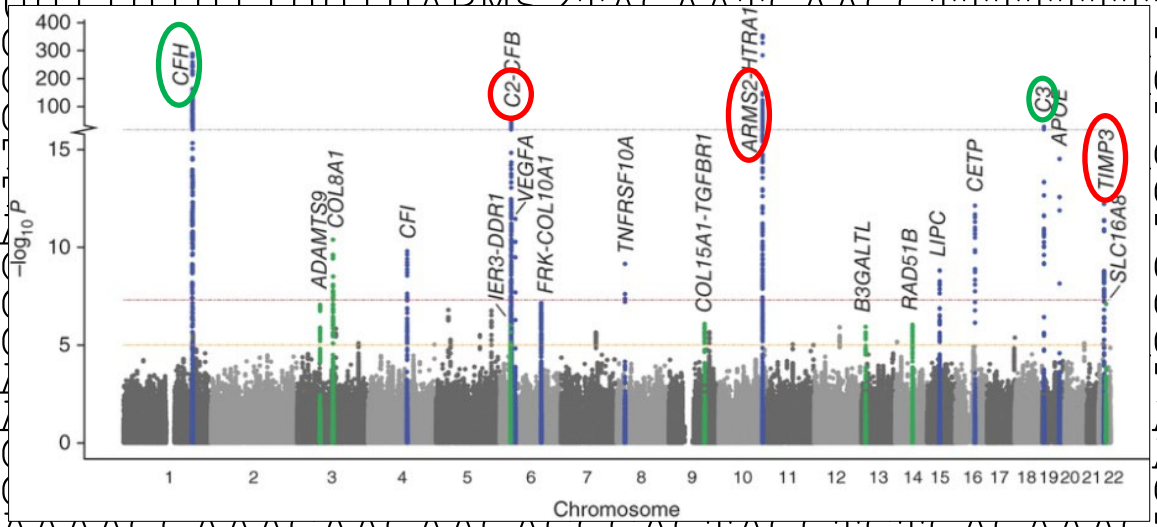
18.700/18.701/6.047/6.878/HST.507/MLCB
Computational Biology: Genomes, Networks, Evolution
Prof. Manolis Kellis

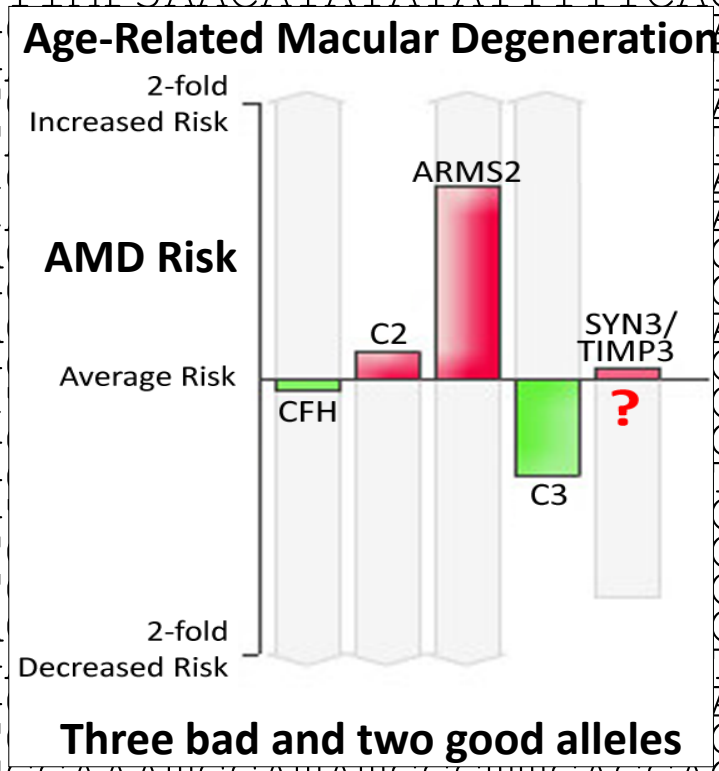
Lecture 18 – Disease Mechanism Circuitry, eQTLs, Mediation, Heritability

- Genetics, Variation, GWAS, PRS/PGS, insights
- From Region to Mechanism / Circuitry
- Quantitative Trait Loci: eQTL/meQTL analysis
- Mediation Analysis + Mendelian Randomization
- Heritability and Systems Genetics

Genetics, Variation, GWAS, PRS, Mechanism

1. Genetics, Variation, GWAS
2. Polygenic scores (PGS)
3. From GWAS to biological insights
4. From Region to Mechanism / Circuitry
5. Quantitative Trait Loci: eQTL/meQTL analysis
6. Mediation Analysis + Mendelian Randomization
7. Heritability and Systems Genetics



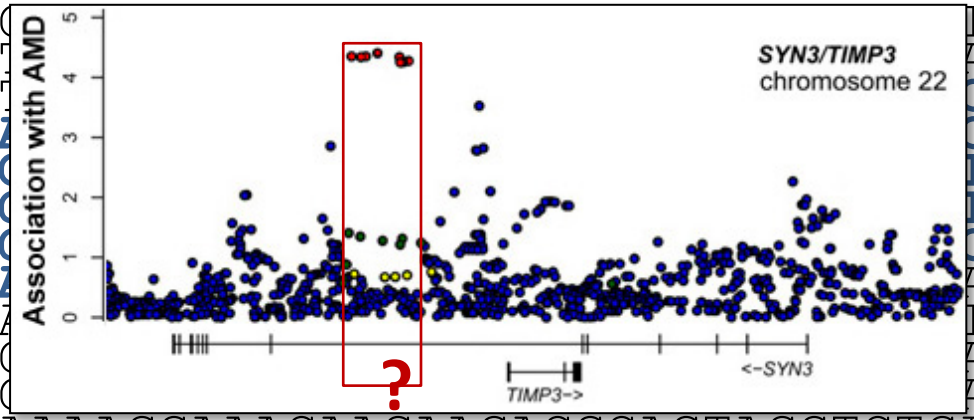


ARMS2

TIMP3/SYN3

C2

ARMS2gene: AAAGCTTCACAGATGATTTCAATGGATACTAGGGACCTCTGTTGCCTCCT
CTGGCAGAGCAGGACTGAGGGGTGGACCCTCCCTGAGACCACCCAACAATTCAGGGTGGA
TTTTTGTGTGACGGGAAAAGACAATGCTCC
AGGGGACAGCACCTTTGTCACCACATTATG
CAGGACCGATGGTAACTGAGGCGGAGGGGA
CCTCGGTGGTTCCTGTGTCCTTCATTTCCA
TTGGTGGAGAAGGAGCCAGTGACAAGCAGA
TCCCAGCTGCTAAAATCCACACTGAGCTCT
CCCAGAGGAGGTTCCAGCAGCCTCAGCACC
AGGCCTTACCCCAGACCTATTGAATCAGAA
TTAACCAGCCTTCAGGTGCTTCTGATGCAT
AGGCCTAGGCAATTCAGCCTTCCTCTGGTT
ATACTTCCAATGCACCTGCTACATGCCAGA
AAAAGGAAACAAGAAGACGCAGTAGGTCTGAGAAAGTGATGGGGTGAGCAGAAACCCAAA
GCTTATAGAAGGCCATCTGAGTGGCCCCTCAAGCCGGTGAATTGGCTTTAGGGTTTACTG



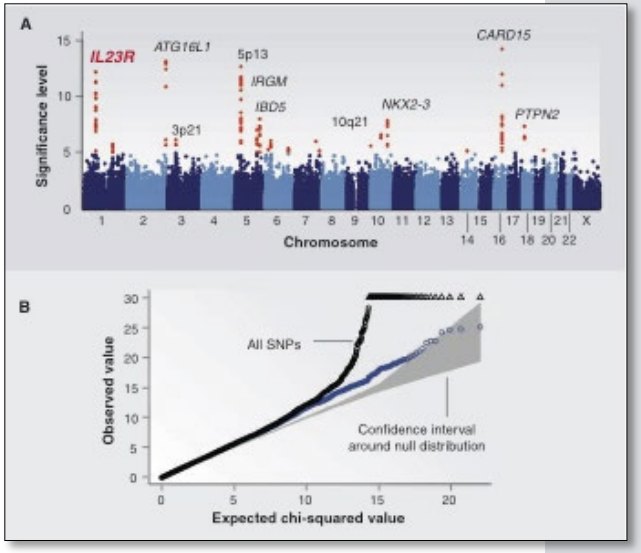
SYN3/TIMP3: ATATATTTTTCAGTGGCAGGAAGTCTTGCCCGAGGTGGGAATGTTACTG
GGTTAATATCTGGGGGAAAGAGAAATATTTTTCCCTTTGTTAGCTGGCTCTGGGCAGCCT
GAAAACTCTTGATCCTCTCTGTCTGCTGCTTGGGACATAATGACCTGCTTTCAATCCCTT
TCAATTACAGGATTTCTGATAGGAATTTGGAAAACAACCTAAATCCCAAAGCTTGGATGG
TAGCCCATGCTTCATTCCACGTCTCTGTACCCAGTTTTTCAAAGAGATTTTTTTTTTTCA

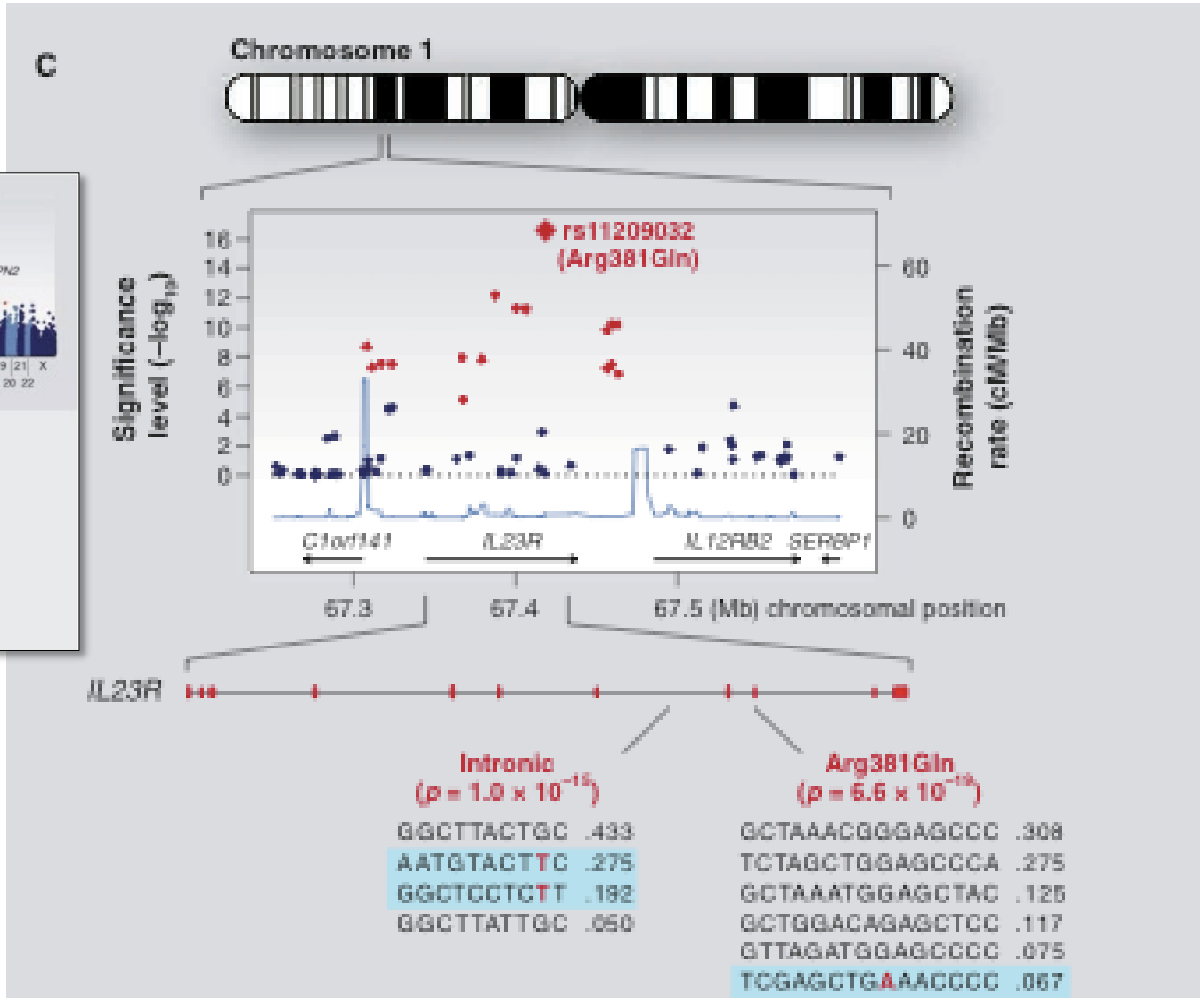
C2gene: TGTTTCCTTGACTGGCAGCTCAGCGGGGCCCTCCCGCTTGGATGTTCCGGGAA
AGTGATGTGGGTAGGACAGGCGGGGCGAGCCGCAGGTGCCAGAACACAGATTGTATAAAA
TGGGCTTCCAGGCAGCCCTGGAAGTCAAGAGAACACTCAGAAATGGGGAGGGAGAAGCAG

Questions for this module:

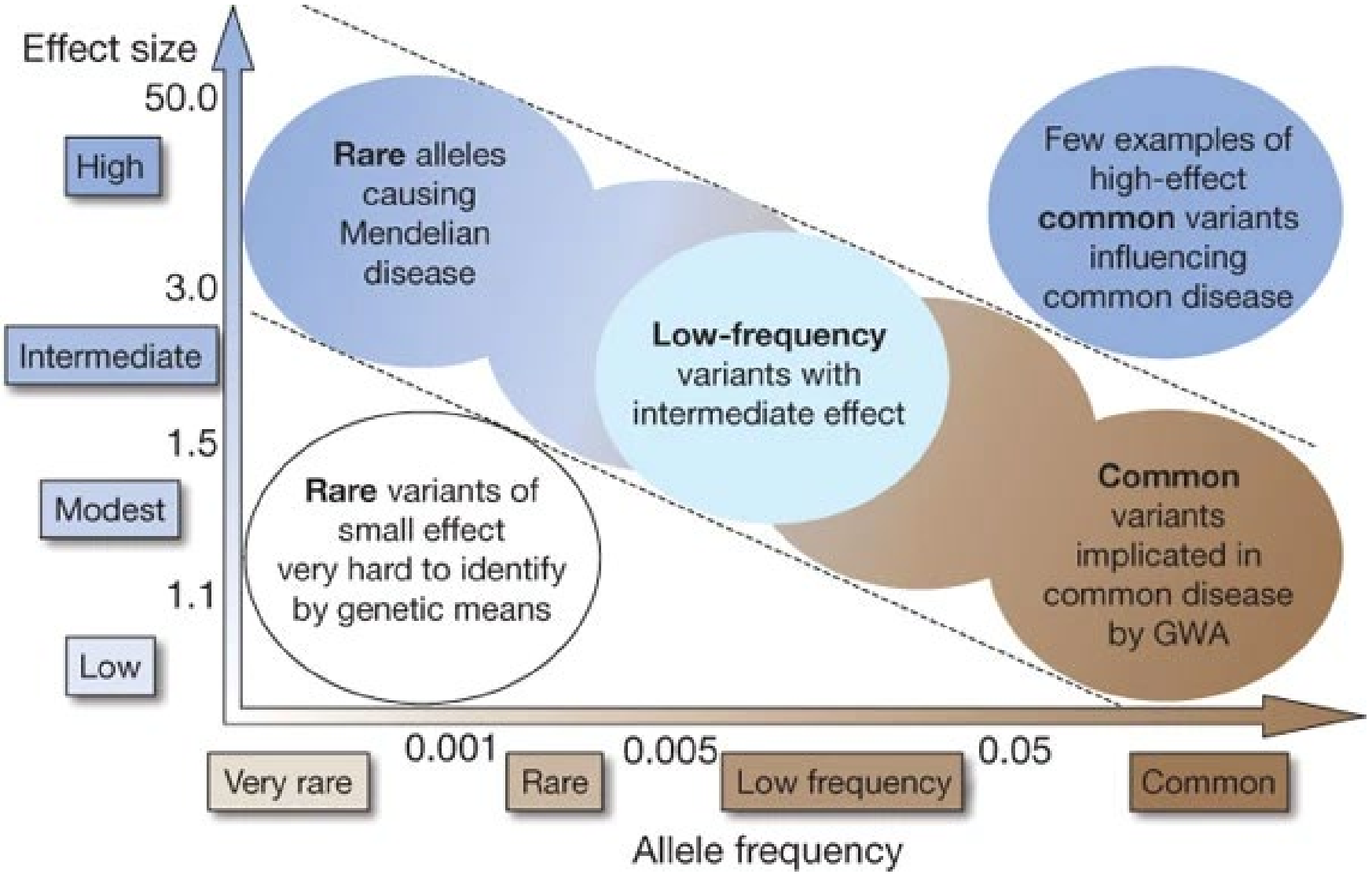
1. **How do we catalogue all genetic variants in the human genome (SNPs)**
2. **How do we systematically associate them with disease (GWAS)**
3. **How do we use GWAS to understand disease mechanism (Function)**
4. **How can we translate these insights into therapeutics (Manipulation)**

Fine Mapping: GWAS locus view



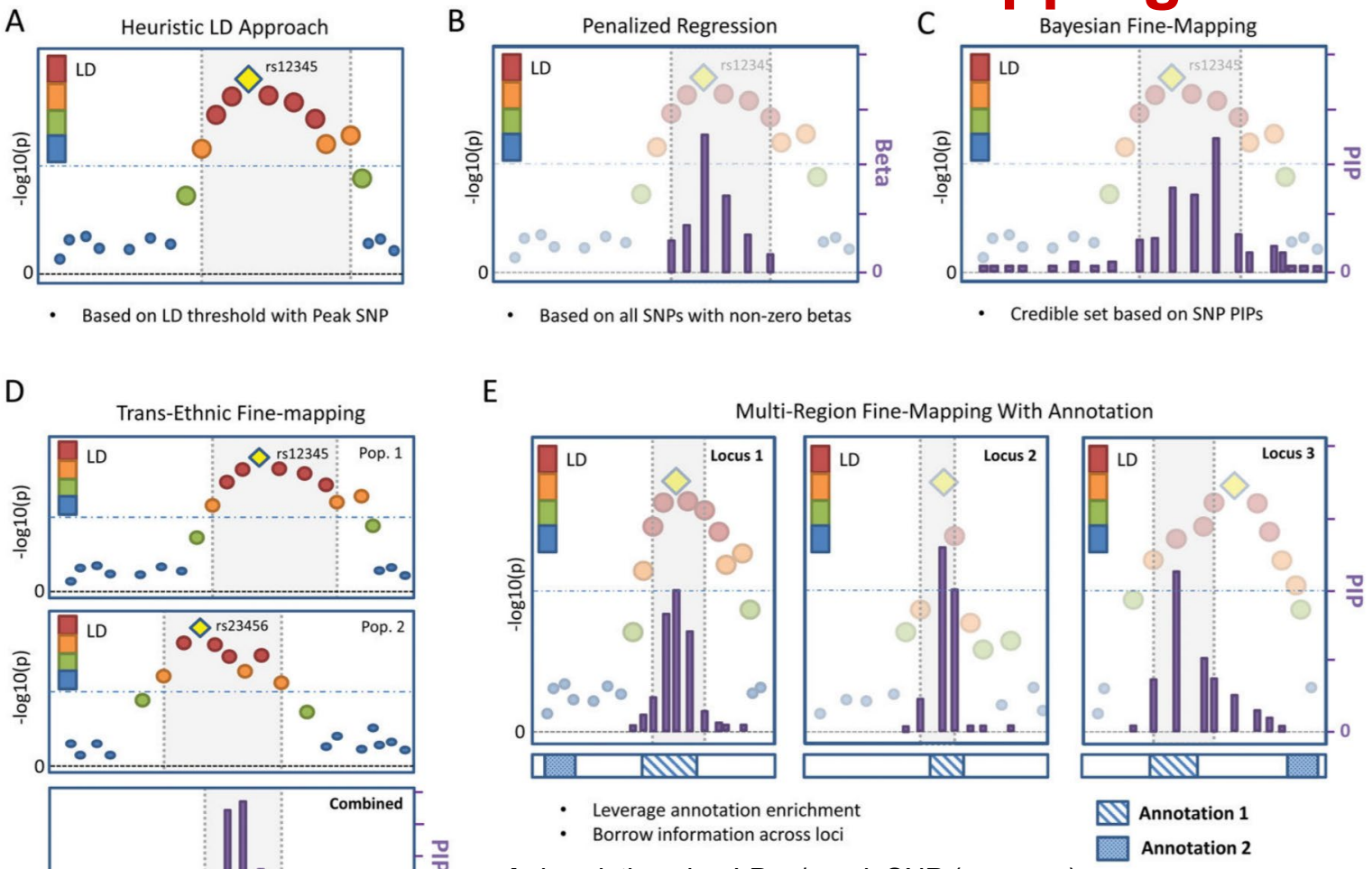


Most common variants have small effects



Manolio et al 2009, PMID: 19812666

GWAS fine-mapping

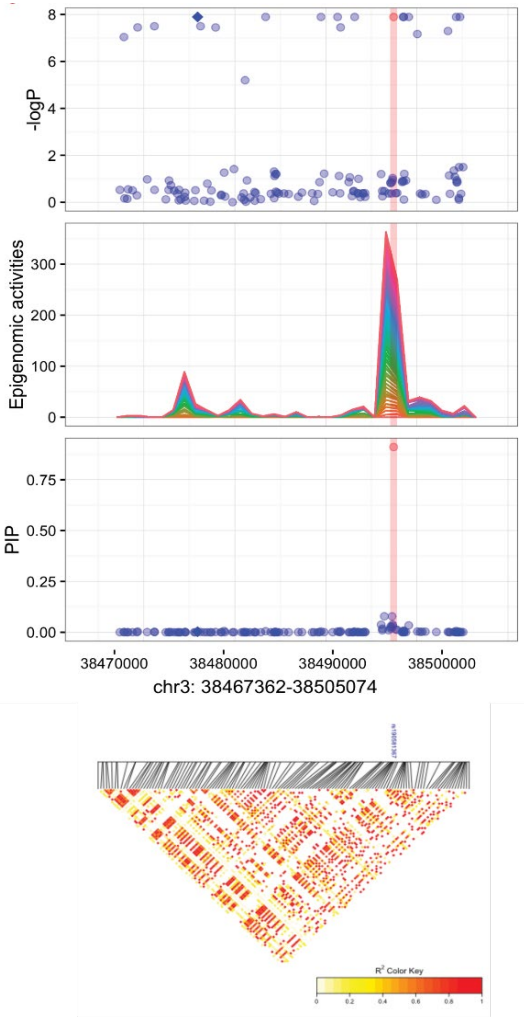


- LocusZoom of marginal SNP associations
- Y-axis: $-\log_{10}(p\text{-values})$
- X-axis: Variant positions
- Gold: peak SNP
- Other=degree LD w/peak SNP (red, orange, green, blue)
- Purple bars=additional variant-level statistics by fine-mapping
- (Penalized regression=Beta; Bayesian: posterior inclusion probabilities (PIPs))
- Light grey=regions selected by fine-mapping

- **A**=heuristic using LD w/ peak SNP (>orange)
- **B**=Penalized regression=Beta not shrunk to zero
- **C**=Bayesian PIPs summed to credible sets using $P_{coverage}$>95%
  (note: peak SNP not always highest PIP ← correlation structure of SNPs in region)
- **D**=2 pops w/ different local LD struct → meta-analysis narrow fine-mapping credible region
- **E**=Anno1 overlap in locus 1 & 2 → predict top-PIP SNP in locus 3 (overlaps anno1)

Fine-mapping disease associations: (1) Epigenomics / functional data (next lecture)

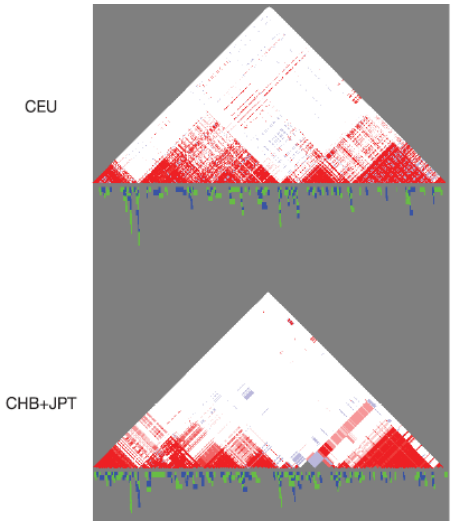
- **Association mapping** refers to identifying variants/gene associated with disease
- This is confounded by LD
- Many variants are strongly correlated to the true causal variant, and will show nearly as strong associations
- Use estimated correlations to explain correlated associations and recover the true underlying effects



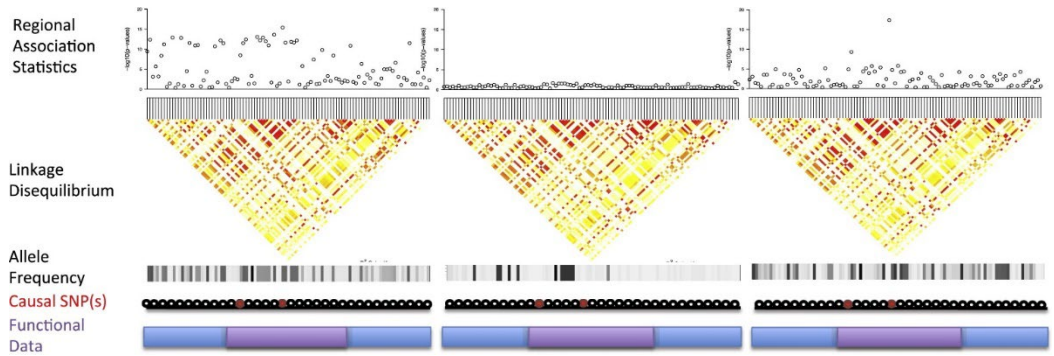
Li and Kellis *BiorXiv* 2016

Fine-mapping disease associations (2) Multi-ethnic analysis

Case 1: LD boundaries differ



Case 2: allele frequencies differ



- Allele frequencies and LD patterns can differ between populations
- Currently, disease associations are biased for discovery in European cohorts
- As we begin conducting association studies in Asia/Africa, there is a pressing need to develop statistical methods which can account for population genetic differences

Kichaev et al. *Am J Hum Genet* 2015

Genetics, Variation, GWAS, PRS, Mechanism

1. Genetics, Variation, GWAS
2. Polygenic scores (PGS)
3. From GWAS to biological insights
4. From Region to Mechanism / Circuitry
5. Quantitative Trait Loci: eQTL/meQTL analysis
6. Mediation Analysis + Mendelian Randomization
7. Heritability and Systems Genetics

Potential of PRS in clinical practice

AHA SCIENTIFIC STATEMENT

Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association

Jack W. O'Sullivan, MBBS, DPhil, Chair; Sridharan Raghavan, MD, PhD; Carla Marquez-Luna, PhD; Jasmine A. Luzum, PharmD, PhD; Scott M. Damrauer, MD, FAHA; Euan A. Ashley, MBChB, DPhil, FAHA; Christopher J. O'Donnell, MD, MPH; Cristen J. Willer, DPhil; Pradeep Natarajan, MD, MMSc, Vice Chair; on behalf of the American Heart Association Council on Genomic and Precision Medicine; Council on Clinical Cardiology; Council on Arteriosclerosis, Thrombosis and Vascular Biology; Council on Cardiovascular Radiology and Intervention; Council on Lifestyle and Cardiometabolic Health; and Council on Peripheral Vascular Disease

"*These observations point to the **possibility of using genetic profiling to inform clinical practice** in significantly larger groups of individuals than for whom monogenic cardiovascular variants are considered. As a result of exponential increases in the proportion of individuals with broad genetic profiling, **cardiovascular PRSs are beginning to enter clinical practice**. Such PRSs may be appropriately considered in select scenarios, given the current evidence base.*"

O'Sullivan, et al. *Circulation* (2022);

Polygenic scores combine effects of disease-associated alleles for each individual

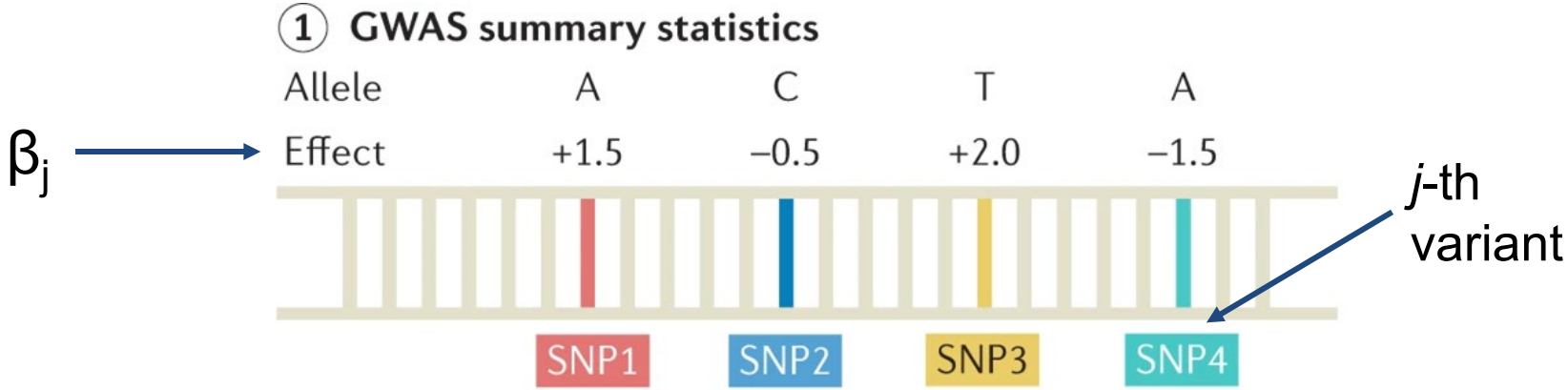
- Polygenic scores (PGS)
  - aka. Genetic risk score (GRS), Polygenic risk score (PRS), etc.
    - “risk” → disease risks
    - “Polygenic” → statement of the genetic architecture of a trait

- Polygenic score := weighted sum of disease-associated alleles

$$\mathrm{PRS}_i = \sum_{j \in J} \beta_j \, G_{ij}$$

i-th individual *G*: genotype

j-th variant β: effect size

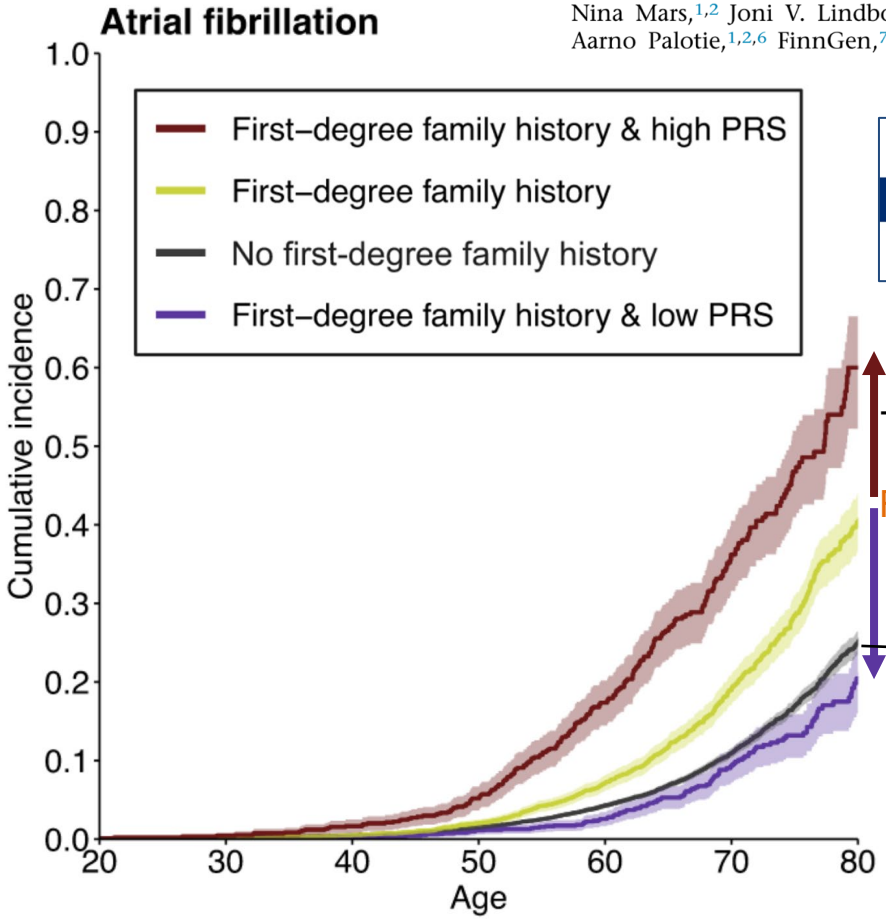


Family history (FH) complements PGS

ARTICLE

Systematic comparison of family history and polygenic risk across 24 common diseases

Nina Mars,[1,2] Joni V. Lindbohm,[3,4,5] Pietro della Briotta Parolo,[1] Elisabeth Widén,[1] Jaakko Kaprio,[1,3] Aarno Palotie,[1,2,6] FinnGen,[7] and Samuli Ripatti[1,3,5,*]



FINNGEN

+ Positive PRS (increased risk)

Family history (high risk) (baseline, no PRS info)

- Negative PRS (low risk)

No family history (low risk)

Family history (high risk) but low PRS!!

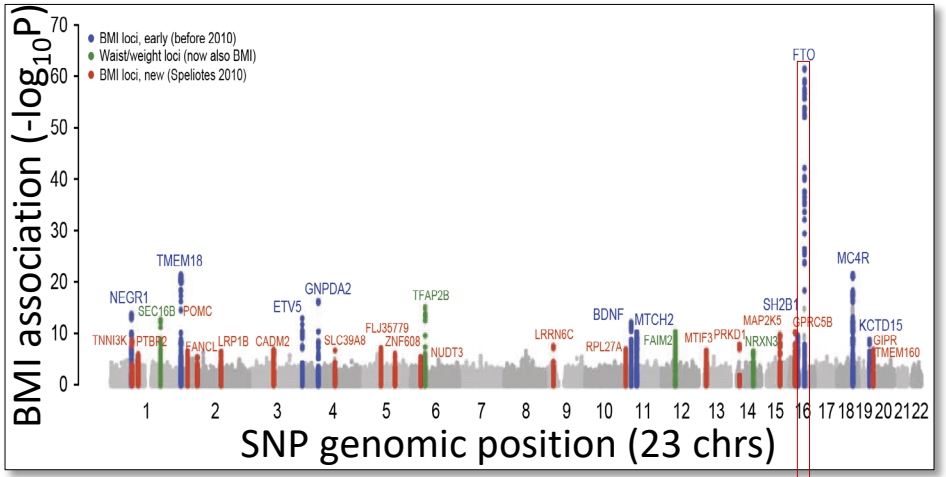
Mars, et al. AJHG 2022

Genetics, Variation, GWAS, PRS, Mechanism

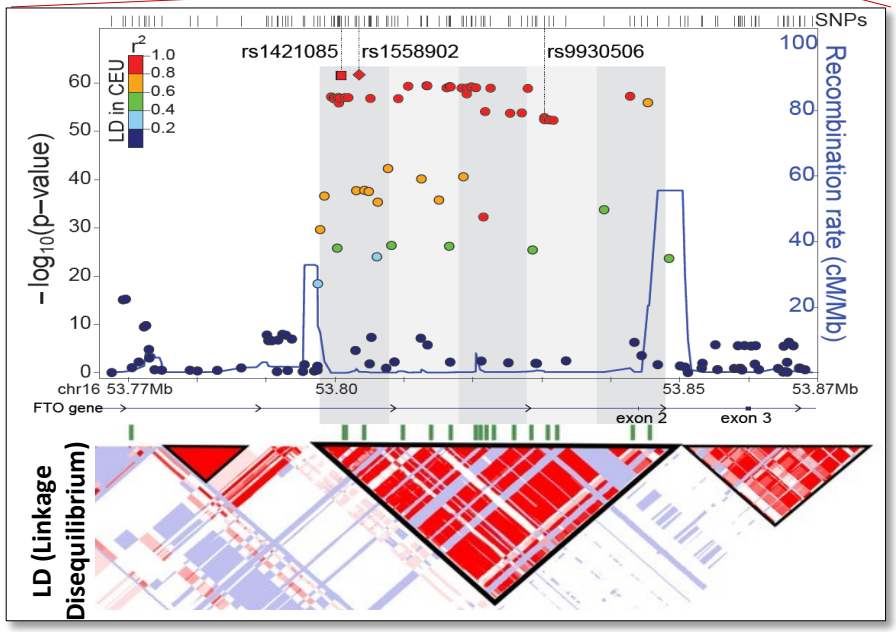
1. Genetics, Variation, GWAS
2. Polygenic scores (PGS)
3. From GWAS to biological insights
4. From Region to Mechanism / Circuitry
5. Quantitative Trait Loci: eQTL/meQTL analysis
6. Mediation Analysis + Mendelian Randomization
7. Heritability and Systems Genetics

Genomic medicine today: challenge and promises

GWAS Manhattan Plot: simple χ^2 statistical test



Speliotes NG 2010



Dina NG 2007, Frayling Science 2007, Claussnitzer NEJM 2015

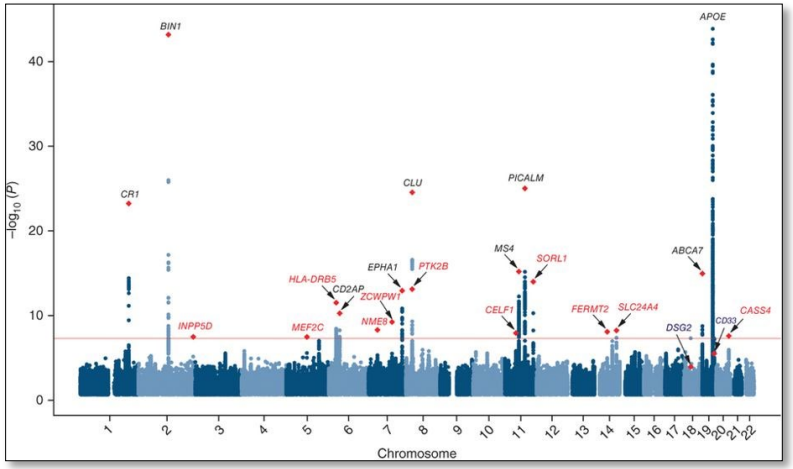
The promise of genetics

- Unbiased, Causal, Uncorrected
- New disease mechanisms
- New target genes
- New therapeutics
- Personalized medicine

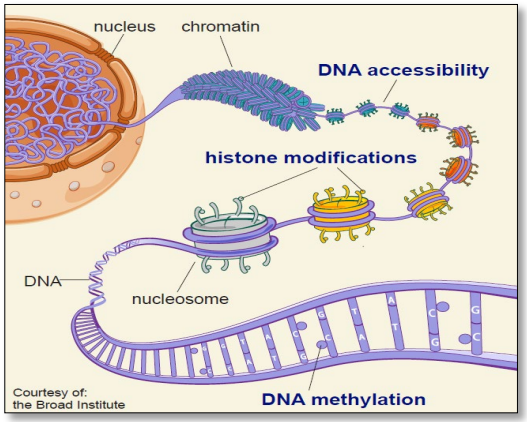
The challenge of mechanism

- **90+% disease hits non-coding**
- Target gene not known
- Causal variant not known
- Cell type of action not known
- Relevant pathways not known
- Mechanism not known

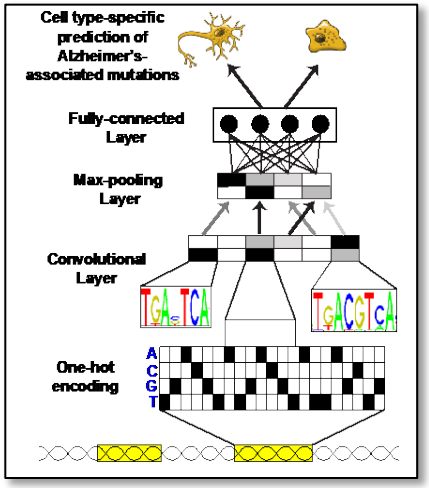
Dissect mechanisms of disease-associated regions



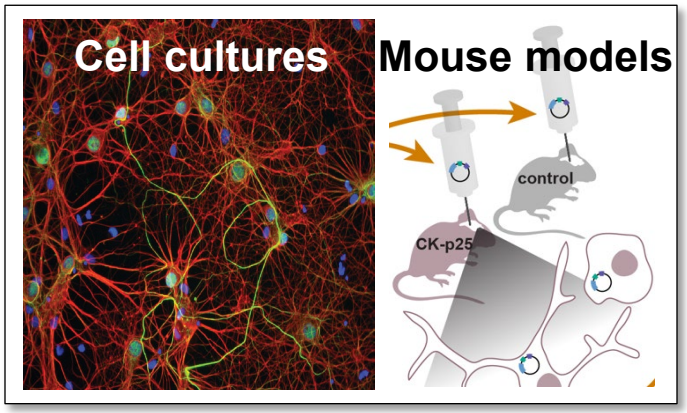
1. Disease genetics reveals common + rare variants/regions



2. Profile RNA + Epigenome in healthy + disease samples



3. Integrate data to predict driver genes, regions, cell types

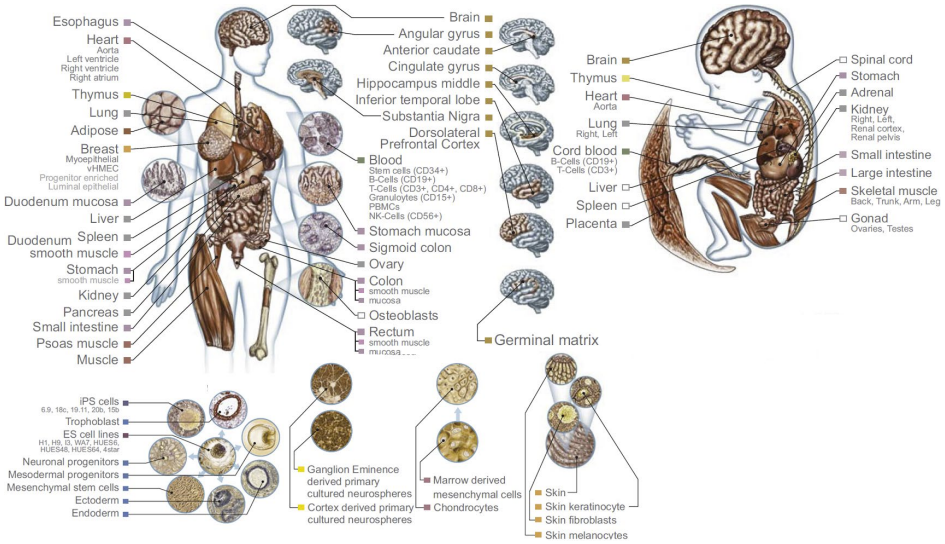


4. Validate predictions in human cells + mouse models

5. Disseminate results

Epigenomic mapping across 100+ tissues/cell types

Diverse tissues and cells



Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood…)

Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood…)

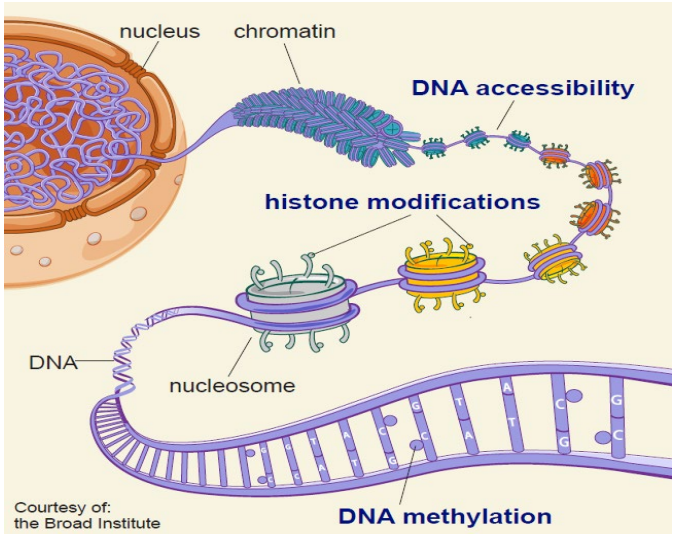
ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench…)

ROADMAP epigenomics PROJECT

ENCODE

X

Diverse epigenomic assays



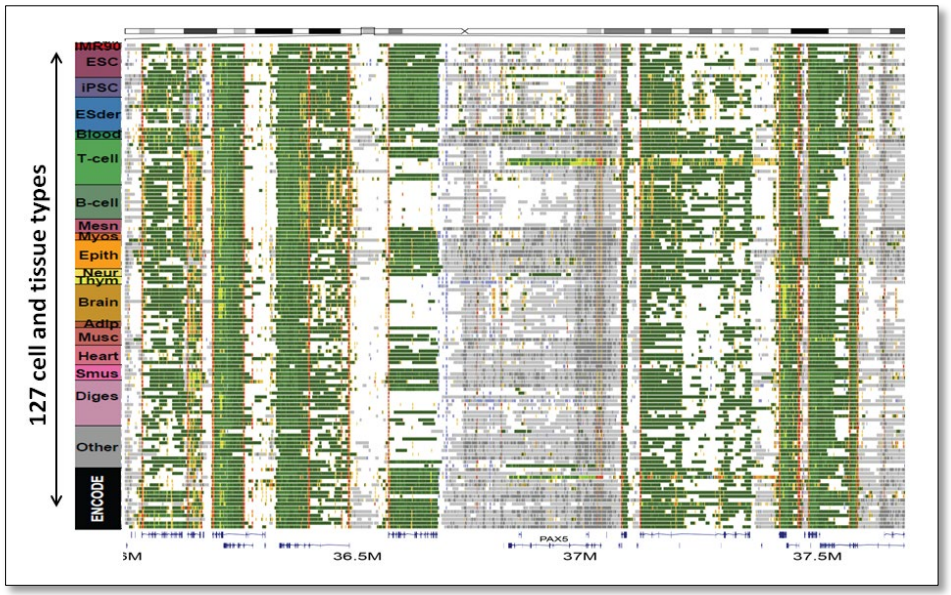
Histone modifications
- **H3K4me3, H3K4me1, H3K36me3**
- **H3K27me3, H3K9me3, H3K27/9ac**
- **+20 more**

Open chromatin:
- **DNA accessibility**

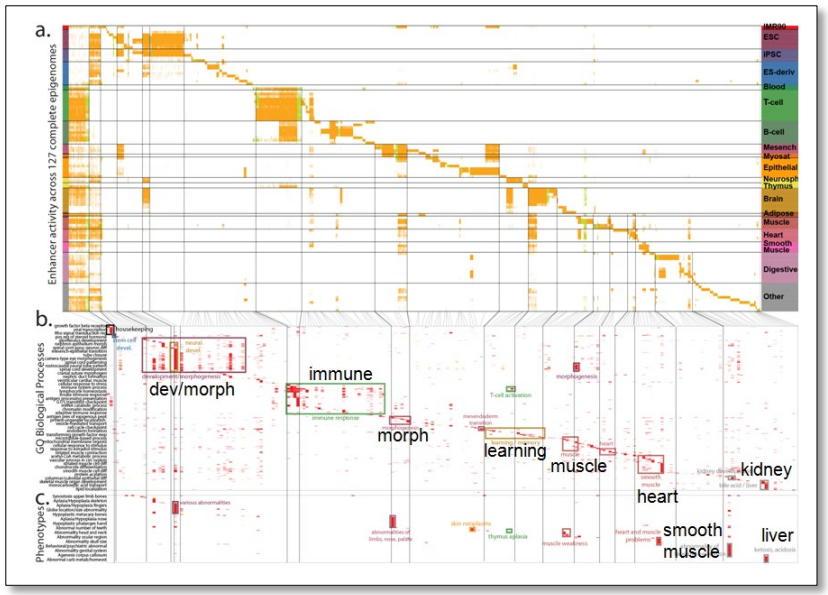
DNA methylation:
- **WGBS, RRBS, MRE/MeDIP**

