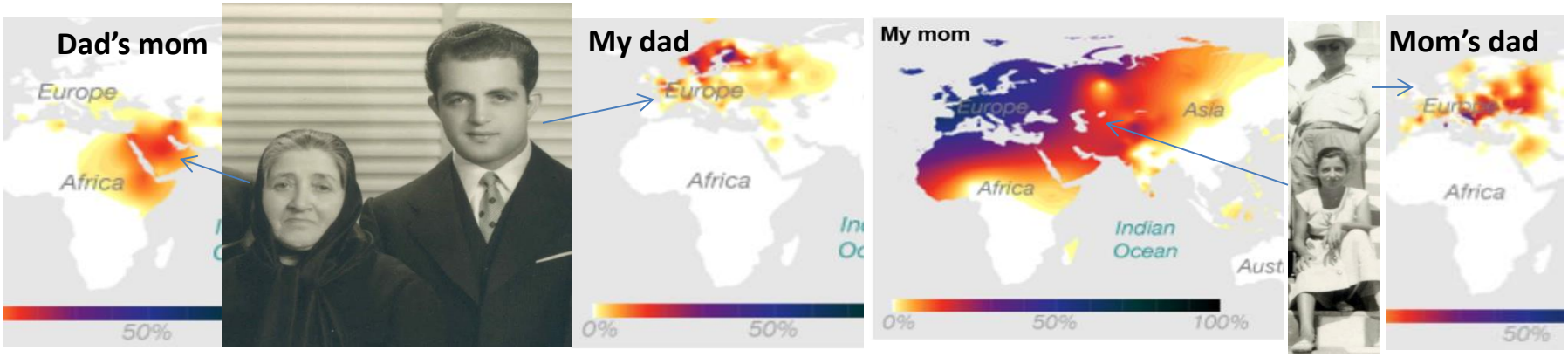
Lecture 20
Personal genomics, disease epigenomics, systems approaches to disease

Predictive Medicine
Molecular Epidemiology
Mendelian Randomization
Polygenic Risk Prediction Models

Personal genomics today: 23 and We

Family Inheritance

Human ancestry



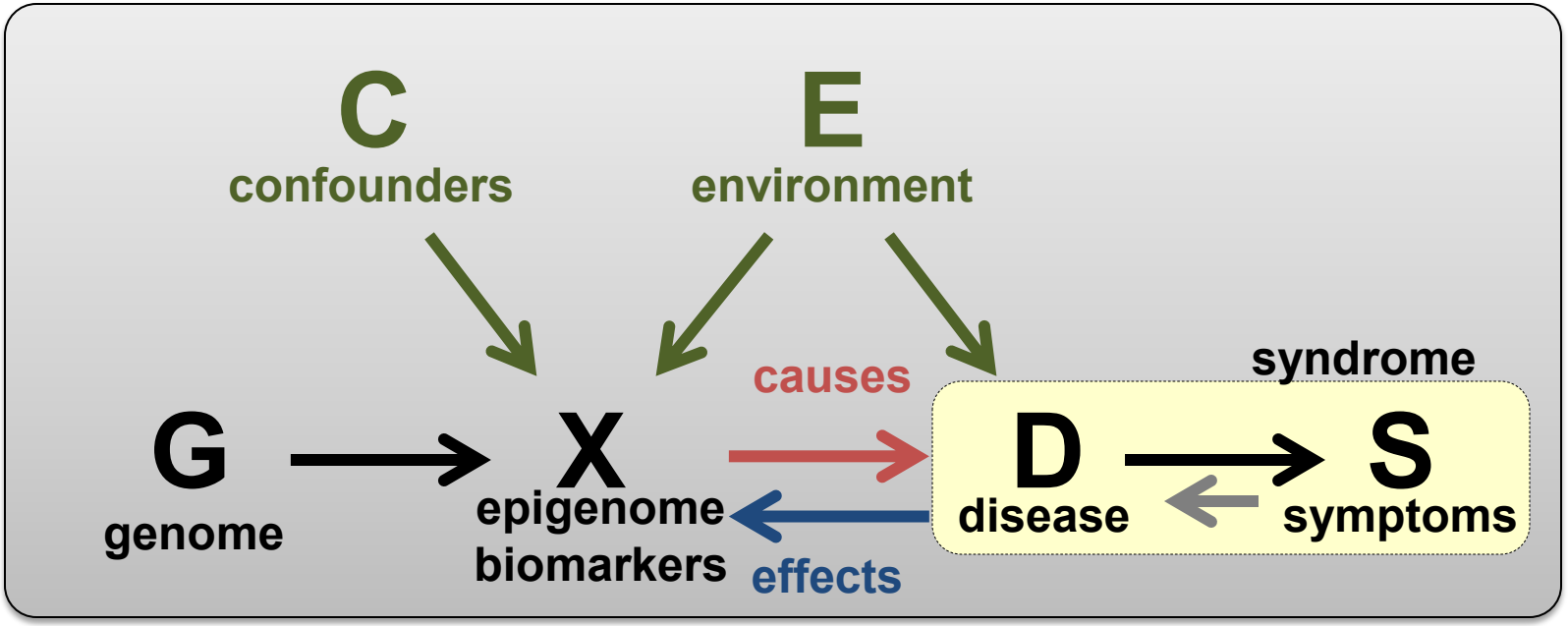
Disease risk

Genomics: Regions → mechanisms → drugs

Systems: genes → combinations → pathways

Goal: Personalized and Predictive Medicine

1. Intro to Epidemiology: basis of human disease
2. Genetic Epidemiology:
   - Genetic basis: GWAS and screening
   - Interpreting GWAS with functional genomics
   - Calculating functional enrichments for GWAS loci
3. Molecular epidemiology
   - meQTLs: Genotype-Epigenome association (cis-/trans-)
   - EWAS: Epigenome-Disease association
4. Resolving Causality
   - Statistical: Mendelian Randomization
   - Application to genotype + methylation in AD
5. Systems Genomics and Epigenomics of disease
   - Beyond single loci: polygenic risk prediction models
   - Sub-threshold loci and somatic heterogeneity in cancer



Epidemiology

The study of the
patterns, **causes**, and **effects**
of health and disease conditions
in defined populations

Epidemiology: Definitions and terms

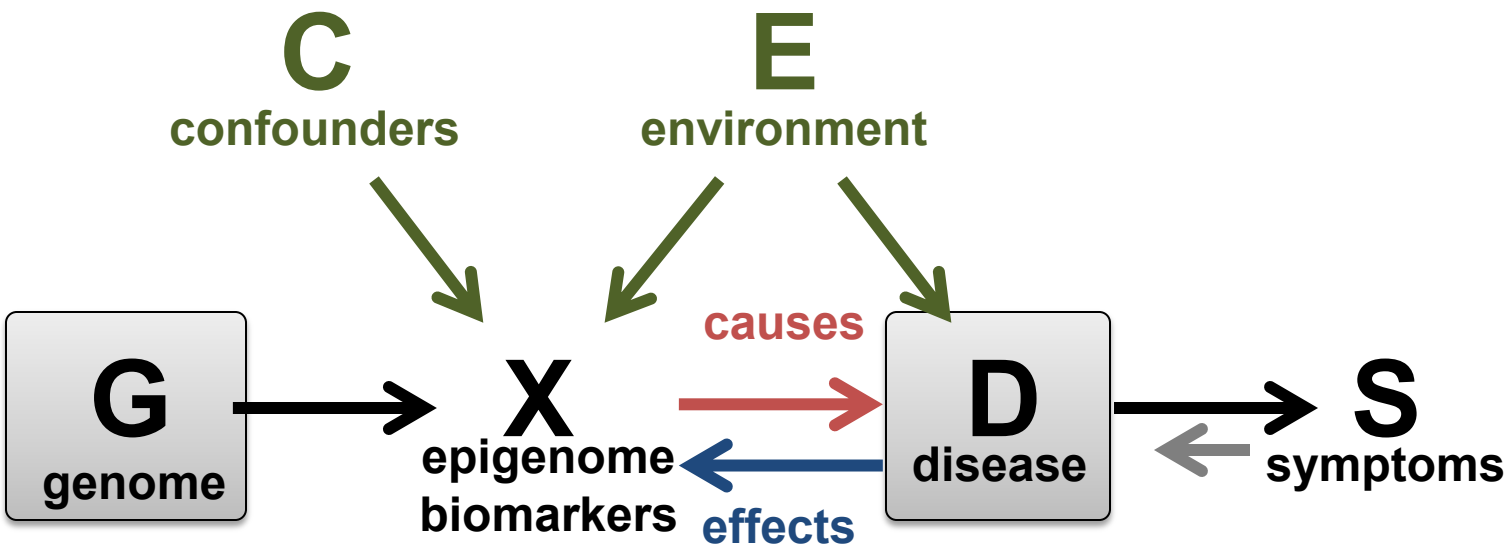
- **Morbidity** level: how sick an individual is
- **Incidence**: # of ***new*** cases / # people / time period
- **Prevalence**: Total # of cases in population
- **Attributable risk**: rate in exposed vs. not exposed
- **Population burden**: yrs of potential life lost (YPLL), quality-/disability-adjusted life year (QALY/DALY)
- **Syndrome:** Co-occurring signs (observed), symptomes (reported), and other phenomena; (often hard to establish causality / risk factors)
- **Prevention challenge:** Determine disease, cause, understand whether, when, and how to intervene

Determining disease causes: study design

- **Principles of experimental design**
  - **Control**: comparison to baseline, placebo effect
  - **Randomization**: Difficult to achieve, ensure mixing
  - **Replication**: control variability in initial sample
  - **Grouping**: understand variation between subgroups
  - **Orthogonality**: all combinations of factors/treatments
  - **Combinatorics**: factorial design $n \times n \times n \times \ldots \times n$ table
- **Challenge of human subjects**
  - Legal and ethical constraints, Review boards
  - Randomization by instrumental variables
  - Clinical trials: blind (patient), double-blind (doctor too)

