferno.lisanwanglab.org/

INFERNO

Acknowledgements

- **Li-San Wang**
- Liming Qu
- Elisabeth Mlynarski
- Fanny Leung
- Pavel Kuksa
- Mitchell Tang
- Zile Katanic
- Rhea Bhatta
- Abha Belorkar

- **Jerry Schellenberg**
- **Casey Brown**
- Mingyao Li
- Edward Lee
- Barbara Engelhardt

- T32 AG000255-18: Training in Age Related Neurodegenerative Diseases (Virginia Lee)

PNGC
Penn Neurodegeneration Genomics Center

http://lisanwanglab.org/~alexaml/

@AlexAmlie

alexamlie@gmail.com

http://lisanwanglab.org/

Perelman School of Medicine
University of Pennsylvania