Gene expression
- **RNA-seq, Exon Arrays**

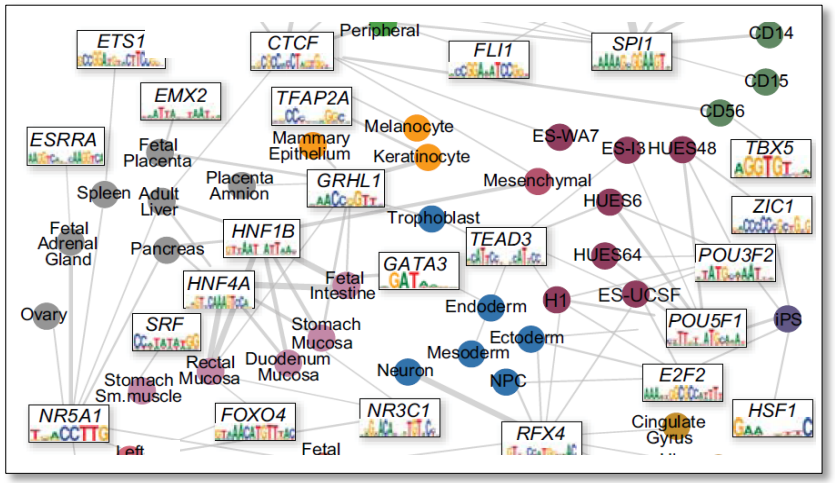
Enhancer modules, regulators, and target genes



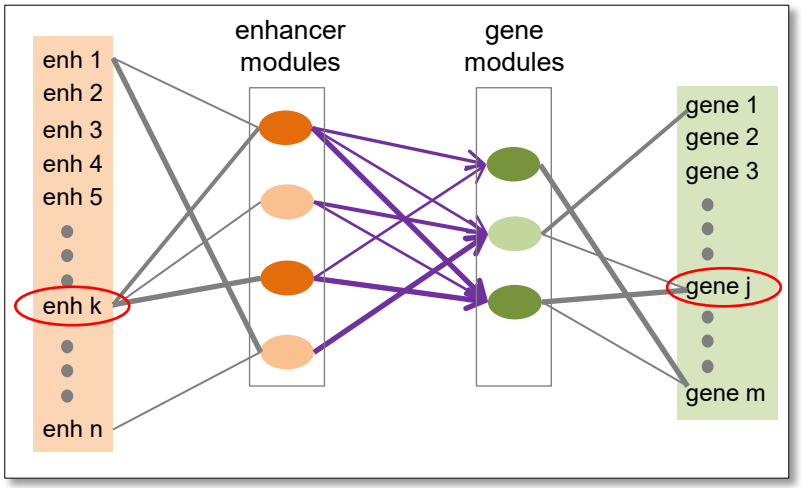
1. *Map chromatin states across 127 tissue/cells*



2. *Group enhancers into modules of common function*



3. *Predict module regulators using motif enrichment*



4. *Predict target genes using common activity*

HaploReg: systematic mining of GWAS variants

Query SNP: rs4684847 and variants with r^2 >= 0.8

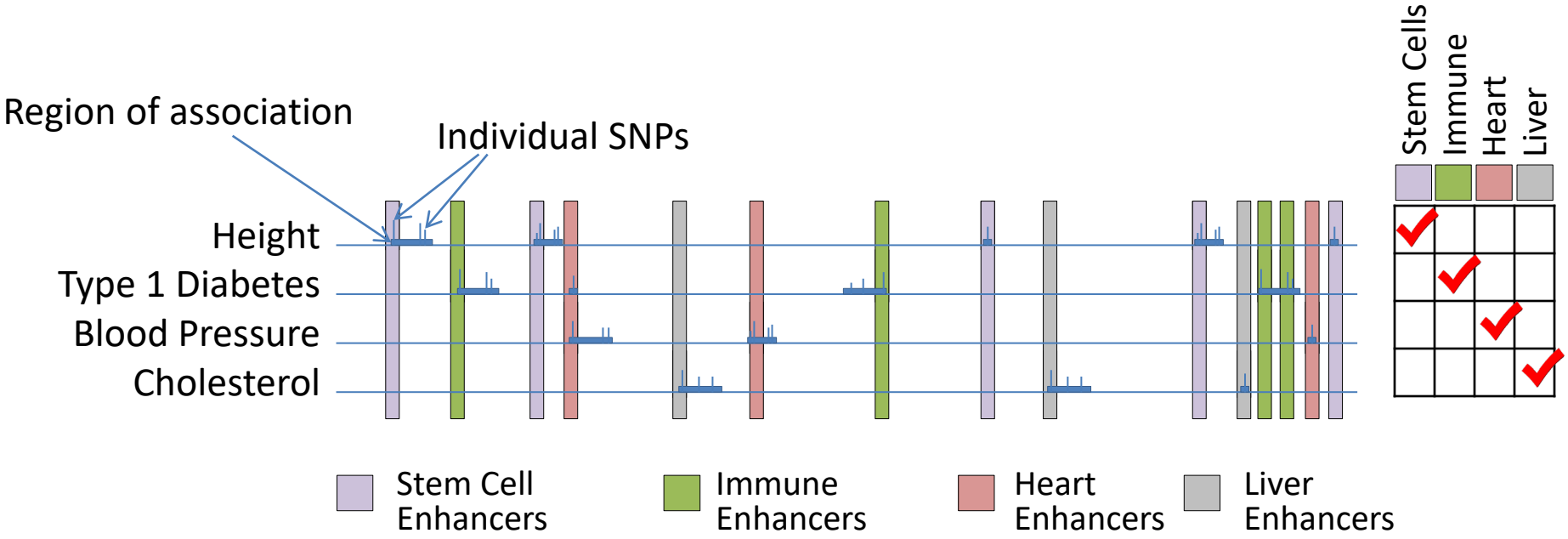
pos (hg19)	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL tissues	Motifs changed	Drivers disrupted	GENCODE genes	dbSNP func annot
chr3:12329783	chr3:12288284	0.95	0.97	rs17036160	C	T	0.01	0.06	0.04	0.12		24 organs	7 organs	4 organs			4 altered motifs		PPARG	intronic
chr3:12336507	chr3:12295008	0.95	0.97	rs11709077	G	A	0.01	0.07	0.04	0.12		LNG	9 organs	15 organs			4 altered motifs		PPARG	intronic
chr3:12344730	chr3:12303231	0.94	0.97	rs11712037	C	G	0.01	0.08	0.04	0.12			6 organs	BLD			AP-1,TCF11::MafG		PPARG	intronic
chr3:12351521	chr3:12310022	0.95	0.97	rs35000407	T	G	0.01	0.07	0.04	0.12		LNG	5 organs				Smad		PPARG	intronic
chr3:12360884	chr3:12319385	0.95	0.97	rs150732434	TG	T	0.01	0.07	0.04	0.12		FAT	7 organs	MUS,VAS	CFOS		Hdx,Sox,TATA		PPARG	intronic
chr3:12365308	chr3:12323809	0.95	0.97	rs13083375	G	T	0.01	0.07	0.04	0.12		BLD	BLD, FAT				Homez,Sox,YY1		PPARG	intronic
chr3:12369401	chr3:12327902	0.95	0.97	rs13064760	C	T	0.01	0.07	0.04	0.12			7 organs				9 altered motifs		PPARG	intronic
chr3:12375956	chr3:12334457	0.95	0.97	rs2012444	C	T	0.01	0.07	0.04	0.12			SKIN, FAT, BLD				7 altered motifs		PPARG	intronic
chr3:12383265	chr3:12341766	0.96	0.99	rs13085211	G	A	0.18	0.10	0.04	0.12			FAT, SKIN				NRSF		PPARG	intronic
chr3:12383714	chr3:12342215	0.96	0.99	rs7638903	G	A	0.18	0.10	0.04	0.12			6 organs	CRVX					PPARG	intronic
chr3:12385828	chr3:12344329	0.95	1	rs11128603	A	G	0.18	0.10	0.04	0.12			CRVX				RXRA		PPARG	intronic
chr3:12386337	chr3:12344838	1	1	rs4684847	C	T	0.01	0.07	0.04	0.12			6 organs						PPARG	intronic
chr3:12388409	chr3:12346910	0.99	1	rs7610055	G	A	0.17	0.09	0.04	0.12			BLD				4 altered motifs		PPARG	intronic
chr3:12389313	chr3:12347814	0.99	1	rs17036326	A	G	0.17	0.09	0.04	0.12			FAT, BL						PPARG	intronic
chr3:12390484	chr3:12348985	0.99	1	rs17036328	T	C	0.17	0.09	0.04	0.12			FAT, CR						PPARG	intronic
chr3:12391207	chr3:12349708	0.99	1	rs6802898	C	T	0.61	0.15	0.04	0.12			FAT, BL						PPARG	intronic
chr3:12391583	chr3:12350084	0.99	1	rs2197423	G	A	0.17	0.09	0.04	0.12		FAT, LIV	8 organ						PPARG	intronic
chr3:12391813	chr3:12350314	0.99	1	rs7647481	G	A	0.17	0.09	0.04	0.12		4 organs	9 organ						PPARG	intronic
chr3:12392272	chr3:12350773	0.99	1	rs7649970	C	T	0.17	0.09	0.04	0.12		5 organs	9 organ						PPARG	intronic
chr3:12393125	chr3:12351626	1	1	rs1801282	C	G	0.01	0.07	0.04	0.12		FAT, LIV	9 organ						PPARG	missense
chr3:12393682	chr3:12352183	0.99	1	rs17036342	A	G	0.17	0.09	0.04	0.12		FAT	9 organs						PPARG	intronic
chr3:12394840	chr3:12353341	0.99	1	rs1899951	C	T	0.61	0.15	0.04	0.12		FAT	9 organs				Mef2		PPARG	intronic
chr3:12395645	chr3:12354146	0.99	1	rs4684848	G	A	0.61	0.15	0.04	0.12		FAT, BLD	9 organs	ADRL,GI,CRVX	5 bound proteins				PPARG	intronic
chr3:12396845	chr3:12355346	0.93	1	rs4135250	A	G	0.17	0.09	0.04	0.13			4 organs	PLCNT					PPARG	intronic
chr3:12396913	chr3:12355414	0.98	1	rs71304101	G	A	0.01	0.07	0.04	0.12			4 organs	PLCNT			Crx,NF-E2		PPARG	intronic
chr3:12396955	chr3:12355456	0.96	1	rs2881654	G	A	0.61	0.15	0.04	0.12			4 organs				7 altered motifs		PPARG	intronic

Adipose_Derived_Mesenchymal_Stem_Cell_Cultured_Cells, CD4+_CD25-_IL17+_PMA-Ionomcyin_stimulated_Th17_Primary_Cells, Muscle_Satellite_Cultured_Cells, Penis_Foreskin_Fibroblast_Primary_Cells_skin01, Penis_Foreskin_Fibroblast_Primary_Cells_skin02, Penis_Foreskin_Keratinocyte_Primary_Cells_skin02, Penis_Foreskin_Keratinocyte_Primary_Cells_skin03, A549_EtOH_0.02pct_Lung_Carcinoma, HeLa-S3_Cervical_Carcinoma, NHEK-Epidermal_Keratinocytes

- **Start with any list of SNPs or select a GWA study**
  - Mine ENCODE and Roadmap epigenomics data for hits
  - Hundreds of assays, dozens of cells, conservation, motifs
  - Report significant overlaps and link to info/browser
- **Try it out: http://compbio.mit.edu/HaploReg**

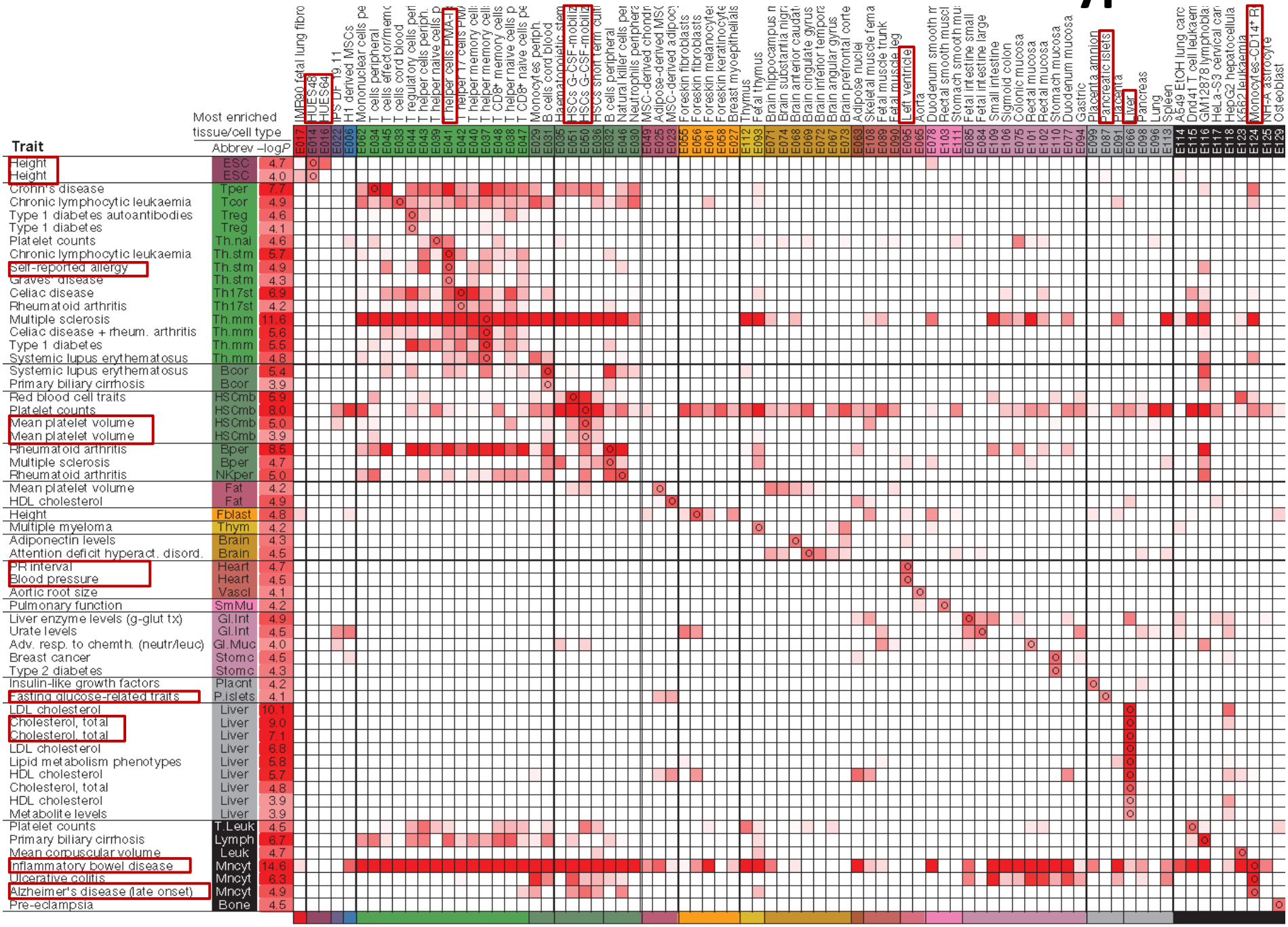
Ward, Kellis NAR 2011

Identifying disease-relevant cell types

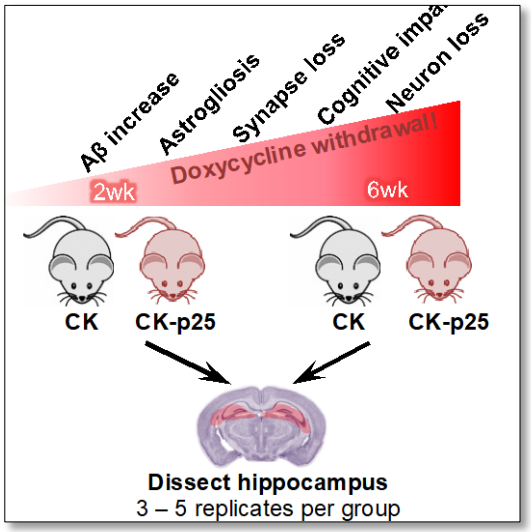


- For every trait in the GWAS catalog:
  - Identify all associated regions at P-value threshold
  - Consider all SNPs in credible interval ($R^2 \geq .8$)
  - Evaluate overlap with tissue-specific enhancers
  - Keep tissues showing significant enrichment ($P<0.001$)
- Repeat for all traits (rows) and all cell types (columns)

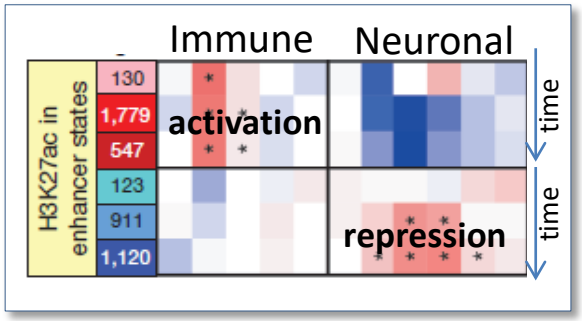
GWAS hits in enhancers of relevant cell types



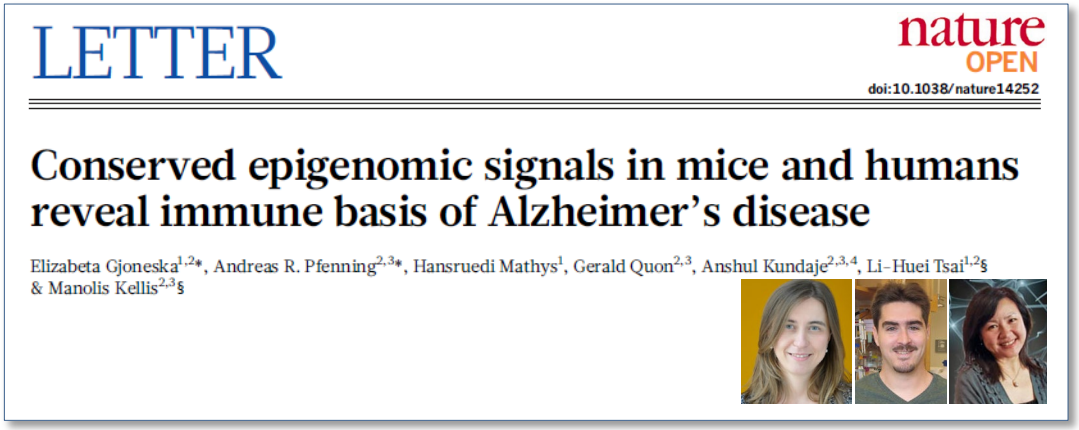
Immune activation + neural repression in human + mouse

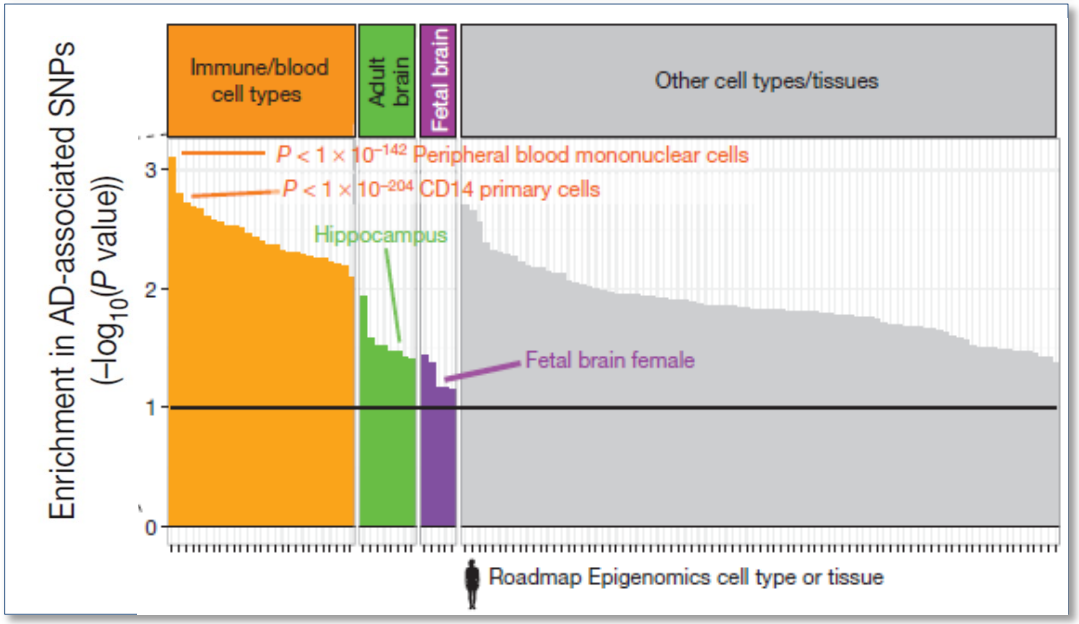


Epigenomics of AD progression



Immune activation precedes neuronal repression

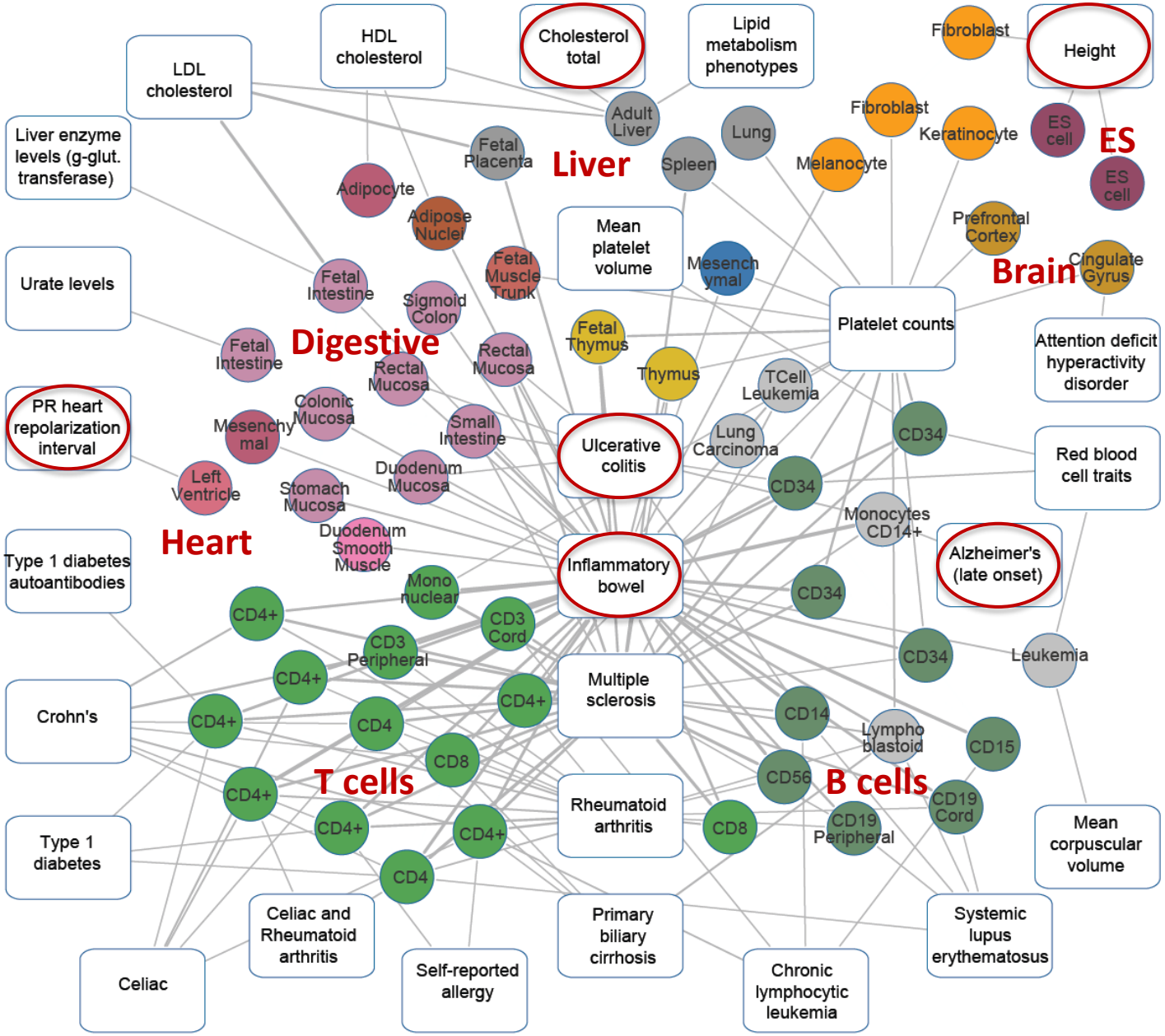




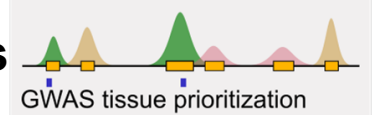
AD variants localize in immune cells, not neuronal

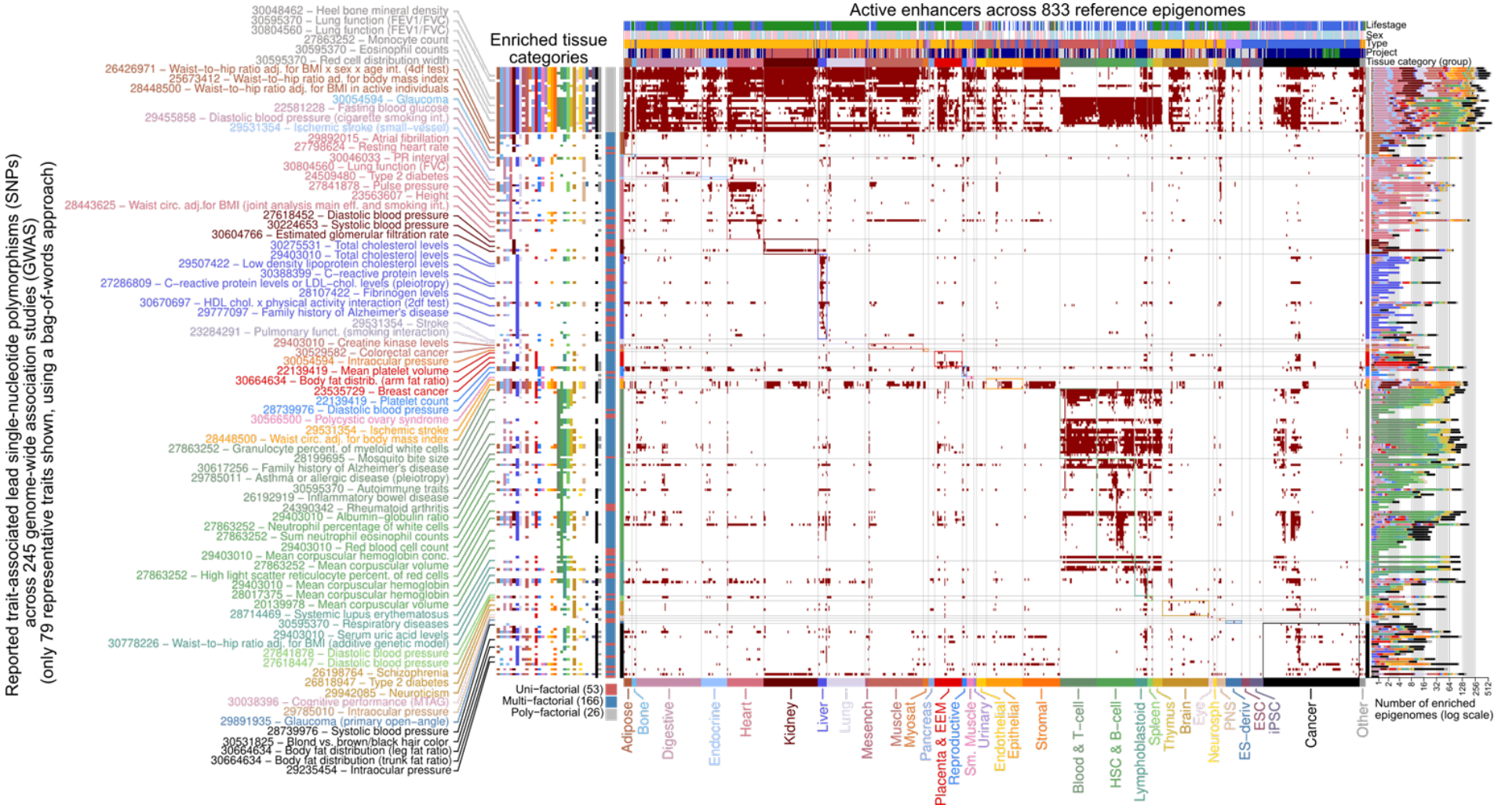
Inflammation as the causal component of Alzheimer's disease

Linking traits to their relevant cell/tissue types



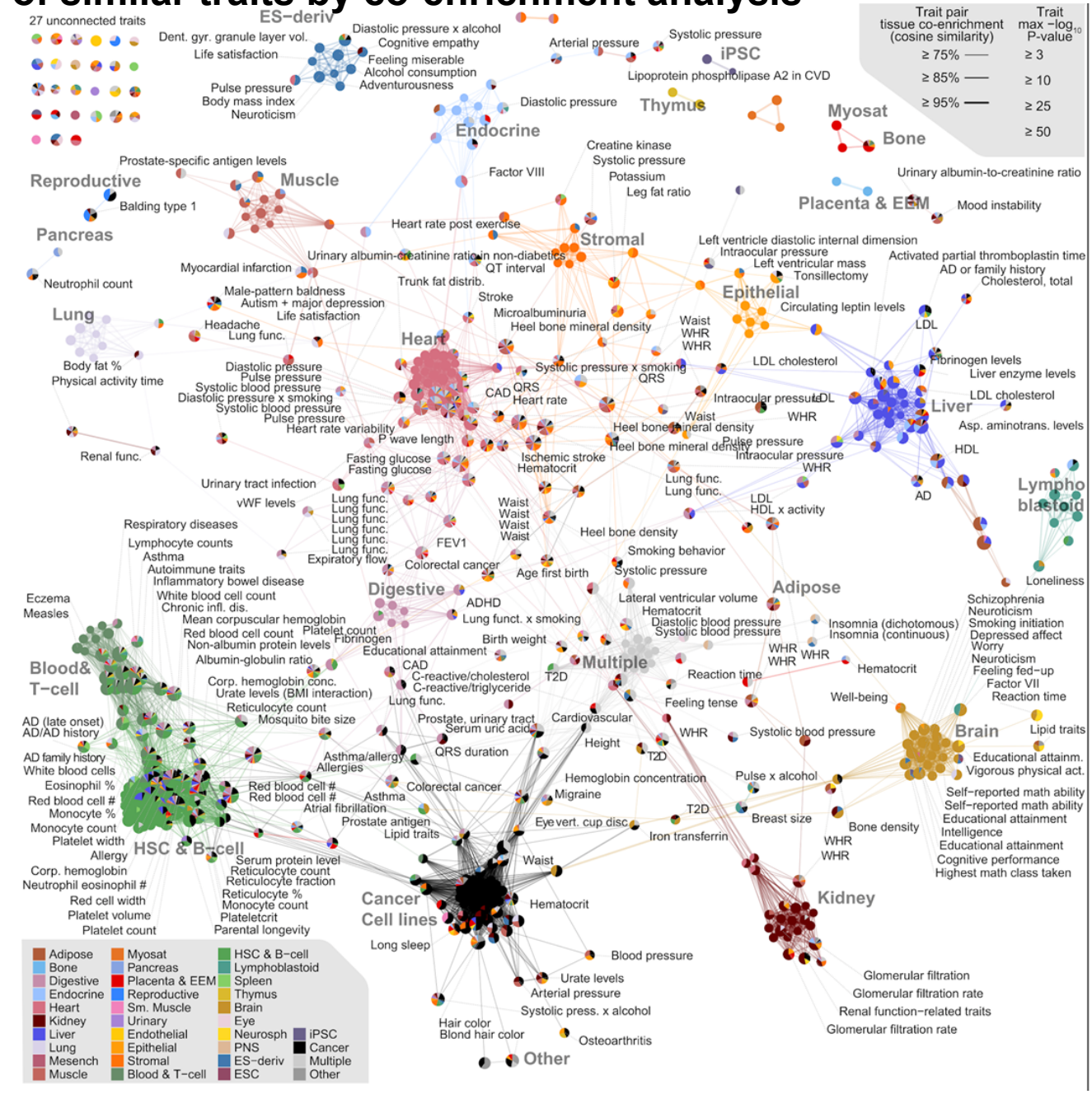
Predict tissues for 200+ traits by epigenome enrichments





Tissue x GWAS enrichments ➔ 245 GWAS-tissue enrichments, most in novel epigenomes (shown)
Enhancer-tree enrichments ➔ 540 GWAS-tissue enrichments, focusing at right level of resolution on tree

GWAS-Tissue enrichment: predict disease cell type/tissue in 500+ traits + network of similar traits by co-enrichment analysis



Range of trait complexity in epigenomic similarity network:

1. Uni-factorial traits (cores):

QT/PR intervals/QRS (heart)
C-reactive protein (liver)
TSH levels (endocrine)
Educational attain. (brain)
Schizophrenia (brain)
Life satisfaction (ES-deriv neur)
Glomer. filtration rate (kidney)
Autoimmune traits (T-cells)
Monocyte count (HSC & B-cell)

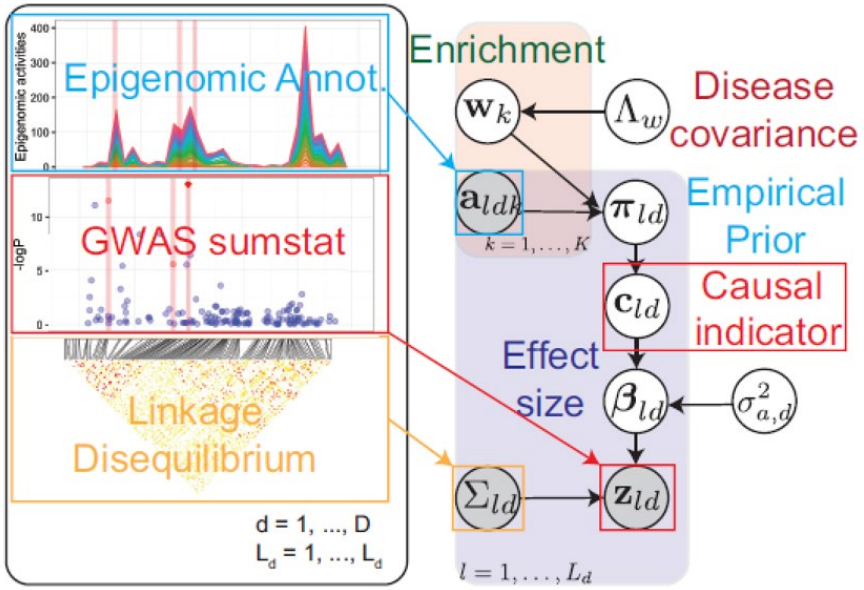
2. Multi-factorial (connect):

CAD (heart, endocrine, liver)
HDL/triglycerides (liver/adipose)
Lung FEV1, FVC (lung, heart, digestive)
Blood pressure (heart with endocrine, endothelial, and liver)
Alzheimer's (immune and brain)
Blood cell fractions (principal blood with liver, digestive, other)

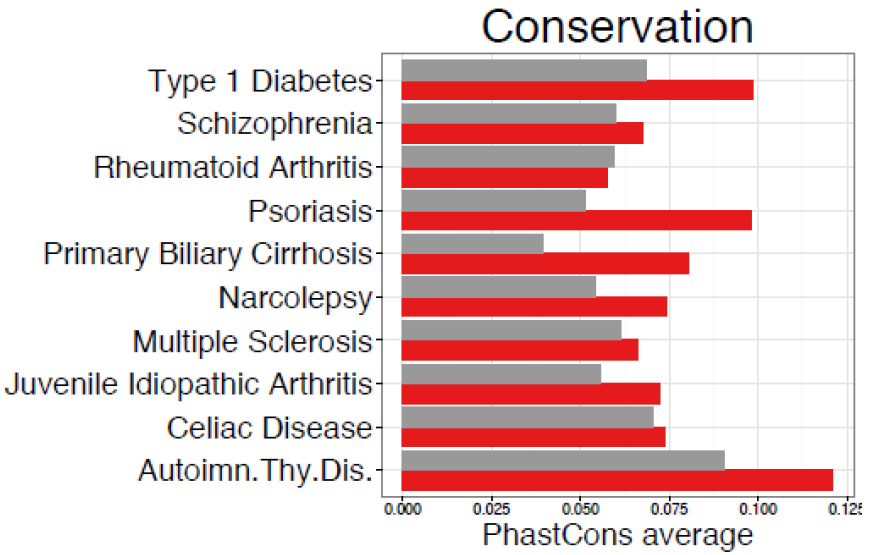
3. Poly-factorial:

Waist-hip ratio measures
Heel bone mineral density

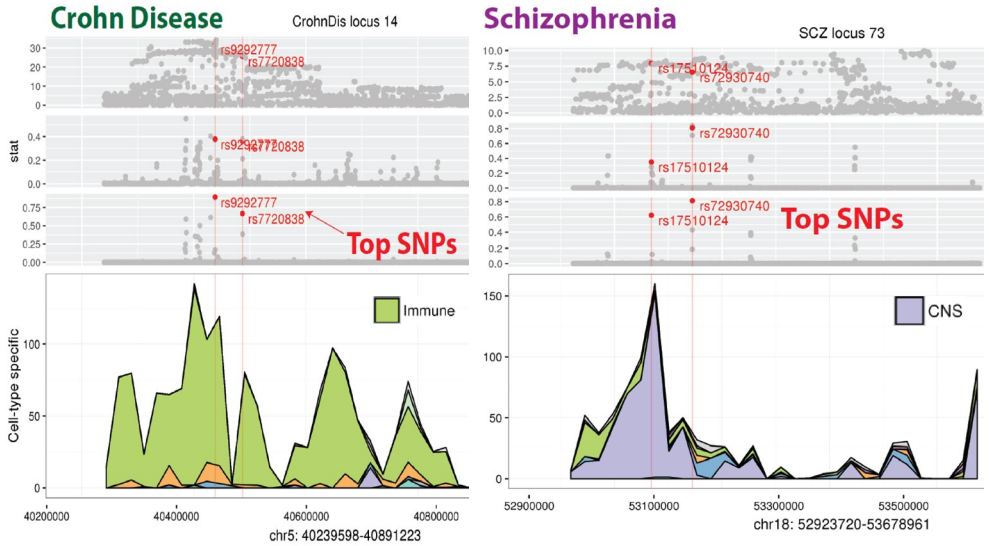
Bayesian fine-mapping: Predict causal variant and cell type



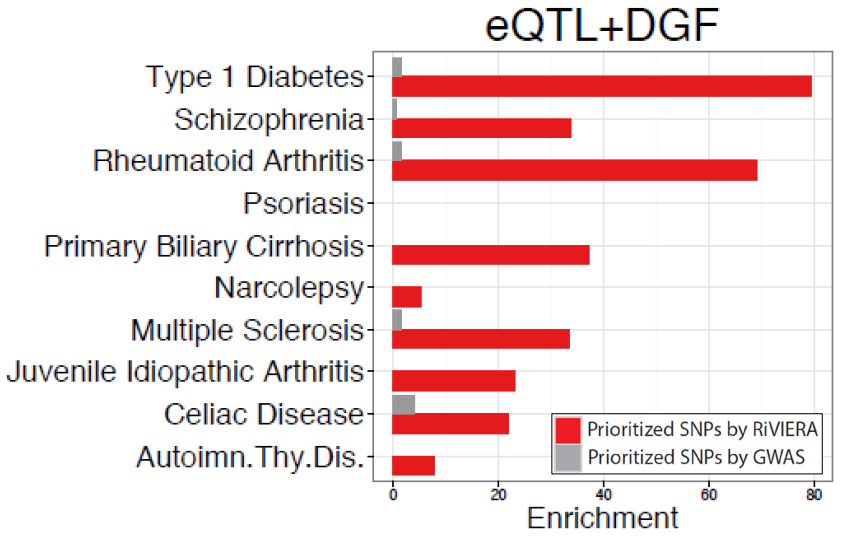
RiVIERA: multi-trait GWAS integration



Capture conserved elements

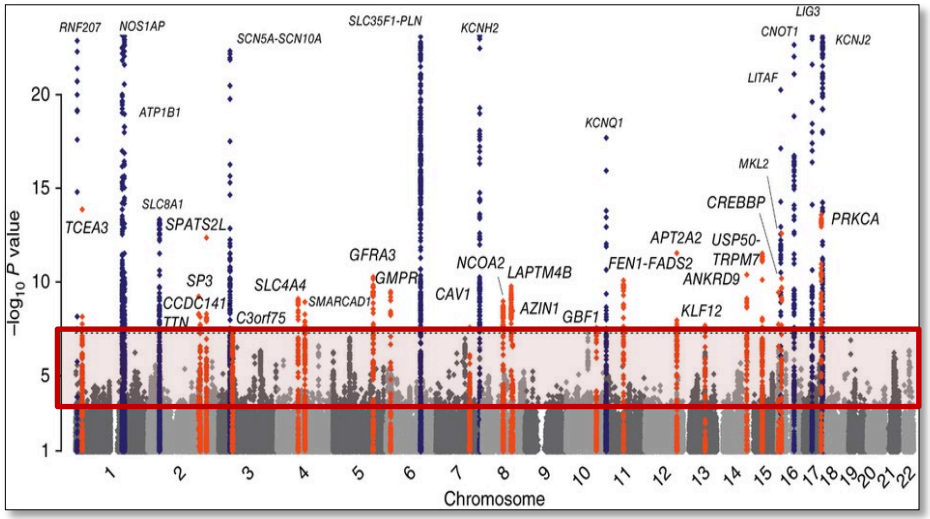


Predict causal variants and cell types



Capture eQTLs from GTEx

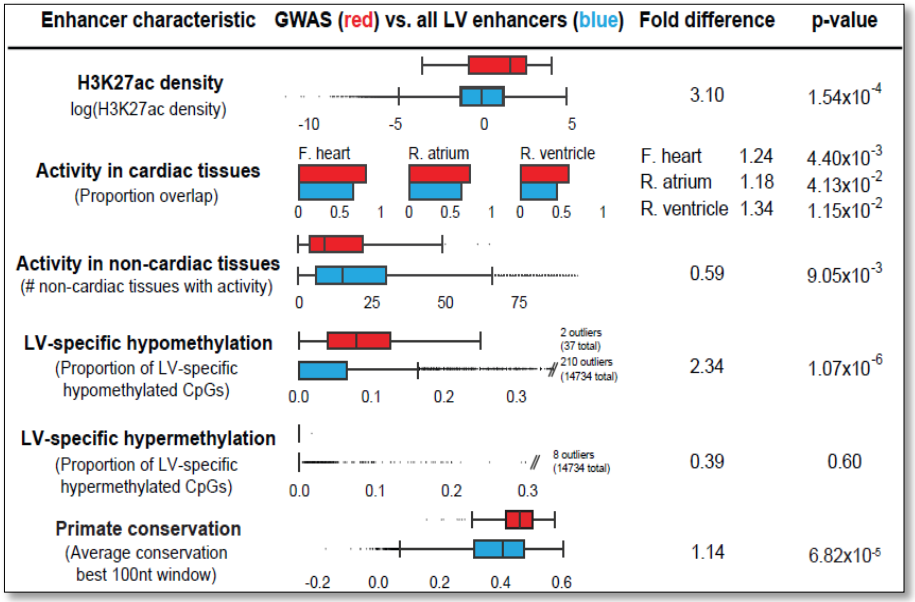
Combine GWAS+Epig to find new target genes/SNPs



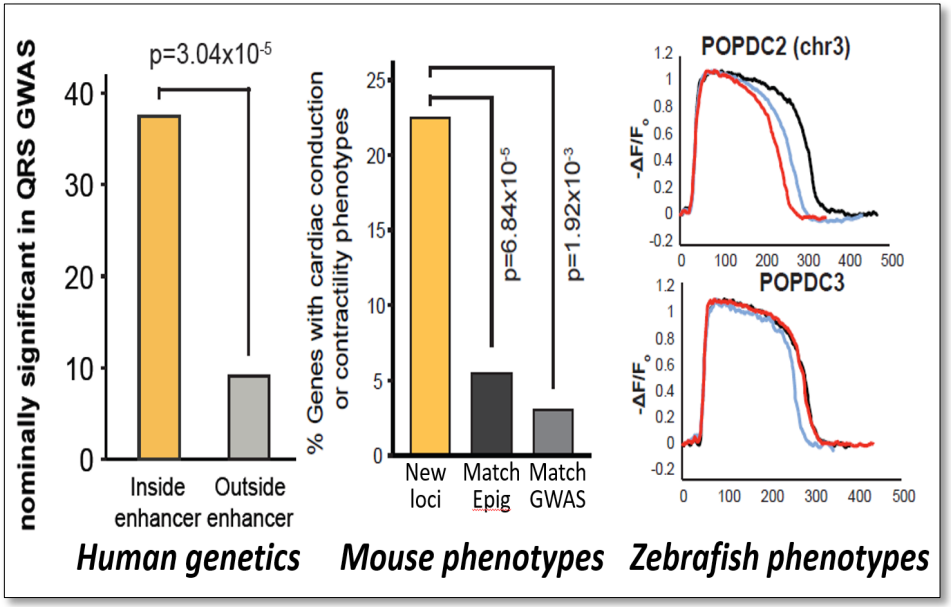
Lead SNP	p-value	Enhancer	1. Luciferase reporter	2. 4C-seq interactions
rs1886512	4.30×10^{-8}	chr13:74,520,000-74,520,400	0.015	No interactions
rs1044503	5.13×10^{-7}	chr14:102,965,400-102,972,000	4.70×10^{-9}	CINP, RCOR1
rs10030238	6.21×10^{-7}	chr4:141,807,800-141,809,600	1.35×10^{-14}	RNF150
		chr4:141,900,800-141,908,000	-	RNF150
rs6565060	1.52×10^{-5}	chr16:82,746,400-82,750,800	5.00×10^{-3}	No interactions
rs3772570	1.73×10^{-5}	chr3:148,733,200-148,738,600	0.67	-
rs3734637	2.23×10^{-5}	chr6:126,081,200-126,081,800	1.06×10^{-4}	HDDC2
rs1743292	6.48×10^{-5}	chr6:105,706,600-105,710,200	3.20×10^{-4}	BVES, POPDC3
		chr6:105,720,200-105,723,000	-	BVES, POPDC3
rs11263841	6.87×10^{-5}	chr1:35,307,600-35,312,200	0.22	GJA4, DLGAP3
rs11119843	7.14×10^{-5}	chr1:212,247,600-212,248,600	0.031	-
rs6750499	7.37×10^{-5}	chr2:11,559,600-11,563,000 (split into two 2kb fragments)	0.54	ROCK2
			3.26×10^{-7}	
rs17779853	7.73×10^{-5}	chr17:30,063,800-30,066,800	4.33×10^{-3}	No interactions

Prioritize sub-threshold loci (<10^{-4})

Validate new enhancers: allelic activity, enh-prom looping

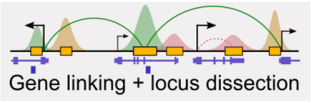


Machine learning predictive features

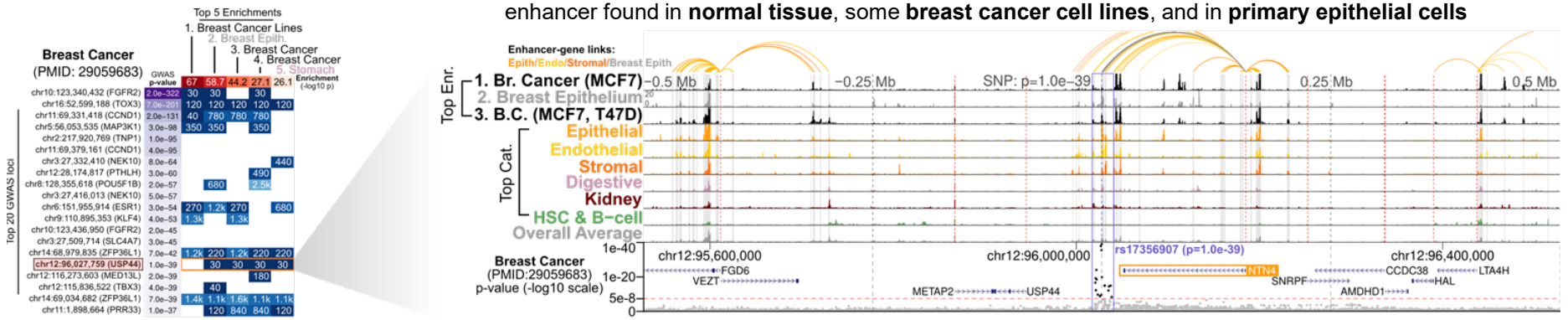


Validate new genes in hum/mou/zb

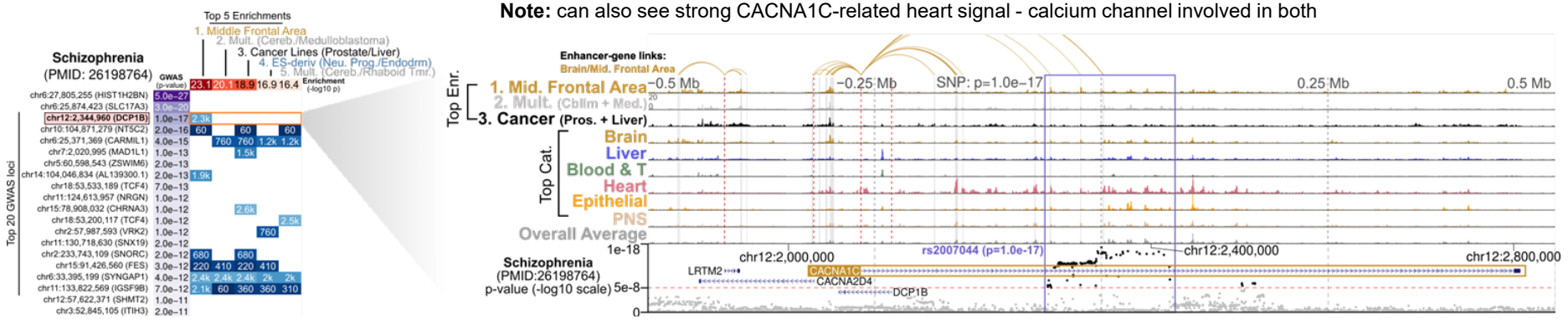
GWAS locus dissection: enriched tissues, driver SNPs, target genes