Goal: Personalized and Predictive Medicine

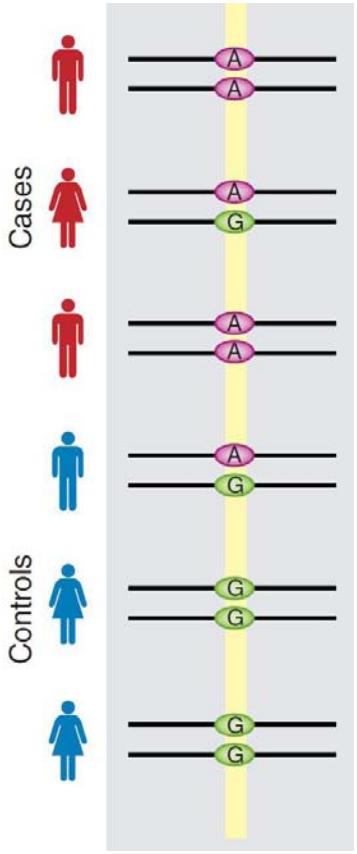
1. Intro to Epidemiology: basis of human disease
2. Genetic Epidemiology:
   - Genetic basis: GWAS and screening
   - Interpreting GWAS with functional genomics
   - Calculating functional enrichments for GWAS loci
3. Molecular epidemiology
   - meQTLs: Genotype-Epigenome association (cis-/trans-)
   - EWAS: Epigenome-Disease association
4. Resolving Causality
   - Statistical: Mendelian Randomization
   - Application to genotype + methylation in AD
5. Systems Genomics and Epigenomics of disease
   - Beyond single loci: polygenic risk prediction models
   - Sub-threshold loci and somatic heterogeneity in cancer

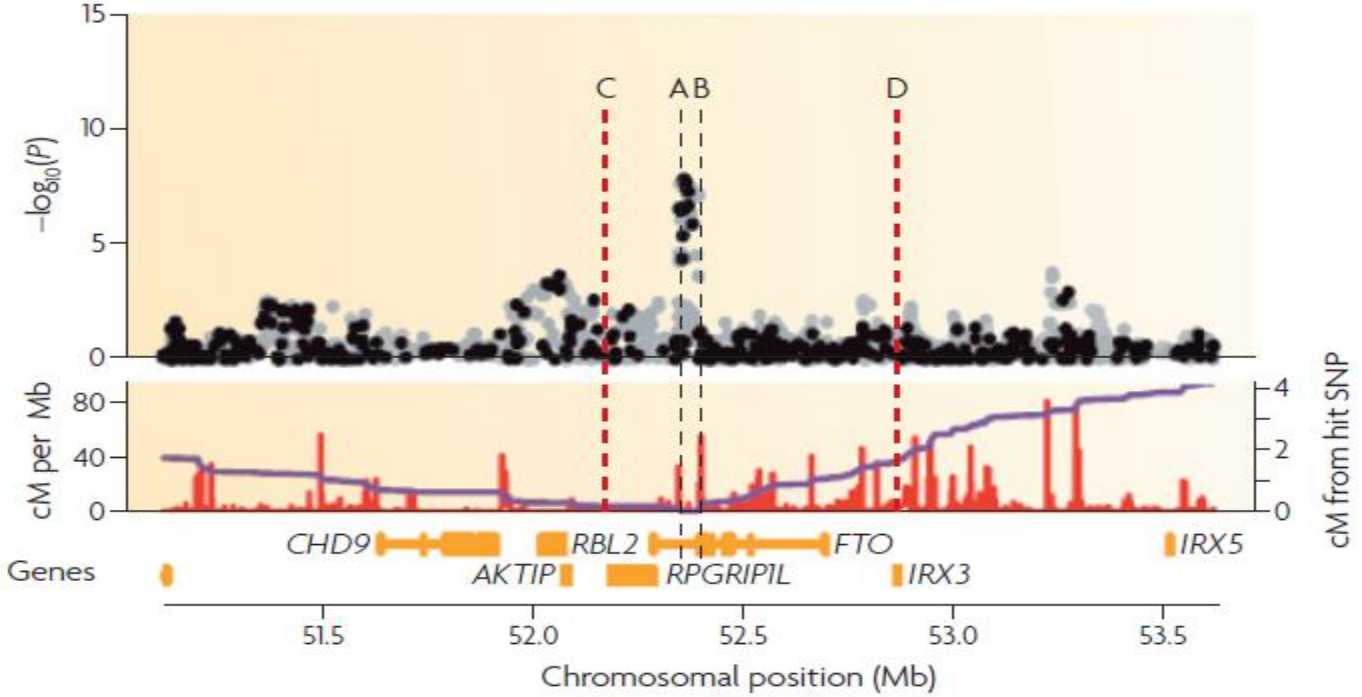


Genetic Epidemiology

Genetic factors contributing to disease

Genome-wide association studies (GWAS)

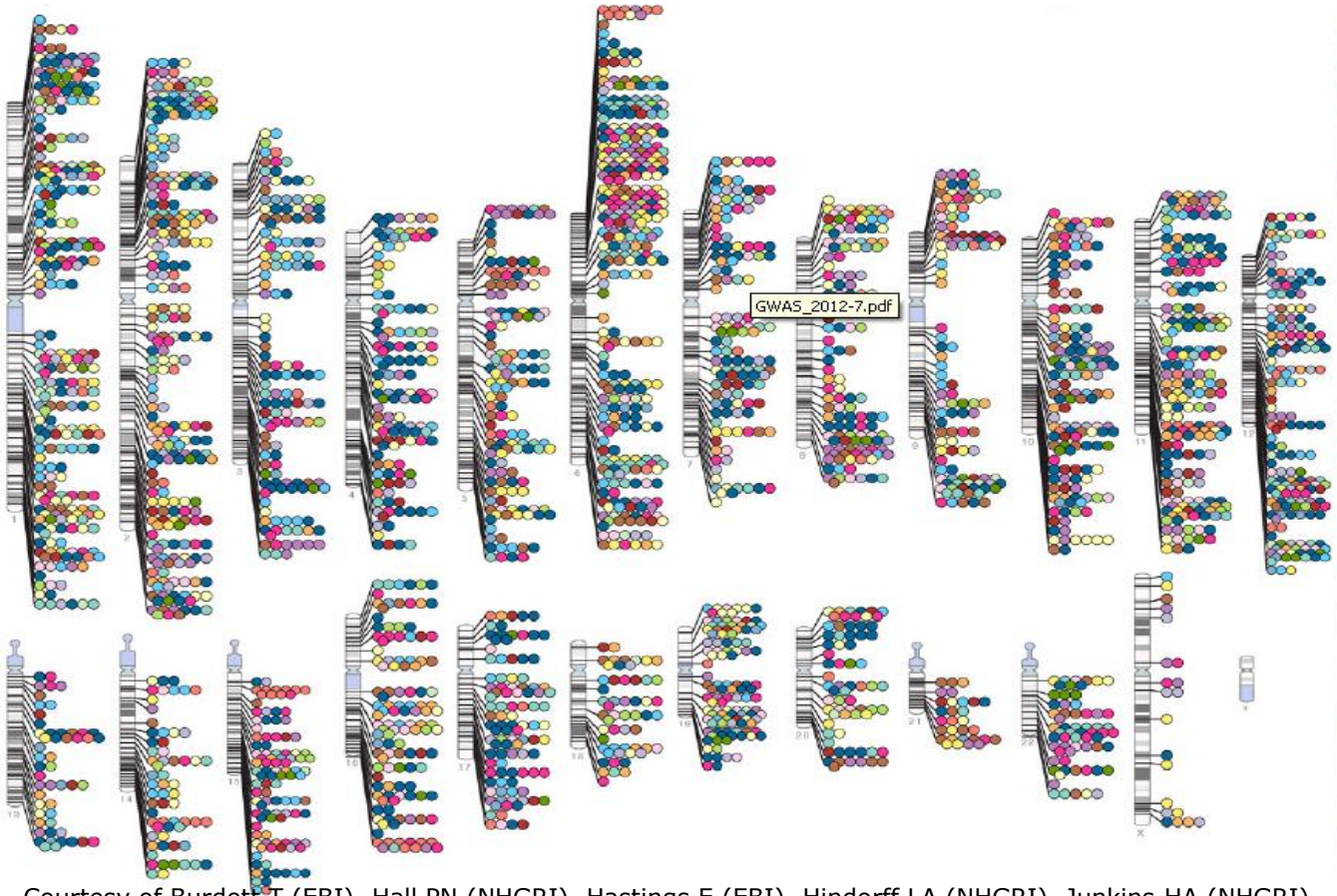




Source: Mccarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., & Hirschhorn, J. N. (2008). "Genome-wide association studies for complex traits: Consensus, uncertainty and challenges." Nat Rev Genet Nature Reviews Genetics, 9(5), 356-369.