Example1: Localized breast cancer signal in USP44 locus links strongly to NTN4 gene (assoc. w/ prognosis, metastasis)
enhancer found in **normal tissue**, some **breast cancer cell lines**, and in **primary epithelial cells**

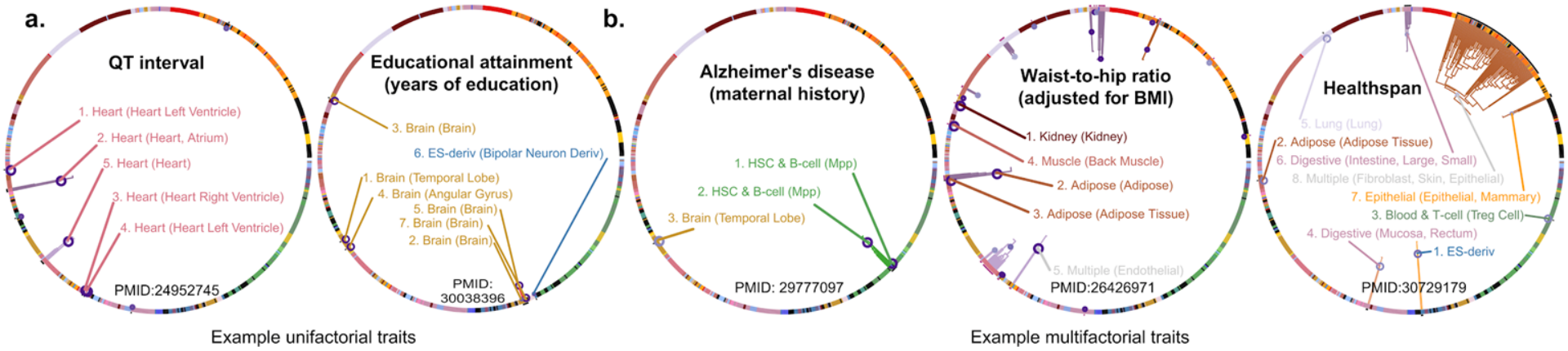


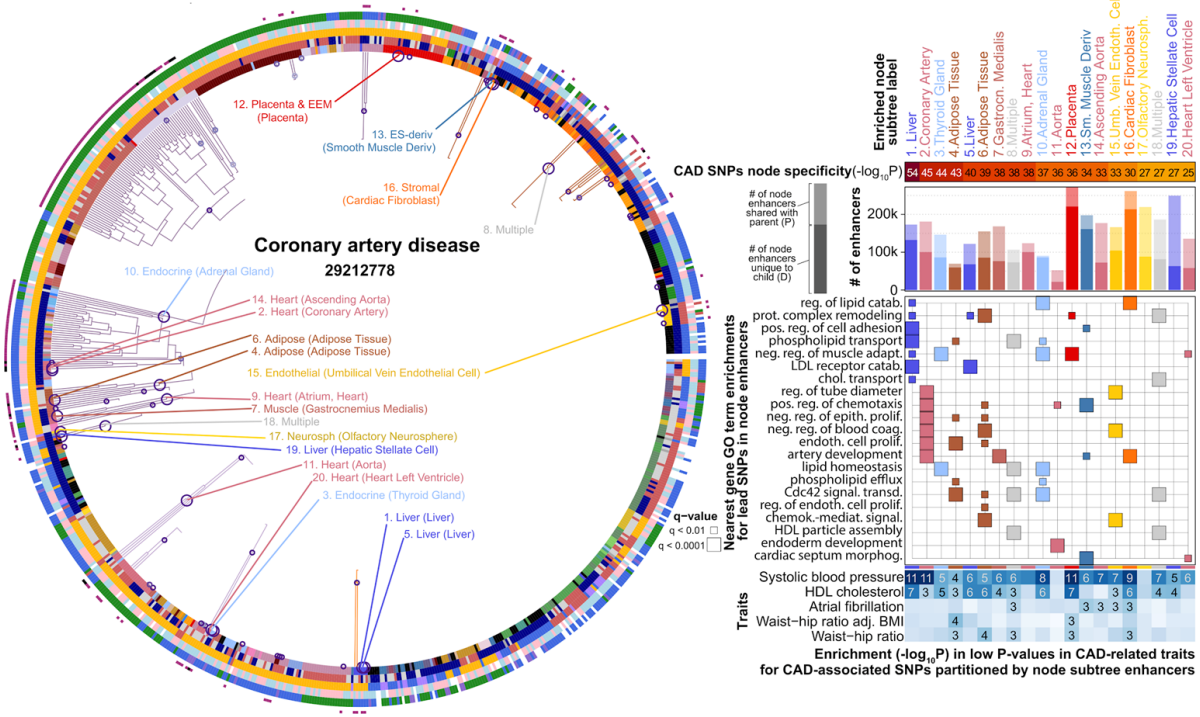
Example 2: Broad schizophrenia signal in the CACNA1C locus in USP44 locus links to CACNA1C gene through multiple enhancers
Note: can also see strong CACNA1C-related heart signal - calcium channel involved in both



Looking at enrichment-prioritized tissues for each GWAS shows **tissue-specific activity of enhancers in loci**.
Locus dissection for a **localized signal (NTN4 - breast cancer)** and a **broad signal (CACNA1C - schizophrenia)**

1. **Range of unifactorial → poly-factorial traits**
2. **Partitioning multi-factorial traits into tissues + pathways of action**

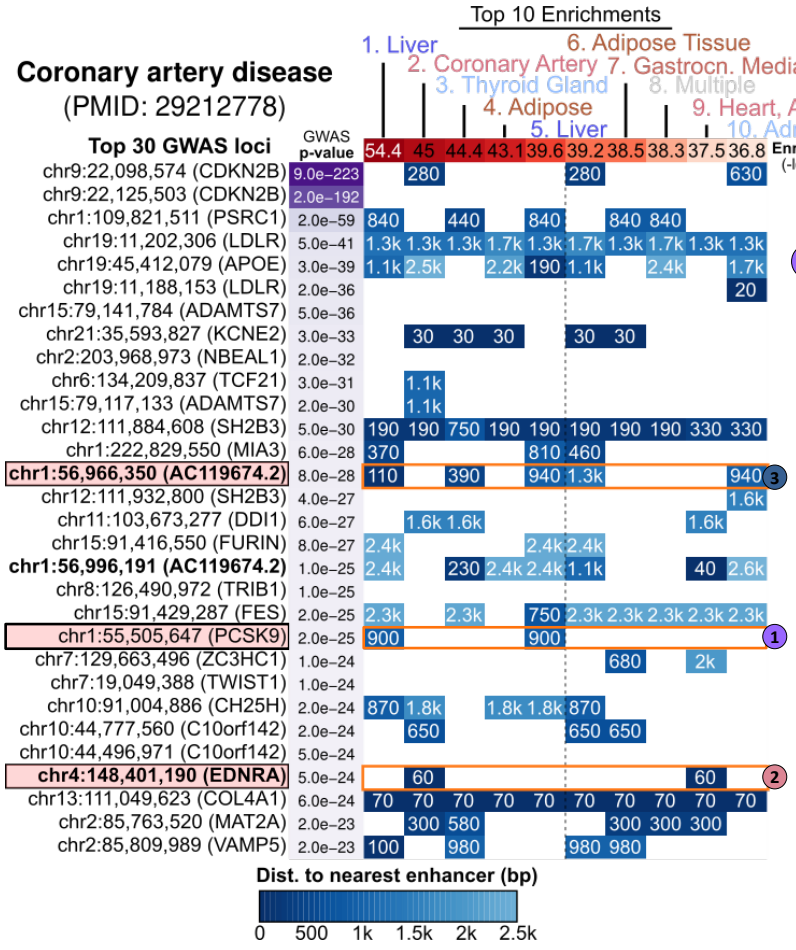




2. CAD is partitioned by top nodes:

Heart	Liver	Adipose	Endocrine
212 SNPs	169 SNPs	122 SNPs	121 SNPs
artery, cardiac, vessel morph.	chol./lipid metab. and transport	adipose tissue innervation	lipid homeostasis
Systolic BP, Atrial fib.	Systolic BP, HDL chol.	Waist-hip-ratio, HDL chol.	Systolic BP, HDL chol.

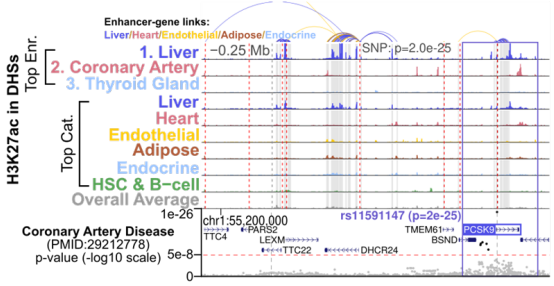
CAD locus analysis illustrates both GWAS-level and locus-level pleiotropy



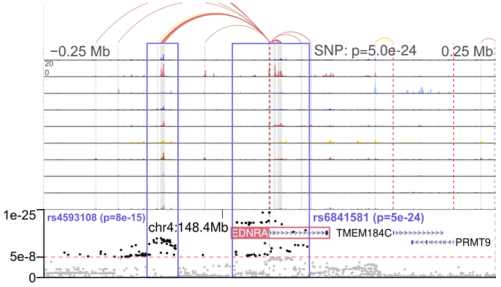
Lead SNP by tissue heatmap:

- SNPs only overlapping heart enhancers (eg. EDNRA, TCF21, ADAMTS7)
- SNPs only overlapping liver (eg. PCSK9)
- SNPs without overlaps (non-enhancer/conditions not captured?)
- SNPs with multiple tissue overlaps (LDLR, APOE, SH2B3, PLPP3)

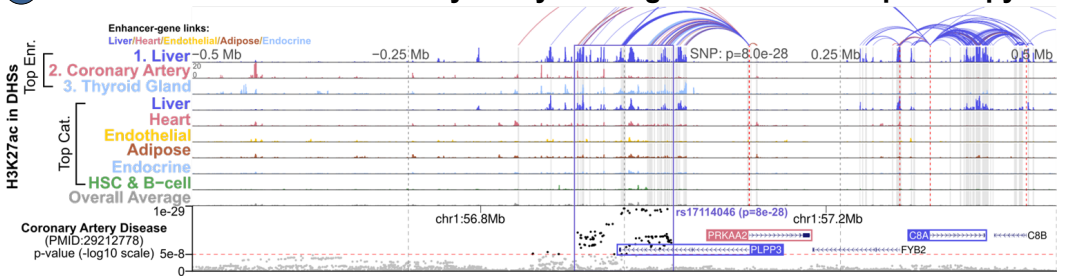
1 **PCSK9: Liver-only** mechanism, mediated through primarily one variant



2 **EDNRA Heart/vasculature-only**, mediated through **multiple enhancers**



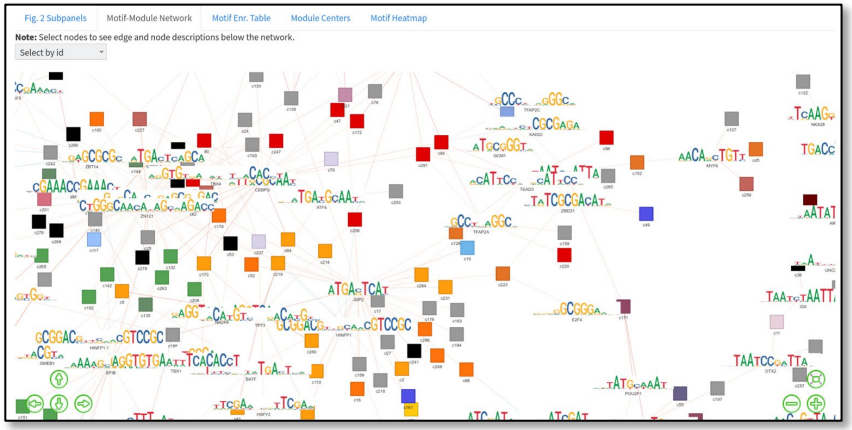
3 **PLPP3: Both liver and coronary artery: multi-gene/multi-tissue pleiotropy**



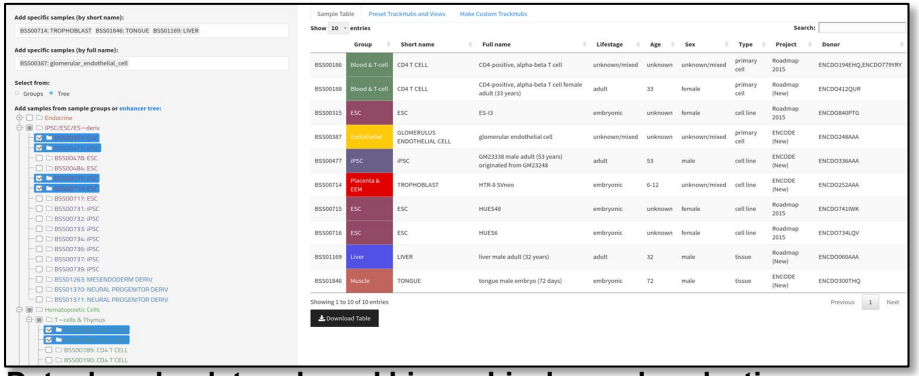
Exploration: Develop browser for: epimap data / gene regulation analysis / GWAS analysis

Interactive browser including:

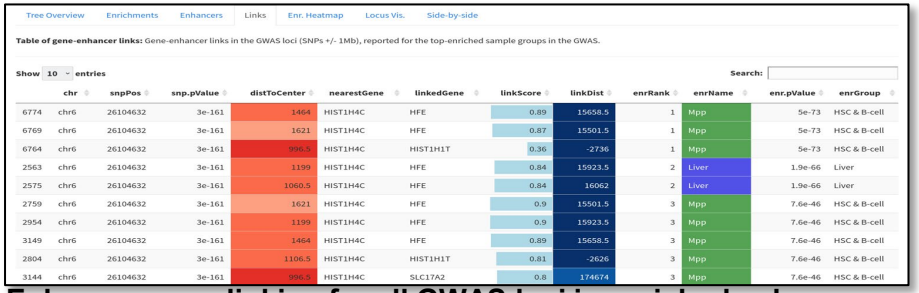
- Custom track hub creation
- Modules-motifs network
- GWAS enrichments
- Per-GWAS locus visualizations



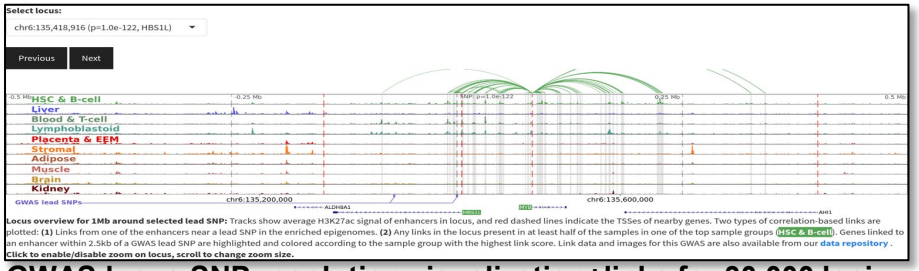
Gene-regulatory circuitry: TF-enhancer regulator linking



Data download: tree-based hierarchical sample selection



Enhancer-gene linking for all GWAS loci in enriched enhancers

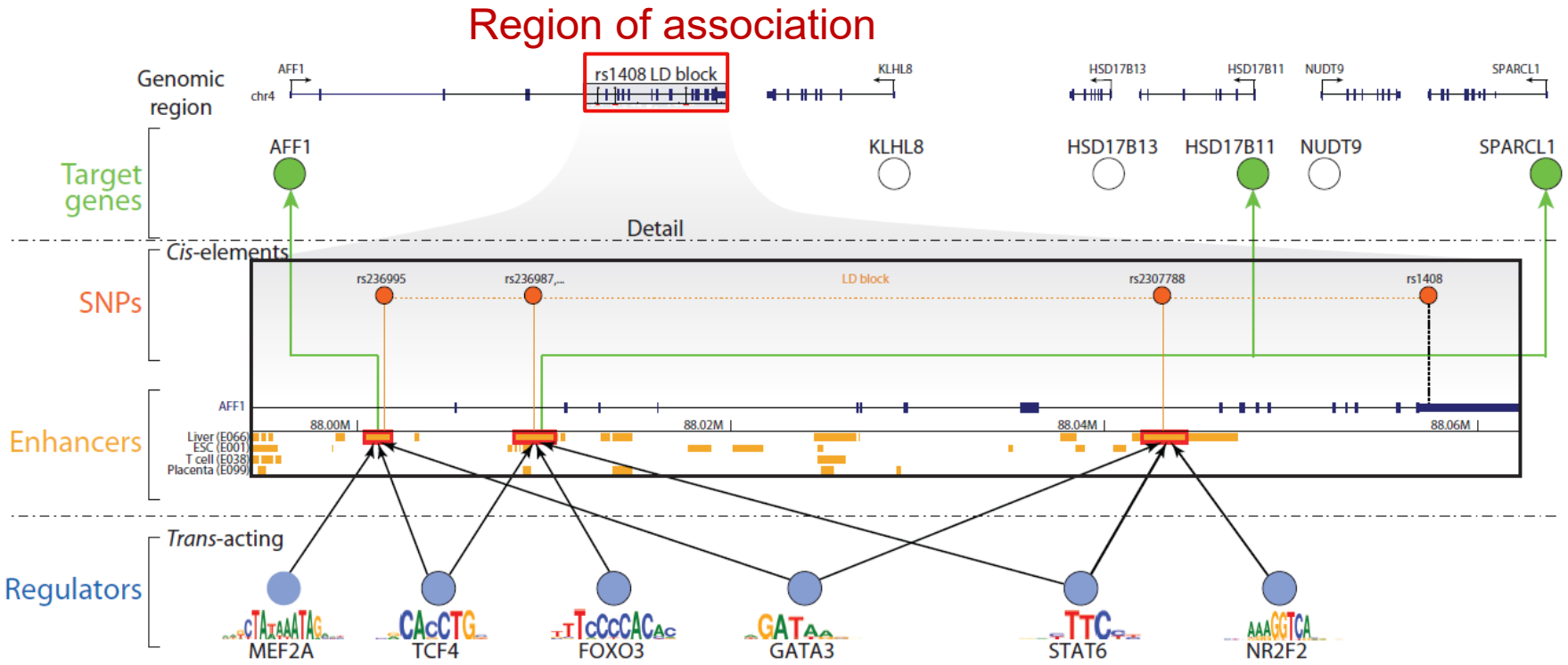


GWAS locus SNP-resolution visualization+links for 30,000 loci

Genetics, Variation, GWAS, PRS, Mechanism

1. Genetics, Variation, GWAS
2. Polygenic scores (PGS)
3. From GWAS to biological insights
4. From Region to Mechanism / Circuitry
5. Quantitative Trait Loci: eQTL/meQTL analysis
6. Mediation Analysis + Mendelian Randomization
7. Heritability and Systems Genetics

Non-coding circuitry helps interpret disease loci



- Expand each GWAS locus using SNP linkage disequilibrium (LD)
  - Recognize **relevant cell types**: tissue-specific enhancer enrichment
  - Recognize **driver TFs**: enriched motifs in multiple GWAS loci
  - Recognize **target genes**: linked to causal enhancers

Quon bioRxiv 467852

The NEW ENGLAND JOURNAL of MEDICINE

FTO Obesity Variant Circuitry and Adipocyte Browning in Humans

Melina Claussnitzer, Ph.D., Simon N. Dankel, Ph.D., Kyoung-Han Kim, Ph.D., Gerald Quon, Ph.D., Wouter Meuleman, Ph.D., Christine Haugen, M.Sc., Viktoria Glunk, M.Sc., Isabel S. Sousa, M.Sc., Jacqueline L. Beaudry, Ph.D., Vijitha Puviindran, B.Sc., Nezar A. Abdennur, M.Sc., Jannel Liu, B.Sc., Per-Arne Svensson, Ph.D., Yi-Hsiang Hsu, Ph.D., Daniel J. Drucker, M.D., Gunnar Mellgren, M.D., Ph.D., Chi-Chung Hui, Ph.D., Hans Hauner, M.D., and Manolis Kellis, Ph.D.

SEPTEMBER 3, 2015 VOL. 373 NO. 10 N Engl J Med 2015;373:895-907.

Mechanistic dissection of a non-coding disease locus

- Identify cell type, causal SNP, regulator, targets, process
- Genome editing demonstrates variant causality
- Adipocyte browning drivers of obesity

Collaborators and contributors

MIT / Broad Institute


Melina Claussnitzer


Gerald Quon


Wouter Meuleman


Nezar Abdennur


Manolis Kellis

U Bergen, Norway


Simon Dankel


Gunnar Mellgren

U. Toronto


Chi-Chung Hui


Kyoung-Han Kim

Munich

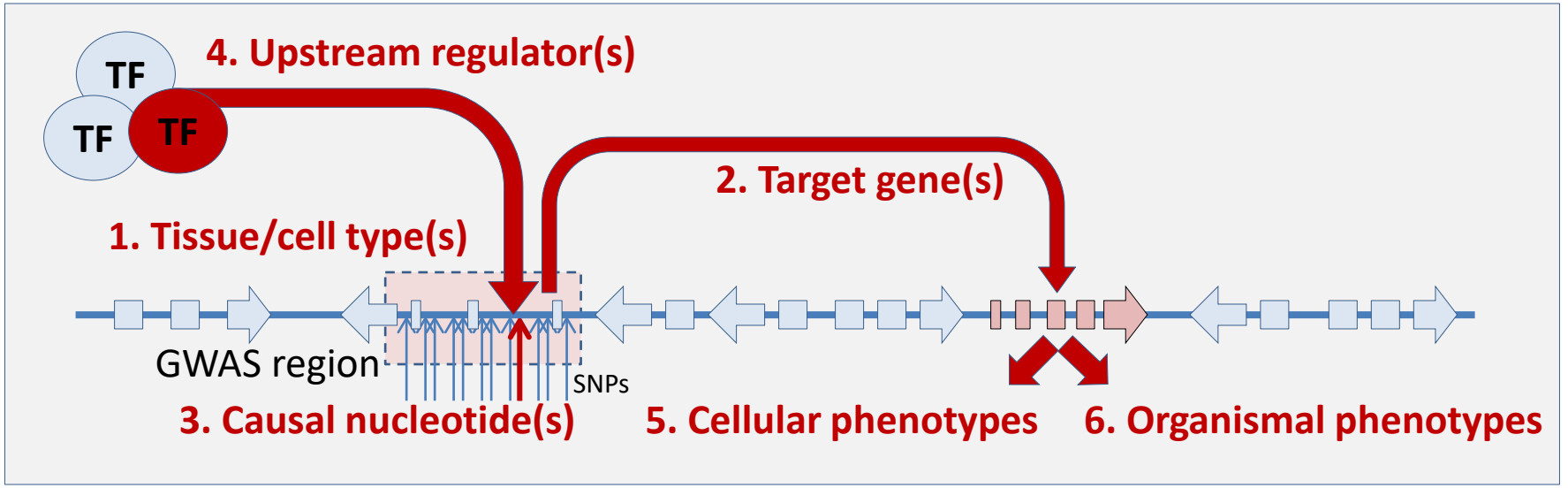

Hans Hauner

Harvard


Yi-Hsiang Hsu

Funding: NIH: NHGRI, Common Fund; Kroner-Fresenius

Dissecting non-coding genetic associations



1. Establish relevant **tissue/cell type**
2. Establish downstream **target** gene(s)
3. Establishing **causal** nucleotide variant
4. Establish upstream **regulator** causality
5. Establish **cellular** phenotypic consequences
6. Establish **organismal** phenotypic consequences

This talk:

Apply these to the FTO locus

The NEW ENGLAND JOURNAL of MEDICINE

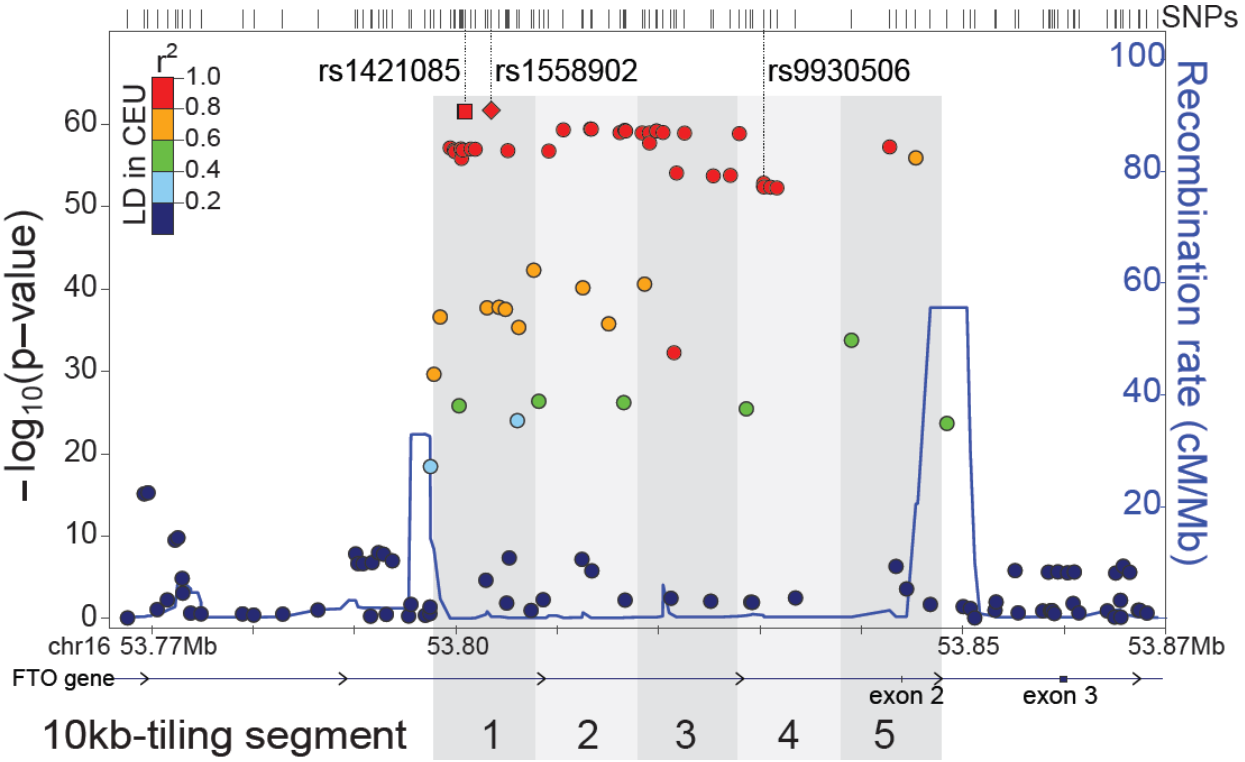
FTO Obesity Variant Circuitry and Adipocyte Browning in Humans

Melina Claussnitzer, Ph.D., Simon N. Dankel, Ph.D., Kyoung-Han Kim, Ph.D., Gerald Quon, Ph.D., Wouter Meuleman, Ph.D., Christine Haugen, M.Sc., Viktoria Glunk, M.Sc., Isabel S. Sousa, M.Sc., Jacqueline L. Beaudry, Ph.D., Vijitha Puviindran, B.Sc., Nezar A. Abdennur, M.Sc., Jannel Liu, B.Sc., Per-Arne Svensson, Ph.D., Yi-Hsiang Hsu, Ph.D., Daniel J. Drucker, M.D., Gunnar Mellgren, M.D., Ph.D., Chi-Chung Hui, Ph.D., Hans Hauner, M.D., and Manolis Kellis, Ph.D.

SEPTEMBER 3, 2015 VOL. 373 NO. 10 N Engl J Med 2015;373:895-907.

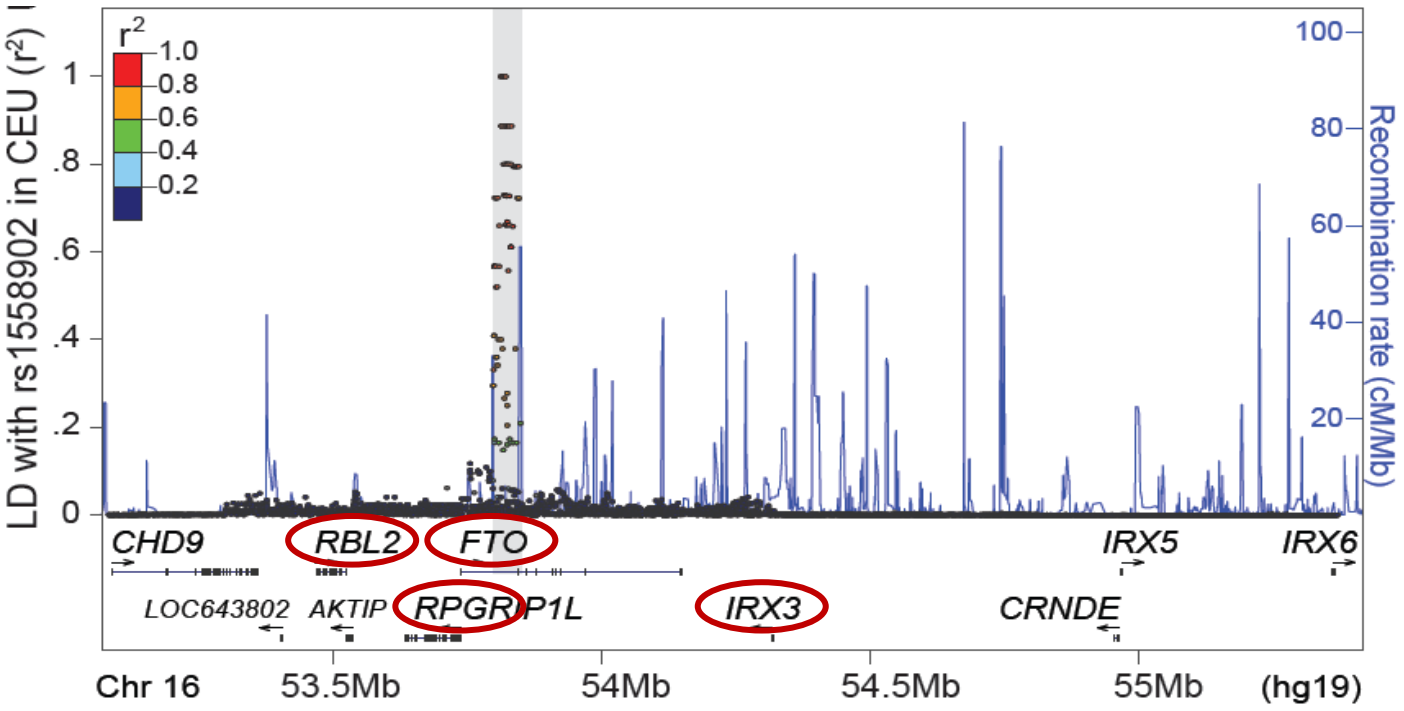
- Implications for obesity therapeutics
- Deep down, a model for dissecting GWAS

FTO region: strongest association with obesity



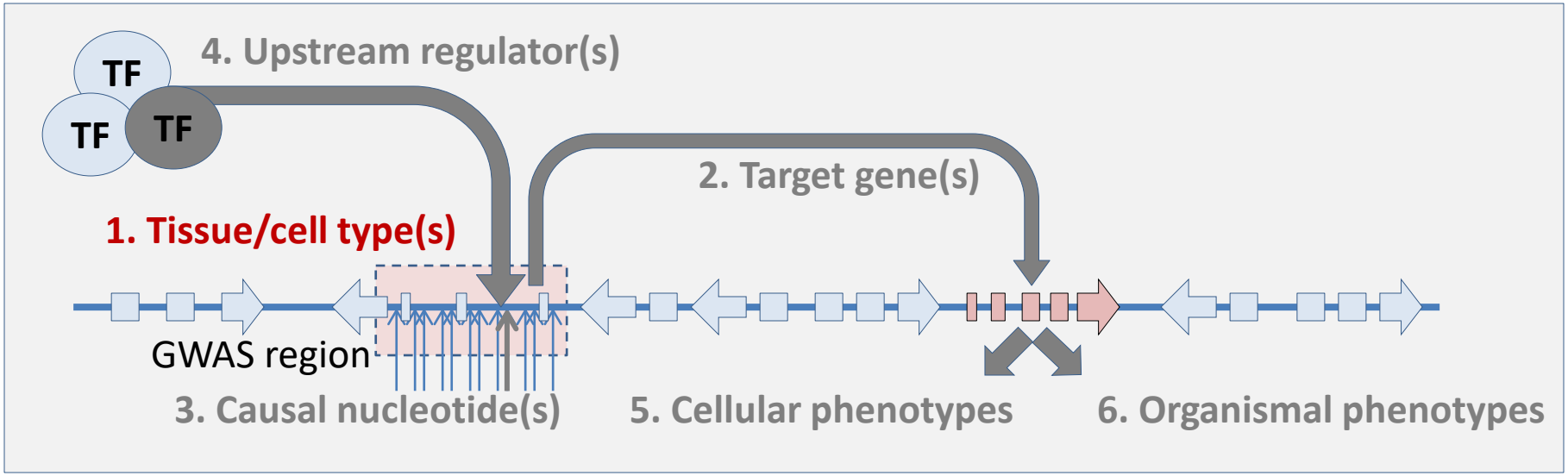
- Associated with **obesity**, Type 2 Diabetes, Cardiovascular traits
- 89 variants in LD, spanning 47kb, intron 1 of FTO gene
- No protein-altering variants: regulatory role? Target? Tissue?

Conflicting proposals of target gene, tissue



- Conflicting predictions: different targets/tissues/species:
  - **FTO** itself: Fischer Nature 09 (Overlap, Mouse **whole-body KO**)
  - **IRX3** in **pancreas**: Ragvin PNAS 10 (4C, Zebrafish KO)
  - **RBL2** in **lymphocytes**: Jowett Diabetes 2010 (Expression levels, eQTL)
  - **RPRGRIP1L** in **brain**: Stratigopoulos JBC 2014 (Leptin signaling, CUX binding)
  - **IRX3** in **brain**: Smemo Nature 14 (4C, Mouse brain DN)

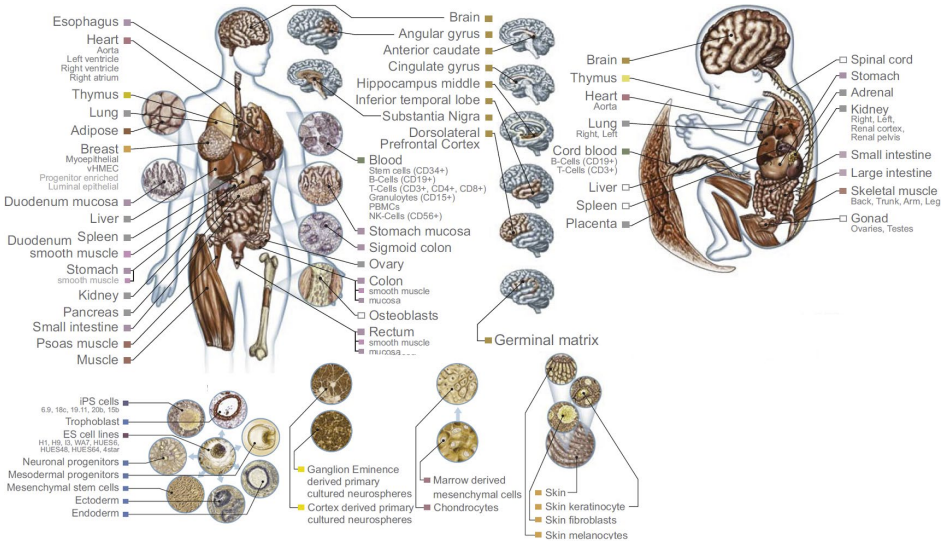
1. Establish relevant tissue/cell type



1. Establish relevant **tissue/cell type**
2. Establishing **causal** nucleotide variant
3. Establish downstream **target** gene(s)
4. Establish upstream **regulator** causality
5. Establish **cellular** phenotypic consequences
6. Establish **organismal** phenotypic consequences

Epigenomic mapping across 100+ tissues/cell types

Diverse tissues and cells



Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood…)

Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood…)

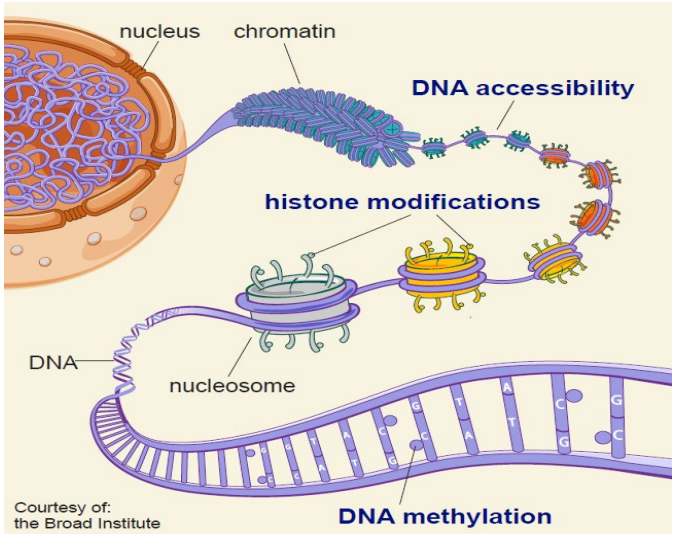
ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench…)

ROADMAP epigenomics PROJECT

ENCODE

X

Diverse epigenomic assays



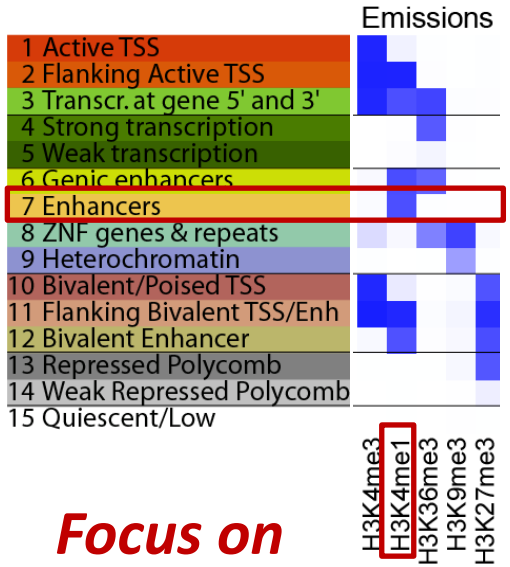
Histone modifications
- **H3K4me3, H3K4me1, H3K36me3**
- **H3K27me3, H3K9me3, H3K27/9ac**
- **+20 more**

Open chromatin:
- **DNA accessibility**

DNA methylation:
- **WGBS, RRBS, MRE/MeDIP**

Gene expression
- **RNA-seq, Exon Arrays**

Chromatin state annotations across 127 epigenomes

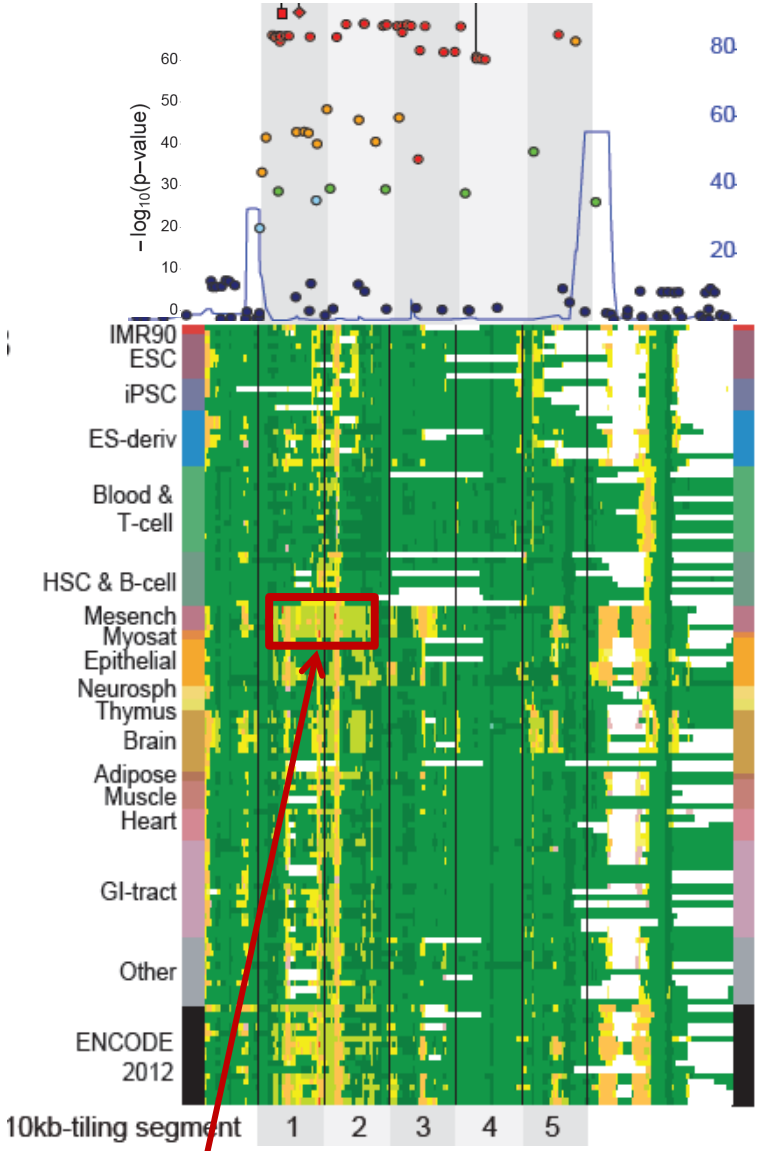


Focus on enhancer states

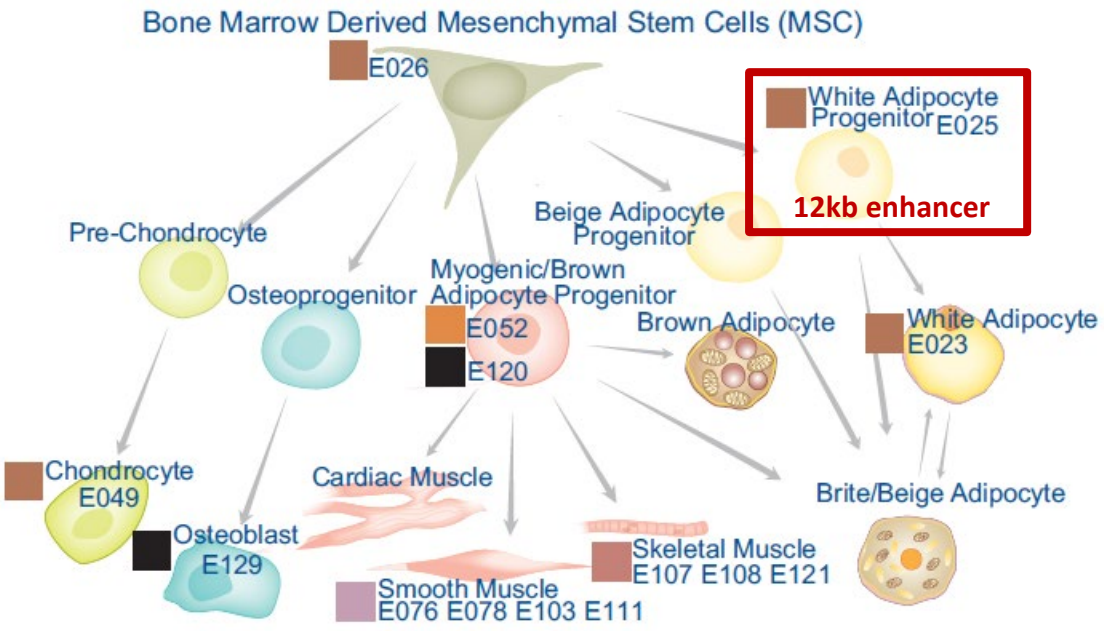
Roadmap Epigenomics, Nature 2015

Tissue-specific annotations of predicted enhancers, promoters, transcribed, repressed, quiescent regions

1. Tissue: Chromatin states predict adipocyte function

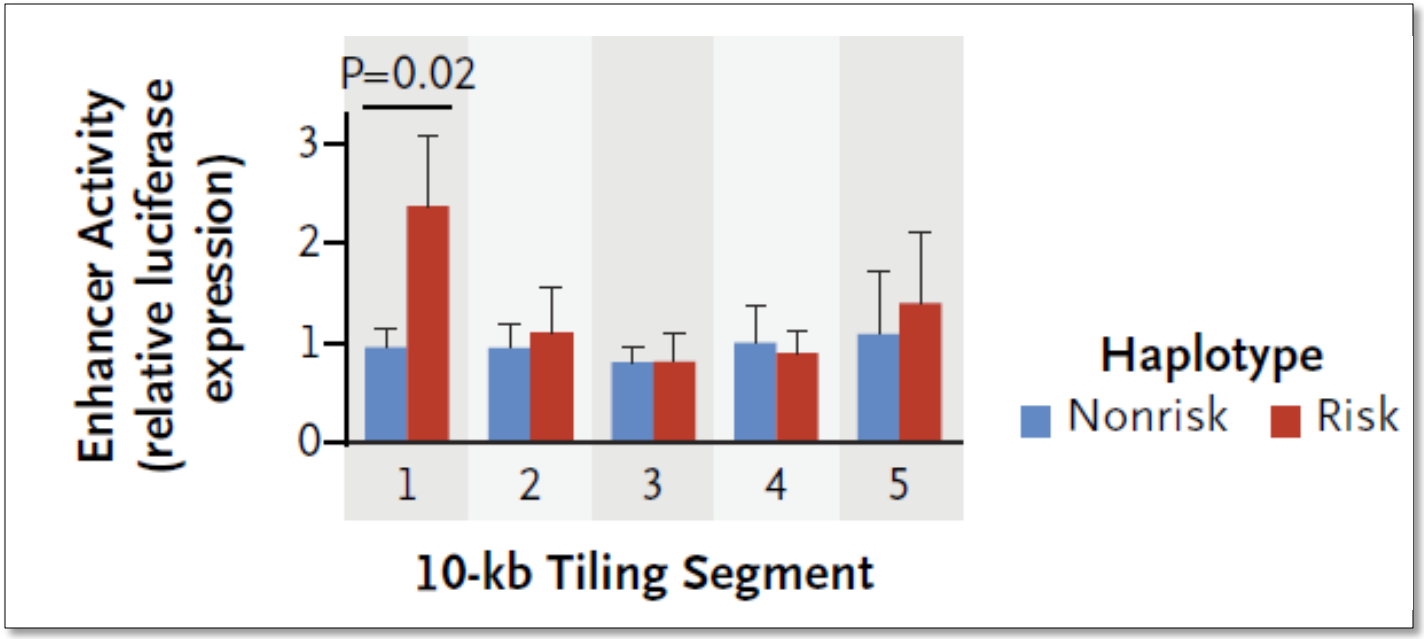


12 kb super-enhancer



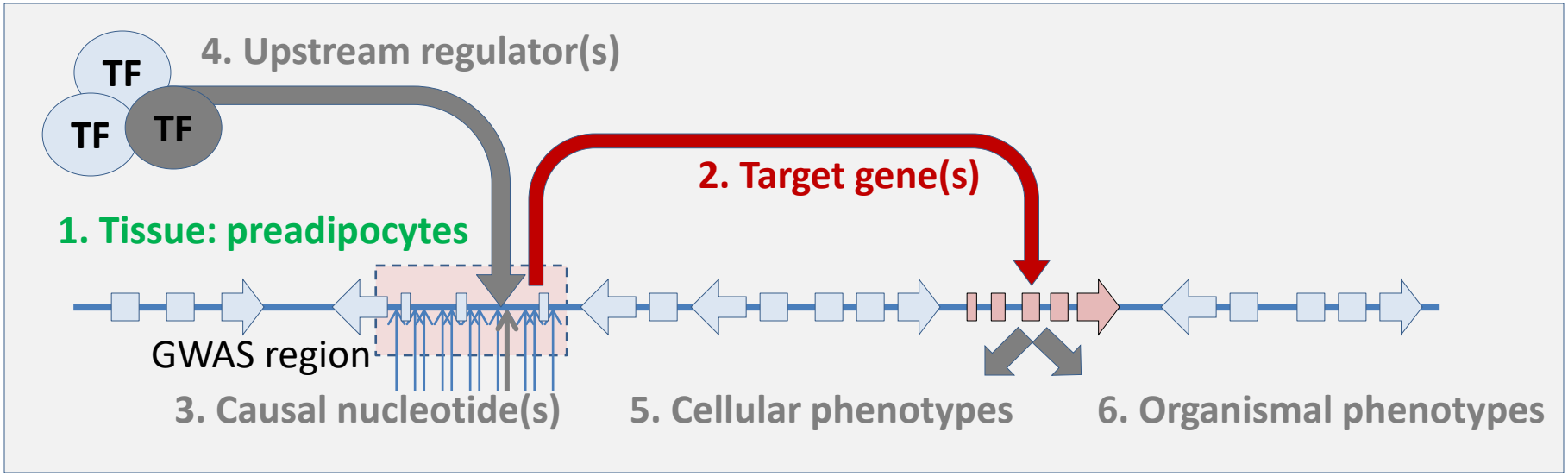
Epigenomic signatures point to progenitors of white/beige adipocytes