- Identify regions that co-vary with the disease
- Risk allele G more frequent in patients, A in controls
- But: large regions co-inherited ➔ find causal variant
- Genetics does not specify cell type or process

All disease-associated genotypes from GWAS

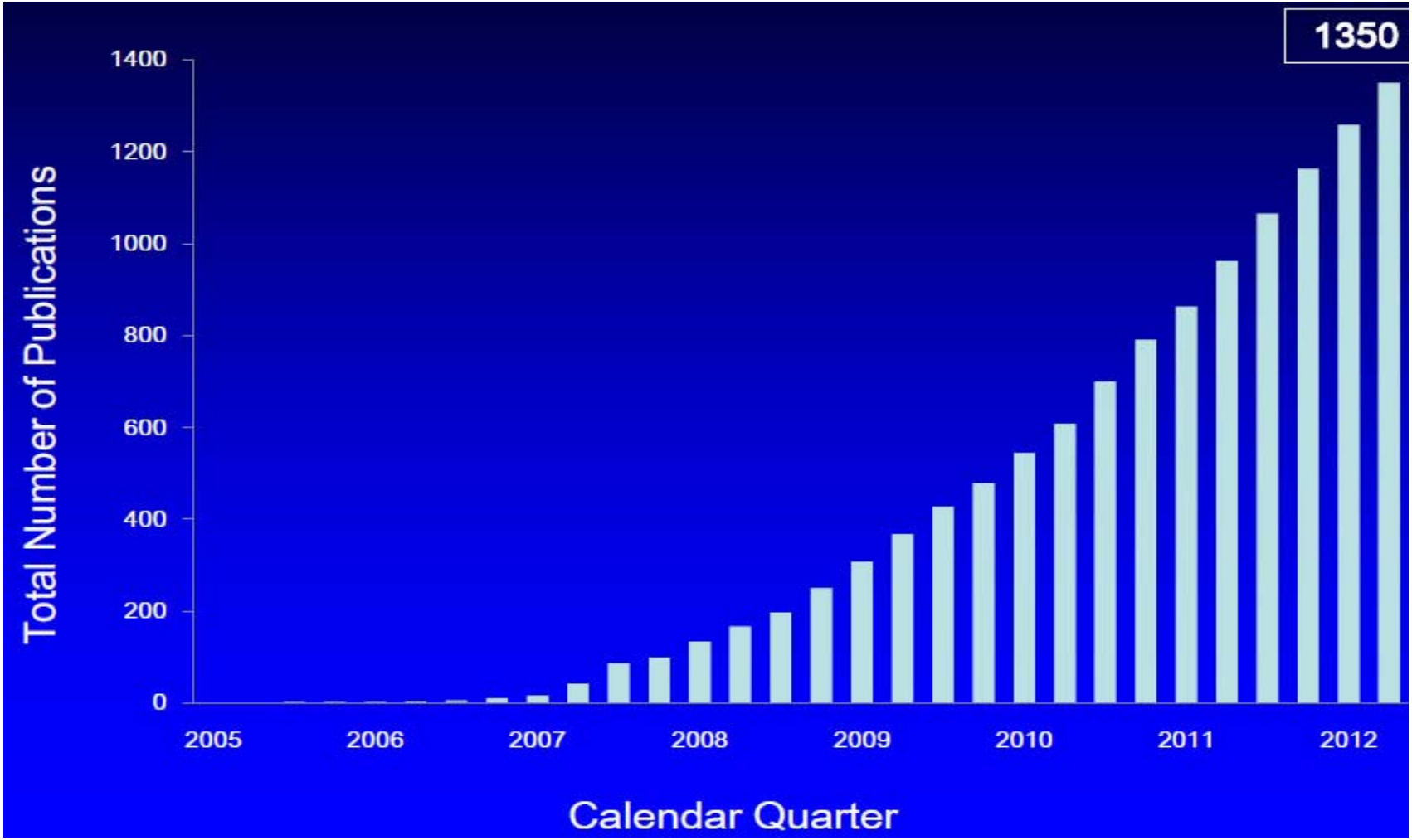


Digestive system disorder
Cardiovascular disorder
Metabolic disorder
Immune system disorder
Neurological disorder
Liver enzyme measurement
Lipid or lipoprotein measurement
Inflammatory marker measurement
Hematological measurement
Body measurement
Cardiovascular measurement
Other measurement
Chemical compound
Biological process
Cancer
Other disease
Other trait
Trait mapping in progress

Courtesy of Burdett T (EBI), Hall PN (NHGRI), Hastings E (EBI), Hindorff LA (NHGRI), Junkins HA (NHGRI), Klemm AK (NHGRI), MacArthur J (EBI), Manolio TA (NHGRI), Morales J (EBI), Parkinson H (EBI) and Welter D (EBI).The NHGRI-EBI Catalog of published genome-wide association studies. Available at: www.ebi.ac.uk/gwas. Used with Permission.

- **1000s of studies, each with 1000s of individuals**
  - Increasing power, meta-analyses reveal additional loci
  - More loci expected, only fraction of heritability explained

More loci on the way: GWAS growth continues



- When to design custom chip: continuously update
- http://www.genome.gov/admin/gwascatalog.txt

Decreasing cost of whole-genome sequencing

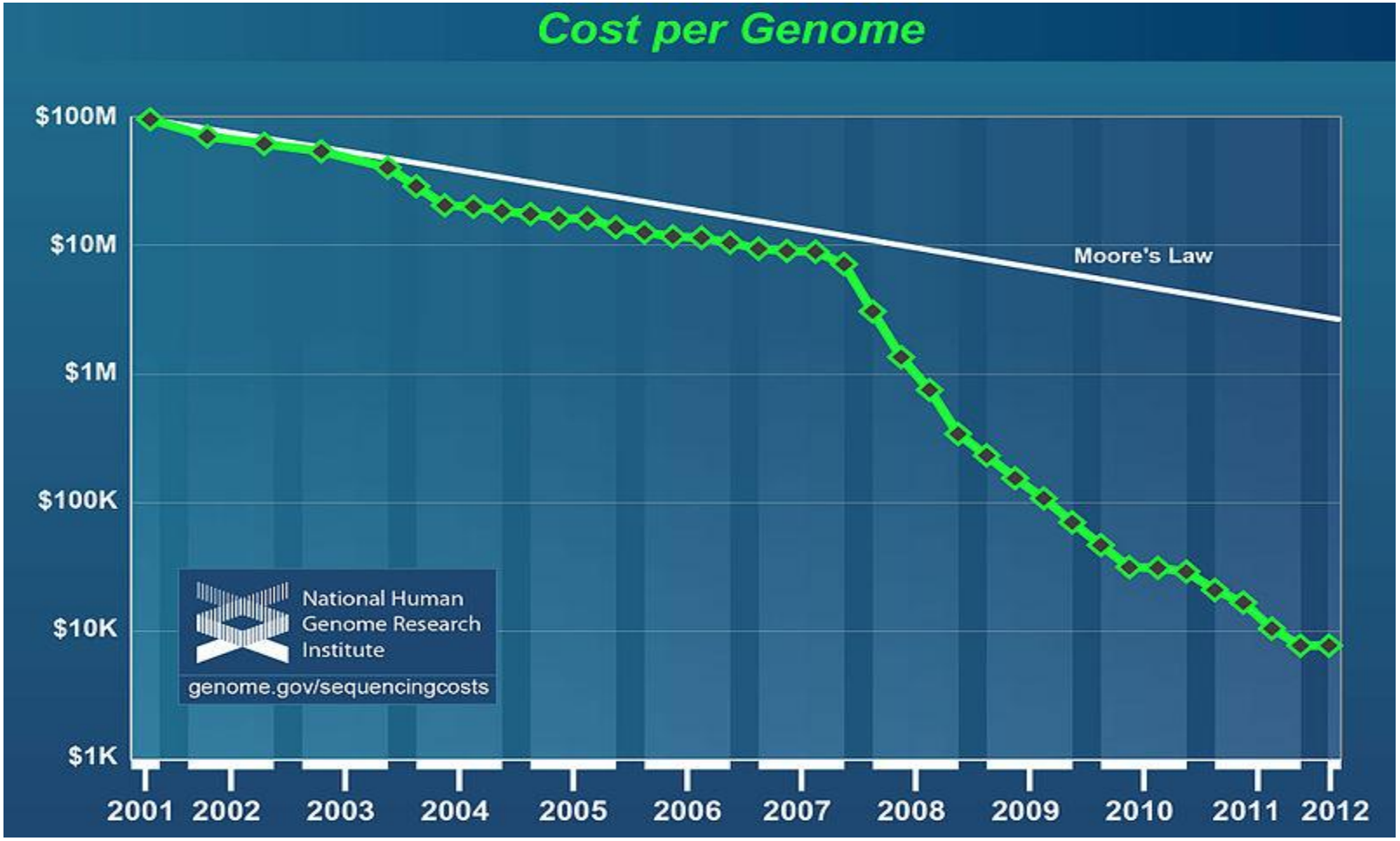


Image by Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Image in the public domain.

- Simply genotype all known variants at >0.1% freq
- Or: sequence complete diploid genome of everyone

Genetic epidemiology: What to test

- **Family risk alleles**, inherited with common trait
  - Specific genes, specific variants, family history
- Monogenic, **actionable**, protein-coding mutations
  - Most understood, highest impact, easiest to interpret
- **All coding** SNPs with known disease association
  - What if not druggable / treatable? Want/need know?
- **All** coding/non-coding **associations** from GWAS
  - Thousands of significant associations (1350 on 6/2012)
- **All common** SNPs, regardless of association
  - HapMap and 1000 Genomes capture common variants
- **Genome**: all SNPs, CNVs, rare/private mutations

Predictive medicine: When to screen

- **Diagnostic testing:** after symptoms, confirm a hypothesis, distinguish between possibilities
- **Predictive risk:** before symptoms even manifest
- **Newborn**: heel pick, store, for early treatment
- **Pre-natal testing:** ulstrasound, maternal serum vs. needles, probes, chorionic villus sampling
- **Pre-conception testing:** common/rare disorders
- **Carrier testing**: specific mutation in family history

- **Genetics vs. biomarkers :** cause vs. consequence?