Enhancer tiling experiments confirm region, cell type



*Risk haplotype shows **increased** activity, gain-of-function*

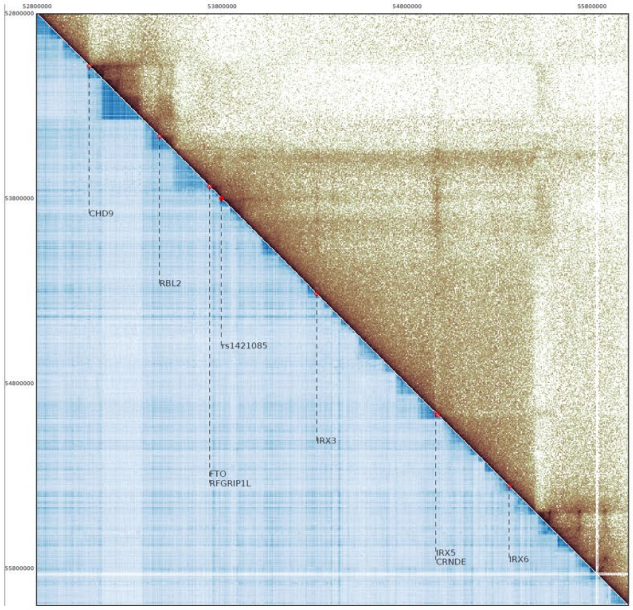
2. Establish downstream target genes



1. Establish relevant **tissue/cell type: pre-adipocytes**
2. Establish downstream **target** gene(s)
3. Establishing **causal** nucleotide variant
4. Establish upstream **regulator** causality
5. Establish **cellular** phenotypic consequences
6. Establish **organismal** phenotypic consequences

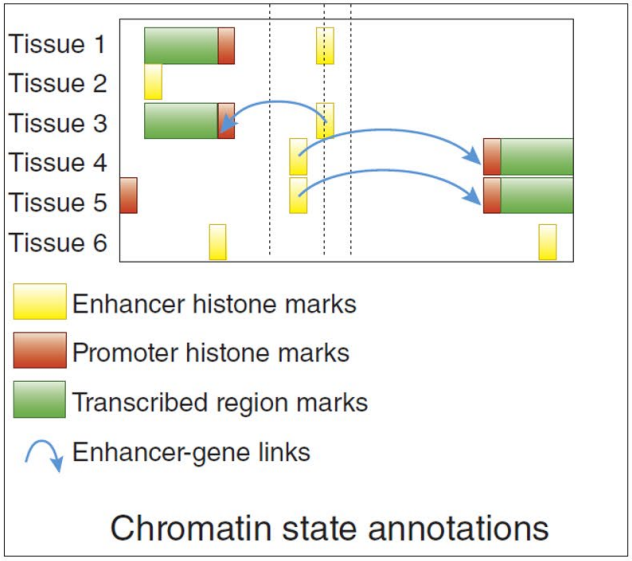
Link enhancers to their target genes 3 lines of evidence:

Physical



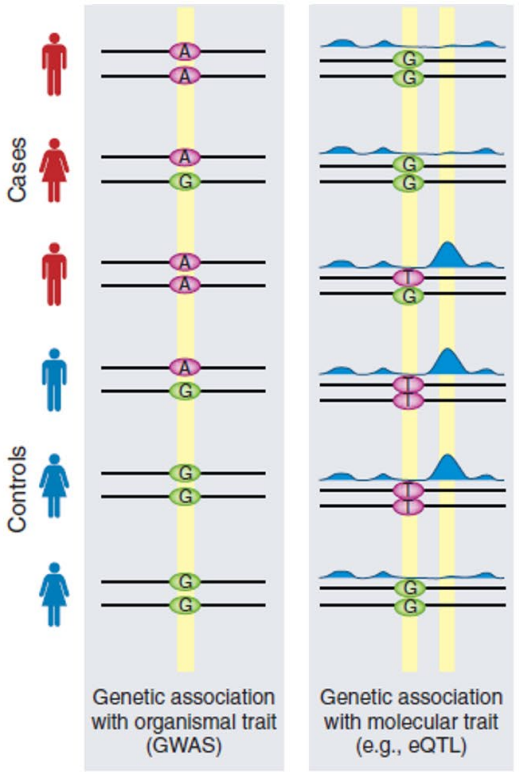
Hi-C: Physical proximity in 3D

Functional



Enhancer-gene activity correlation

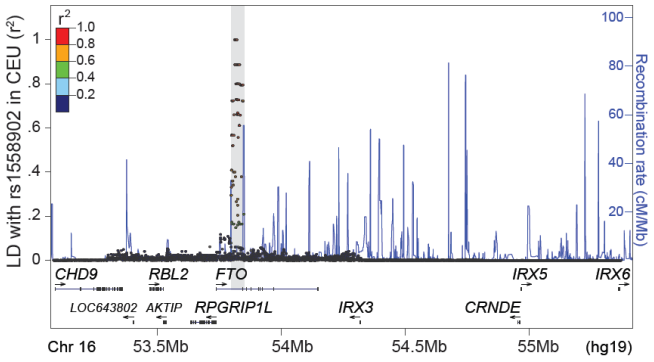
Genetic

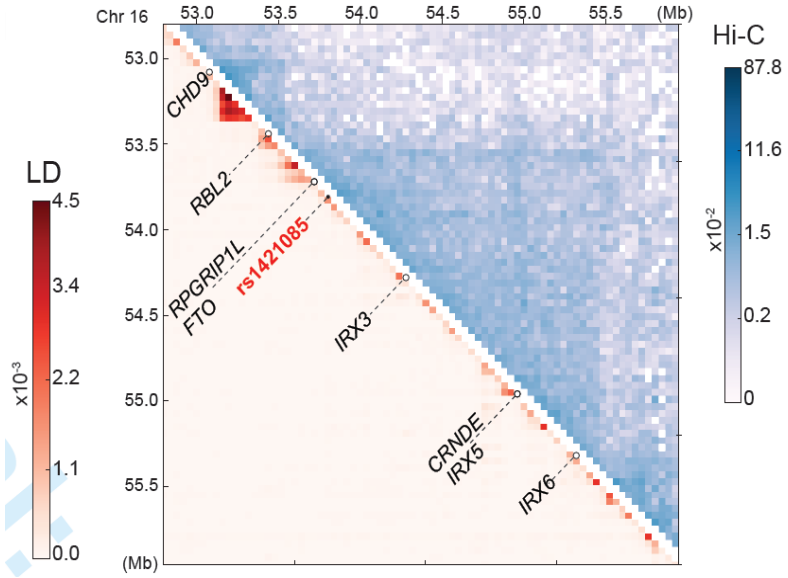


eQTL evidence: SNP effect on expression

Complementary evidence at physical, functional, genetic level

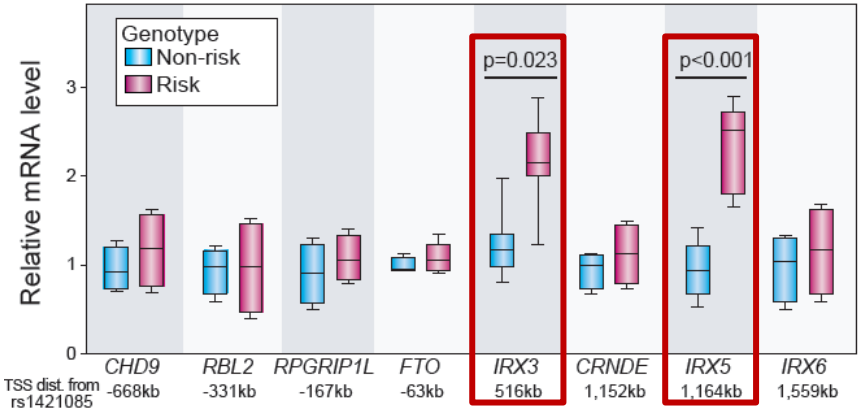
2. Targets: 3D folding and expr. genetics indicate IRX3+IRX5





Dixon, Nature 2012

Cohort of **20 homozygous risk** and **18 homozygous non-risk** individuals:
Genotype-dependent expression?

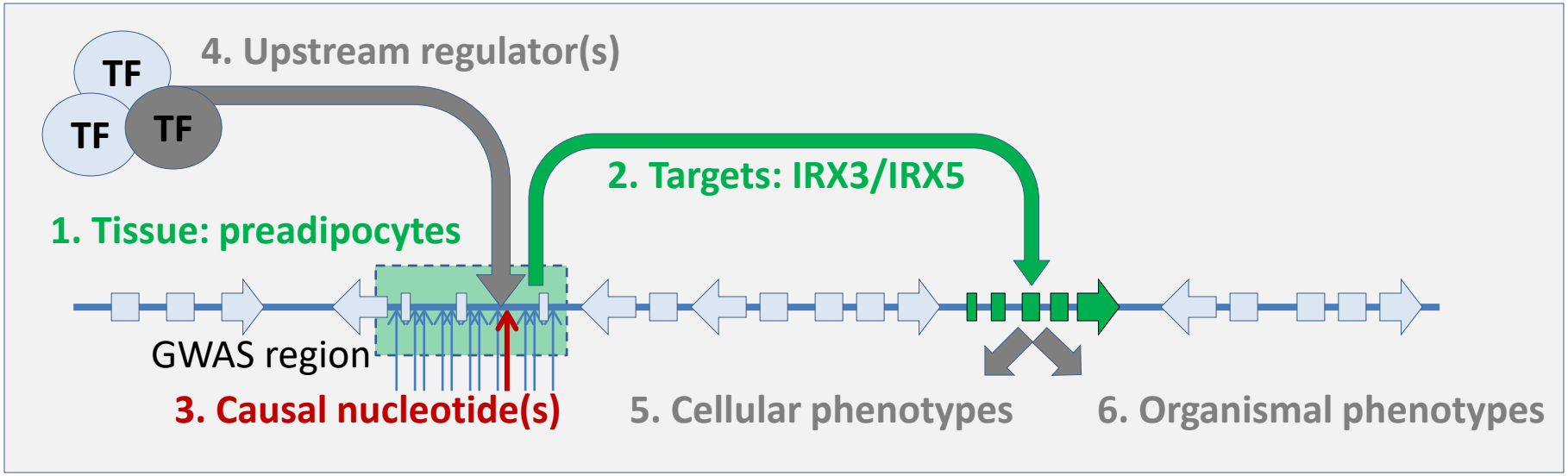


eQTL targets: IRX3 and IRX5

Topological domains span 2.5Mb
Implicate **8 candidate genes**

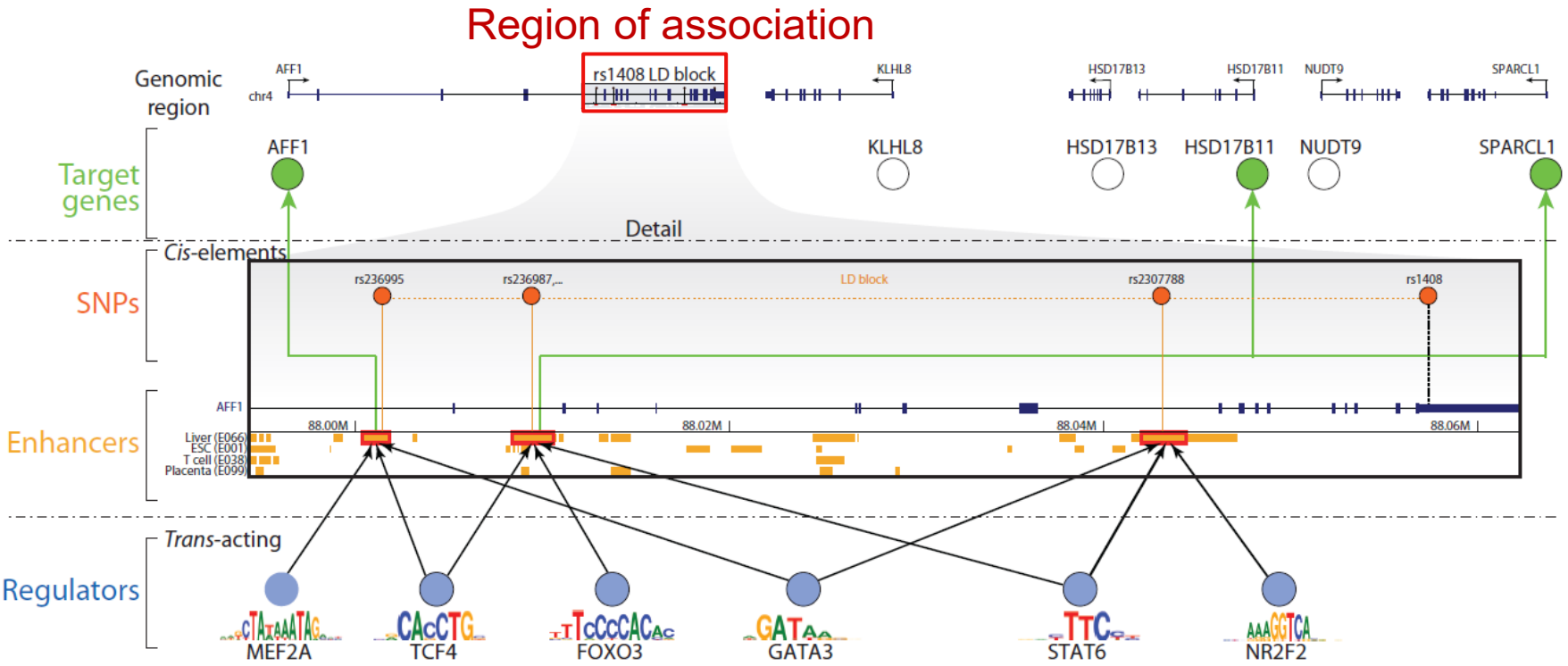
(1) Risk allele: increased expression
(gain-of-function by loss of repressor)
(2) Action in early adipocyte differentiation
(eQTL is not visible in whole-adipose tissue)

3. Establish causal variant



1. Establish relevant **tissue/cell type:** **pre-adipocytes**
2. Establish downstream **target** gene(s): **IRX3 and IRX5**
3. Establishing **causal** nucleotide variant
4. Establish upstream **regulator** causality
5. Establish **cellular** phenotypic consequences
6. Establish **organismal** phenotypic consequences

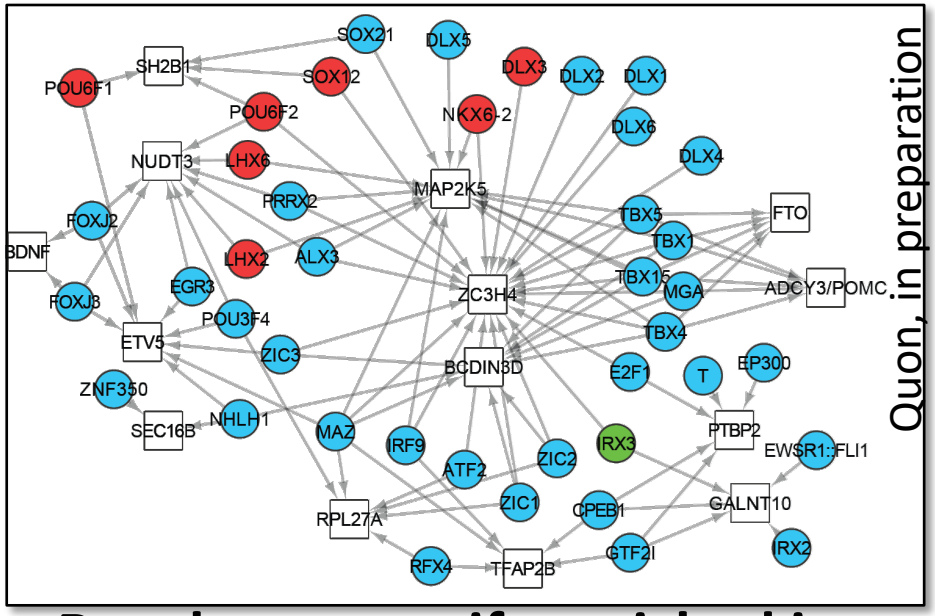
Non-coding circuitry helps interpret disease loci



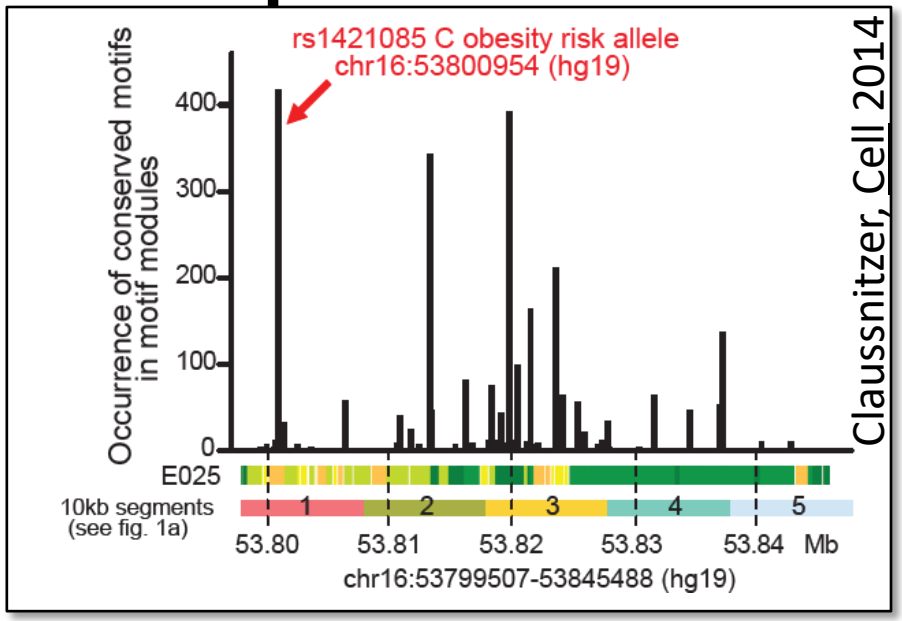
Study multiple GWAS loci to find commonalities/enrichments:

- Epigenomics: narrow down regulatory **regions**, relevant cell types
- Comparative genomics: prioritize **SNPs** over conserved nucleotides
- Regulatory genomics: match **motifs** to predict driver TFs/regulators
- Functional genomics: predict **target genes** in common pathways

Motif enrichment + conservation: predict causal SNP



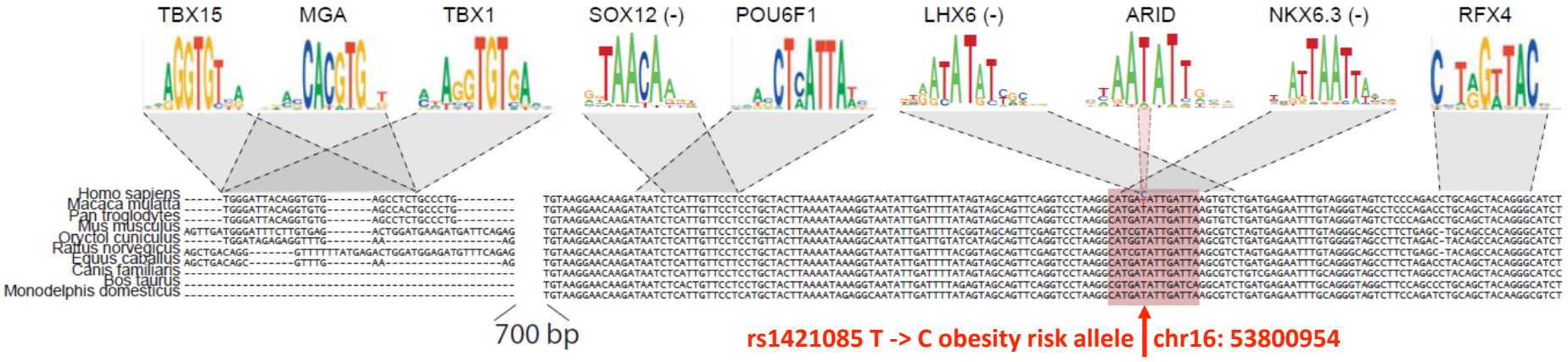
Quon, in preparation



Claussnitzer, Cell 2014

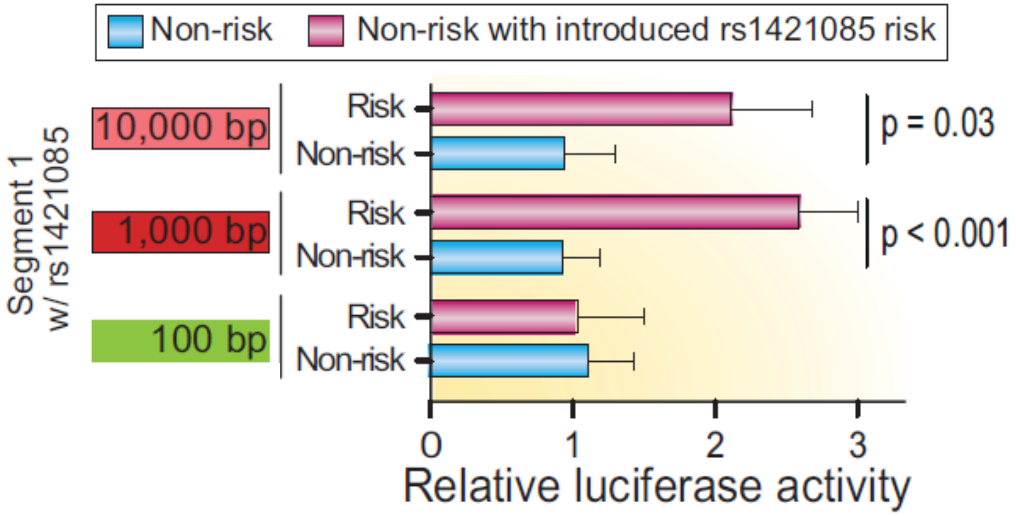
Regulatory motifs enriched in BMI GWAS hits

Regulatory motif combinations conserved across mammals



Causal nucleotide rs1421085: risk alters T to C, abolishes AT-rich motif

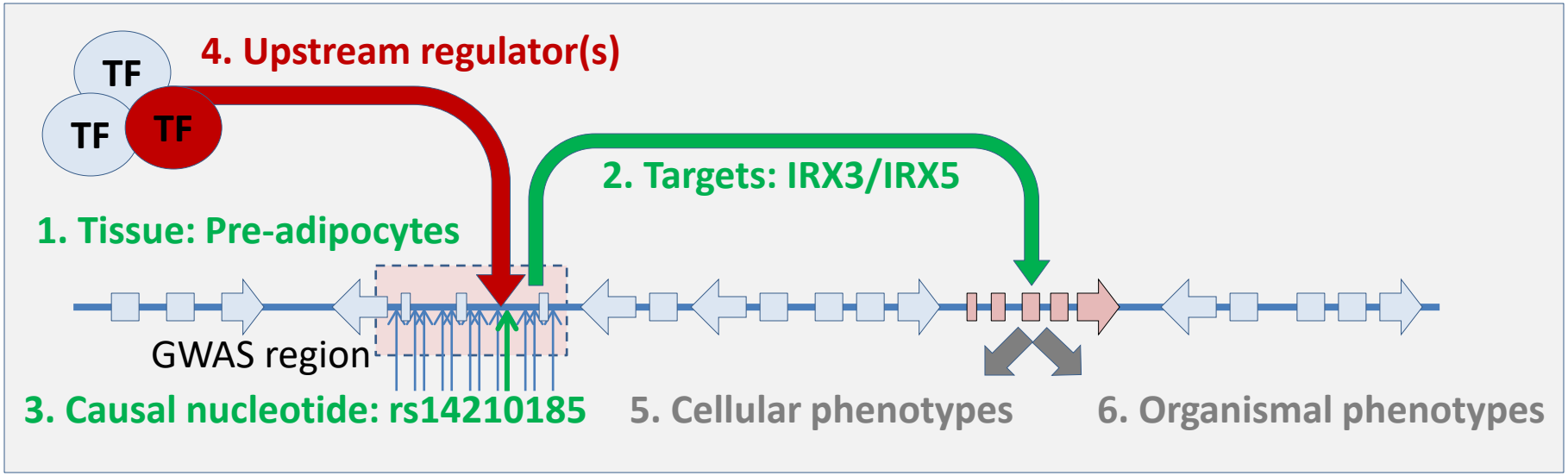
Single-nucleotide alteration alters enhancer activity



SNP recapitulates risk haplotype de-repression

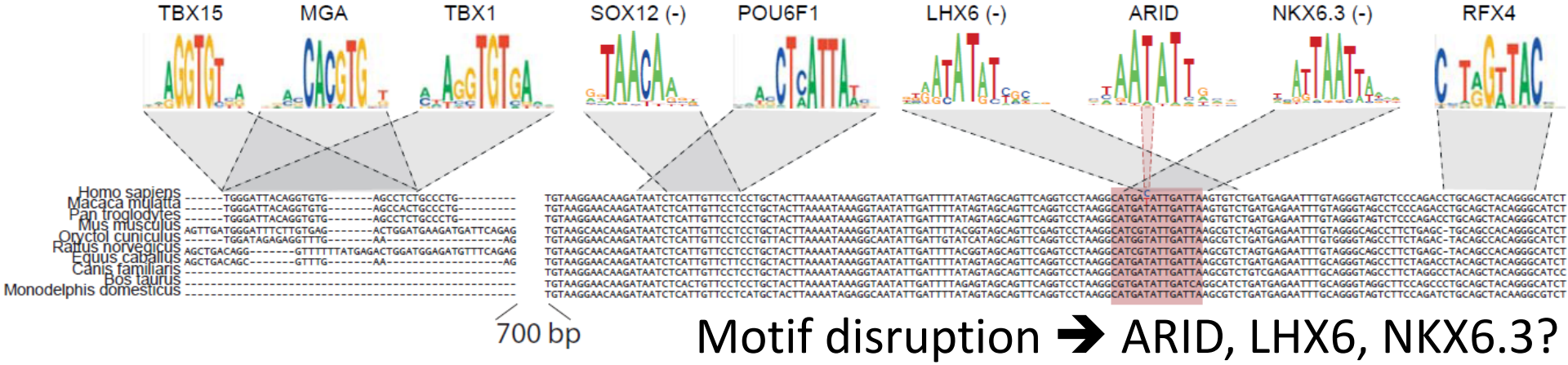
Acts at 10kb and 1kb (not at 100bp)

4. Establish upstream regulator

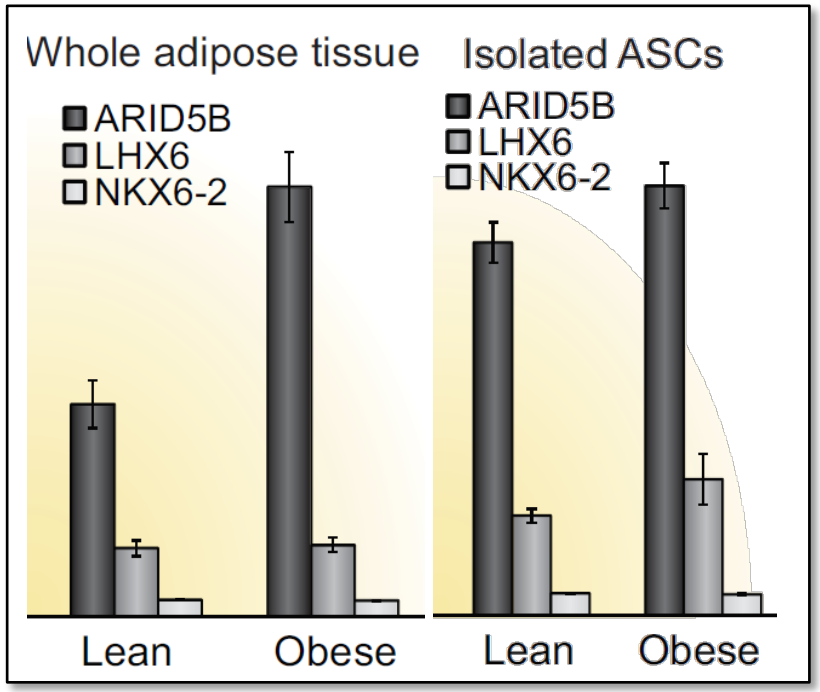


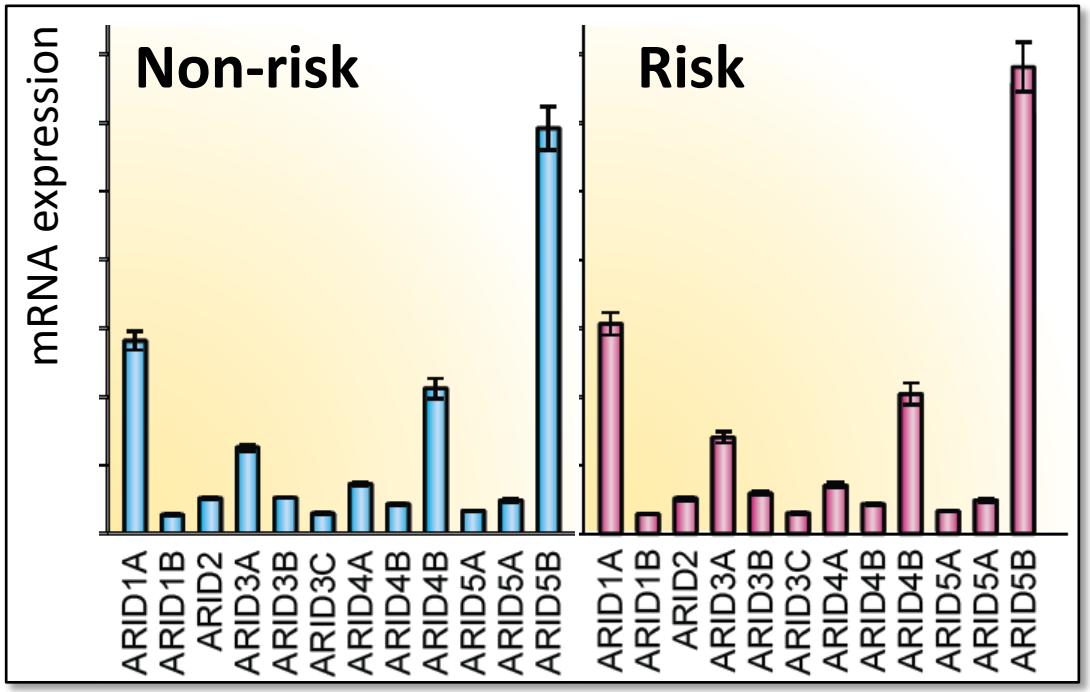
1. Establish relevant **tissue/cell type:** **pre-adipocytes**
2. Establish downstream **target** gene(s): **IRX3 and IRX5**
3. Establishing **causal** nucleotide variant: **rs1421085**
4. Establish upstream **regulator** causality
5. Establish **cellular** phenotypic consequences
6. Establish **organismal** phenotypic consequences

Regulator expression implicates ARID5B repressor



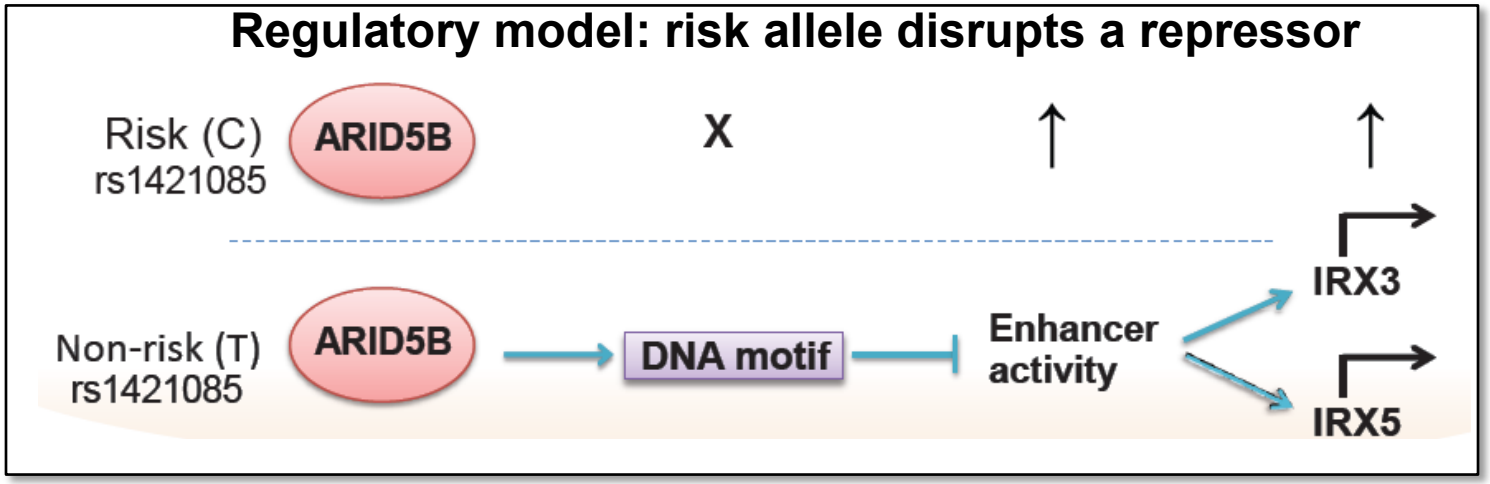
Motif disruption ➔ ARID, LHX6, NKX6.3?





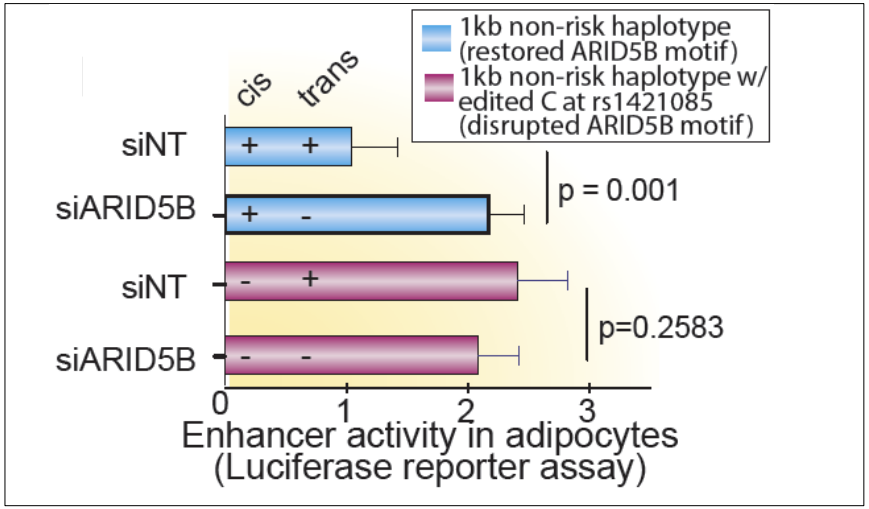
- Adipose/ASC expression suggests ARID family
- Highest expression implicates ARID5B upstream regulator

Causality and epistasis of ARID5B repression

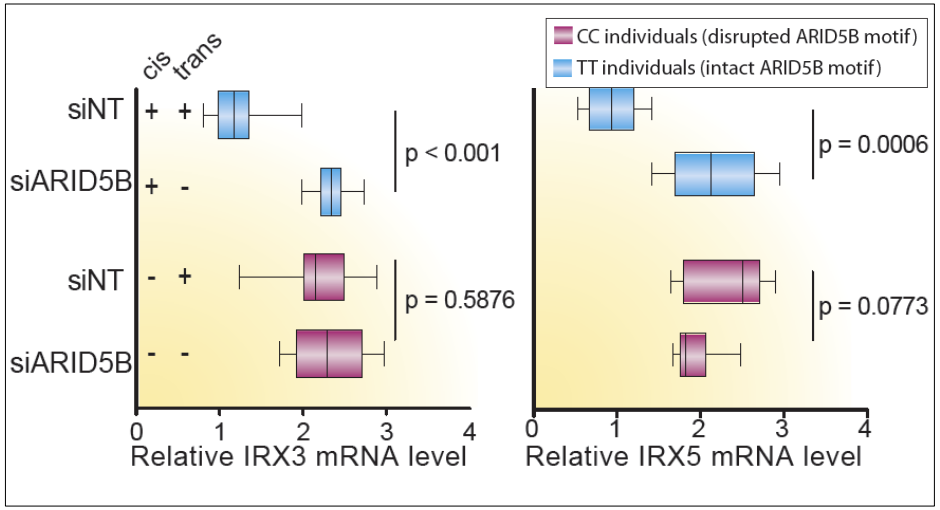


Cis/trans conditional analysis

Enhancer activity

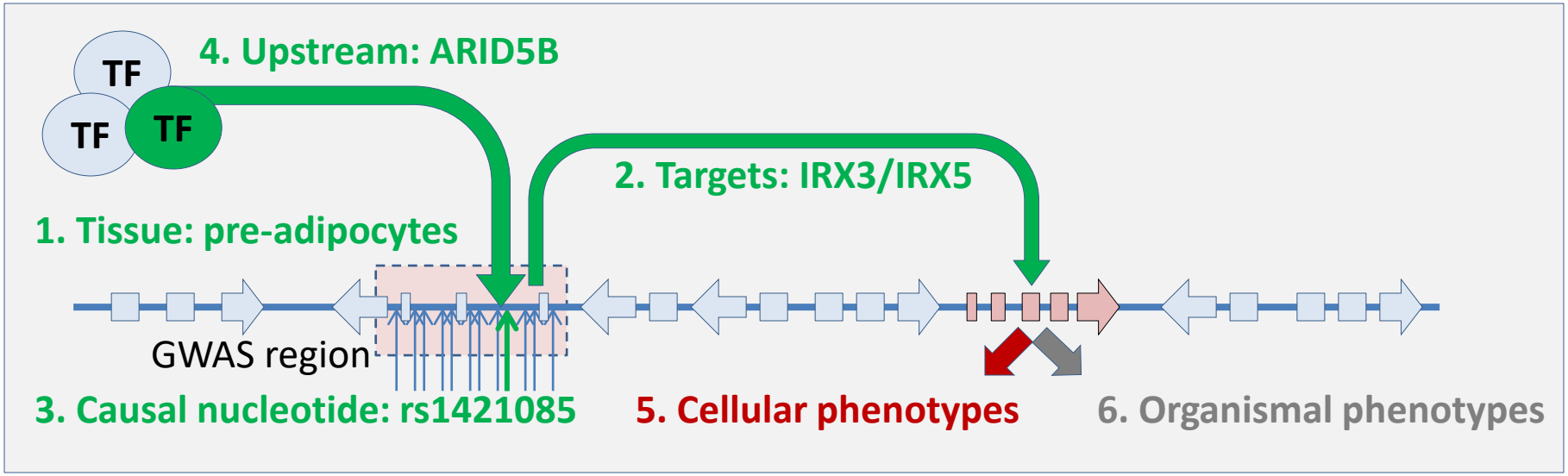


IRX3/5 expression

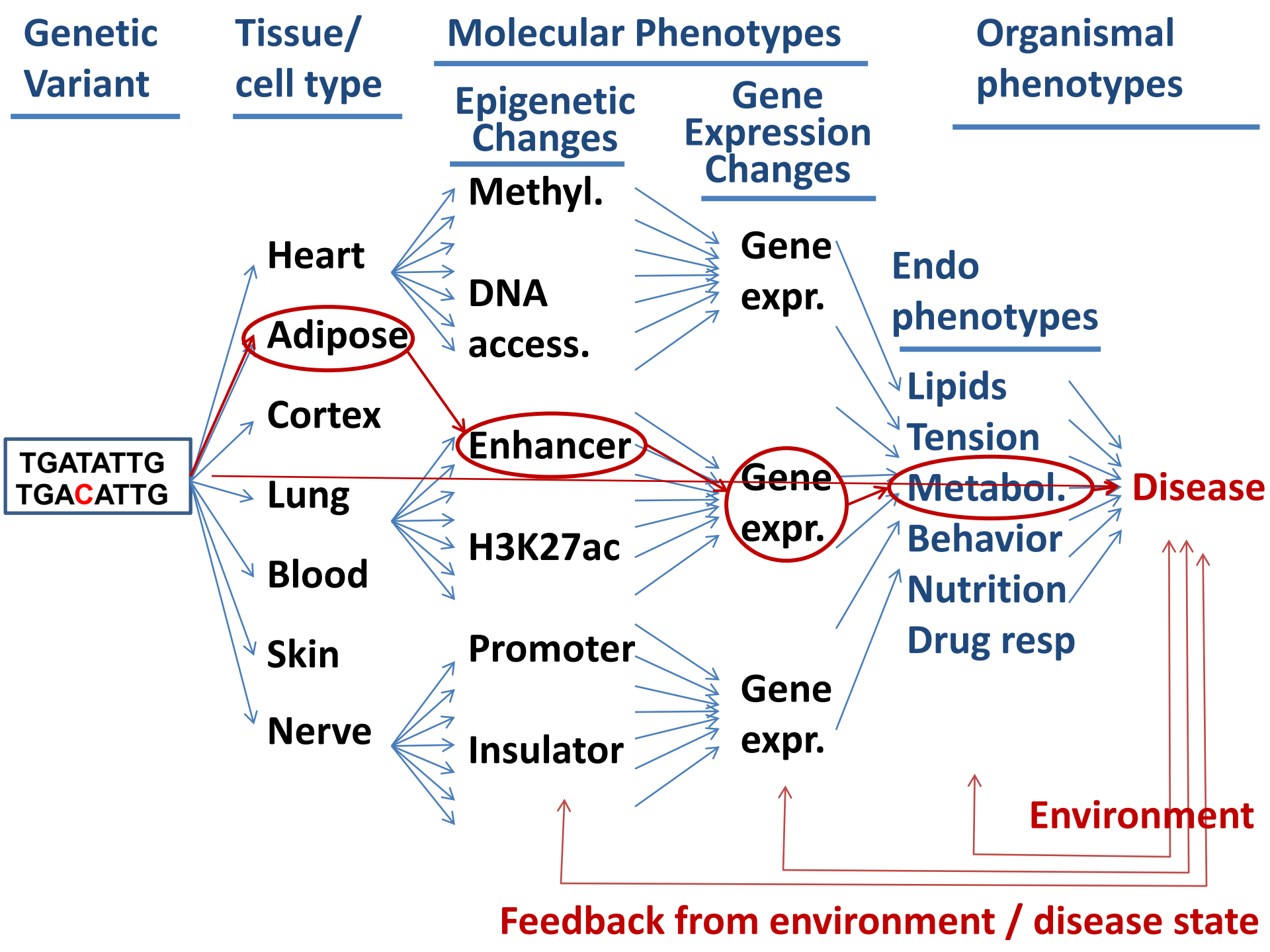


- ***Repression of enhancer, IRX3 and IRX5 all require both TF and motif***
- ***Disrupting motif (CC), or repressing ARID5B (siRNA) ➔ de-repression***

5. Establish cellular phenotypic consequences

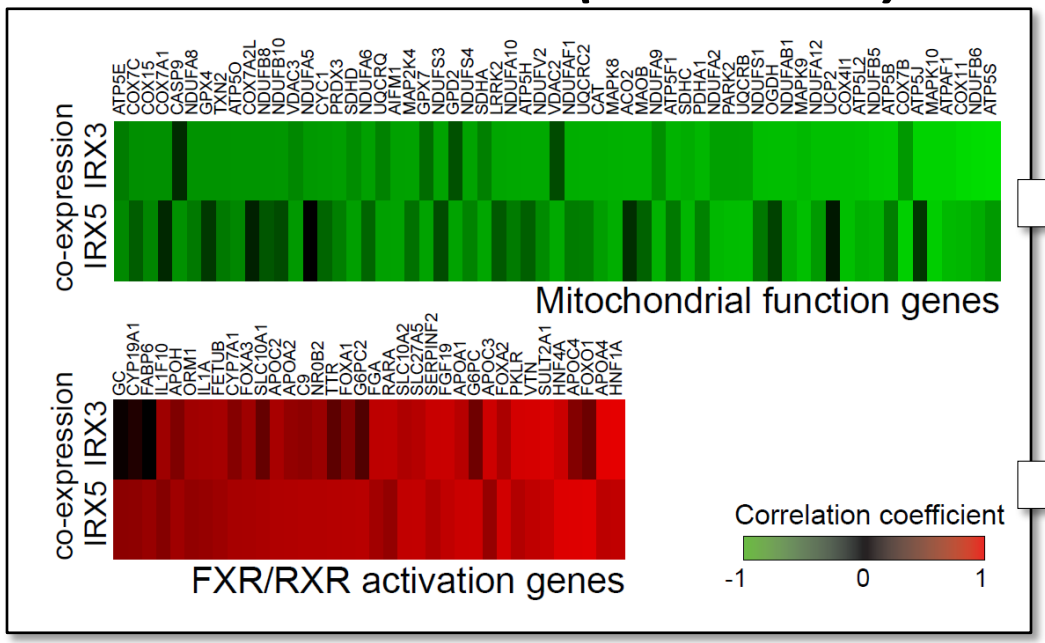


1. Establish relevant **tissue/cell type:** **pre-adipocytes**
2. Establish downstream **target** gene(s): **IRX3 and IRX5**
3. Establishing **causal** nucleotide variant: **rs1421085**
4. Establish upstream **regulator** causality: **ARID5B**
5. Establish **cellular** phenotypic consequences
6. Establish **organismal** phenotypic consequences

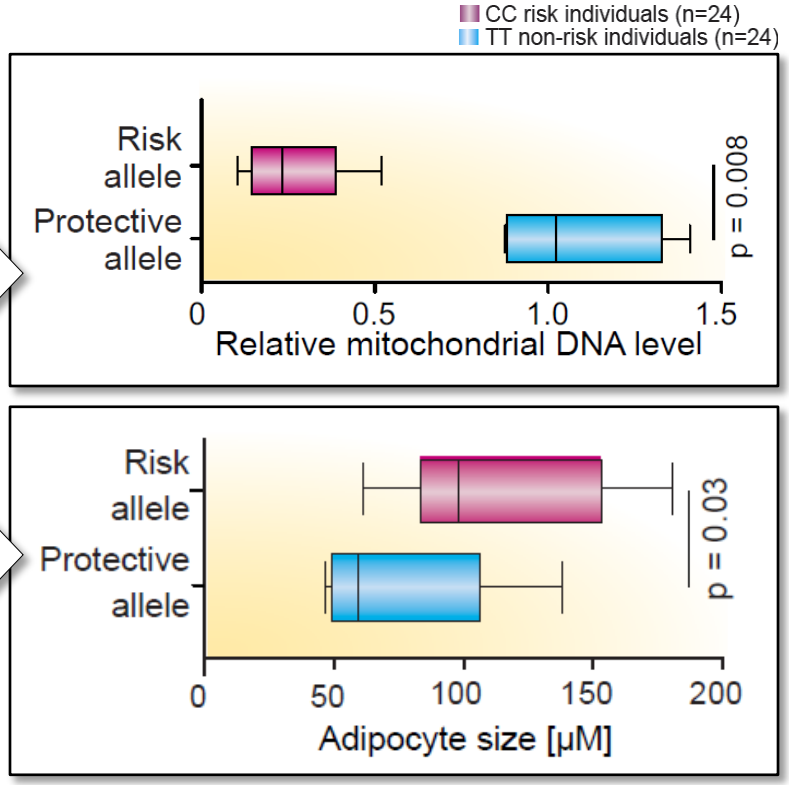

Genetic Variant
Tissue/ cell type
Molecular Phenotypes
Epigenetic Changes
Gene Expression Changes
Organismal phenotypes
TGATATTG
TGACATTG
Heart
Adipose
Cortex
Lung
Blood
Skin
Nerve
Methyl.
DNA access.
Enhancer
H3K27ac
Promoter
Insulator
Gene expr.
Gene expr.
Gene expr.
Endo phenotypes
Lipids
Tension
Metabol.
Behavior
Nutrition
Drug resp
Disease
Environment
Feedback from environment / disease state

Expression analysis to recognize target processes

Search for genes co-expressed with IRX3 and IRX5 (n=20 indiv.)