Goal: Personalized and Predictive Medicine

1. Intro to Epidemiology: basis of human disease
2. Genetic Epidemiology:
   - Genetic basis: GWAS and screening
   - Interpreting GWAS with functional genomics
   - Calculating functional enrichments for GWAS loci
3. Molecular epidemiology
   - meQTLs: Genotype-Epigenome association (cis-/trans-)
   - EWAS: Epigenome-Disease association
4. Resolving Causality
   - Statistical: Mendelian Randomization
   - Application to genotype + methylation in AD
5. Systems Genomics and Epigenomics of disease
   - Beyond single loci: polygenic risk prediction models
   - Sub-threshold loci and somatic heterogeneity in cancer

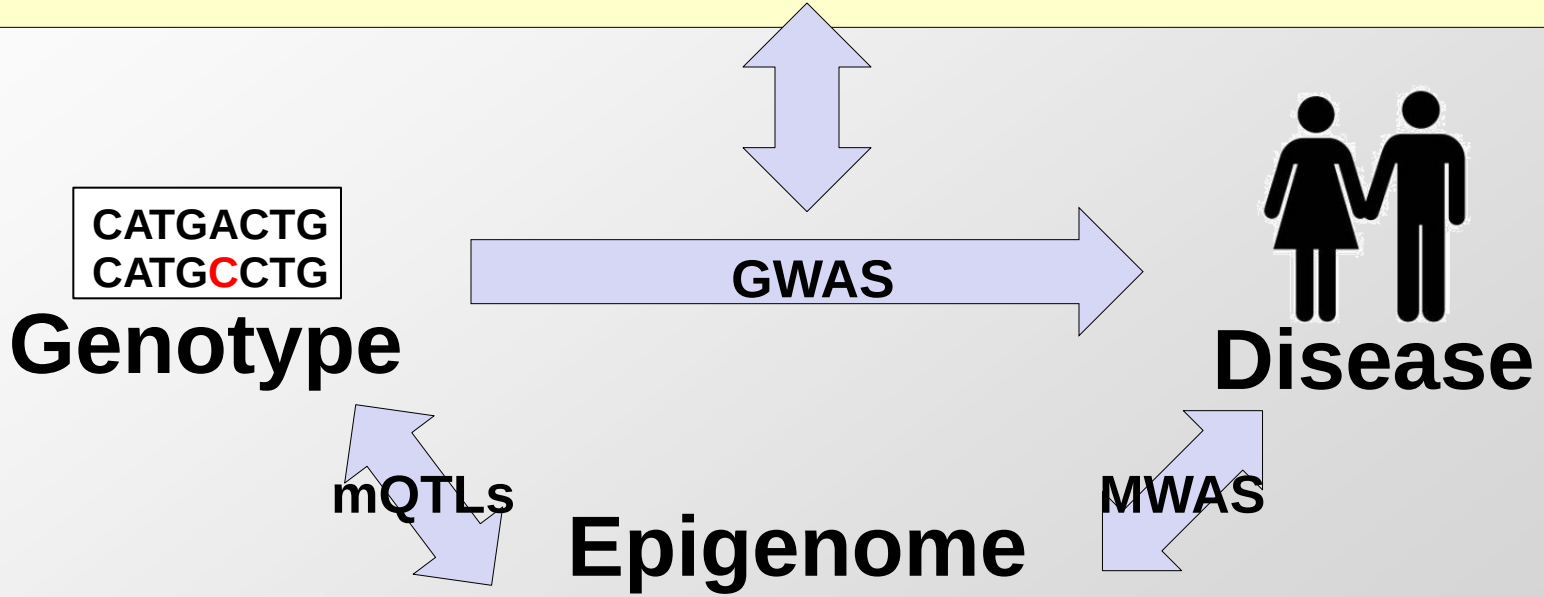
Interpreting disease associations

Functional genomics of GWAS

Interpreting disease-association signals

(1) Interpret variants using Epigenomics

- Chromatin states: Enhancers, promoters, motifs
- Enrichment in individual loci, across 1000s of SNPs in T1D

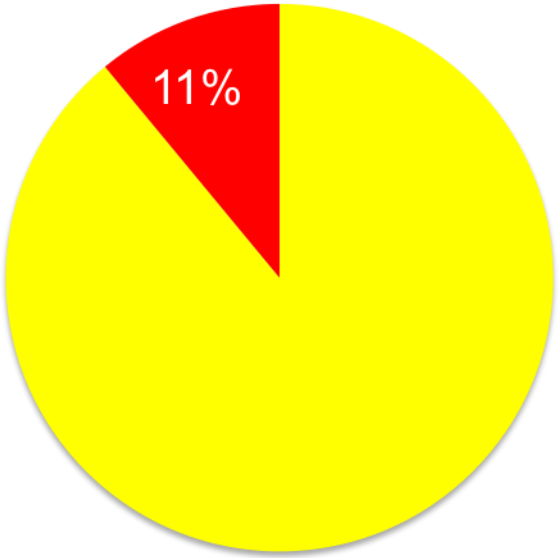


(2) Epigenome changes in disease

- Intermediate molecular phenotypes associated with disease
- Variation in brain methylomes of Alzheimer's patients

Complex disease: strong non-coding component

Monogenic / Mendelian Disease

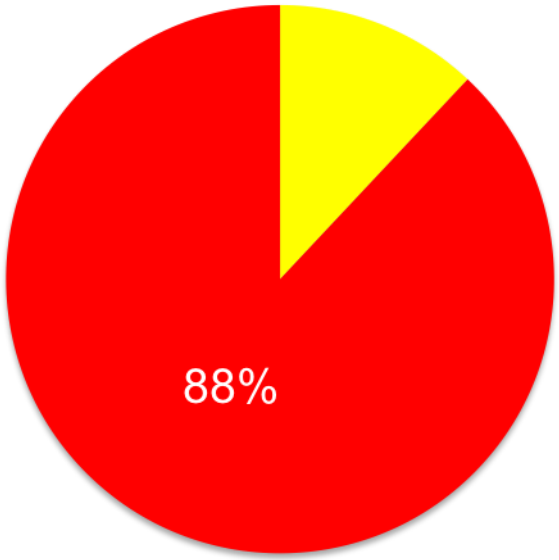


Human Genetic Mutation Database
April 2010 release

Coding

Non-coding

Polygenic / Complex Disease

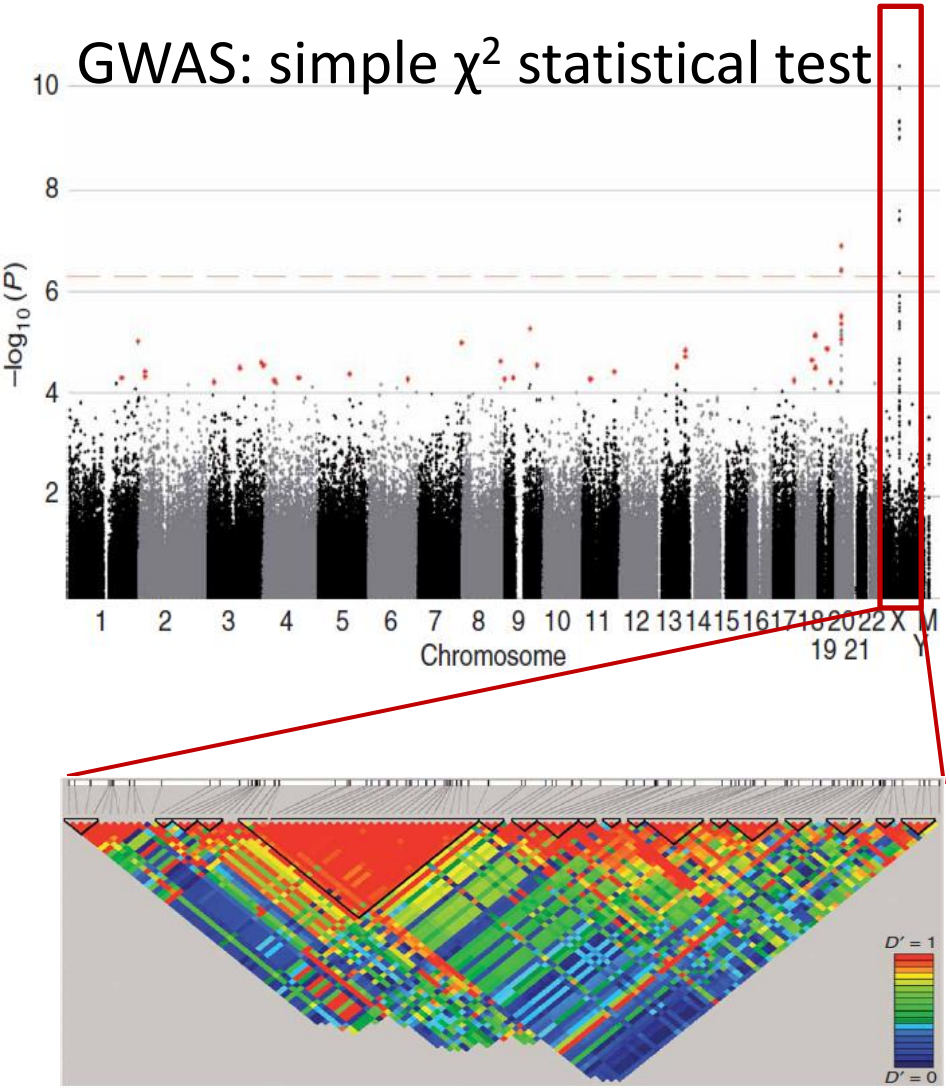


Catalog of GWAS studies
Hindorff et al. PNAS 2009

Slide credit: Benjamin Raby

Genomic medicine: challenge and promises

GWAS: simple χ^2 statistical test



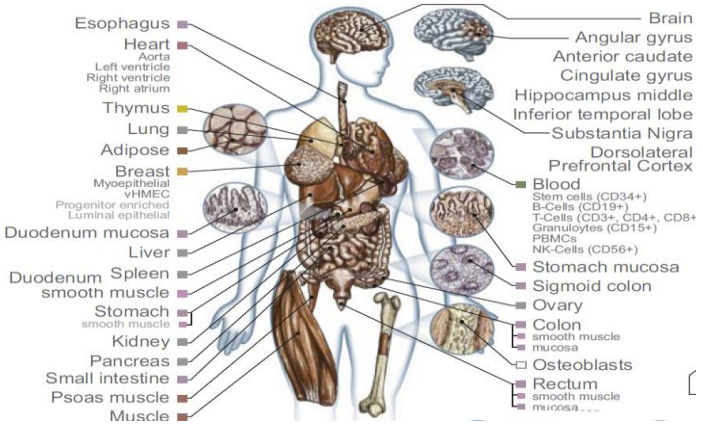
Courtesy of Macmillan Publishers Limited. Used with permission
Source: Hillmer, A. M., Brockschmidt, F. F., Hanneken, S., Eigelshoven, S., Steffens, M., Flaquer, A., . . . Nöthen, M. M. (2008). "Susceptibility variants for male-pattern baldness on chromosome 20p11." Nature Genetics Nat Genet, 40(11), 1279-1281. doi:10.1038/ng.228

Hillmer Nature Genetics 2008

1. The promise of genetics
   - Disease mechanism
   - New target genes
   - New therapeutics
   - Personalized medicine

2. The challenge
   - **90+% disease hits non-coding**
   - Cell type of action not known
   - Causal variant not known
   - Mechanism not known

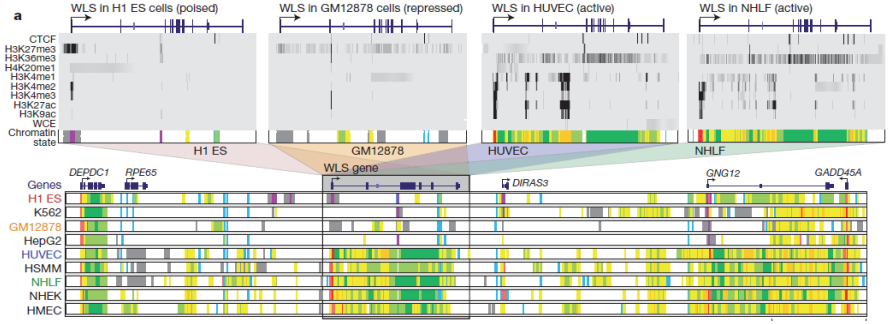
Genomic medicine: challenge and promises



Courtesy of NIH Roadmap Epigenomics Mapping Consortium. Used with permission.

3. The remedy
   - Annotation of non-coding genome (ENCODE/Roadmap)
   - Linking of enhancers to regulators and target genes
   - New methods for utilizing them

Roadmap Epigenomics, Nature 2015



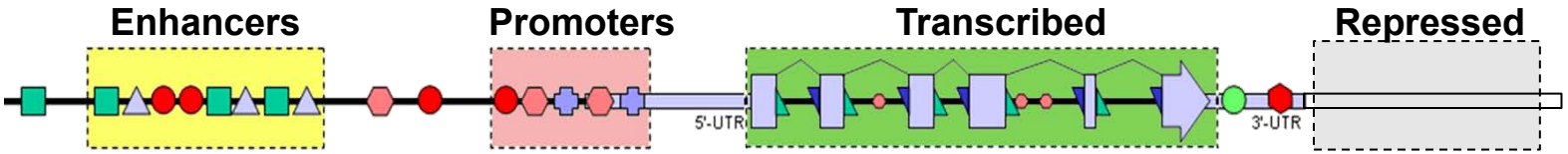
Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Ernst, J. et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 473(7345), 43-49.

Ernst, Nature 2011

4. The deliverables
   - Relevant cell type
   - Target genes
   - Causal variant
   - Upstream regulator
   - Relevant pathways
   - Intermediate phenotypes

This talk: From loci to mechanisms

Building a reference map of the regulatory genome

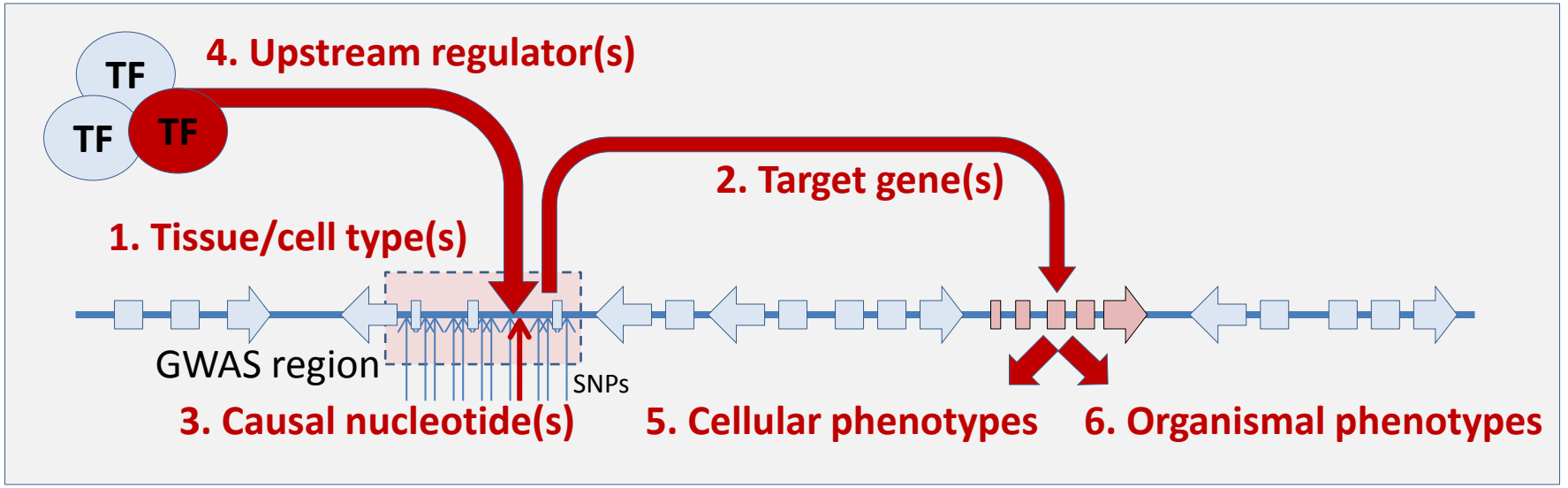


Regions: Enhancers, promoters, transcribed, repressed
Cell types: Predict tissues and cell types of epigenomic activity
Target genes: Link variants to their target genes using eQTLs, activity, Hi-C
Nucleotides: Regulatory consequence of mutation: Conservation, PWMs
Regulators: Upstream regulators whose activity is disrupted by mutation

Application to GWAS, hidden heritability, and Cancer

GWAS hits	CATGCCTG CGTGTCTA	• 93% top hits non-coding → Mechanism? Cell type? • Lie in haplotype blocks → Causal variant(s)?
'Hidden' heritability	CATGCCTG CGTGTCTA	• Many variants, small effects → Pathway-level burden/load • Many false positives → Prioritize w/ regulatory annotations
Cancer mutations	CATGCCTG CATCCCTG	• Loss of function → Protein-coding variants, convergence • Gain of function → Regulatory variants, heterogeneity

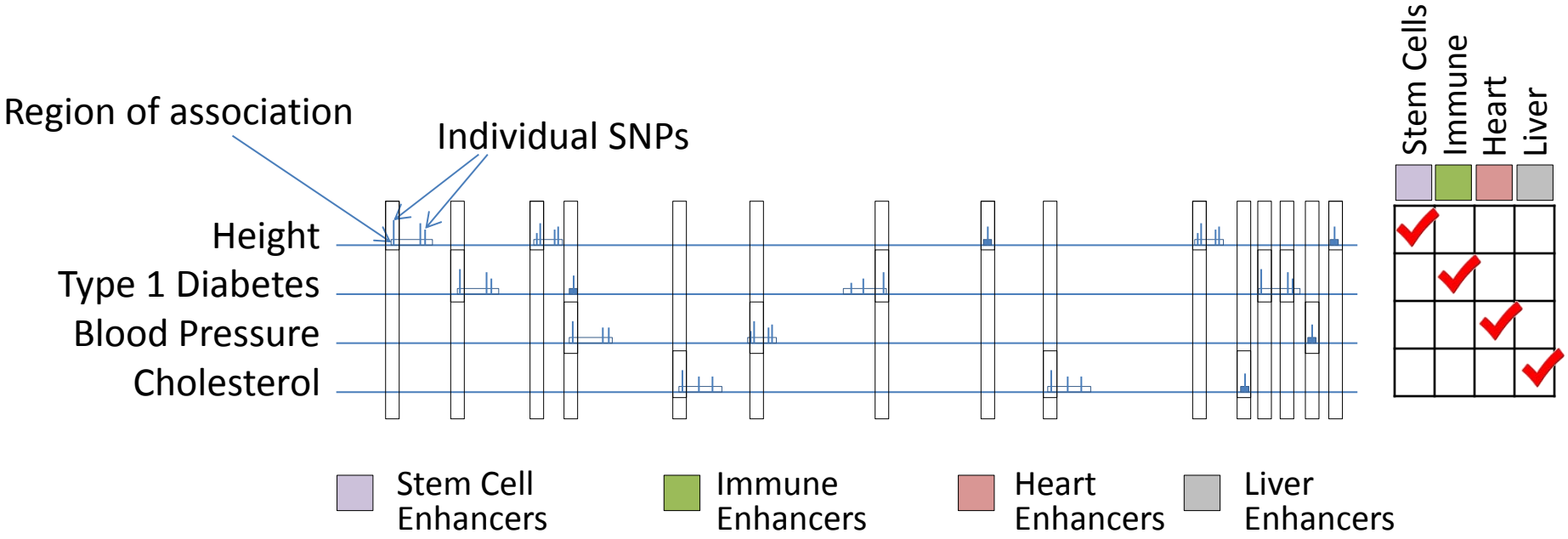
Dissecting non-coding genetic associations