Reflected in cellular phenotypes

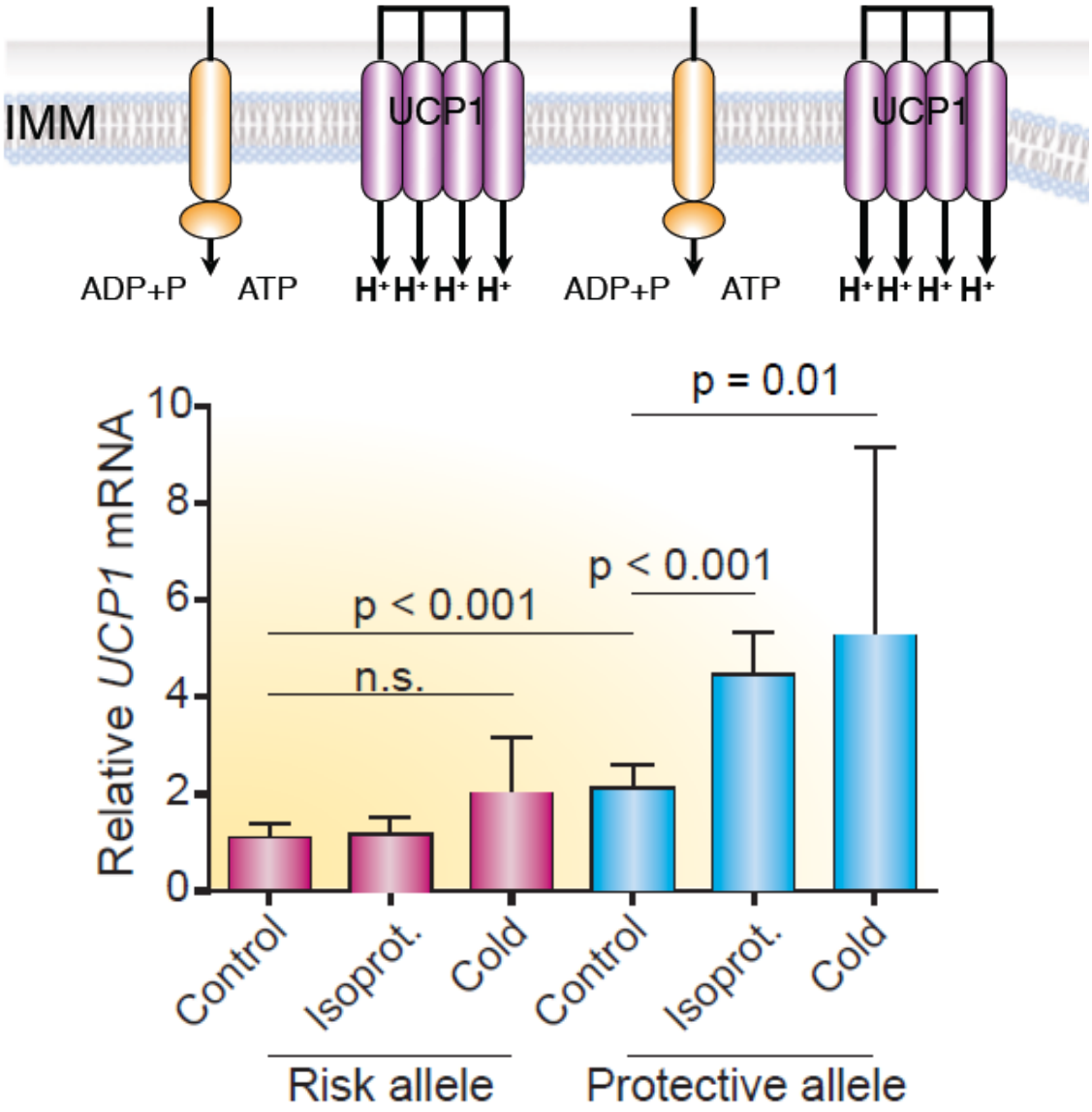


Negative correlation: mitochondria
Positive correlation: lipid storage

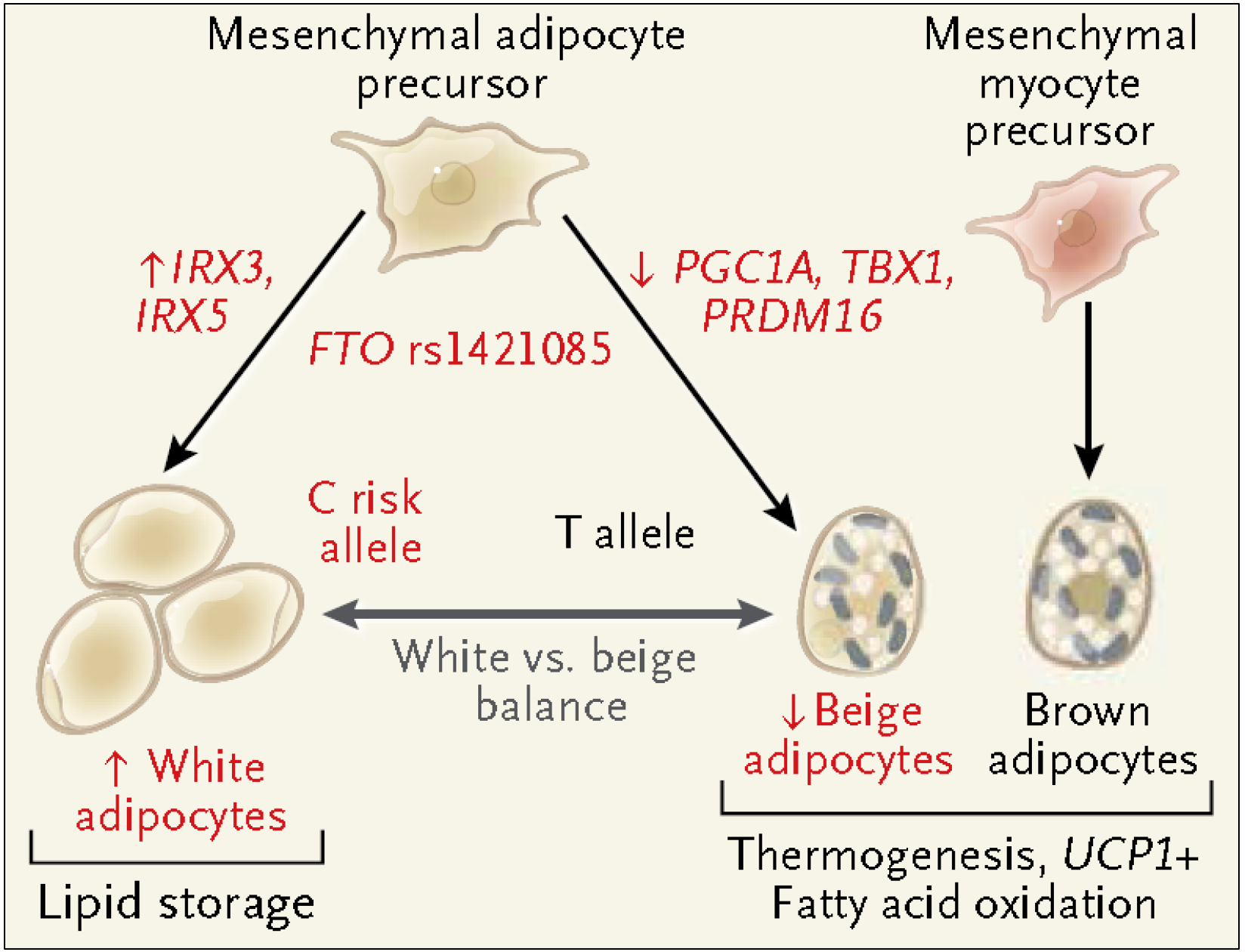
Risk carriers: increased mito
Non-risk: increased adipocytes

Risk allele: shift from dissipation to storage

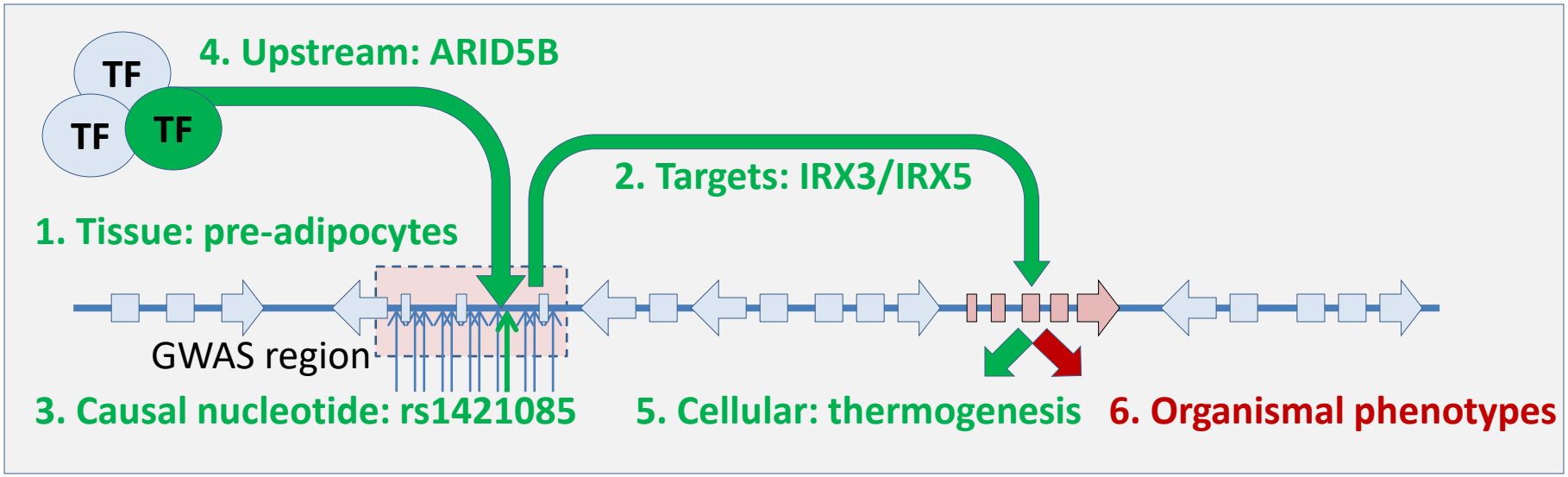
Risk individuals show disrupted thermogenesis



Mechanistic cellular model for FTO obesity locus

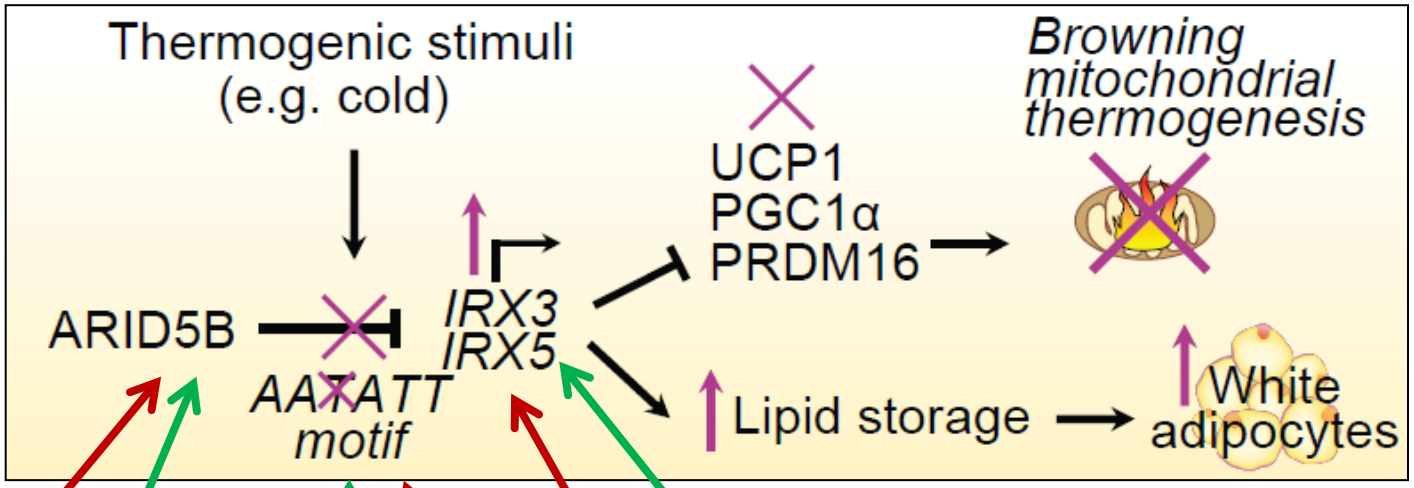


6. Manipulate circuitry to impact organism level



1. Establish relevant **tissue/cell type:** **pre-adipocytes**
2. Establish downstream **target** gene(s): **IRX3 and IRX5**
3. Establishing **causal** nucleotide variant: **rs1421085**
4. Establish upstream **regulator** causality: **ARID5B**
5. Establish **cellular** phenotypic consequences: **thermogenesis**
6. Establish **organismal** phenotypic consequences

Dissected circuitry: entry points for intervention



ARID5B KD (obesity)

ARID5B OE (anti-obesity)

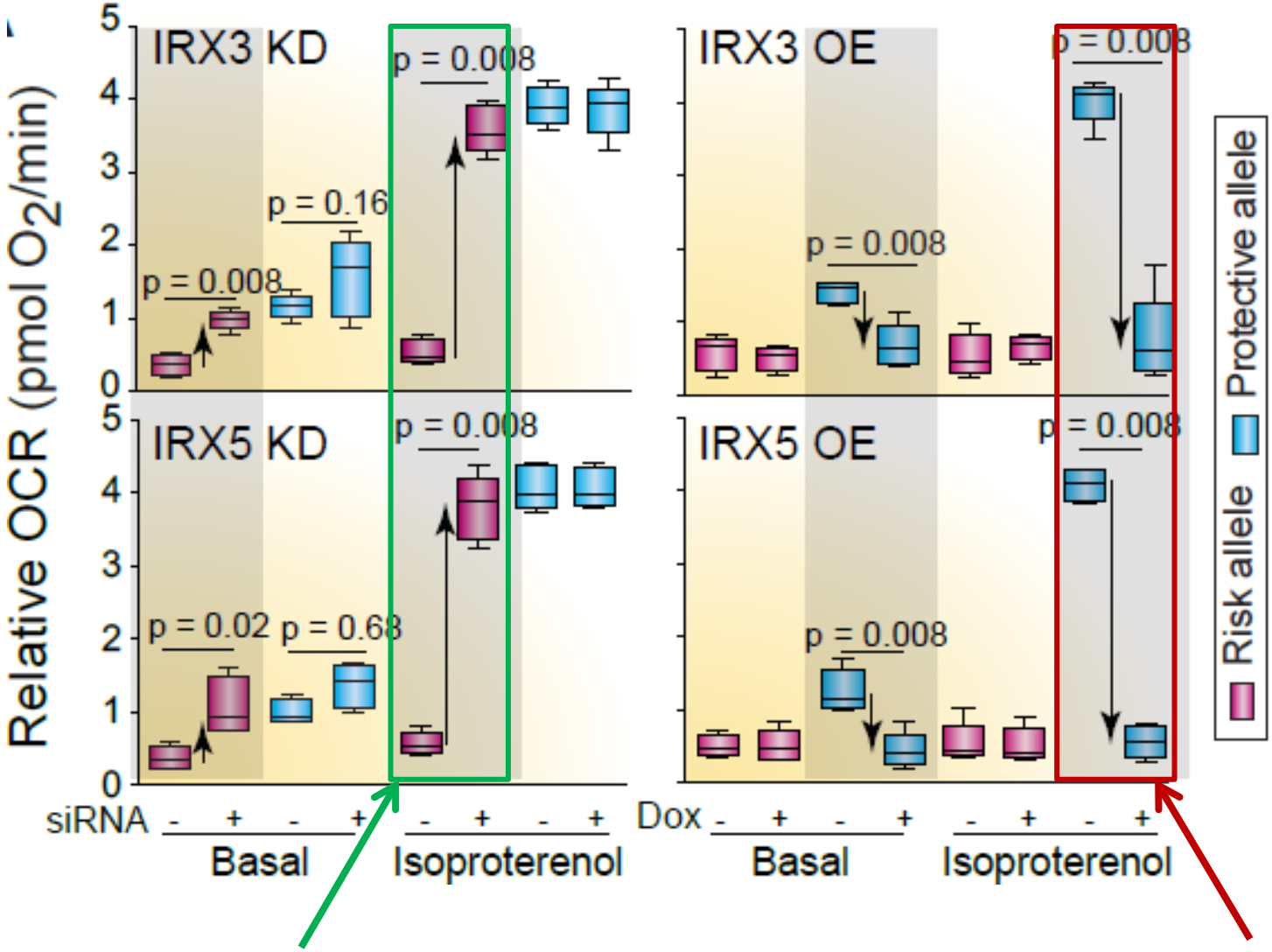
C-to-T motif rescue (anti-obesity phenotypes)

T-to-C motif disruption (pro-obesity phenotypes)

IRX3, IRX5 overexpression (pro-obesity phenotypes)

IRX3, IRX5 knock-down (anti-obesity phenotypes)

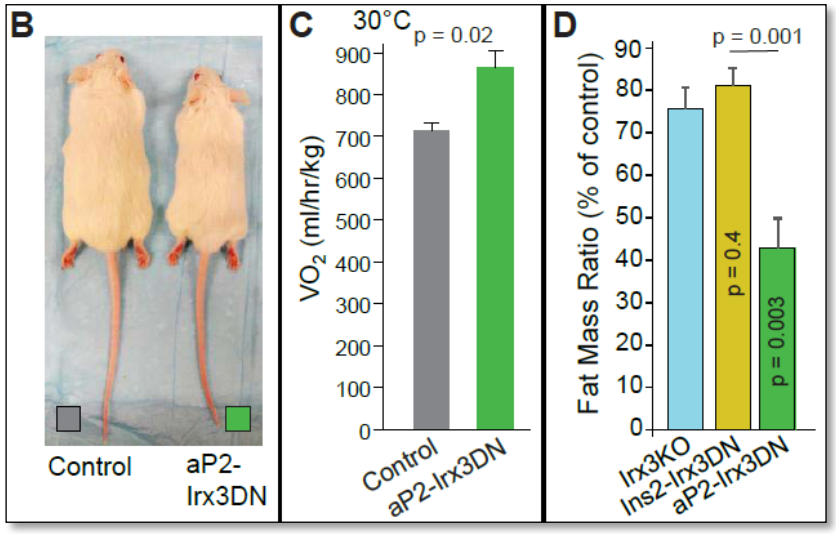
IRX3+IRX5 expression impacts energy utilization



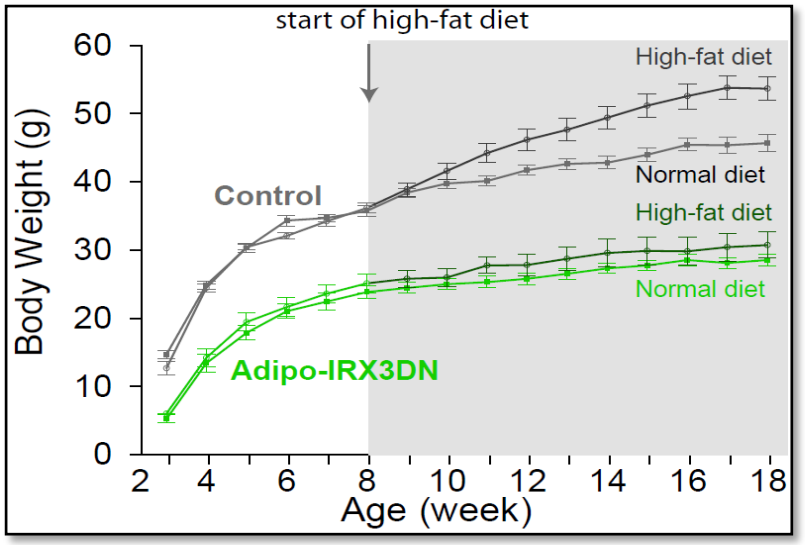
***Risk individuals*: IRX3/5 repression restores respiration,thermogenesis**

***Non-risk*: IRX3/5 overexpression disrupts respiration,thermogenesis**

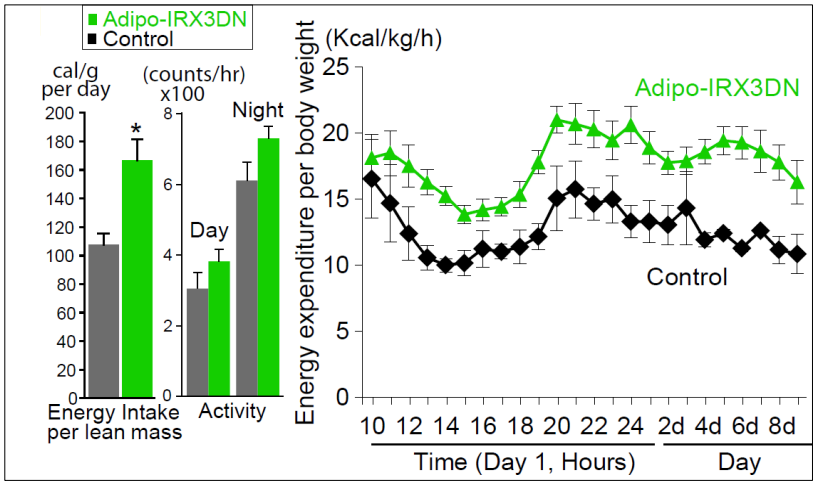
Irx3 adipose repression: anti-obesity phenotypes in mice



54% reduced body weight



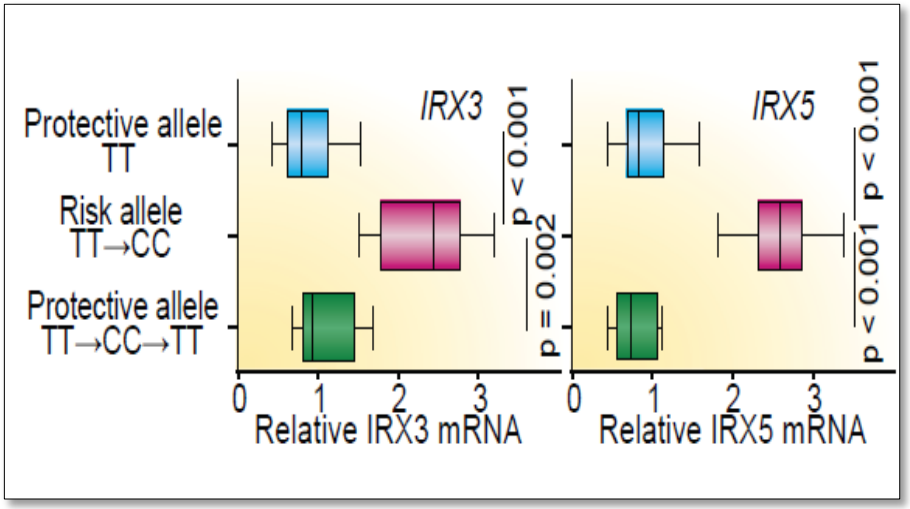
Resistance to high-fat diet



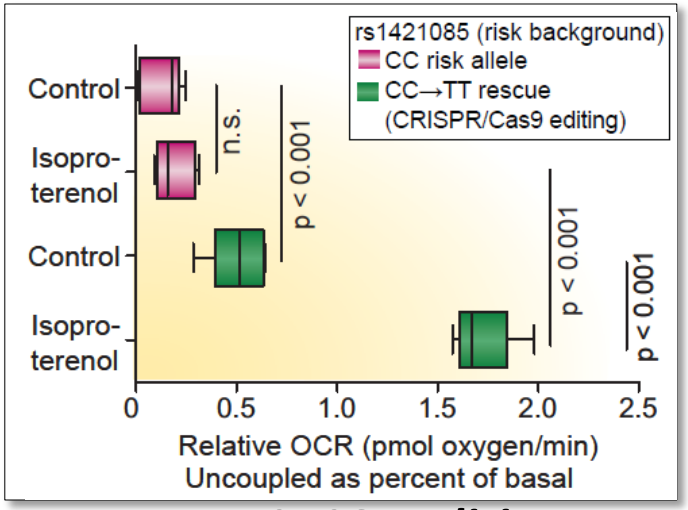
Increased energy dissipation

- *No reduction in appetite*
- *No increase in exercise*
- *In thermoneutral conditions*
- *Day and night (not exercise)*

Single-nucleotide editing reverses thermogenesis in humans



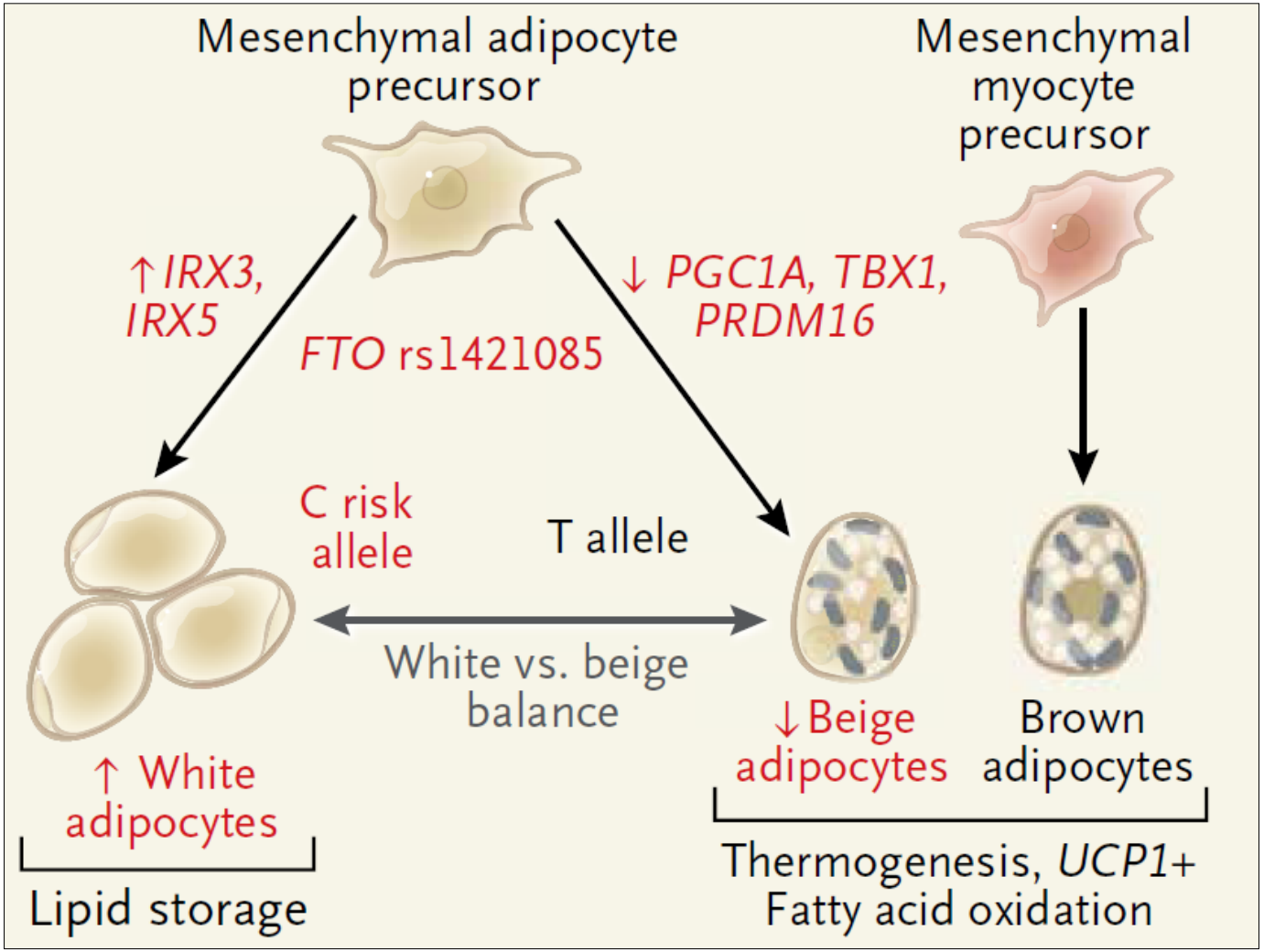
rs1421085 editing alters IRX3+IRX5 expression (500,000 and 1 million nucleotides away!)



rs1421085 editing restores thermogenesis

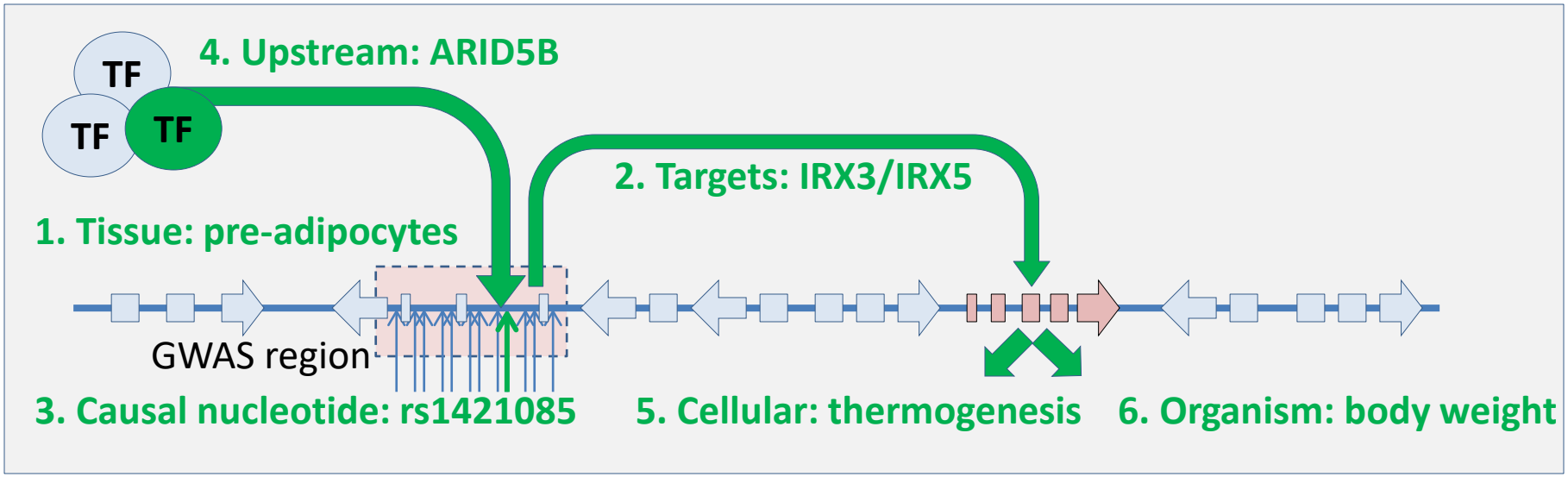
rs1421085 causality: C-to-T editing rescues IRX3/IRX5 expression, thermogenesis

Model: beige ⇔ white adipocyte development



Expand therapeutic focus from brain to adipocytes

FTO obesity locus as a model for dissecting GWAS



1. Establish relevant **tissue/cell type:** **pre-adipocytes**
2. Establish downstream **target** gene(s): **IRX3 and IRX5**
3. Establishing **causal** nucleotide variant: **rs1421085**
4. Establish upstream **regulator** causality: **ARID5B**
5. Establish **cellular** phenotypic consequences: **thermogenesis**
6. Establish **organismal** phenotypic consequences: **body weight**

Scaling up dissection efforts to 100s of variants

	LeadSNP	NumSNPs	NumExonS	MAF	pval	OddRatio	Study
	rs9268839	1	0	45%	1.E-250	2.28	RheumatoidArthritis_24390342
	rs1260326	8	1	39%	2.E-239	8.70	Cholesterol_24097068
	rs12143842	7	0	24%	1.E-213	3.50	QT_24952745
	rs1532085	9	0	40%	1.E-188	9.35	Cholesterol_24097068
	rs1367117	3	2	32%	1.E-182	8.40	Cholesterol_24097068
	rs629301	11	5	24%	2.E-170	7.46	Cholesterol_24097068
	rs2981579	11	0	40%	2.E-170	1.27	BreastCancer_23535729
	rs2476601	2	1	9%	9.E-170	1.80	RheumatoidArthritis_24390342
	rs11209026	27	1	7%	8.E-161	2.01	CrohnIBDUC_23128233
	rs12678919	84	5	13%	1.E-149	6.45	Cholesterol_24097068
	rs4420638	6	0	19%	1.E-149	5.08	Cholesterol_24097068
	rs6927022	1	1	47%	5.E-133	1.44	CrohnIBDUC_23128233
	rs3934467	27	0	22%	3.E-129	2.74	QT_24952745
#14	rs1558902	89	0	42%	5.E-120	2.56	BMI_20935630
	rs3803662	19	0	26%	2.E-114	1.24	BreastCancer_23535729
	rs7759938	31	0	32%	8.E-110	8.33	Menarche_25231870
	rs2954029	22	0	47%	1.E-107	13.16	Cholesterol_24097068
	rs11742570	53	0	40%	2.E-82	1.20	CrohnIBDUC_23128233
	rs2131925	254	9	34%	3.E-74	15.15	Cholesterol_24097068
	rs12916	19	2	40%	5.E-74	1.47	Cholesterol_24097068
	rs4299376	9	0	31%	3.E-73	12.66	Cholesterol_24097068
	rs12994997	72	7	48%	4.E-70	1.23	CrohnIBDUC_23128233
	rs10401969	10	2	9%	1.E-69	8.26	Cholesterol_24097068
	rs6426833	3	0	46%	2.E-68	1.27	CrohnIBDUC_23128233
	rs9533090	6	0	49%	5.E-68	10.00	BoneMineralDensity_22504420
	rs11153730	20	0	50%	2.E-67	1.65	QT_24952745
	rs10453225	81	0	32%	6.E-66	11.11	Menarche_25231870
	rs1883025	3	0	25%	2.E-65	14.29	Cholesterol_24097068
	rs614367	2	0	15%	2.E-63	1.21	BreastCancer_23535729
	rs1366594	5	0	46%	4.E-61	12.50	BoneMineralDensity_22504420
	rs16857031	1	0	13%	7.E-61	2.37	QT_24952745
	rs2153127	14	0	48%	6.E-59	12.50	Menarche_25231870

Top 895 SNPs

Avg:	26.8	1.2	29%	1.E-08	11.41
Median	14	0	30%	2.E-11	1.78
Stdev	35.7	2.8	13%	8.E-08	13.60

Across 11 well-powered association studies:

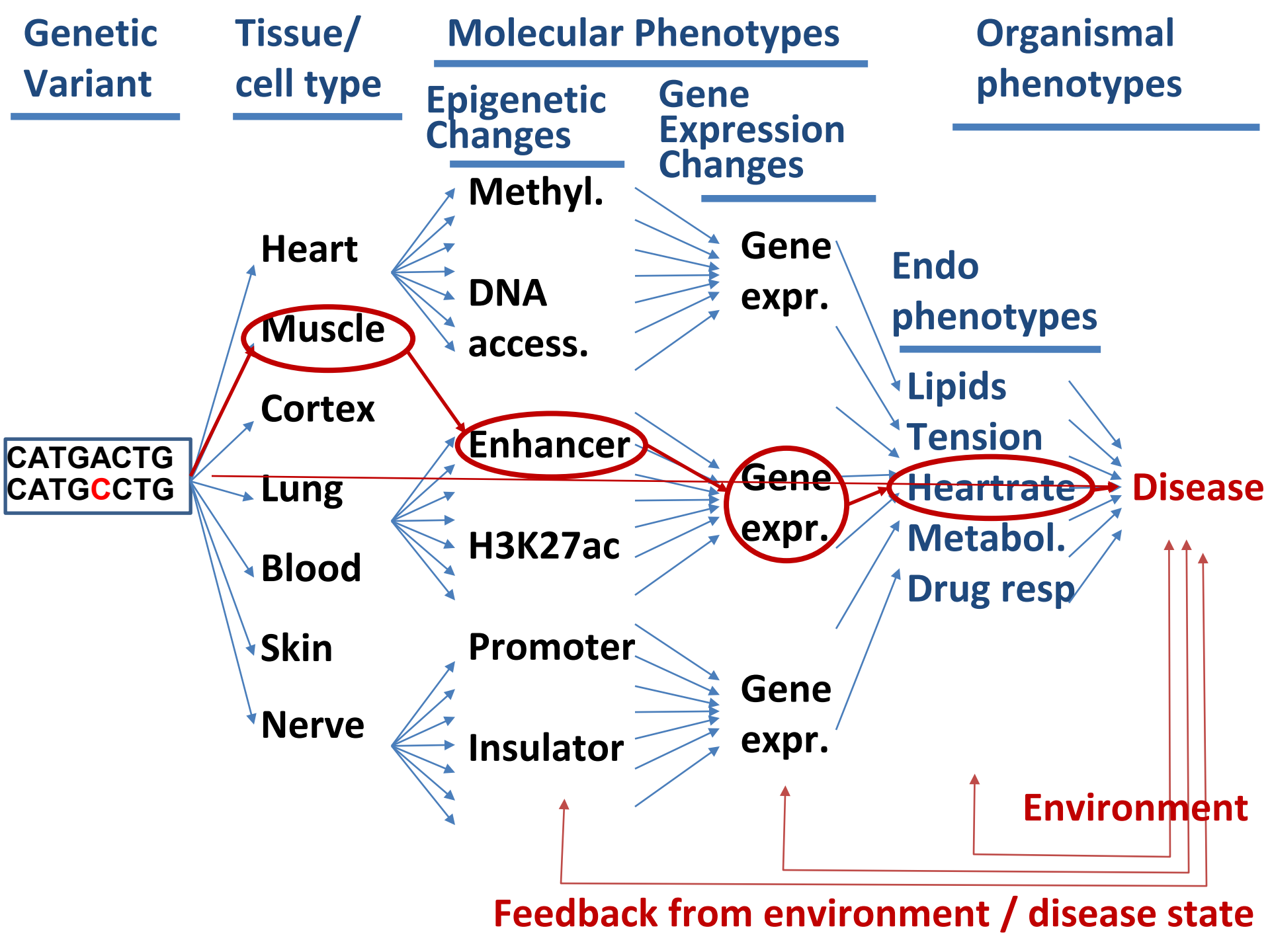
- BMI
- Bone Mineral Density
- Bipolar
- BreastCancer
- Cholesterol
- CrohnIBDUC
- Height
- Menarche
- QT
- RheumatoidArthritis
- Schizophrenia

895 associated loci

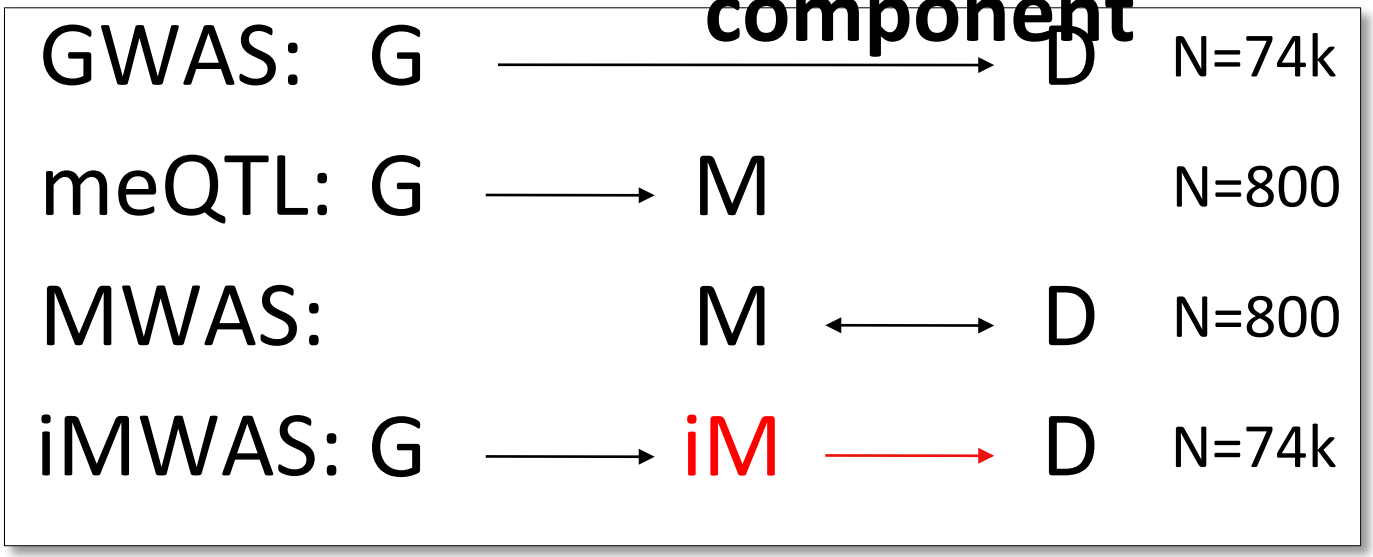
572 (64%) have no protein-coding variants

Genetics, Variation, GWAS, PRS, Mechanism

1. Genetics, Variation, GWAS
2. Polygenic scores (PGS)
3. From GWAS to biological insights
4. From Region to Mechanism / Circuitry
5. Quantitative Trait Loci: eQTL/meQTL analysis
6. Mediation Analysis + Mendelian Randomization
7. Heritability and Systems Genetics


Genetic Variant
Tissue/ cell type
Molecular Phenotypes
Epigenetic Changes
Gene Expression Changes
Organismal phenotypes
CATGACTG
CATGCCTG
Heart
Muscle
Cortex
Lung
Blood
Skin
Nerve
Methyl.
DNA access.
Enhancer
H3K27ac
Promoter
Insulator
Gene expr.
Gene expr.
Gene expr.
Endo phenotypes
Lipids
Tension
Heartrate
Metabol.
Drug resp
Disease
Environment
Feedback from environment / disease state

Imputed MWAS: increased power, genetic component



Learn G→D directly (complex phenotype)

Learn G→M (simpler phenotype)

M□D (no causality)

Apply G→M to get iM
iM→D (causality)

Key Idea:

- Learn G→M model (ROSMAP n=800) Fewer indiv. Simpler phenotype
- Impute methylation iM for GWAS cohort (n=74k)
- iMWAS between genotype-driven M and AD phenotype (n=47k)

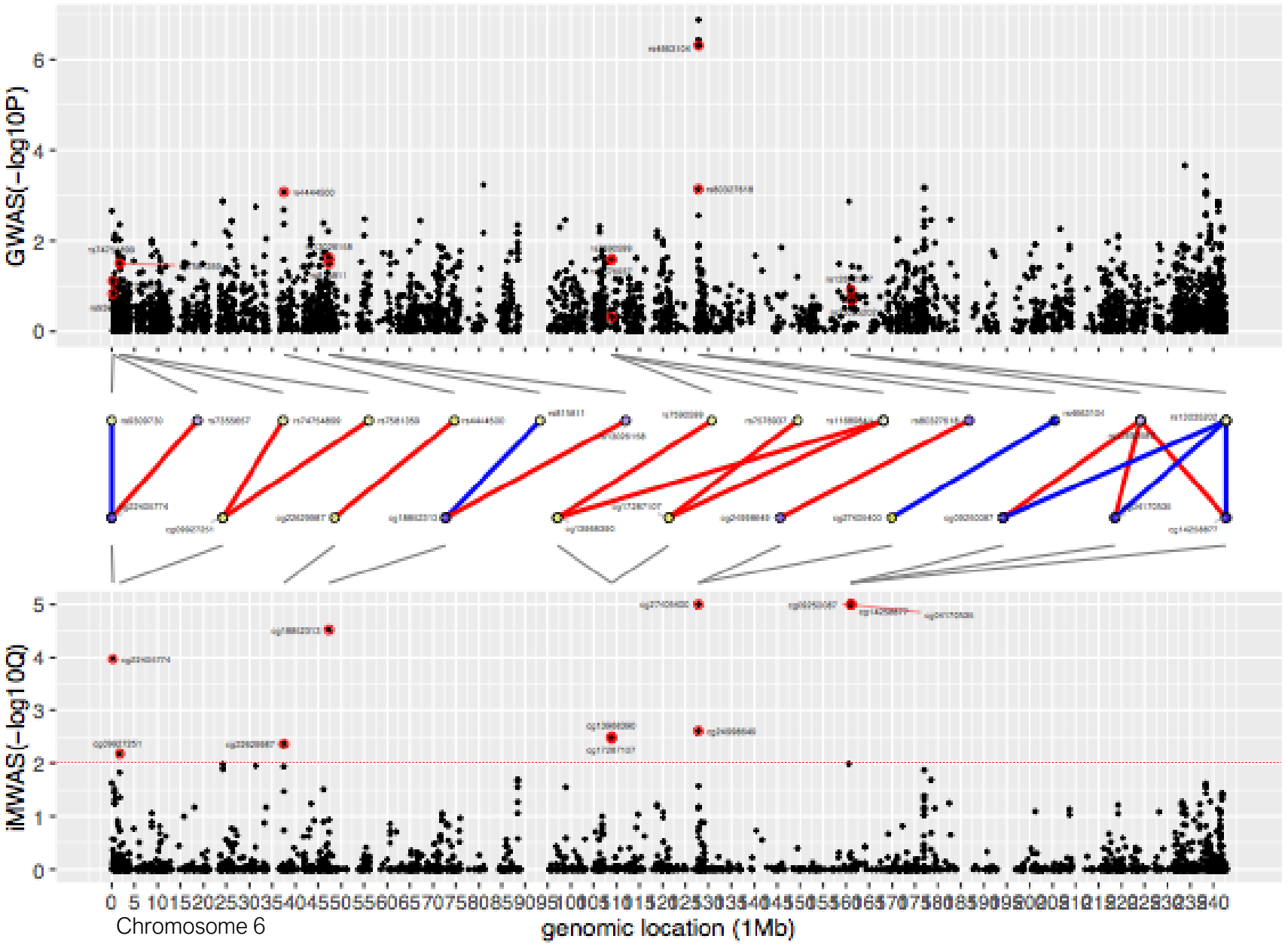
Advantage:

- Much larger GWAS cohorts (>>MWAS): increased power
- Genetic component of methyl. variation

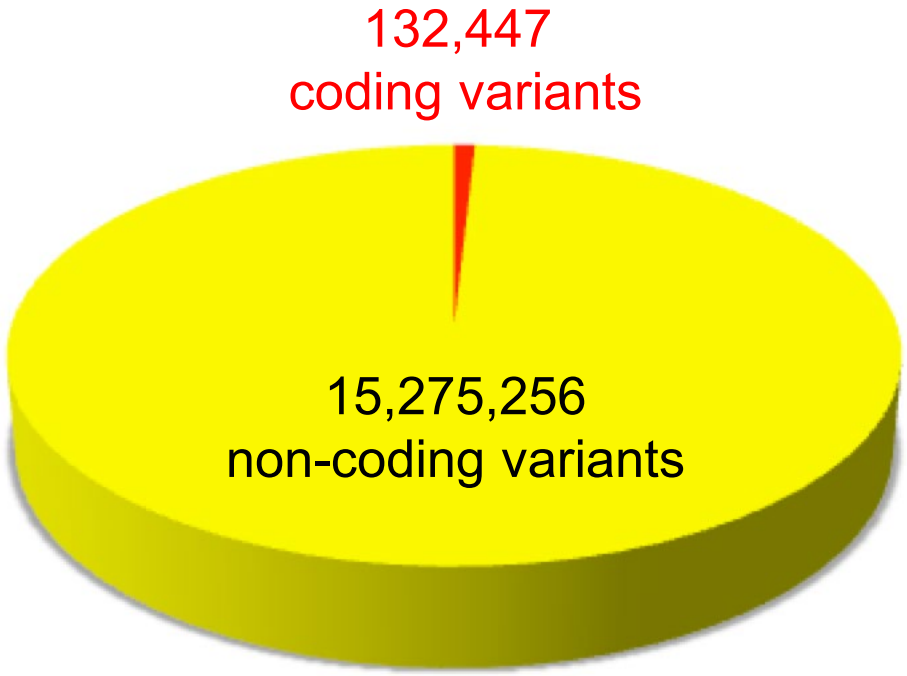
Logistical challenge:

- Summary stats, not full genotypes ❾inear model, impute stats direct

iMWAS results: new loci, multiple contributing SNPs



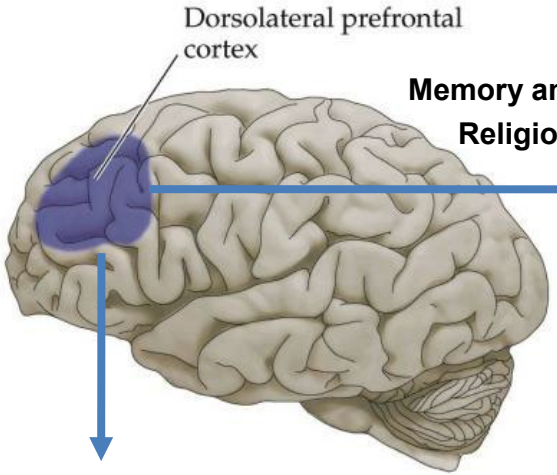
Motivations for eQTL mapping studies: Most genetic variation is non-coding



1000 Genomes Data from initial 179 individuals sequenced

1000 Genomes Project Consortium. Nature. 2010; 467:1061-73.

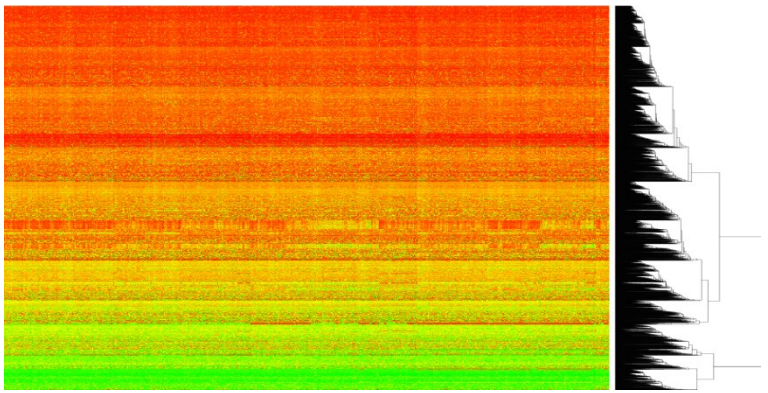
Methylation in 750 Alzheimer patients/controls

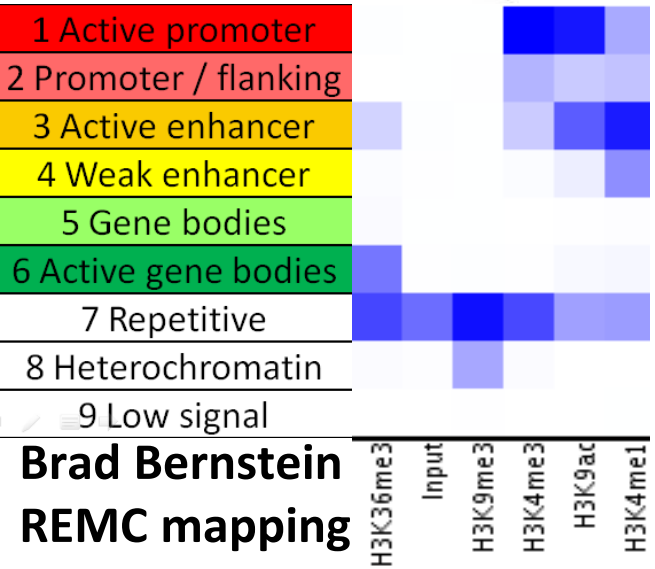


Memory and Aging Project
Religious Order Study

750 individuals (~50% w/AD)

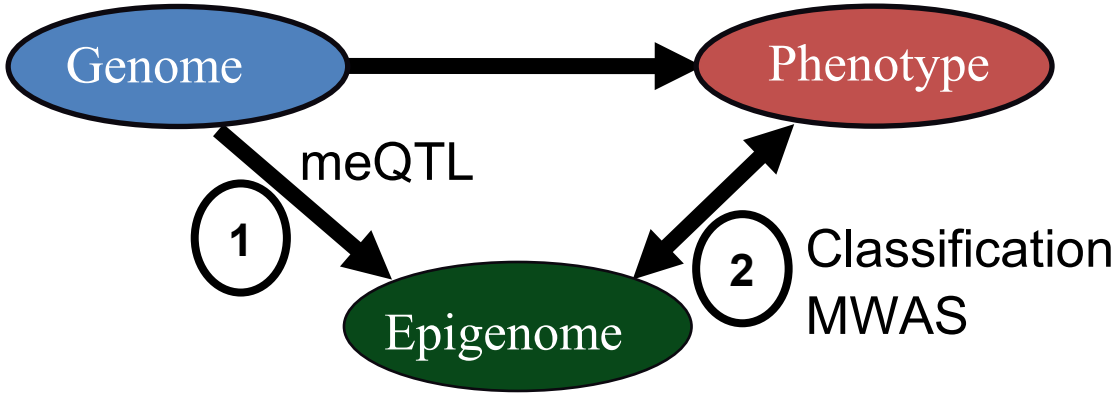
486,000 methylation probes





Brad Bernstein
REMC mapping

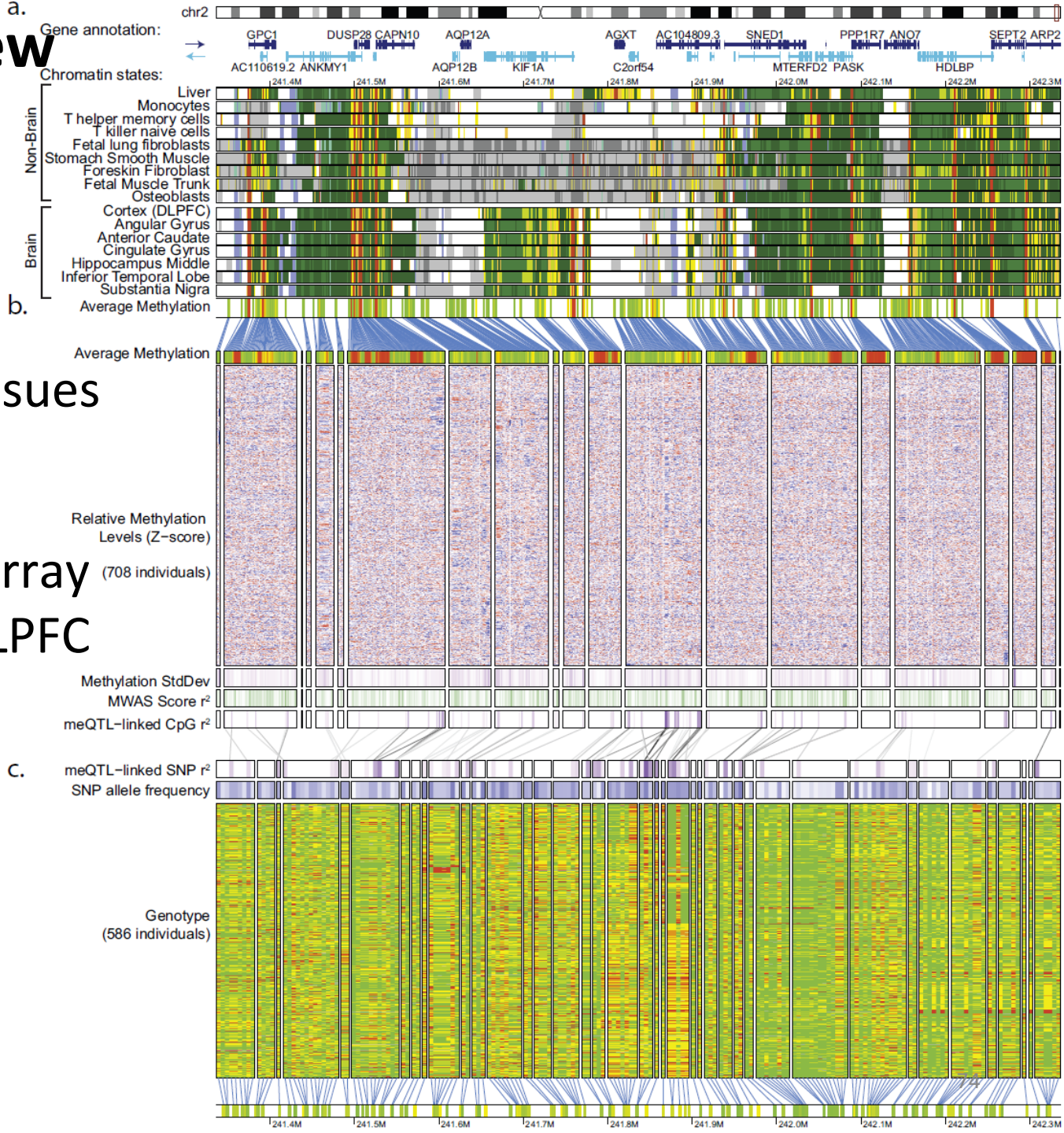
Philip deJager, Epigenomics Roadmap



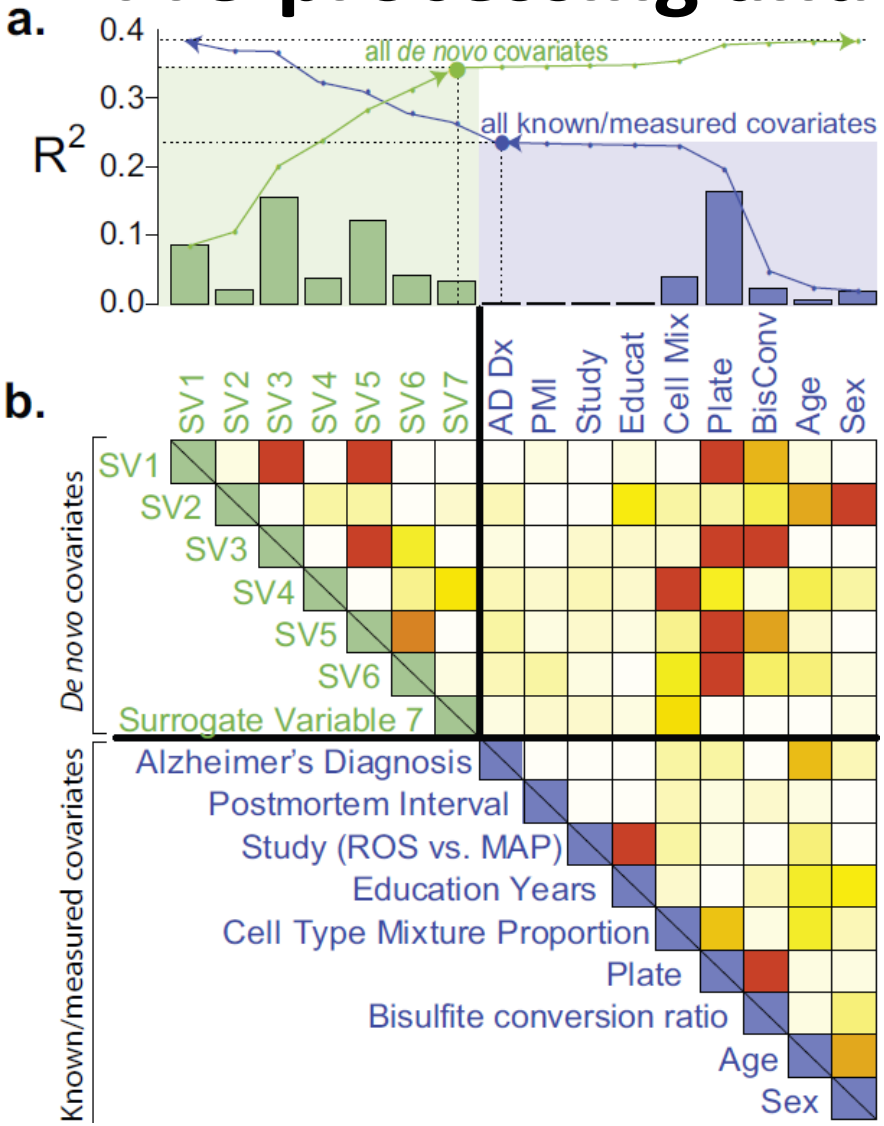
- Patients followed for 10+ years with cognitive evaluations
- Brain samples donated post-mortem methylation/genotype
- Seek predictive features: SNPs, QTLs, mQTLs, regulation

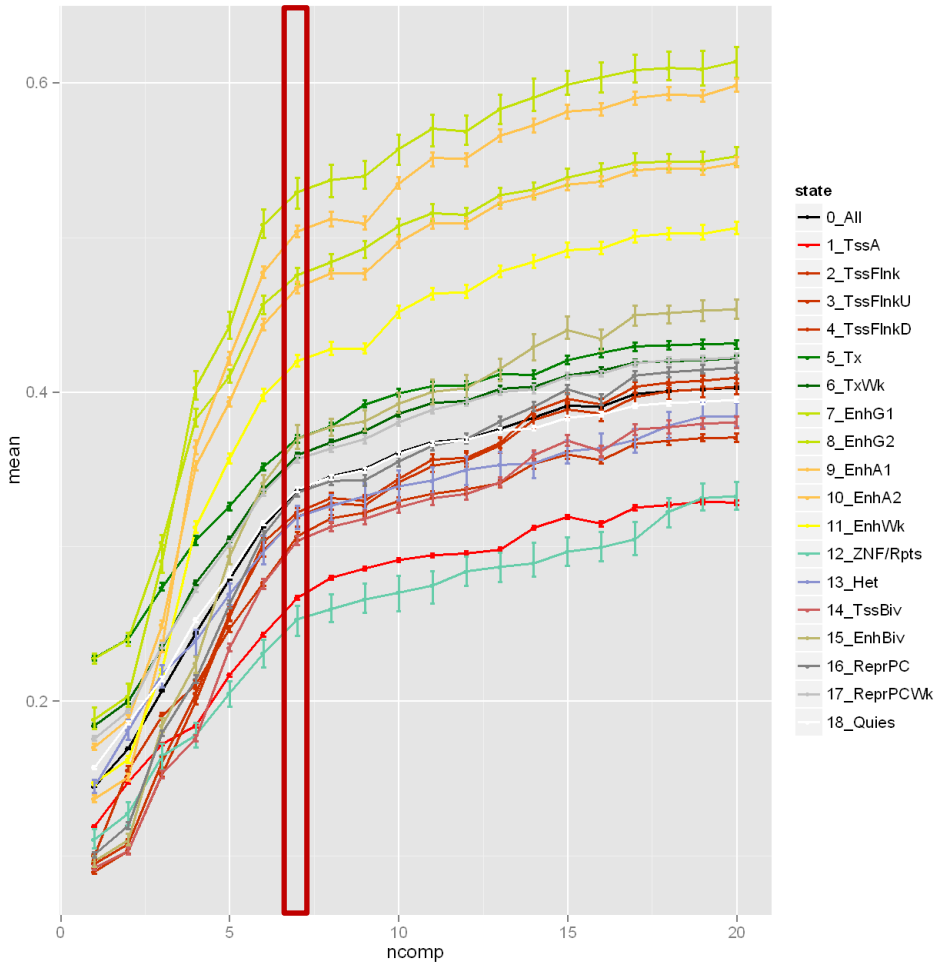
Dataset overview

- Chromatin state
  - 18 states
  - 6 marks
  - DLPFC
  - Joint w/ 127 tissues
- Methylation level
  - 450k Illumina array
  - Brain Cortex DLPFC
  - 708 individuals
- Genotype
  - 620k SNPs
  - 586 individuals
  - Blood



Pre-processing and covariate elimination





- Eliminate 7 *de novo* co-variates, and 8 known co-variates
- Correlate with Plate, Cell Mixture, Conversion, Sex, age

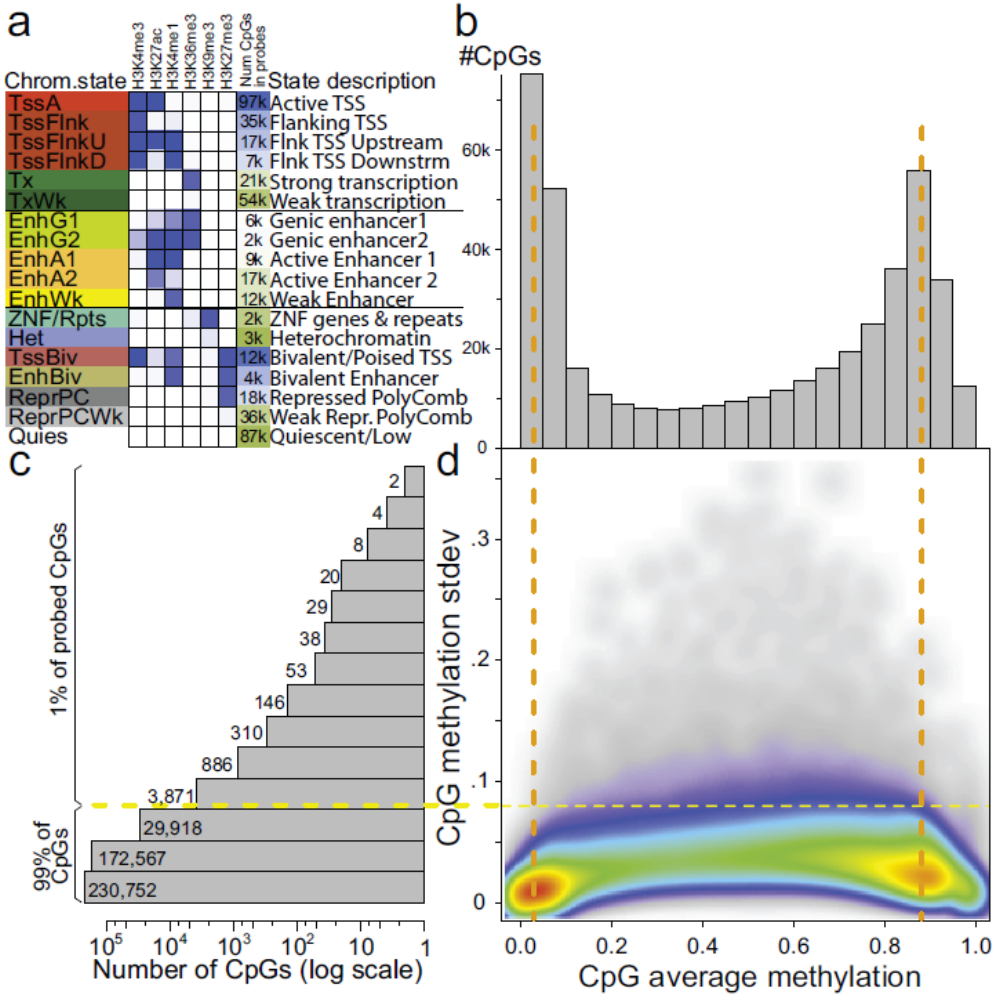
Most methylation probes are high or low, with little variability

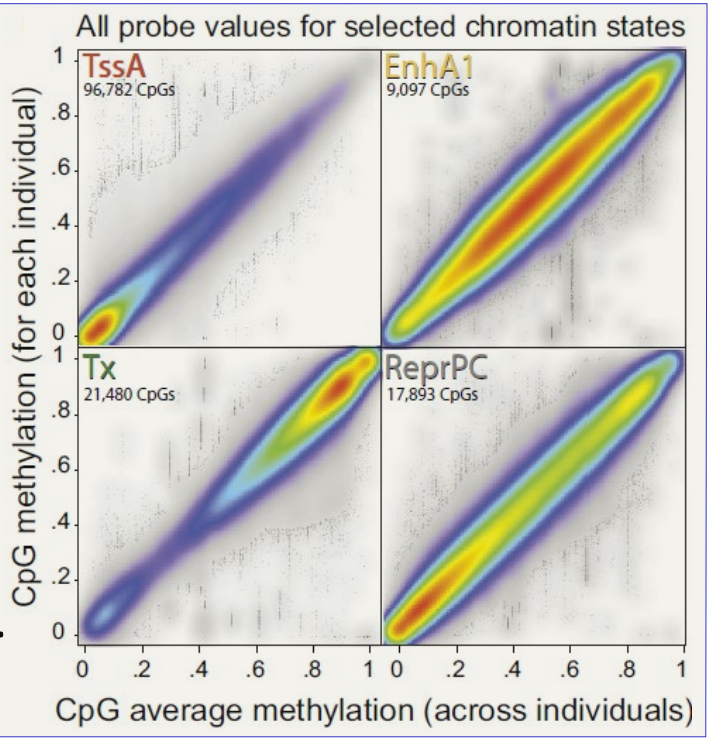
a. Chromatin state definitions

b. Distribution of CpG **avg** methylation levels (in Illumina 450k array)
   - Average methylation across 708 individuals

c. Distribution of CpG methylation **variance** across individuals
   - Log: **Very few probes** show high variance

d. 2D distribution: average vs. variance
   - Highest variance intermediate-methylation

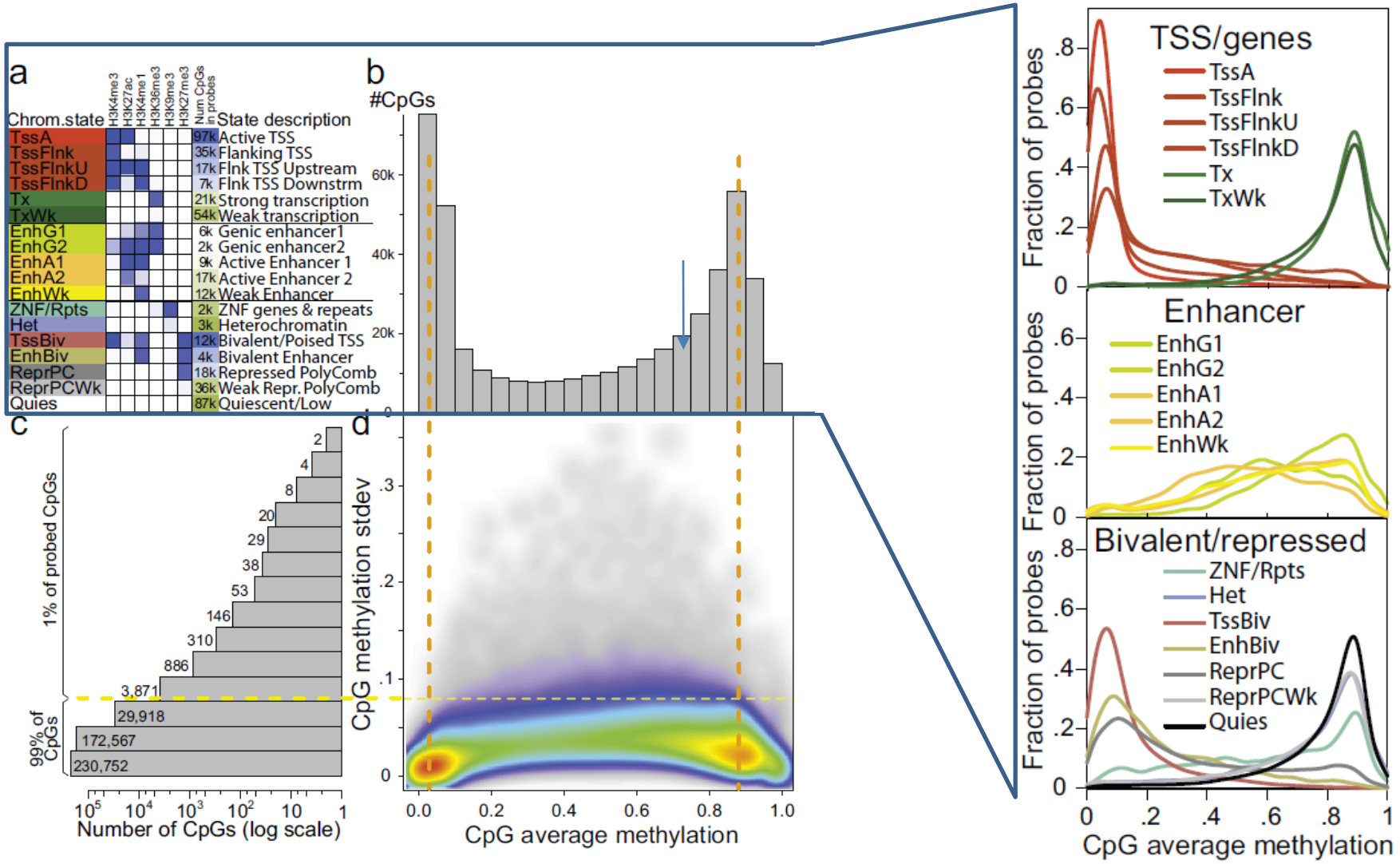




- However: Intermediate methylation is not just an artifact of averaging bimodal levels between individ.

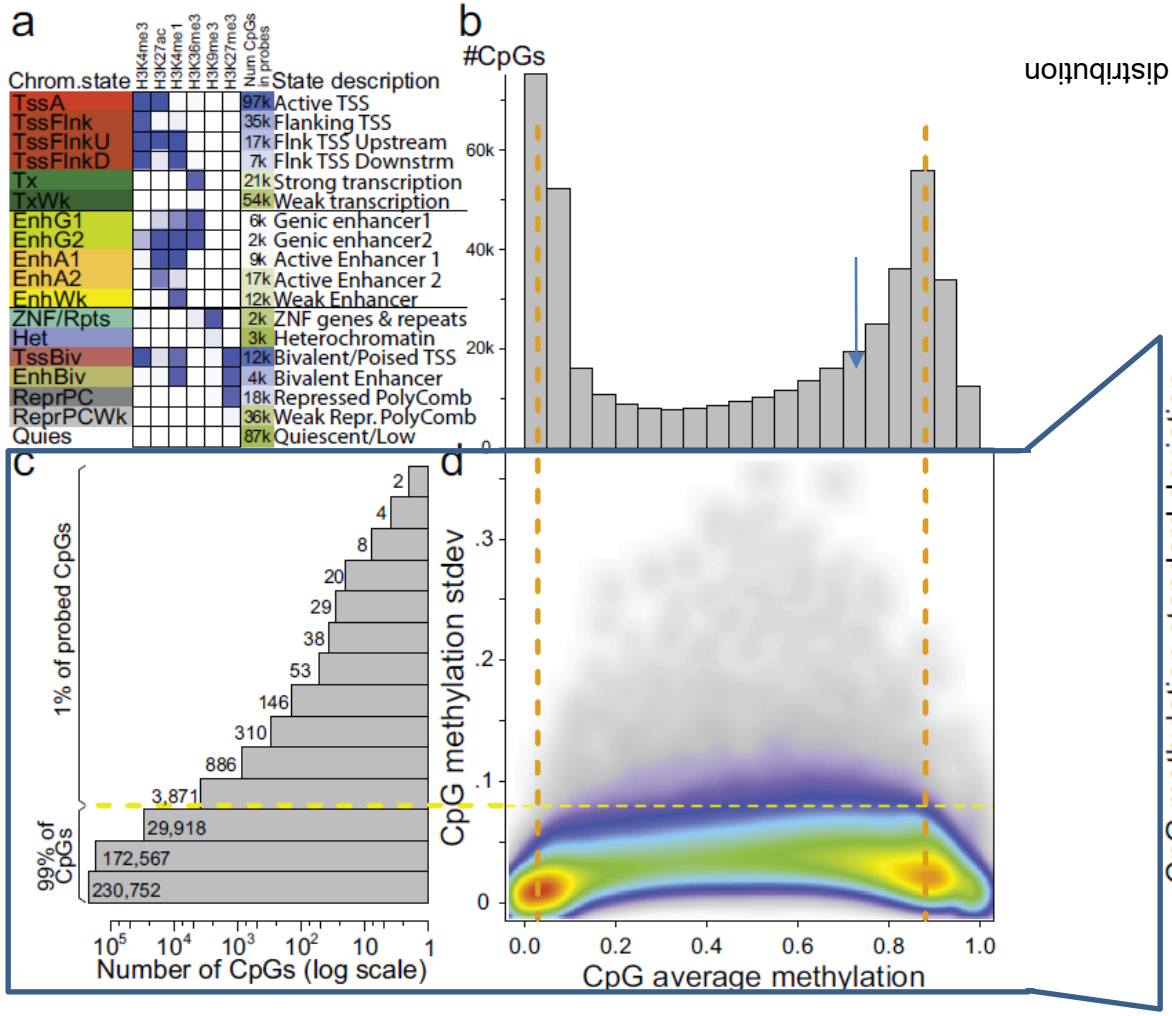
➢ Intermediate methylation is truly intermediate

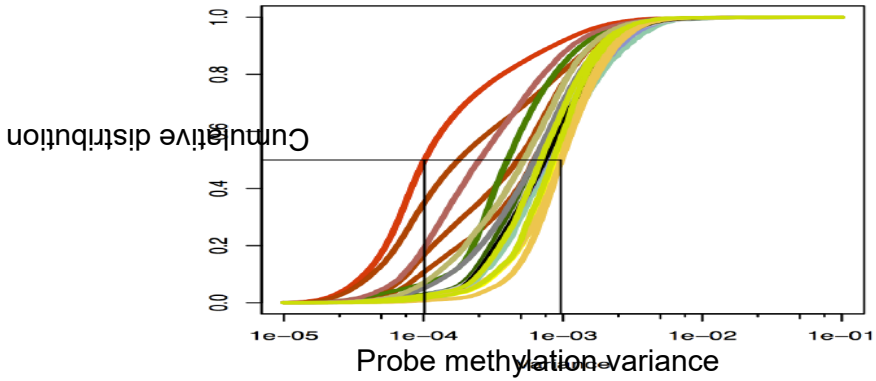
Enhancer regions show intermediate methylation

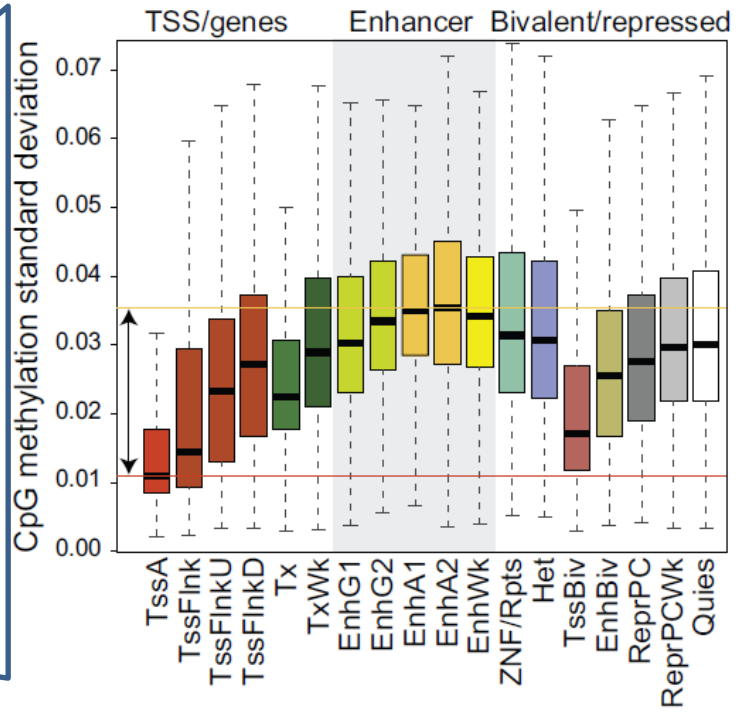


- Enhancer states: Intermediate (EnhG1/G1/A1/A2/Wk)
- Active states: Promoters: low. Tx: high.
- Repressed states: TssBiv/EnhBiv/ReprPC: low. Quies/ReprPCWk: high

Enhancers are most variable, promoters least

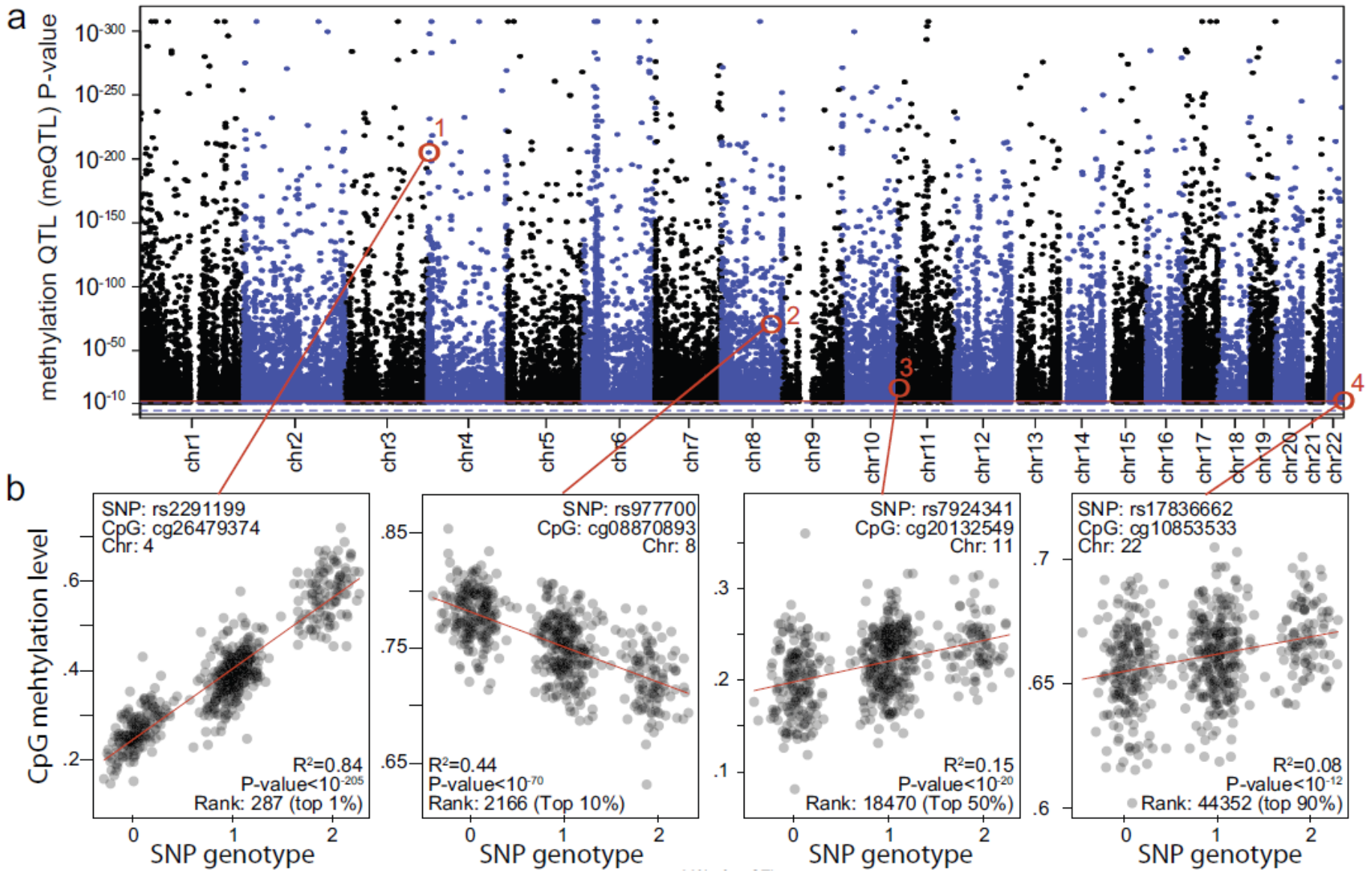






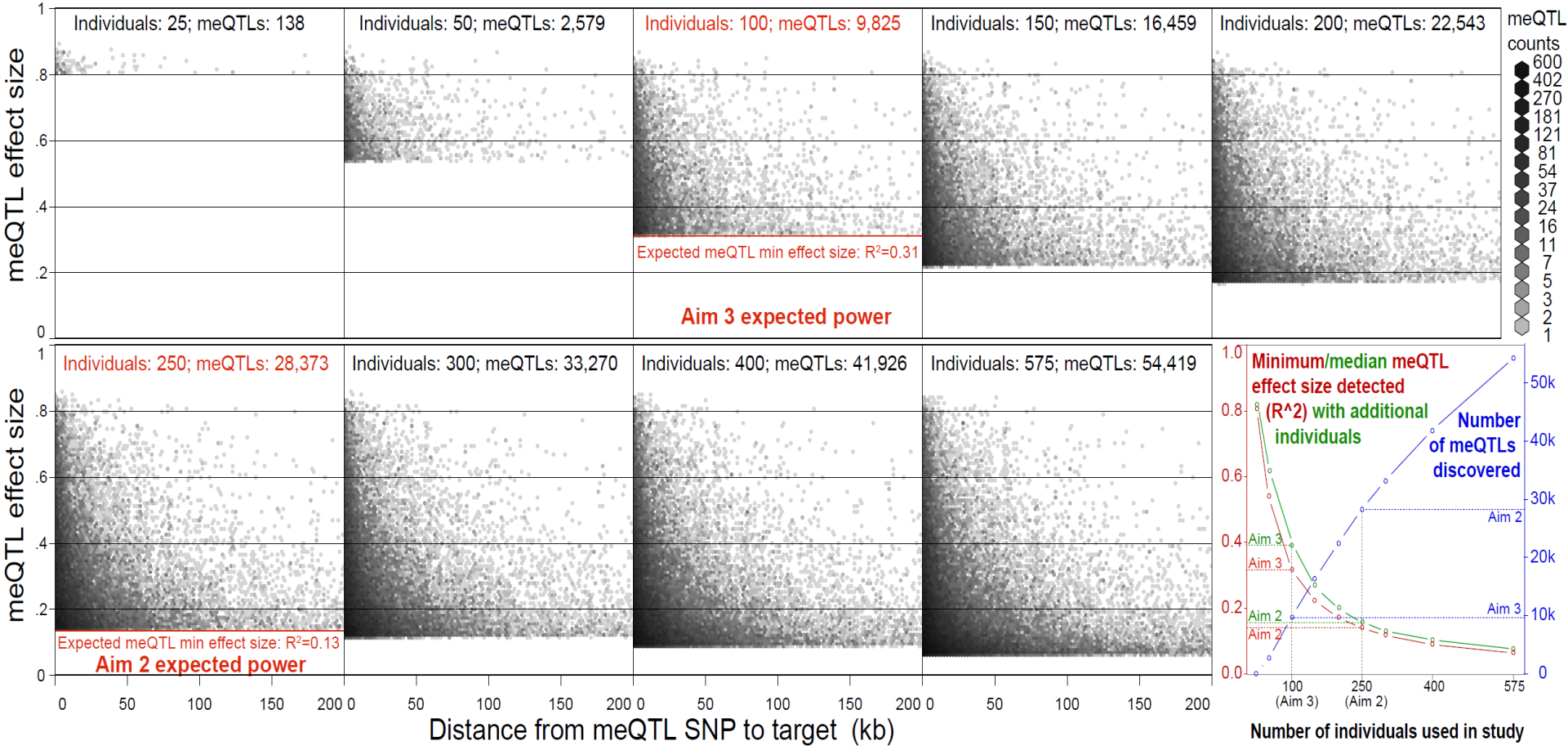
- Chromatin states vary 10-fold in methylation variance, 3-fold in stdev
- Active states: EnhA > EnhWk > EnhG > TxWk > TssFlnk >> TssA
- Repressed states: Quies > ReprPC > EnhBiv >> TssBiv

Discover 50,000 methylation QTLs after Bonferroni



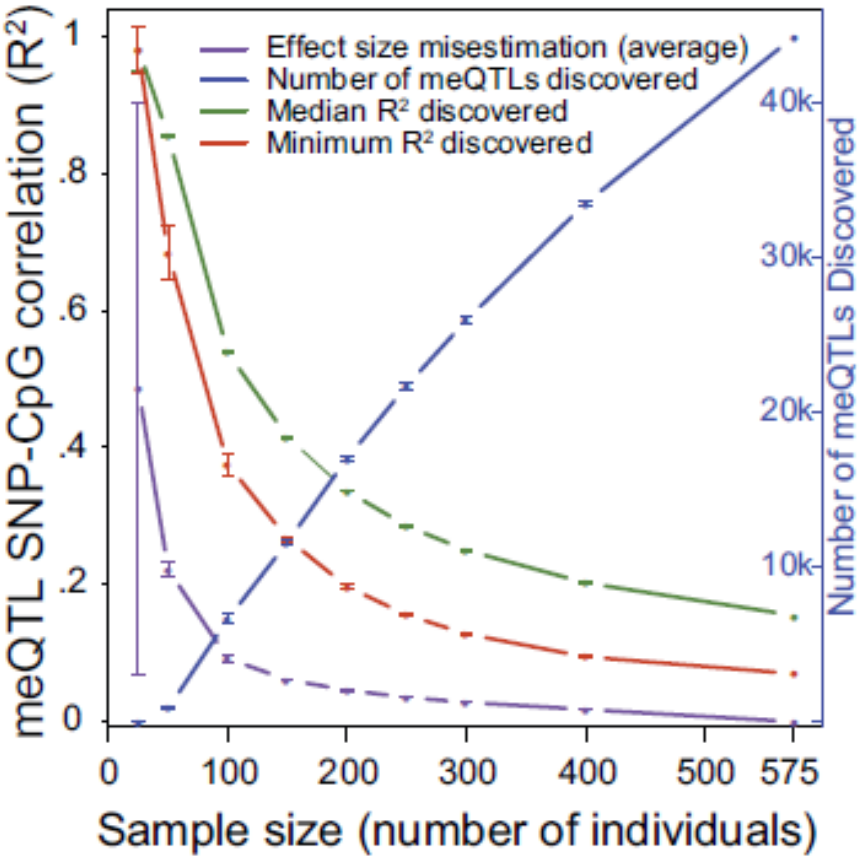
- Overlay meQTL discovery plot

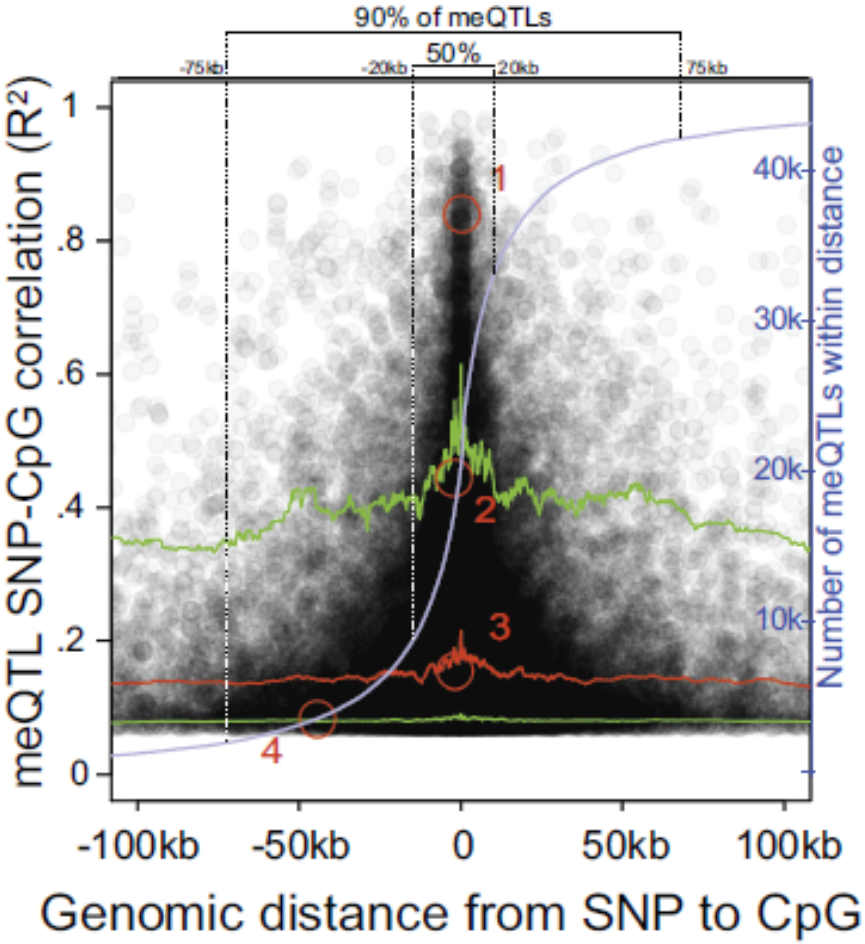
meQTL discovery vs. distance vs. cohort size



- Vary: (1) distance from CpG; (2) effect size; (3) cohort size
- Strongest effects within 20 kb of tested CpGs
- Expectation for 100, 150, 200 individuals (if searching a 1Mb region)

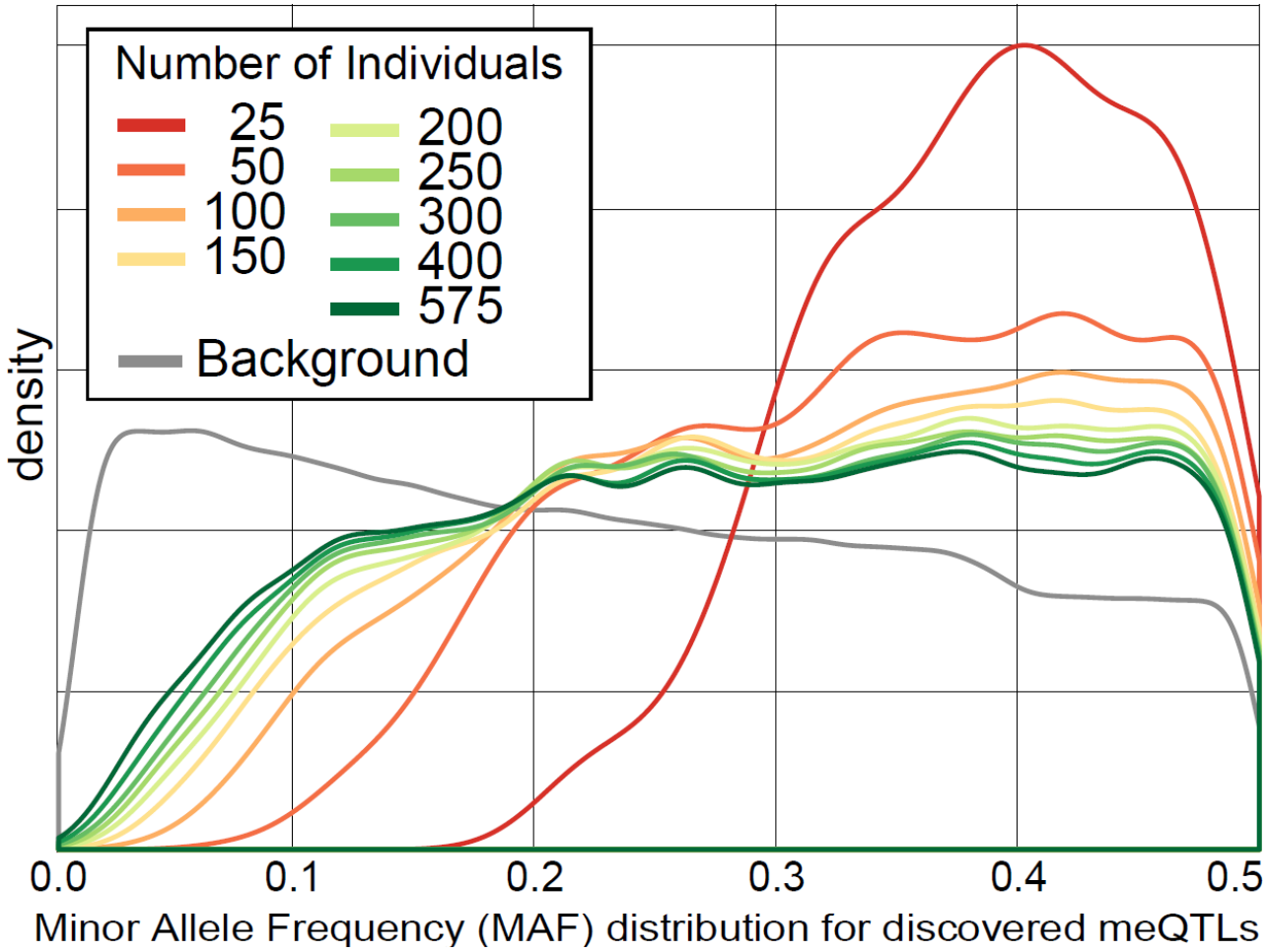
Selection of the number of individuals





- More individuals ❾ linearly more meQTLs, but smaller effect size
- Strongest effects concentrated within 20 kb of tested CpGs ❾ can be used to increase power for smaller sample sizes.