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

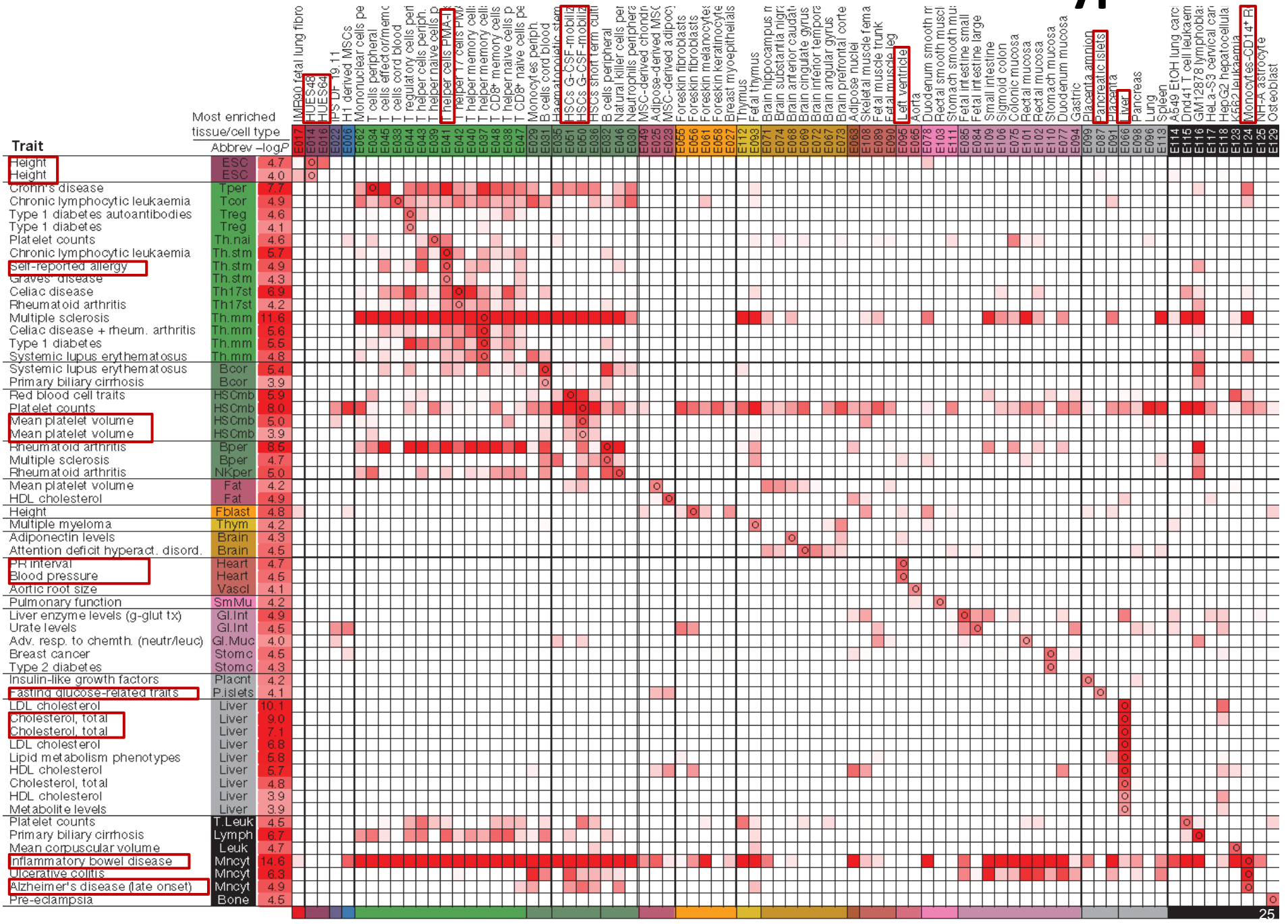
Using epigenomic maps to predict disease-relevant tissues

Identifying disease-relevant cell types

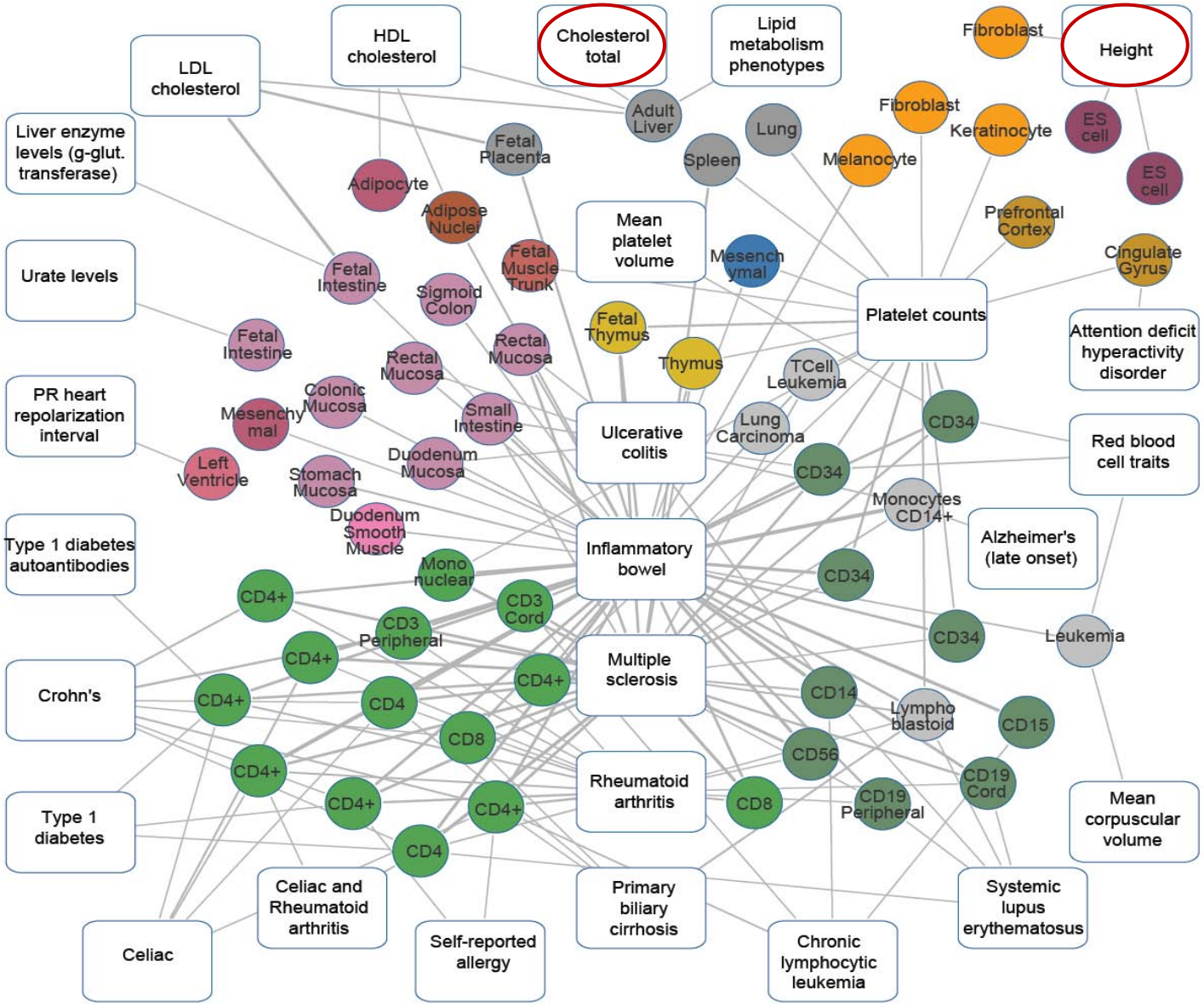


- For every trait in the GWAS catalog:
  - Identify all associated regions at P-value threshold
  - Consider all SNPs in credible interval ($R^2 \geq .8$)
  - Evaluate overlap with tissue-specific enhancers
  - Keep tissues showing significant enrichment ($P<0.001$)
- Repeat for all traits (rows) and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

GWAS hits in enhancers of relevant cell types



Linking traits to their relevant cell/tissue types



LETTER

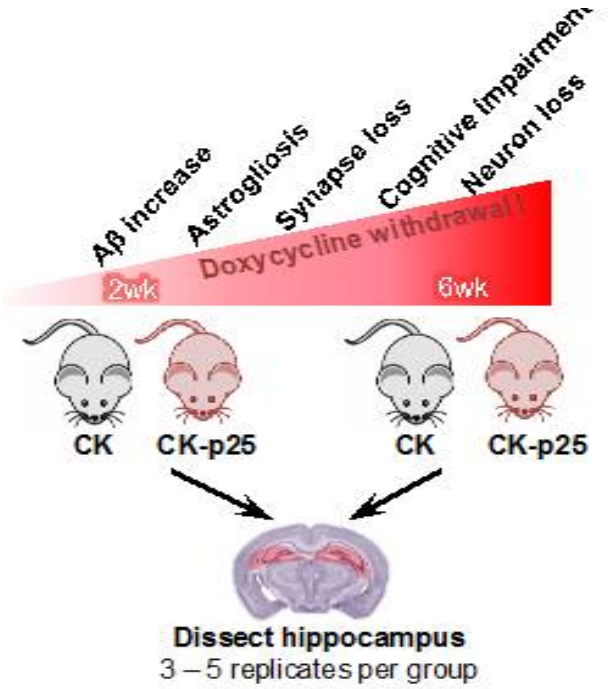
OPEN

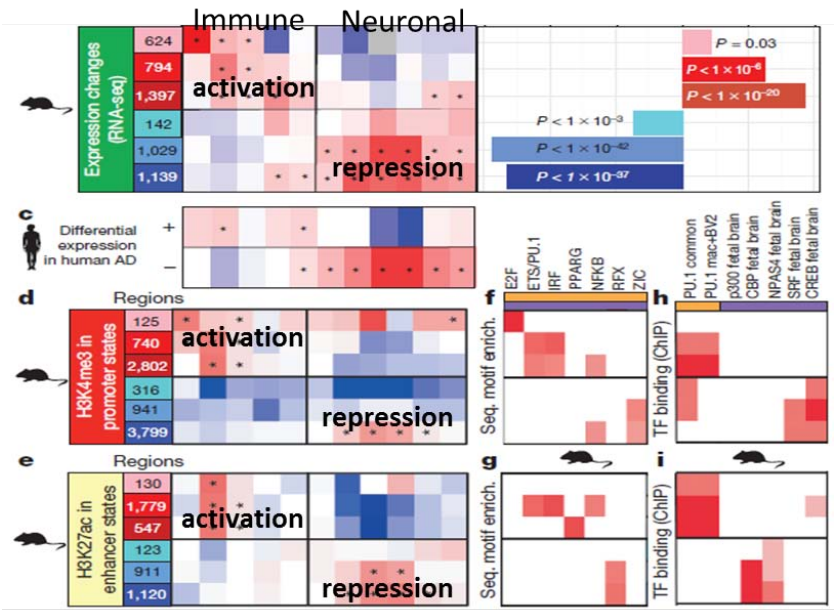
doi:10.1038/nature14252

Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease

Elizabeta Gjoneska[1,2]*, Andreas R. Pfenning[2,3]*, Hansruedi Mathys[1], Gerald Quon[2,3], Anshul Kundaje[2,3,4], Li-Huei Tsai[1,2]§ & Manolis Kellis[2,3]§