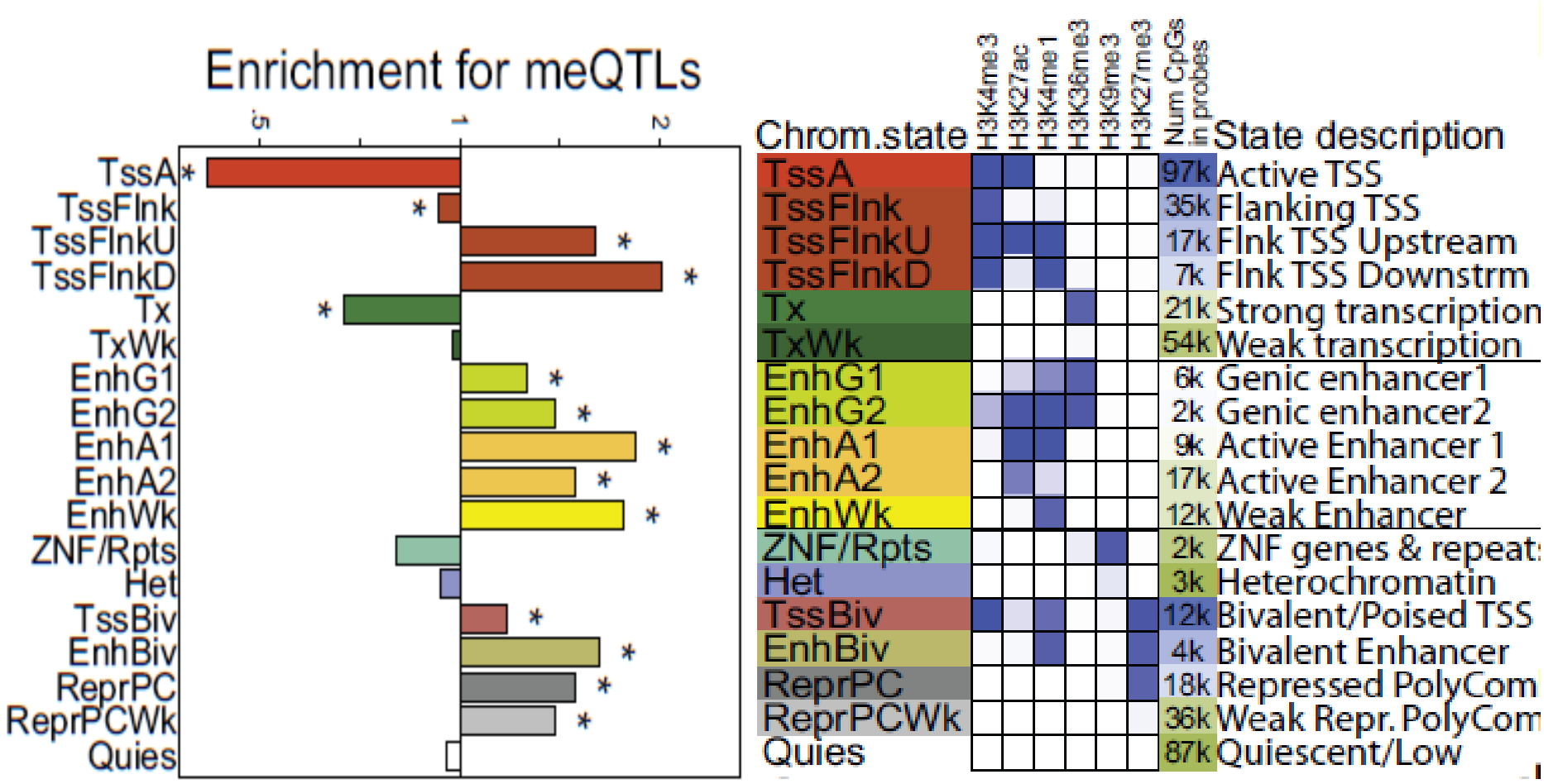
of individuals ⇔MAF of meQTL SNPs



Minor Allele Frequency (MAF) of discovered meQTL SNPs. Discovery power is greater for high-MAF SNPs, resulting in skewed distributions. Thus, we expect the majority of meQTLs to have both alleles represented in samples of 20 individuals (40 chromosomes). For

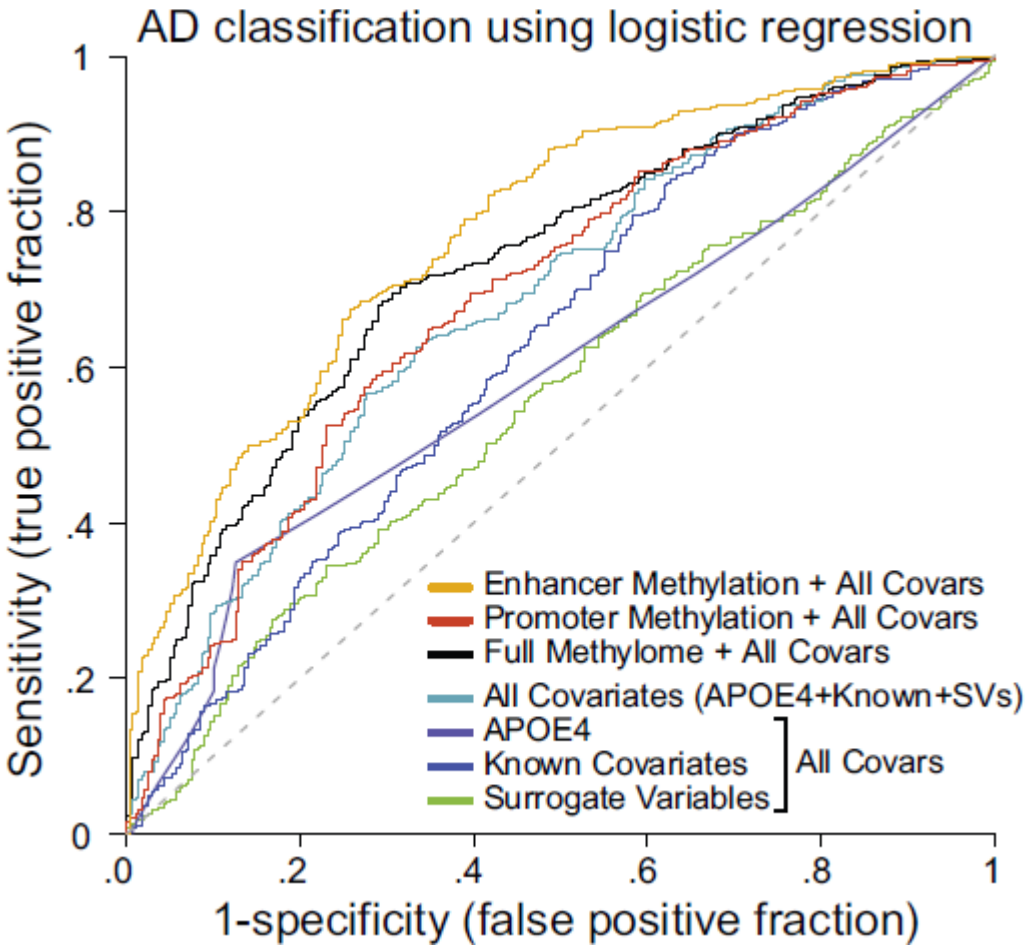
- Focusing on 100-150 individuals, MAF > 0.1, as expected
- Large number of SNPs never probed even with 600 indiv

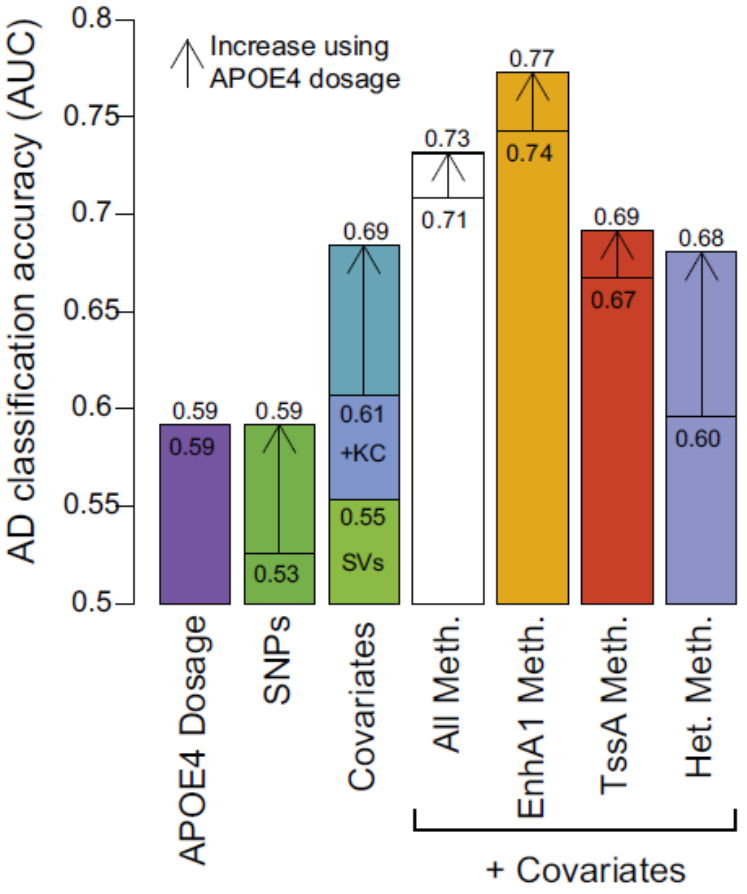
meQTL probes are enriched in enhancers + TssFlnk



- Prioritize EnhA, EnhWk, TssFlnk regions for meQTLs
- Profile variation in H3K27ac directly (ChIP-seq component)

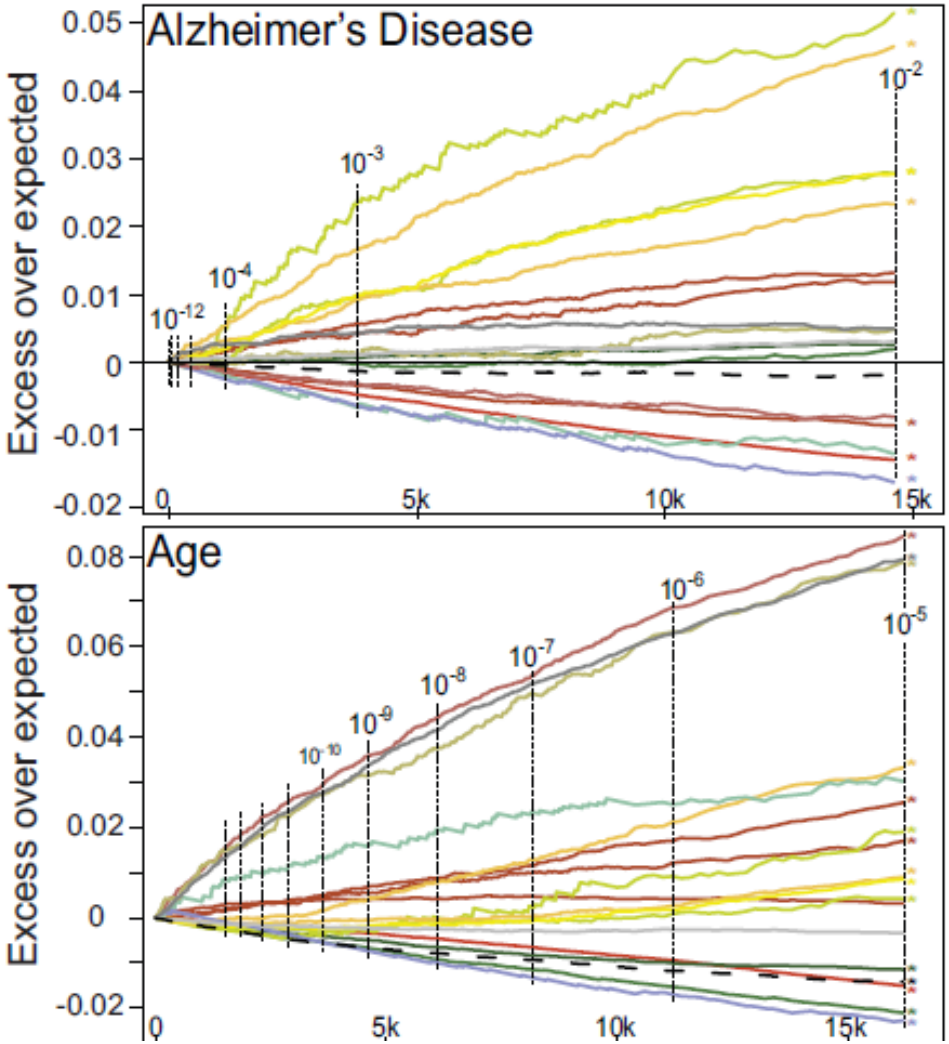
Enhancer variation correlated with AD diagnosis

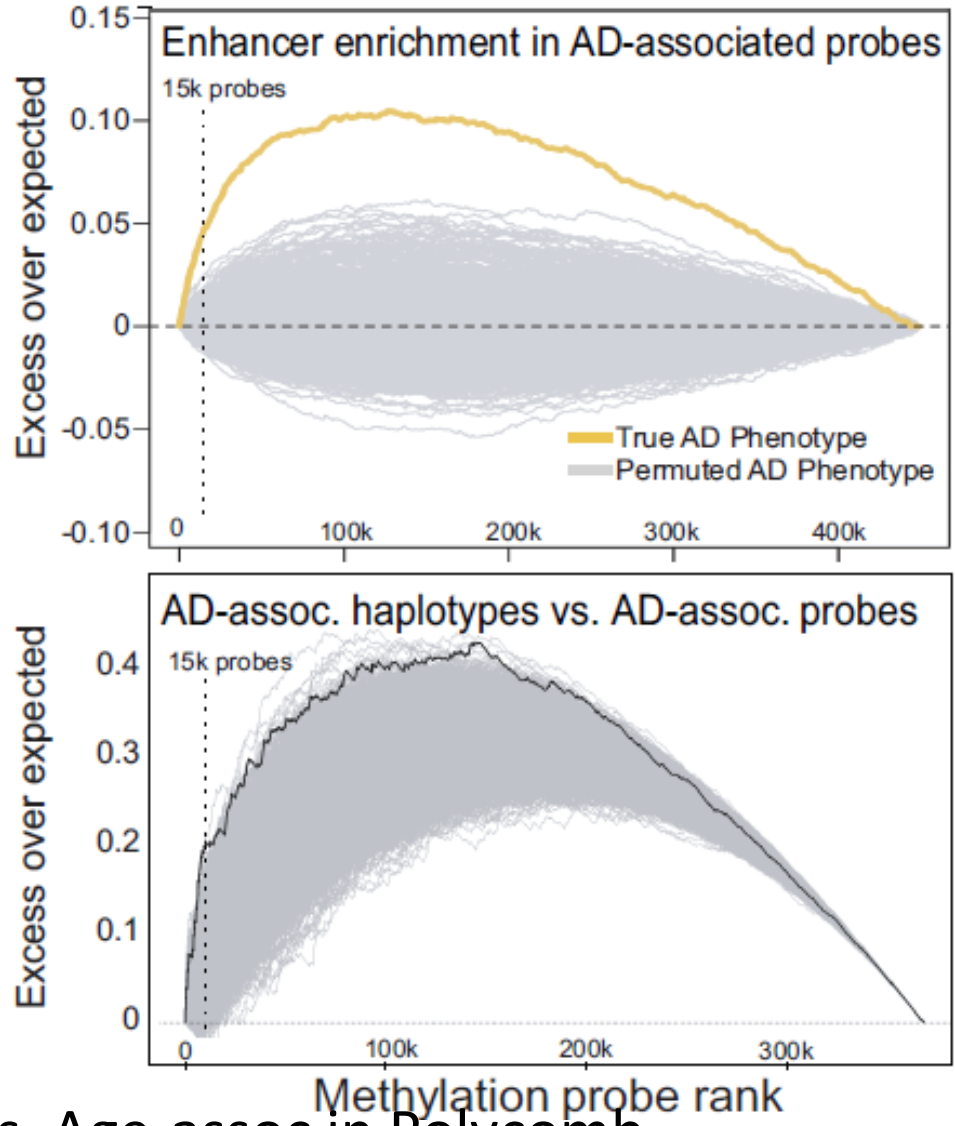




- Enhancer variation is actually biologically meaningful (not just an artifact of meaningless variation)
- Enhancers > all methylation > Promoters > APOE4 >> SNPs

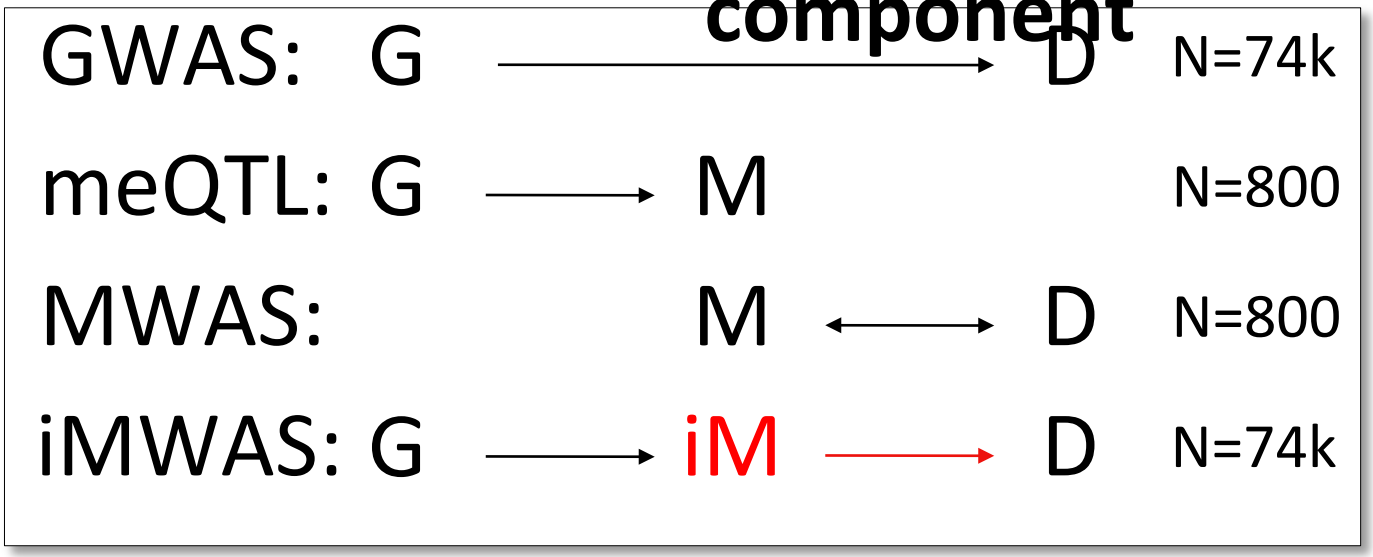
Functional enrichments persist across 1000 probes





- AD-associated probes in enhancers. Age-assoc in Polycomb
- 10,000 phenotype permutations ❾ Statistical significance
- AD top 1k GWAS enrichment persists across 100k+ probes

Imputed MWAS: increased power, genetic component



Learn G→D directly (complex phenotype)

Learn G→M (simpler phenotype)

M□D (no causality)

Apply G→M to get iM
iM→D (causality)

Key Idea:

- Learn G→M model (ROSMAP n=800) Fewer indiv. Simpler phenotype
- Impute methylation iM for GWAS cohort (n=74k)
- iMWAS between genotype-driven M and AD phenotype (n=47k)

Advantage:

- Much larger GWAS cohorts (>>MWAS): increased power
- Genetic component of methyl. variation

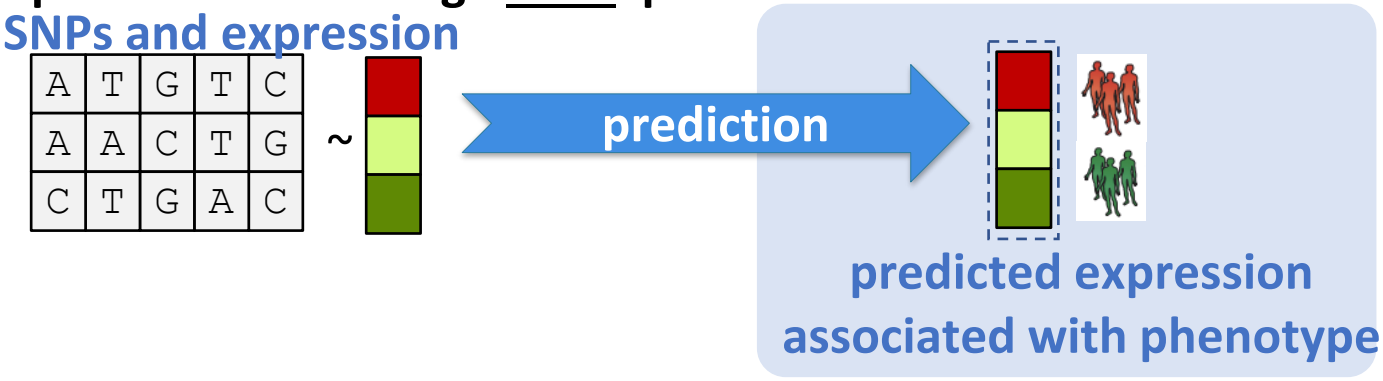
Logistical challenge:

- Summary stats, not full genotypes ❾inear model, impute stats direct

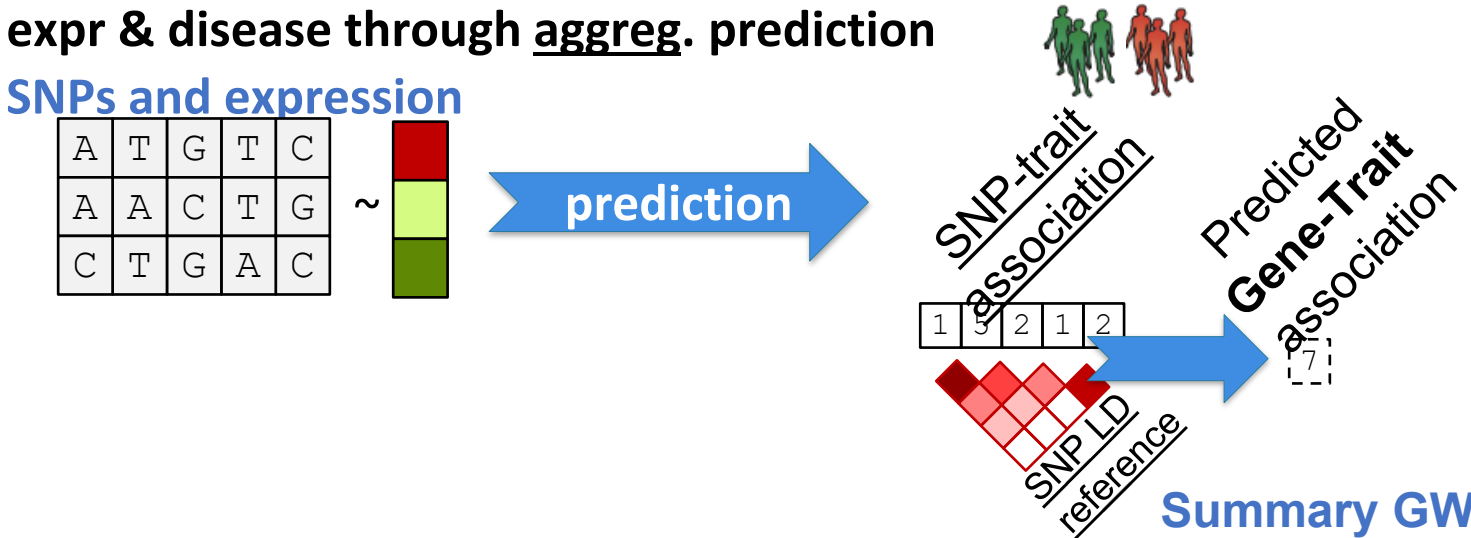
Multiple ways to integrate eQTLs w/ GWAS

Integrate expr & disease through measurement

Integrate expr & disease through indiv. prediction



Integrate expr & disease through aggreg. prediction



Genetics, Variation, GWAS, PRS, Mechanism

1. Genetics, Variation, GWAS
2. Polygenic scores (PGS)
3. From GWAS to biological insights
4. From Region to Mechanism / Circuitry
5. Quantitative Trait Loci: eQTL/meQTL analysis
6. Mediation Analysis + Mendelian Randomization
7. Heritability and Systems Genetics

Mendelian Randomization

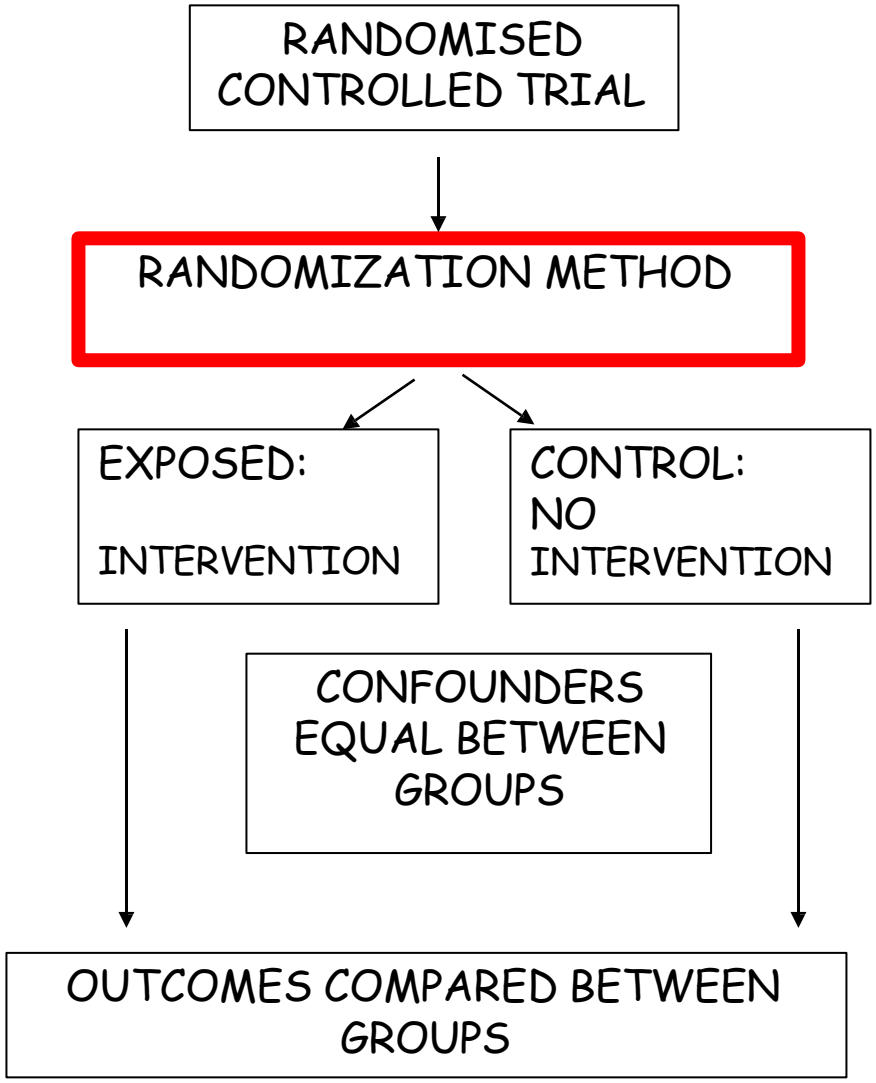
Mediation + Causality inference

Mendelian Randomization

- Problems with observational data
- Randomized controlled trials
- Mendelian Randomization (MR):
  - How it works
  - Core assumptions
  - Calculating causal effect estimates
- MR example
- Limitations of MR

RCTs: the Gold Standard in Inferring Causality

Randomization makes causal inference possible



The Need for Observational Studies

- **Randomized Controlled Trials (RCTs):**
  - Not always ethical or practically feasible eg anything toxic
  - Expensive, requires experimentation in humans
  - Impractical for long follow up times
  - Should only be conducted on interventions that show very strong observational evidence in humans

- **Observational studies:**
  - Association between environmental exposures and disease measured in observational designs (non-experimental)
    eg case-control studies or cohort studies
  - Reliably assigning causality in these types of studies is ***very limited***

The Wide Applicability of MR

- **Traditional Observational Epidemiological Studies**
- **Behavior Genetics and the Social Sciences**
- **Molecular Studies**
- **Pharmacogenomics**

Mendelian Randomization

- Problems with observational data
- Randomized controlled trials
- Mendelian Randomization (MR):
  - How it works
  - Core assumptions
  - Calculating causal effect estimates
- MR example
- Limitations of MR

How does Mendelian randomization work?

What does MR do?

- **Assess causal relationship between two variables**
- **Estimate magnitude of causal effect**

How does it do this?

By harnessing Mendel's laws of inheritance

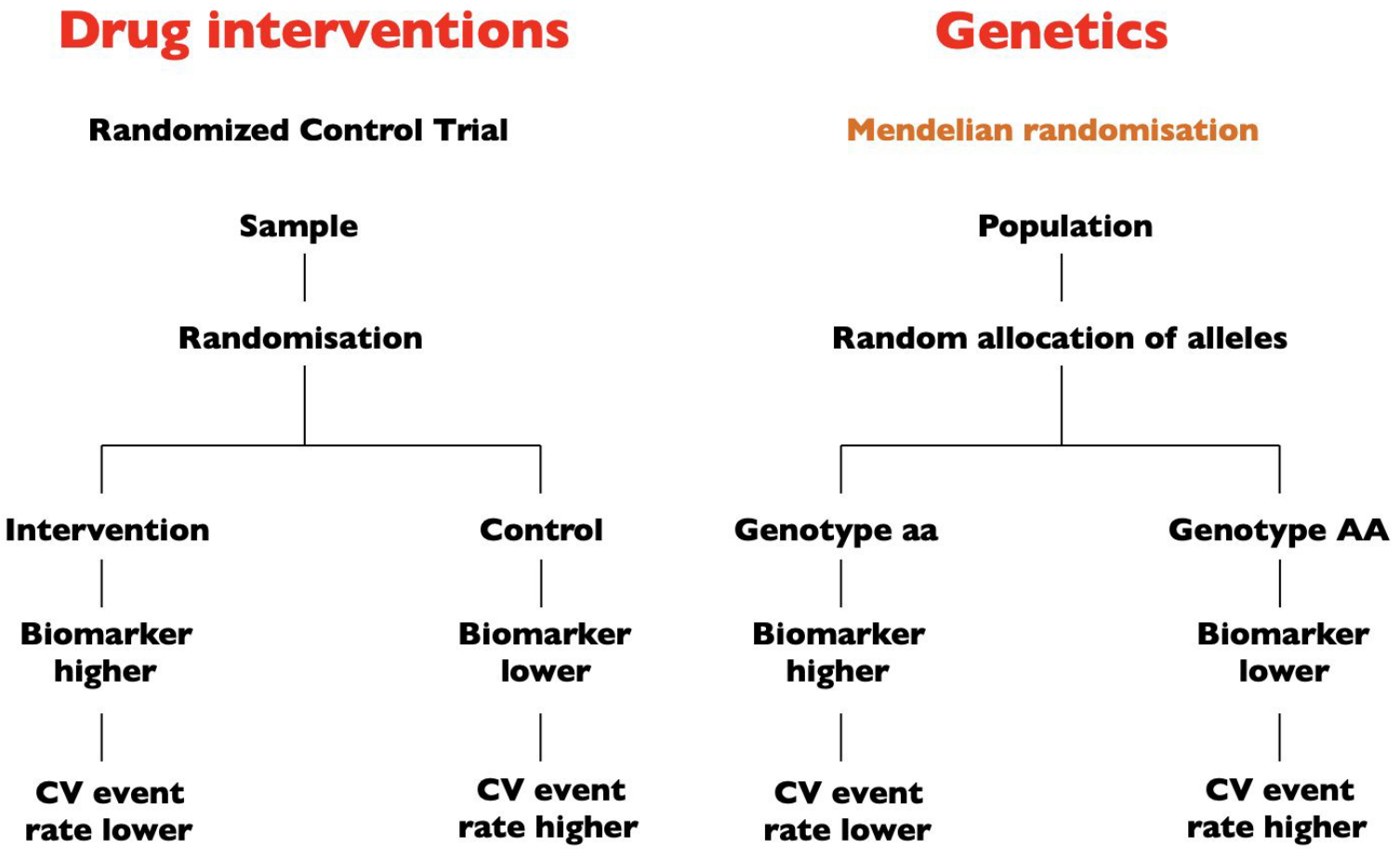
Mendel's Laws of Inheritance

Mendel in 1862

1. **Segregation:** alleles separate at meiosis and a randomly selected allele is transmitted to offspring

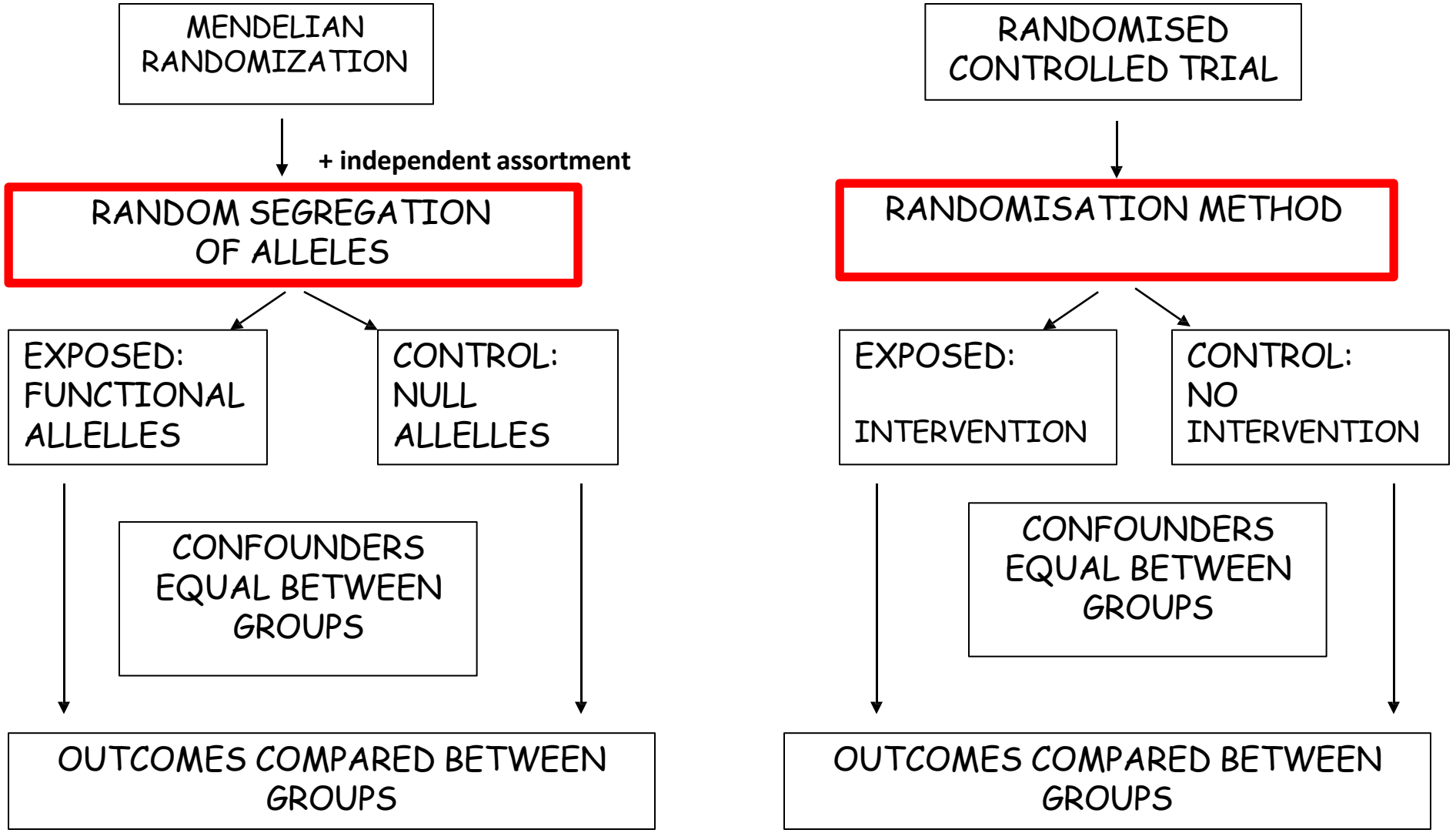
2. **Independent assortment:** alleles for separate traits are transmitted independently of one another

Treat genetics as randomized assignment variable

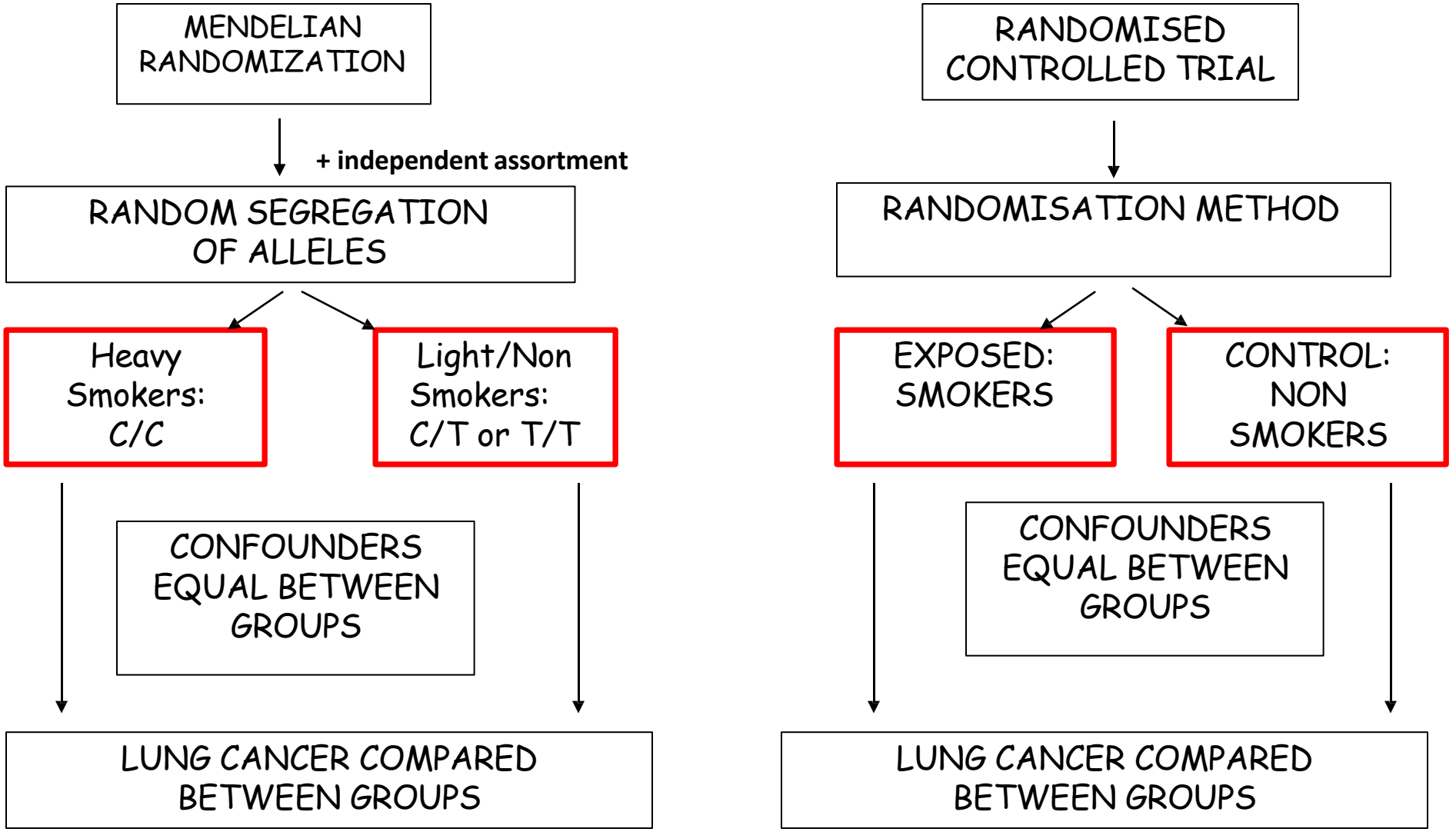


Slide courtesy of John Danesh
Hingorani et al, *Lancet* 2005

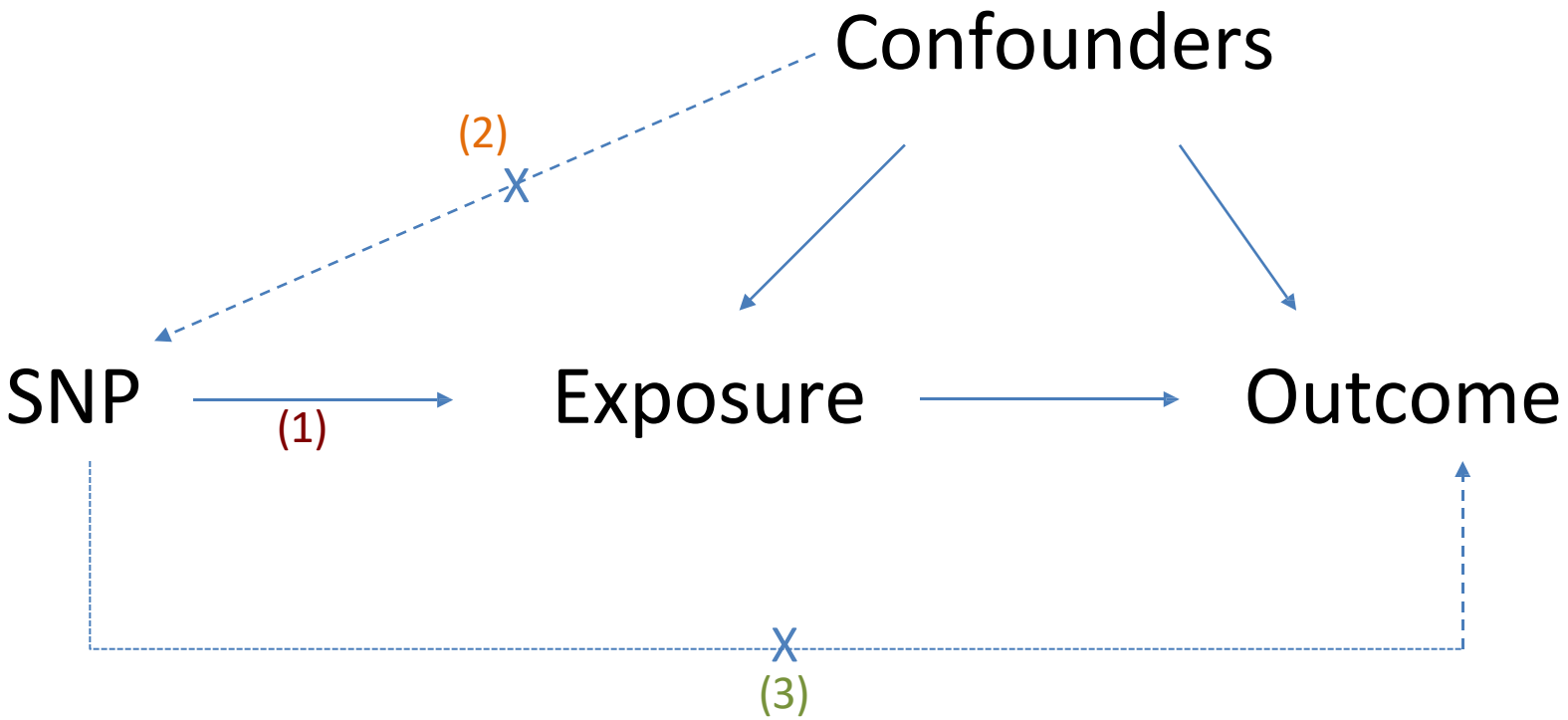
Mendelian randomization and RCTs



Mendelian randomization: Smoking and Lung Cancer



Mendelian Randomization: 3 Core Assumptions



(1) SNP is associated with the exposure

(2) SNP is NOT associated with confounding variables

(3) SNP ONLY associated with outcome through the exposure

Why are genetic associations special?