Immune activation + neural repression in human + mouse



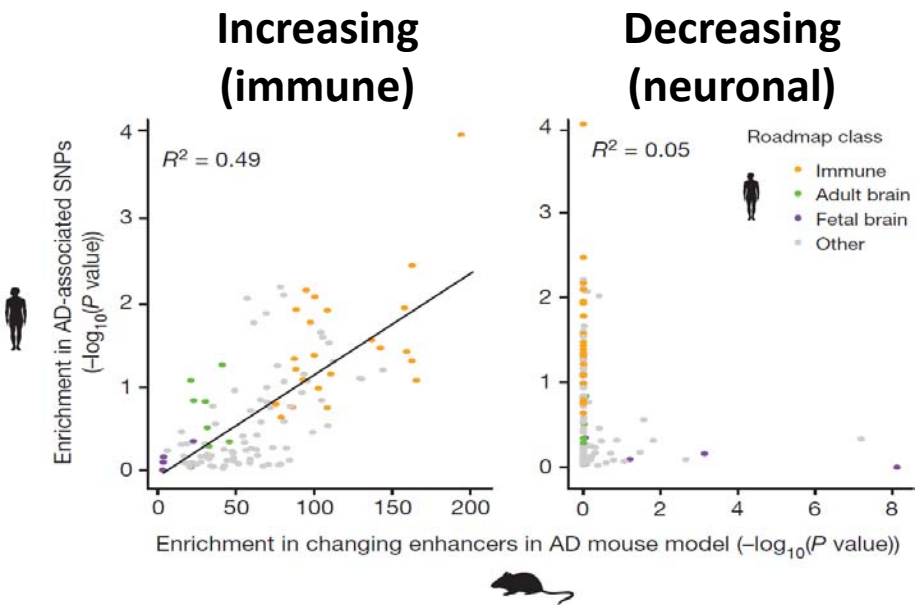


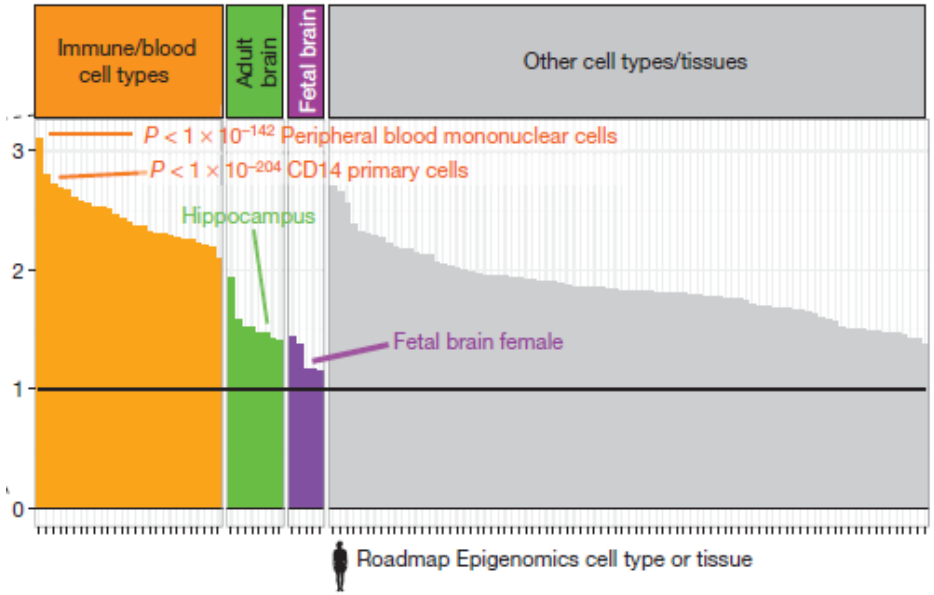
Source: Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G.,Kundaje, A., Tsai, L., & Kellis, M. (2015). "Conserved Epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease." Nature, 518 (7539), 365-369. doi:10.1038/nature14252

Sample mouse brain epigenomics during neurodegeneration

Two contrasting signatures of immune activation vs. neural repression

Genetic evidence for immune vs. neuronal components





Source: Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L., & Kellis, M. (2015). "Conserved Epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease." Nature, 518(7539), 365-369. doi:10.1038/nature14252

Only increasing (immune) enhancers enriched in AD-associated SNPs

Neuronal cell types are depleted for AD-associated SNPs

Indicates immune cell dysregulation is <u>causal</u> component

Microglial cells: resident immune cells of adult brain

Macrophages: infiltrate brain in neurodegeneration

Using epigenomic annotations for fine-mapping disease regions

LD: both a blessing & a curse

Linkage Equilibrium

$r^2 = 0$

0.6	A	B	0.75
0.2	A	b	0.05
0.15	a	B	0
0.05	a	b	0.2

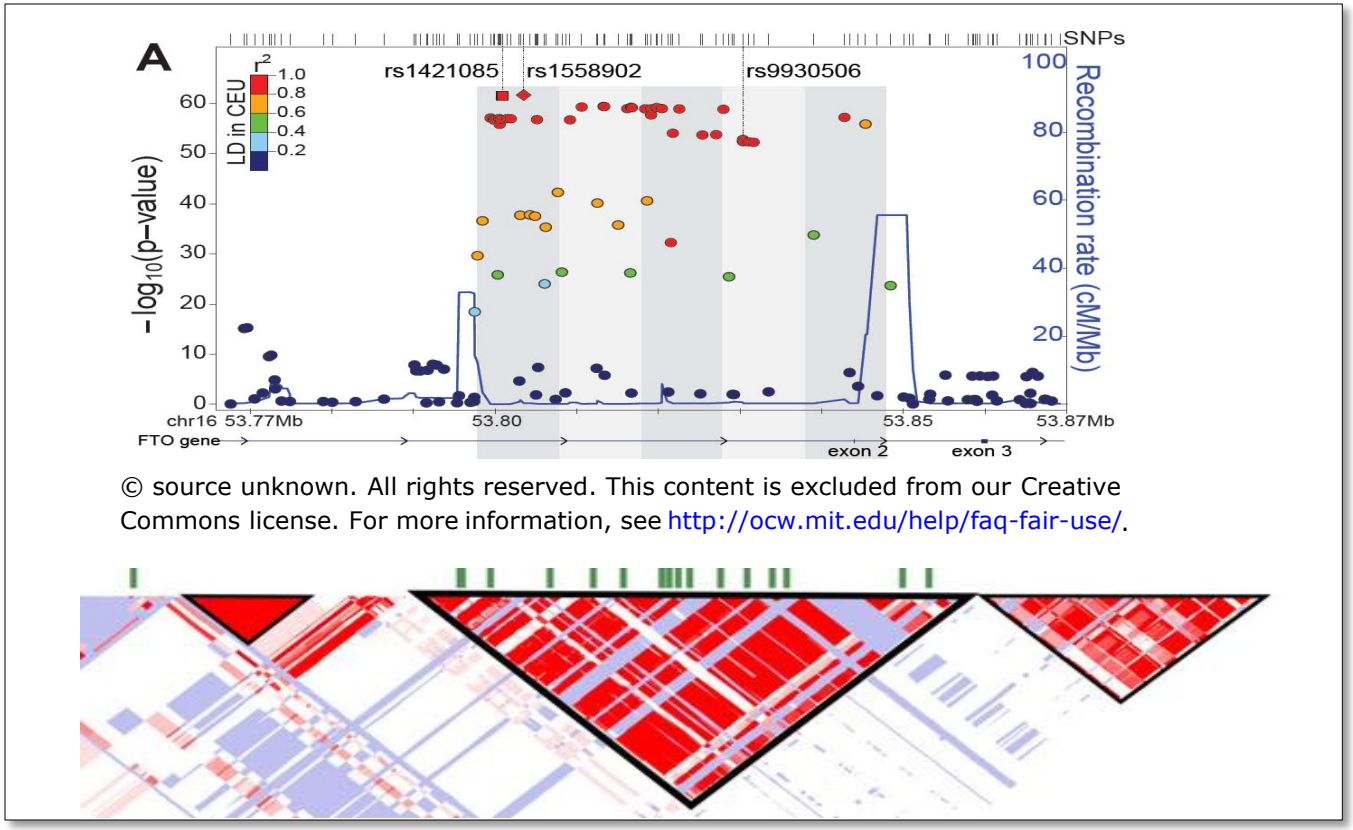
Linkage Disequilibrium

$r^2 = 0.75$

$$pA = 0.8, pa = 0.2, pB = 0.75, pb = 0.25$$

Observation: LD blocks in which there is no evidence for historical recombination