- Robustness to confounding due to Mendel's laws:
  - Law of segregation: inheritance of an allele is random and independent of environment etc
  - Law of independent assortment: genes for different traits segregate independently (assuming not in LD)

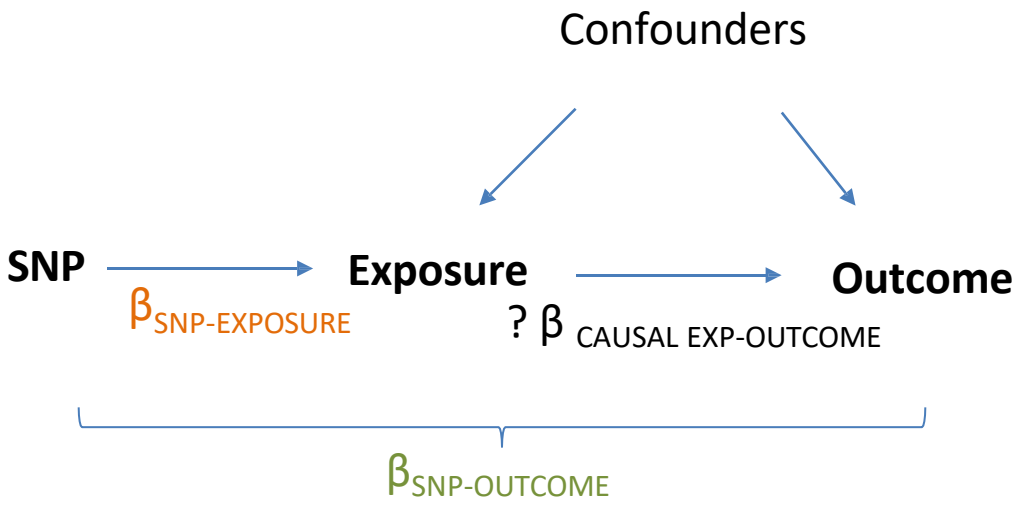
- The direction of causality is known – always from SNP to trait
- Genetic variants are **potentially** very good instrumental variables
- Using genetic variants as IVs is a special case of IV analysis, known as Mendelian randomization

Mendelian Randomization

- Problems with observational data
- Randomized controlled trials
- Mendelian Randomization (MR):
  - How it works
  - Core assumptions
  - Calculating causal effect estimates
- MR example
- Limitations of MR

Calculating causal effect estimates

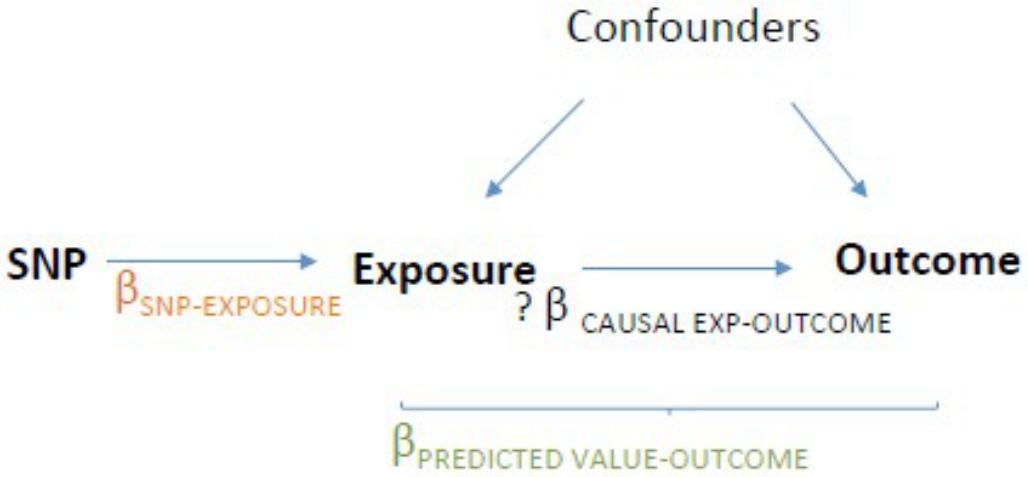
Calculating Causal Effect Estimates

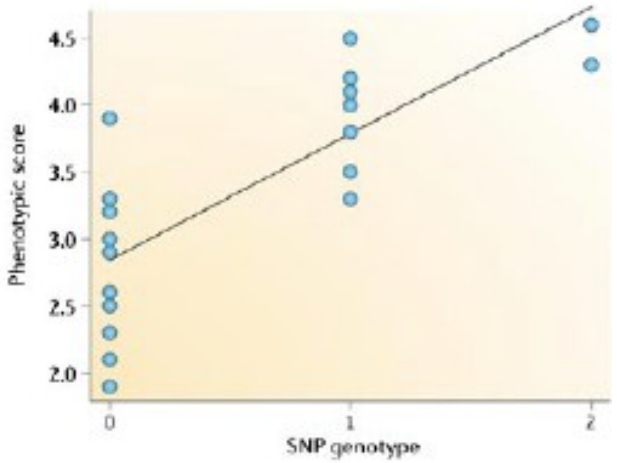


After SNP identified robustly associated with exposure of interest:

- Wald Estimator
- Two-stage least-squares (TSLS) regression

Calculating Causal Effect Estimates





Two-stage Least Squares (2SLS):

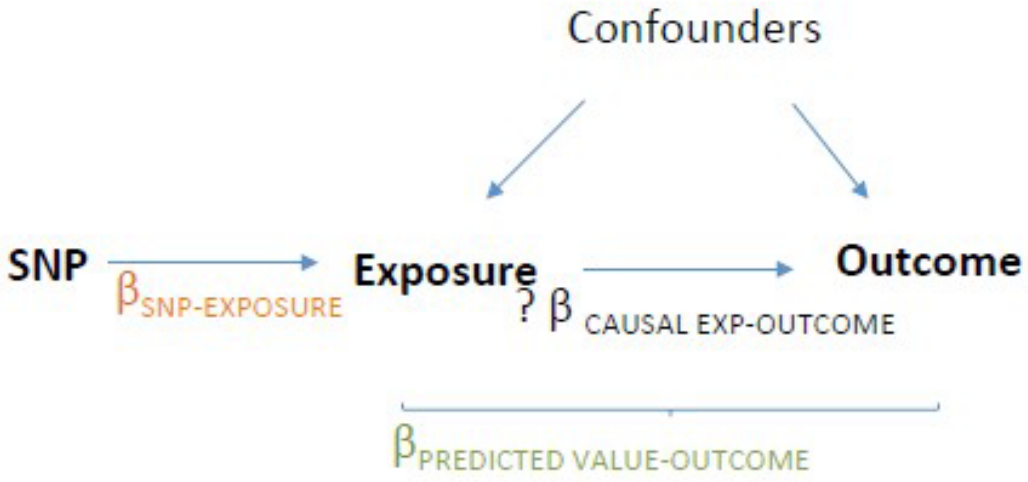
(1) Regress exposure on SNP & obtain predicted values

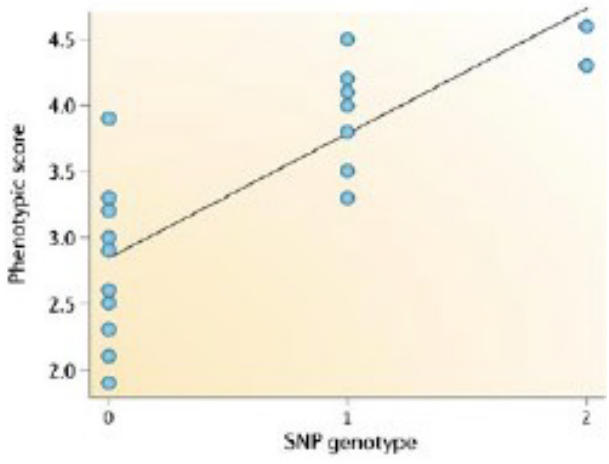
(2) Regress outcome on **predicted** exposure (from 1st stage regression)

(3) Adjust standard errors

*Needs to be done in the one sample ("Single sample MR")

Calculating Causal Effect Estimates





Two-stage Least Squares (2SLS):

(1) Regress exposure on SNP & obtain predicted values

(2) Regress outcome on **predicted** exposure (from 1st stage regression)

(3) Adjust standard errors

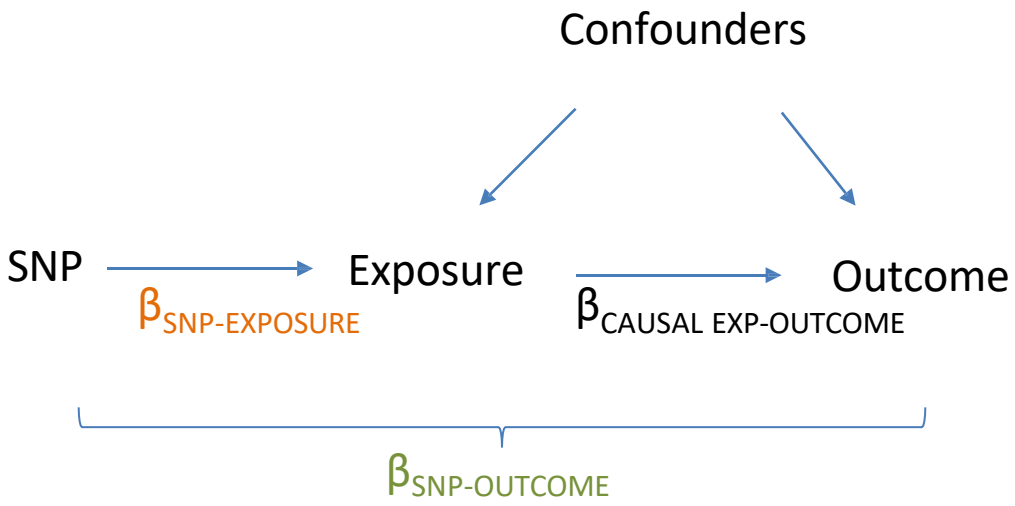
This gives you: difference in outcome per unit change in (genetically-predicted) exposure

Genetically determined exposure → "randomized" → can ascribe causality

(if assumptions are met)

*Needs to be done in the one sample ("Single sample MR")

Calculating Causal Effect Estimates



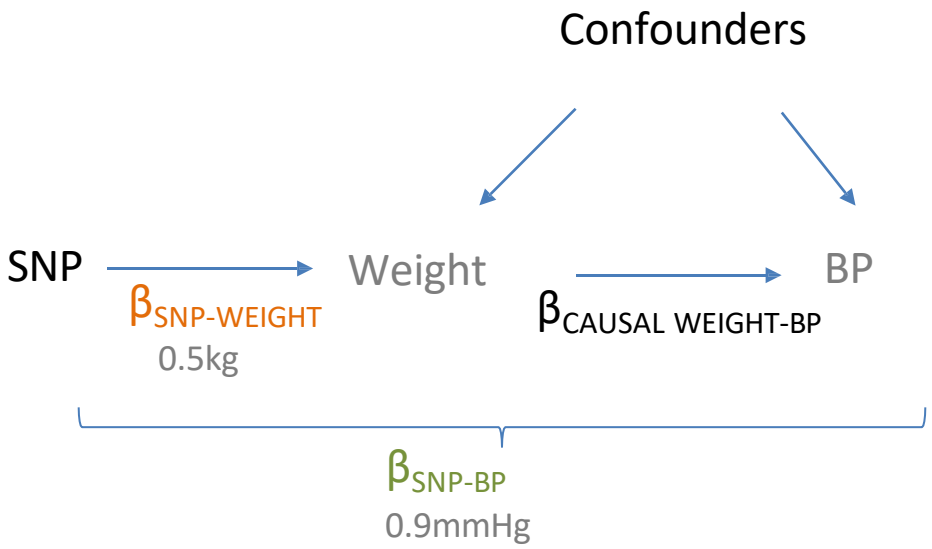
Causal effect by Wald Estimator* :

$$\frac{\hat{\beta}_{SNP\text{-}OUTCOME}}{\hat{\beta}_{SNP\text{-}EXPOSURE}}$$

$$\beta_{SNP\text{-}OUTCOME} = \beta_{CAUSAL\ EXP\text{-}OUTCOME} \times \beta_{SNP\text{-}EXPOSURE}$$

*Can be used in different samples ("Two sample MR")

Calculating Causal Effect Estimates



Causal effect by Wald Estimator* :

$$\frac{\hat{\beta}_{SNP\text{-}OUTCOME}}{\hat{\beta}_{SNP\text{-}EXPOSURE}}$$

= change in outcome per unit change in exposure

BP and weight:

$$\frac{0.9\ \text{mmHg/allele}}{0.5\ \text{kg/allele}}$$

=1.8 mmHg/kg

*Can be used in different samples ("Two sample MR")

MR can also be performed using just the results from GWAS

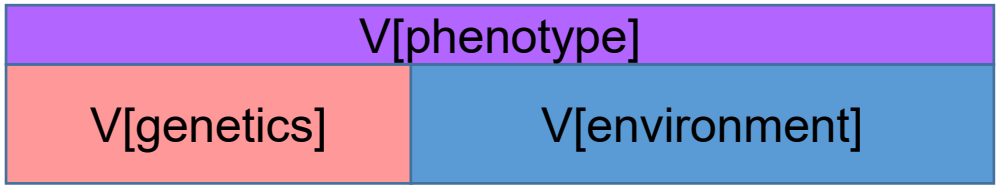
- Also known as two-sample MR, SMR, or MR with summary data etc
- Advantages:
  - The data is readily available, non-disclosive, free, open source
  - The exposure and outcome might not be measured in the same sample
  - The sample size of the outcome variable, key to statistical power, is not limited by requiring overlapping measures of the exposure
- Disadvantages:
  - Some extensions of MR not possible, e.g. non-linear MR, use of GxE for negative controls, various sensitivity analyses

Genetics, Variation, GWAS, PRS, Mechanism

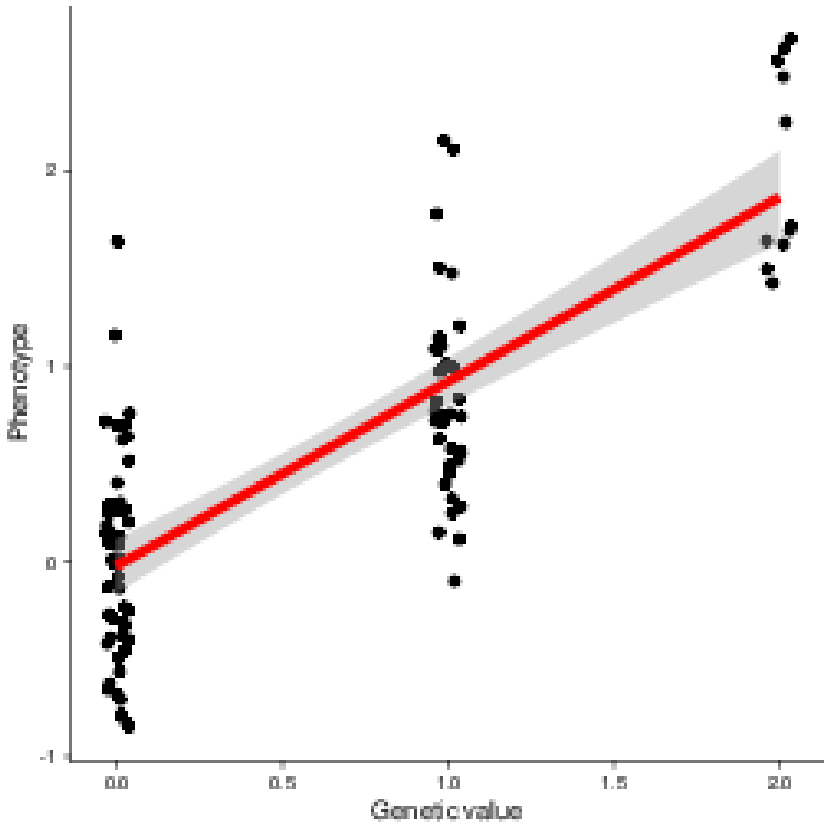
1. Genetics, Variation, GWAS
2. Polygenic scores (PGS)
3. From GWAS to biological insights
4. From Region to Mechanism / Circuitry
5. Quantitative Trait Loci: eQTL/meQTL analysis
6. Mediation Analysis + Mendelian Randomization
7. Heritability and Systems Genetics

Components of phenotypic variance

- Assume p (phenotype) = g (genetic) + e (environment)
- Then, V[p] = V[g] + V[e] + ~~2Cov(G,E)~~
  (assume no gene-environment interactions)

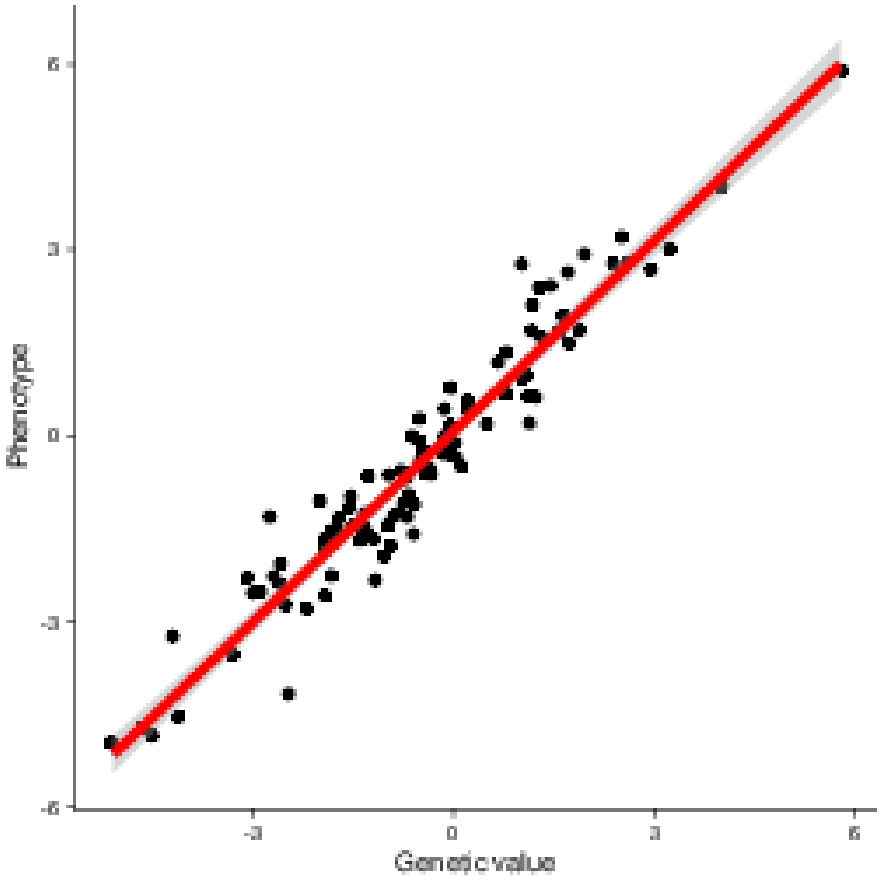


- Example: one causal variant
- Three possible **genetic values** in the population
- Intuition: V[g] is the variance of mean phenotype across different genetic values
- V[e] is the variance of phenotype for the same genetic value



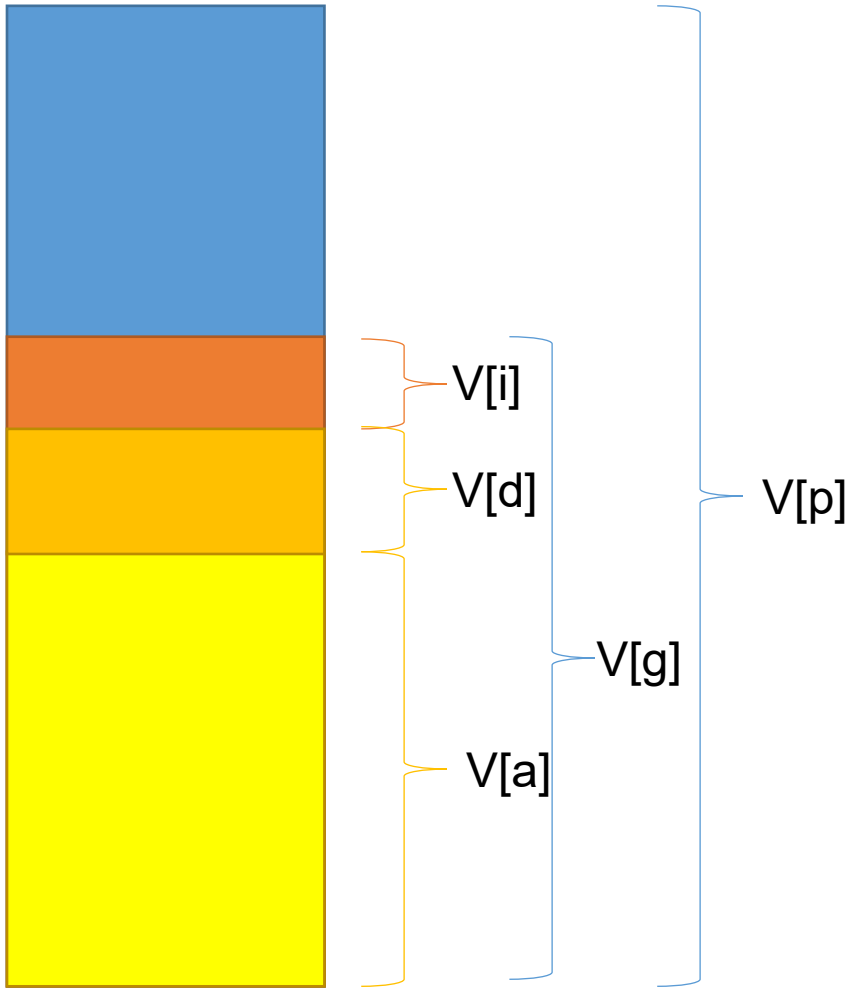
Components of genetic variance

- Assume V[g] = V[a] (additive) + V[d] (dominance) + V[i] (interactions)
- The additive component corresponds to a linear model
- As we add more causal variants, phenotypes become closer to Gaussian
- We could further decompose interactions
- We could include variance due to *de novo* mutations



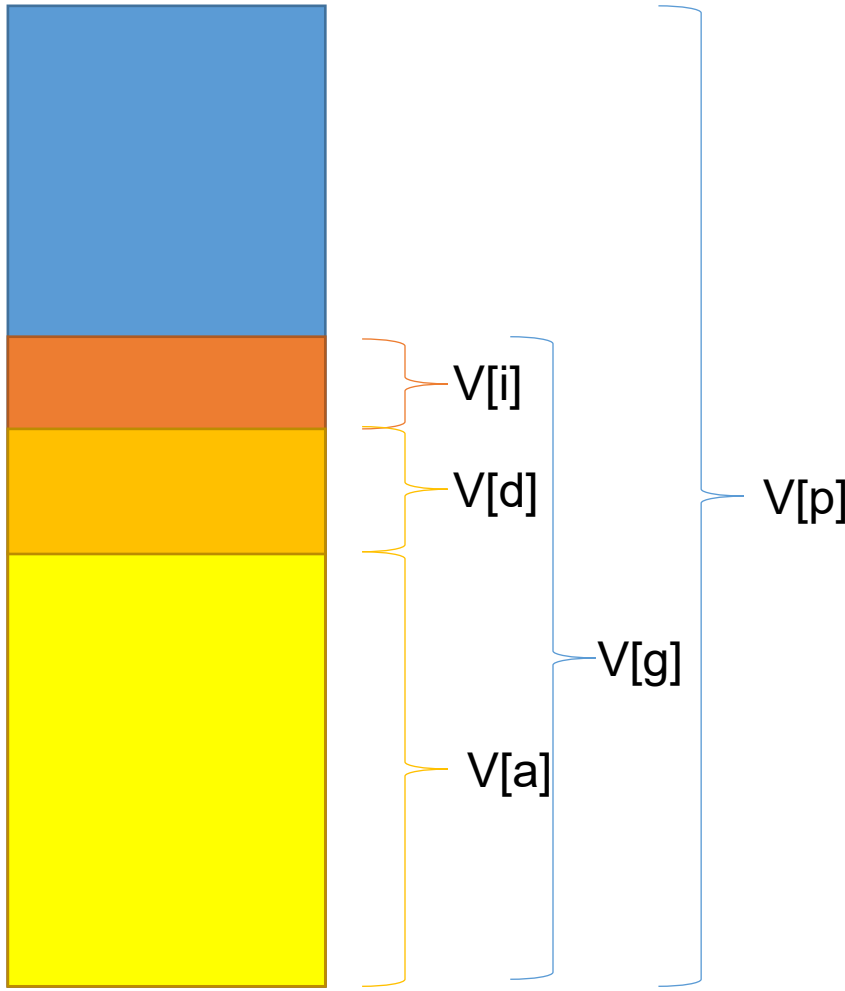
Heritability is a ratio of variances

- V[p] = V[g] + V[e]
- V[g] = V[a] + V[d] + V[i]
- **Broad sense heritability**
  H^2 = V[g] / V[p]
- Broad sense captures all genetic factors
- **Narrow sense heritability**
  h^2 = V[a] / V[p]
- Narrow sense captures only additive effects
- Ongoing debate about the relative importance of additive vs. other effects in disease, selection, etc.



Why study heritability?

- Quantify the importance of genetics vs. environment in traits of interest
- Learn about *genetic architecture*: how many causal variants, effect sizes, allele frequencies
- Narrow sense heritability is the fundamental parameter needed for phenotype prediction (and is the theoretical best possible prediction performance with a linear model)



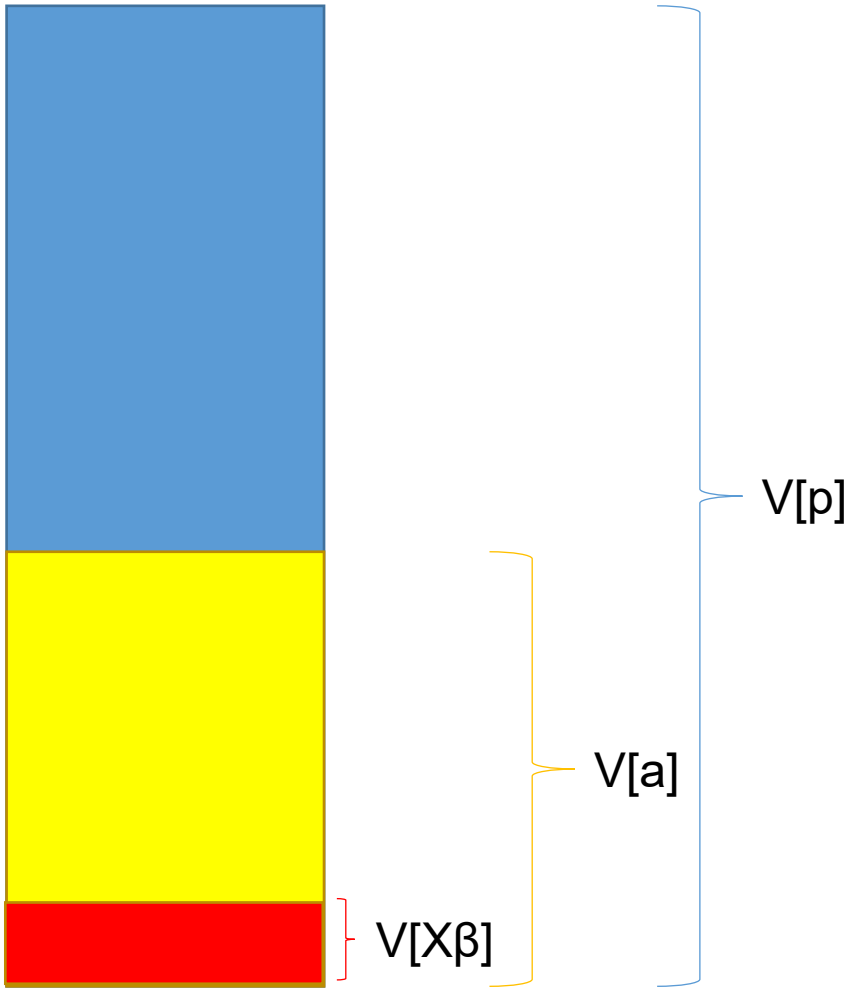
Estimating heritability in relatives

$$p = g + e$$
$$E[p_i\, p_j] = h^2\, E[g_i\, g_j]$$

- Intuition: heritability relates phenotypic correlations to genotypic correlations
- If two individuals have the same allele at each of the causal variants, they will have the same phenotype
- **Haseman-Elston regression:** fit linear regression of phenotypic correlations against genotypic correlations
- Derive genotypic correlation from family relationships: monozygotic twins share 100% of genome, siblings share 50%, etc.
- Example (height): $h^2 = 0.73$

Estimating heritability from GWAS

- Linear model $g = X\beta$
- We can estimate SNP effect sizes β from GWAS
- The variance explained by each SNP depends on effect size and MAF
- $V[X_j \beta_j] = 2 f_j (1 - f_j) \beta_j^2$
- If we do this with genome-wide significant SNPs, we usually $h^2_{GWAS} < h^2$
- Example (height): 253,288 samples; 697 genome-wide significant loci; $h^2_{GWAS}=0.16$, $h^2 = 0.73$
- Known as the **missing heritability problem**

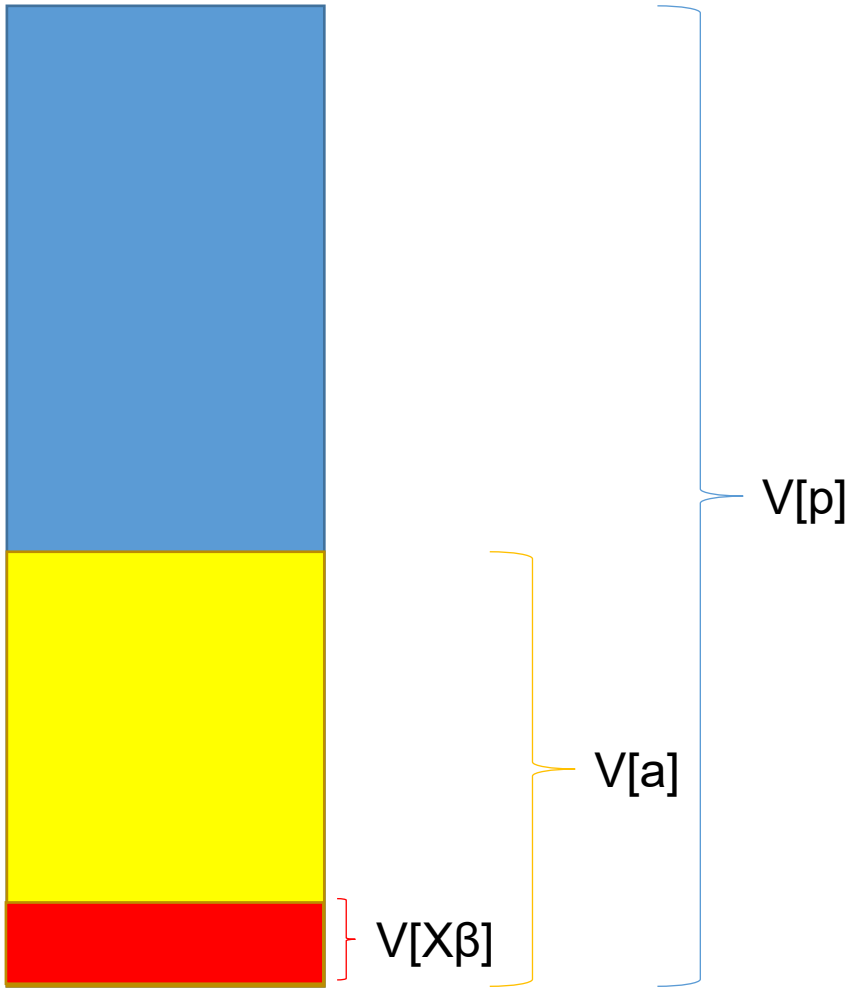


Sources of missing heritability

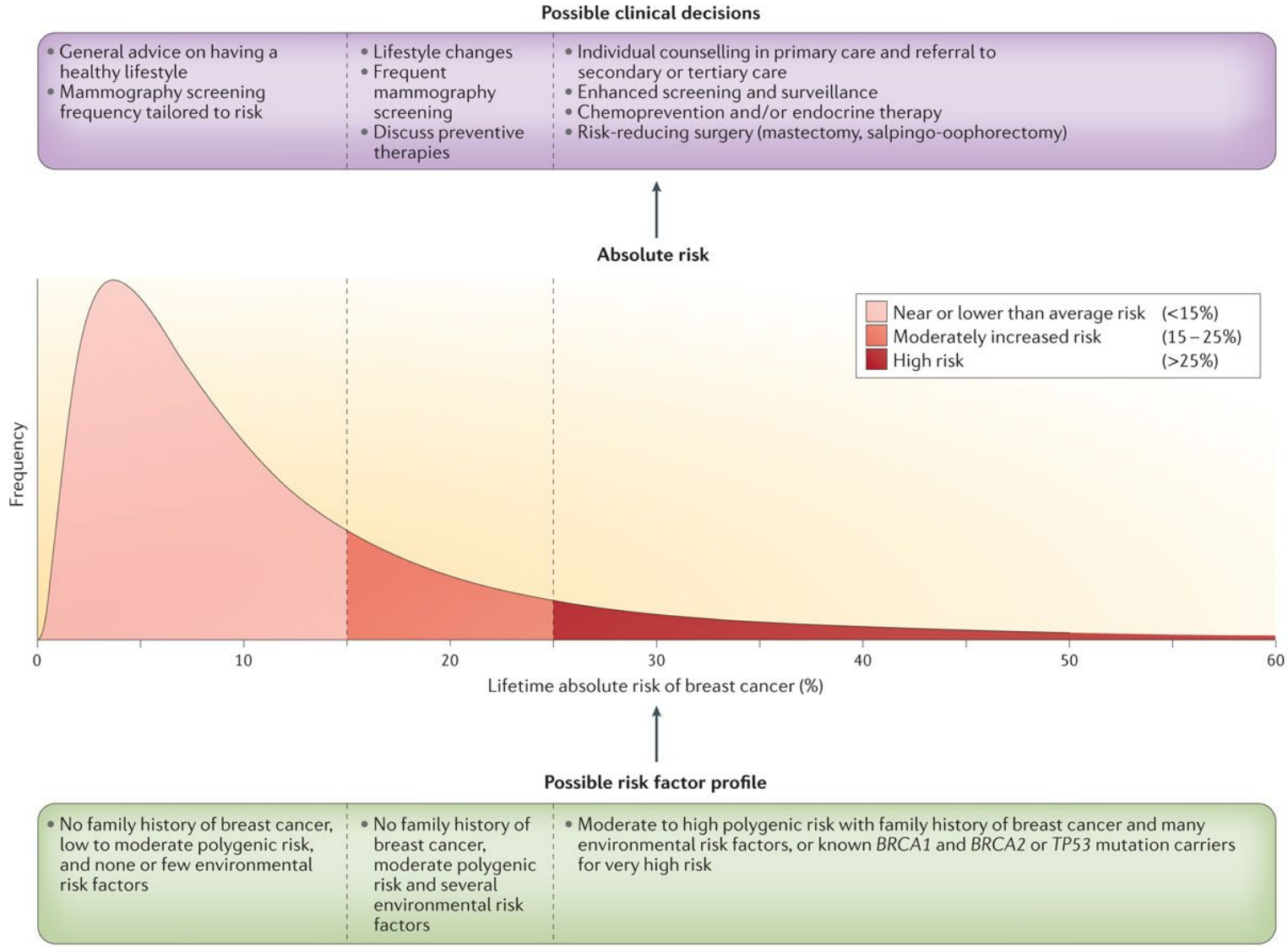
Ongoing debate about several possible explanations for the missing heritability problem.

1. Many common variants, small effects
2. Unobserved rare variants, large effects
3. Wrong model assumptions

Each has very different implications for the future of human genetics studies.

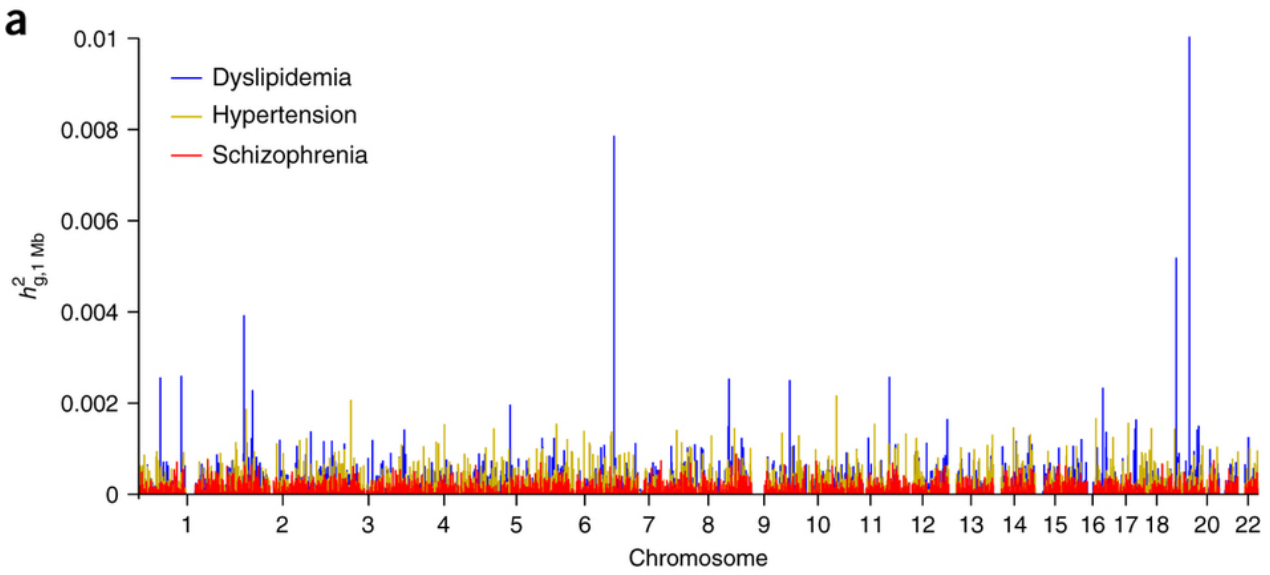


Estimate absolute risk combining genetic and environmental risk factors



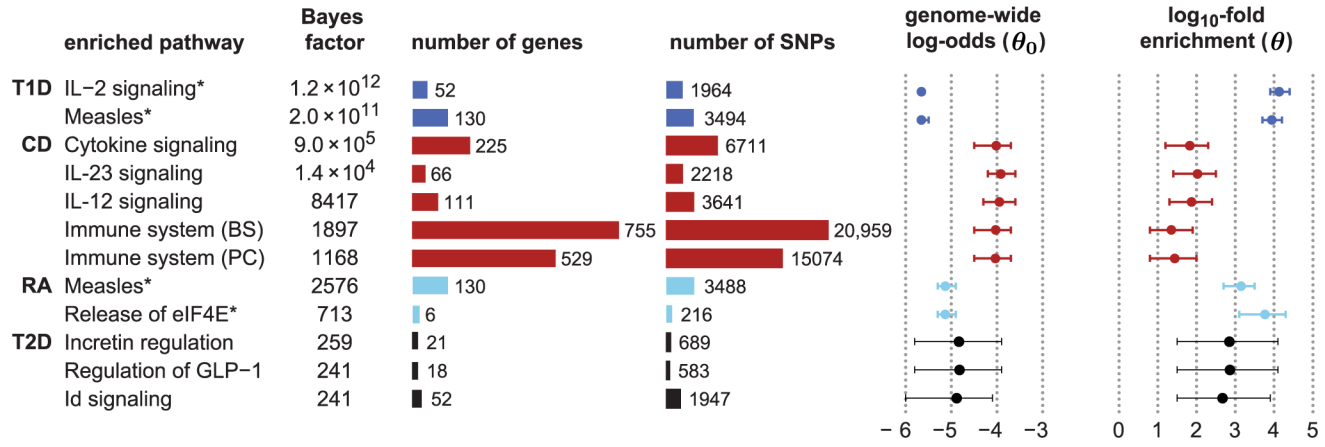
Chatterjee *et al.* Nature Reviews Genetics (2016)

Partitioning heritability



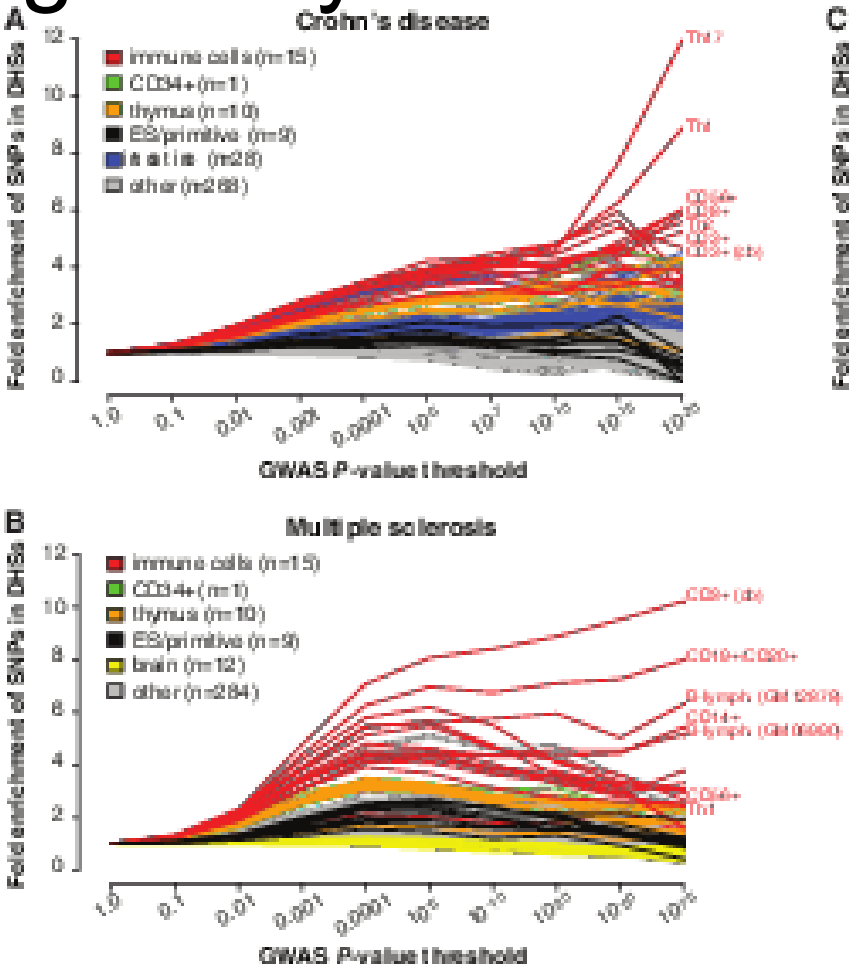
- Fit a model with one component per 1MB window (Loh 2015)
- Bound cumulative heritability explained to estimate number of regions
- Most of the genome explains non-zero heritability

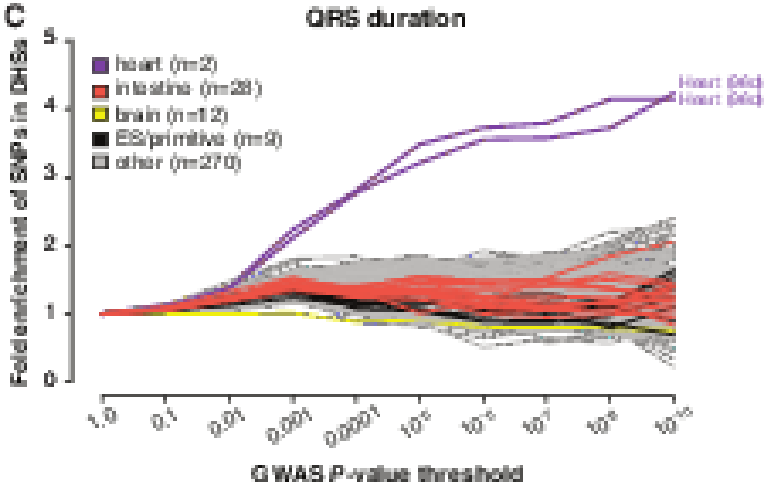
Bayesian variable selection



- Directly fitting the underlying linear model is ill-posed: we have n < p so there are infinitely many solutions
- Idea: use **spike and slab** prior to force many effects to be exactly 0 and regularize the problem (one solution)
- Inference goal: estimate the effect sizes and the level of sparsity (Carbonetto 2013)

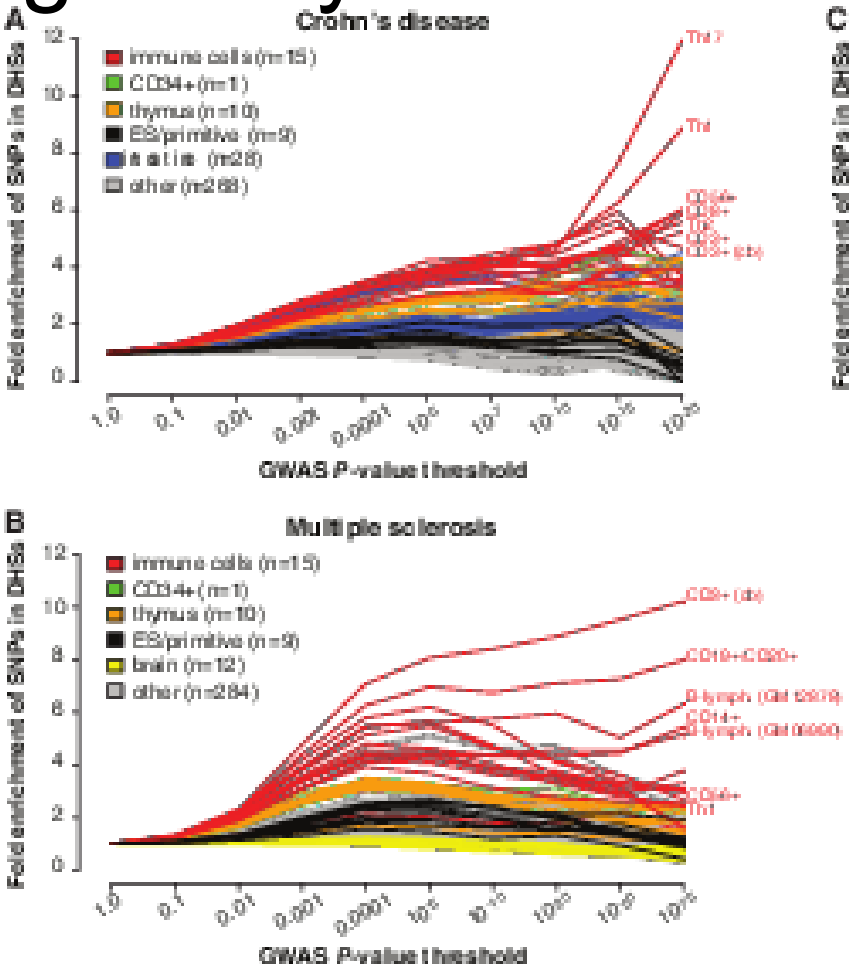
Regulatory enrichments

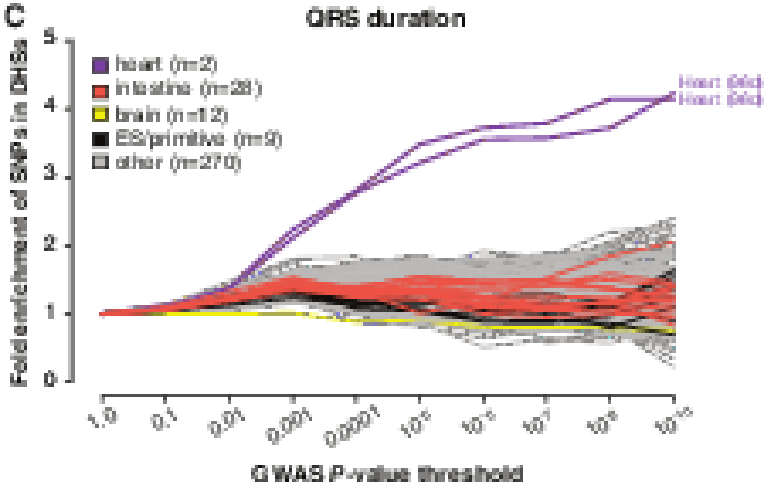




- Weakly associated variants overlap accessible chromatin more often than expected by chance (Maurano 2012)
- Same trend observed in other predicted regulatory elements: histone peaks, ChromHMM segments, super enhancer clusters

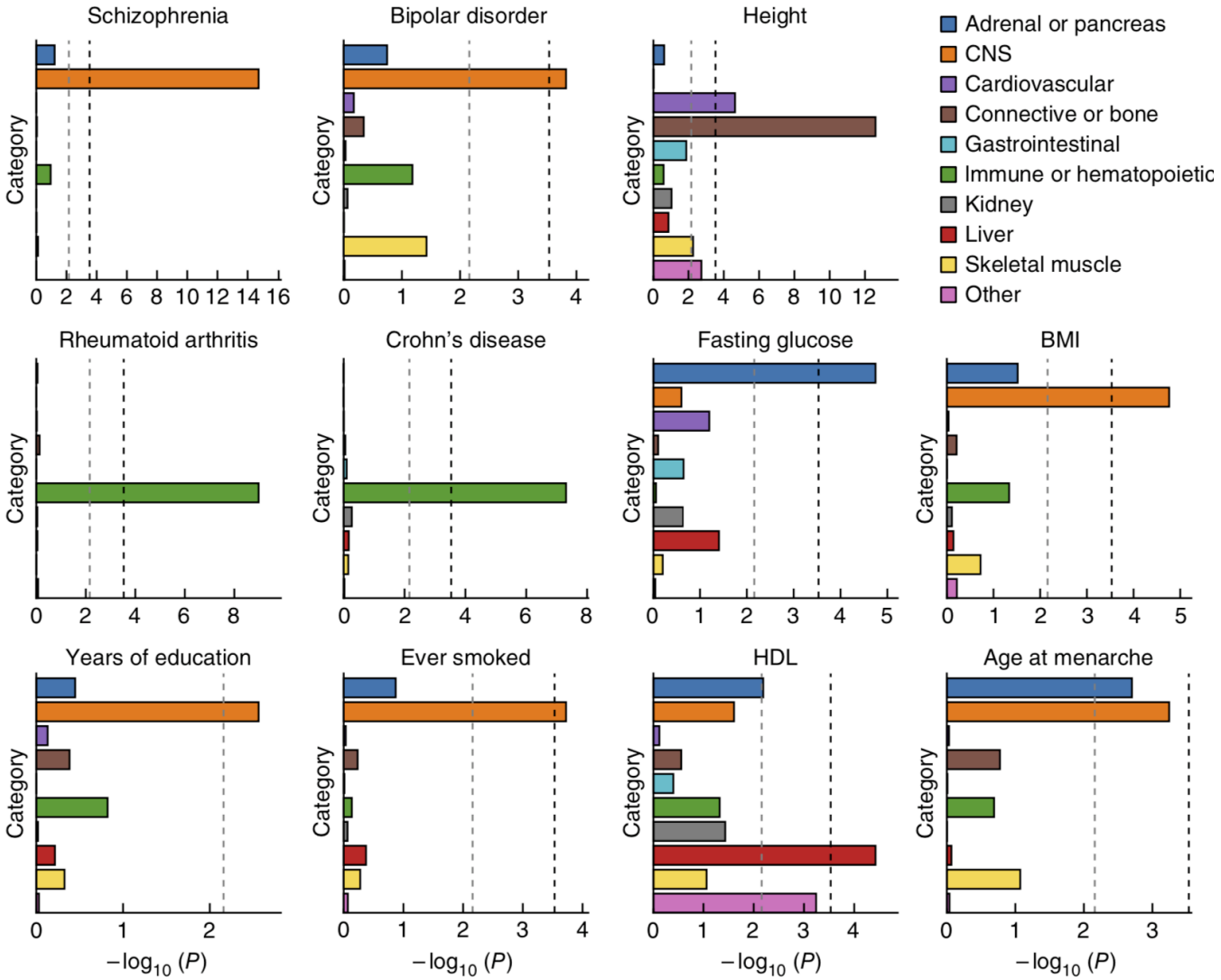
Regulatory enrichments





- Weakly associated variants overlap accessible chromatin more often than expected by chance (Maurano 2012)
- Same trend observed in other predicted regulatory elements: histone peaks, ChromHMM segments, super enhancer clusters

Stratified LDSC partitions heritability of complex trait GWAS summary



Finucane *et al.* (2015)

Genetics, Variation, GWAS, PRS, Mechanism

1. Genetics, Variation, GWAS
2. Polygenic scores (PGS)
3. From GWAS to biological insights
4. From Region to Mechanism / Circuitry
5. Quantitative Trait Loci: eQTL/meQTL analysis
6. Mediation Analysis + Mendelian Randomization
7. Heritability and Systems Genetics