Causal variant not known in most GWAS regions

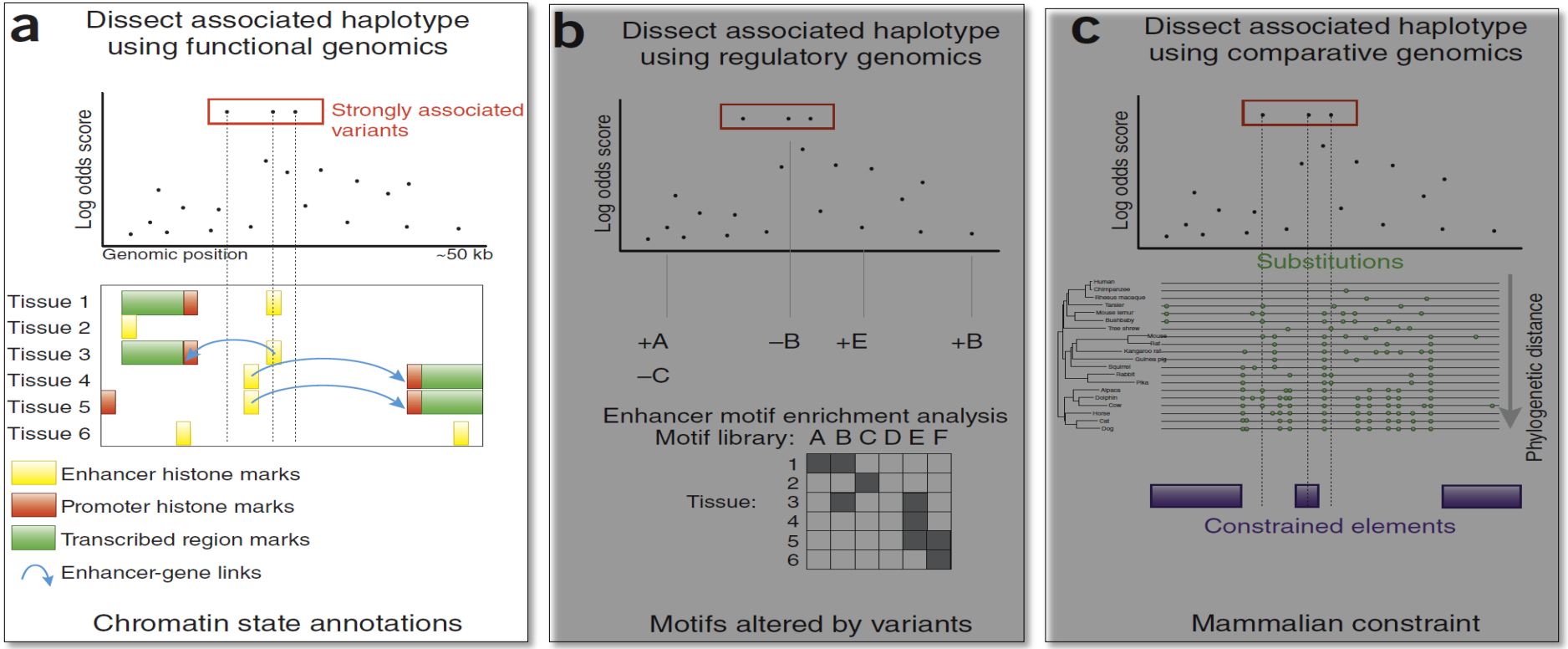


Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Smemo, S., Tena, J. J., Kim, K., Gamazon, E. R., Sakabe, N. J.,Gómez-Marín, C., . . . Nóbrega, M. A. (2014). "Obesity-associated variants within FTO form long-range functional connections with IRX3." Nature, 507(7492), 371-375. doi:10.1038/nature13138

LD (Linkage disequilibrium): large regions co-inherited in blocks
Blessing for initial mapping (few tags), curse for fine-mapping

Use functional annotations to predict causal variant(s)

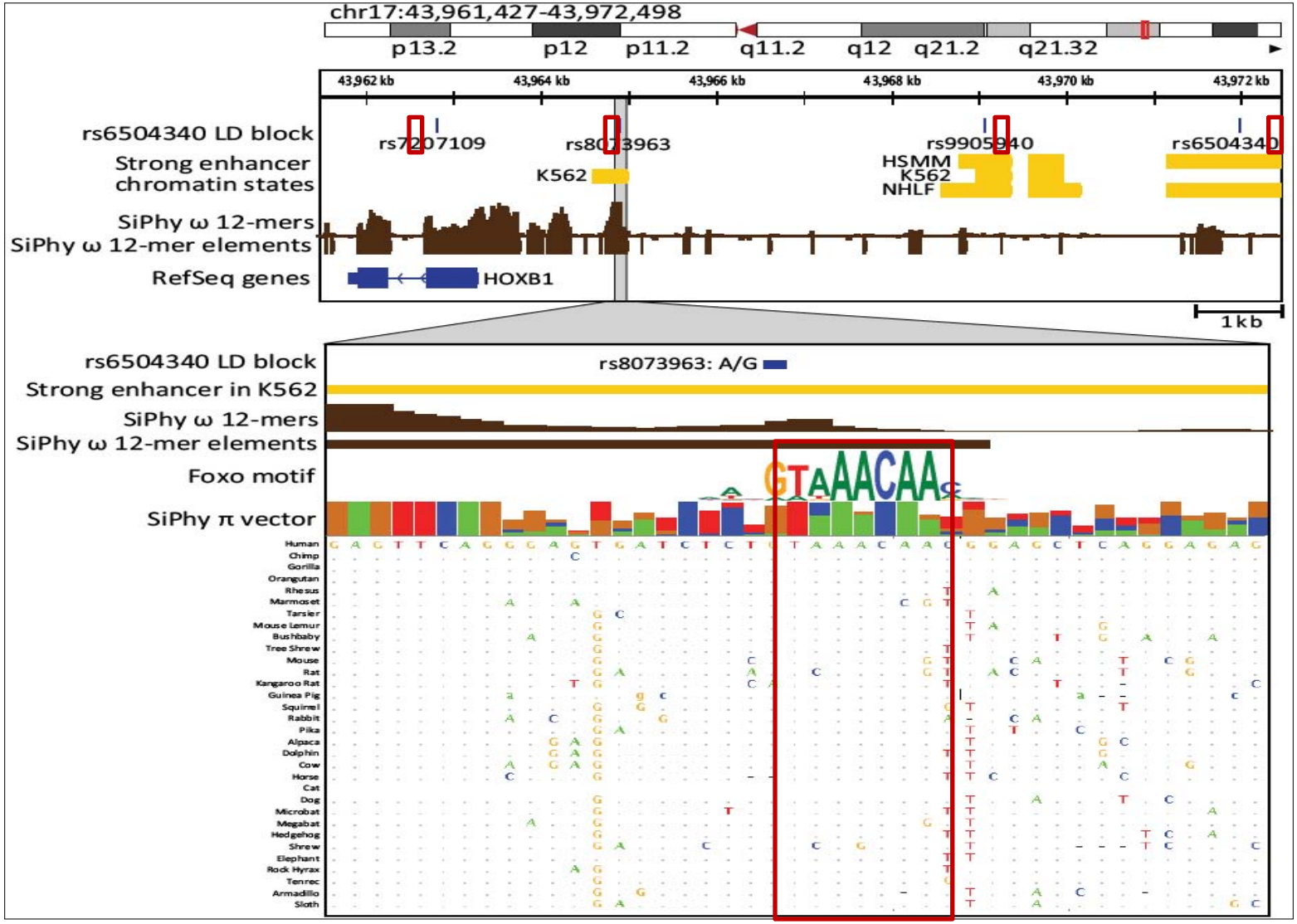
Multiple lines of evidence for fine-mapping



Ward and Kellis, Nature Biotechnology 2012

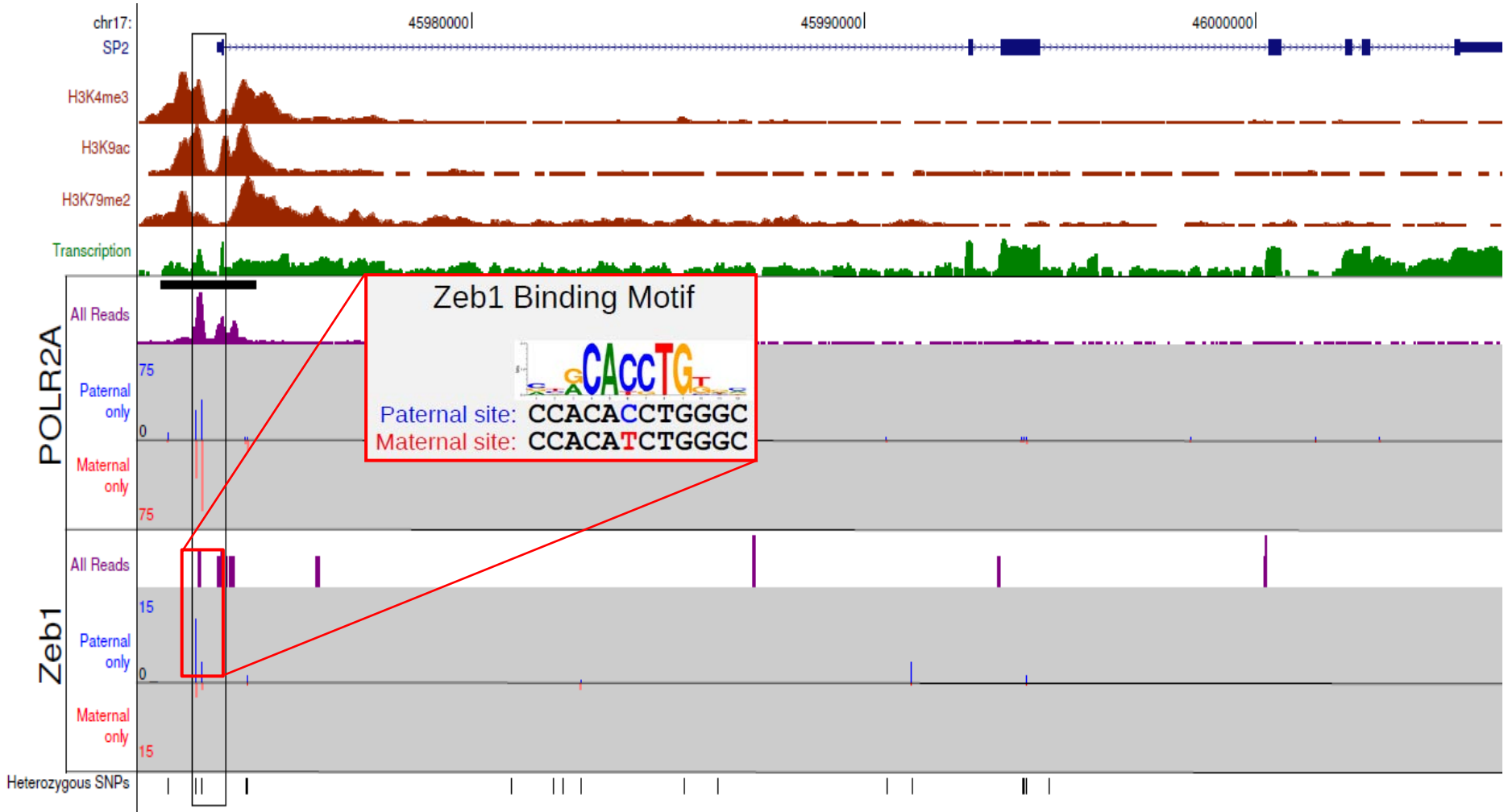
- Epigenomic information: enhancers & linking (target genes)
- Motif information: causal variants & upstream regulators
- Evolutionary conservation: causal variants & conserved motifs

Detect SNPs that disrupt conserved regulatory motifs



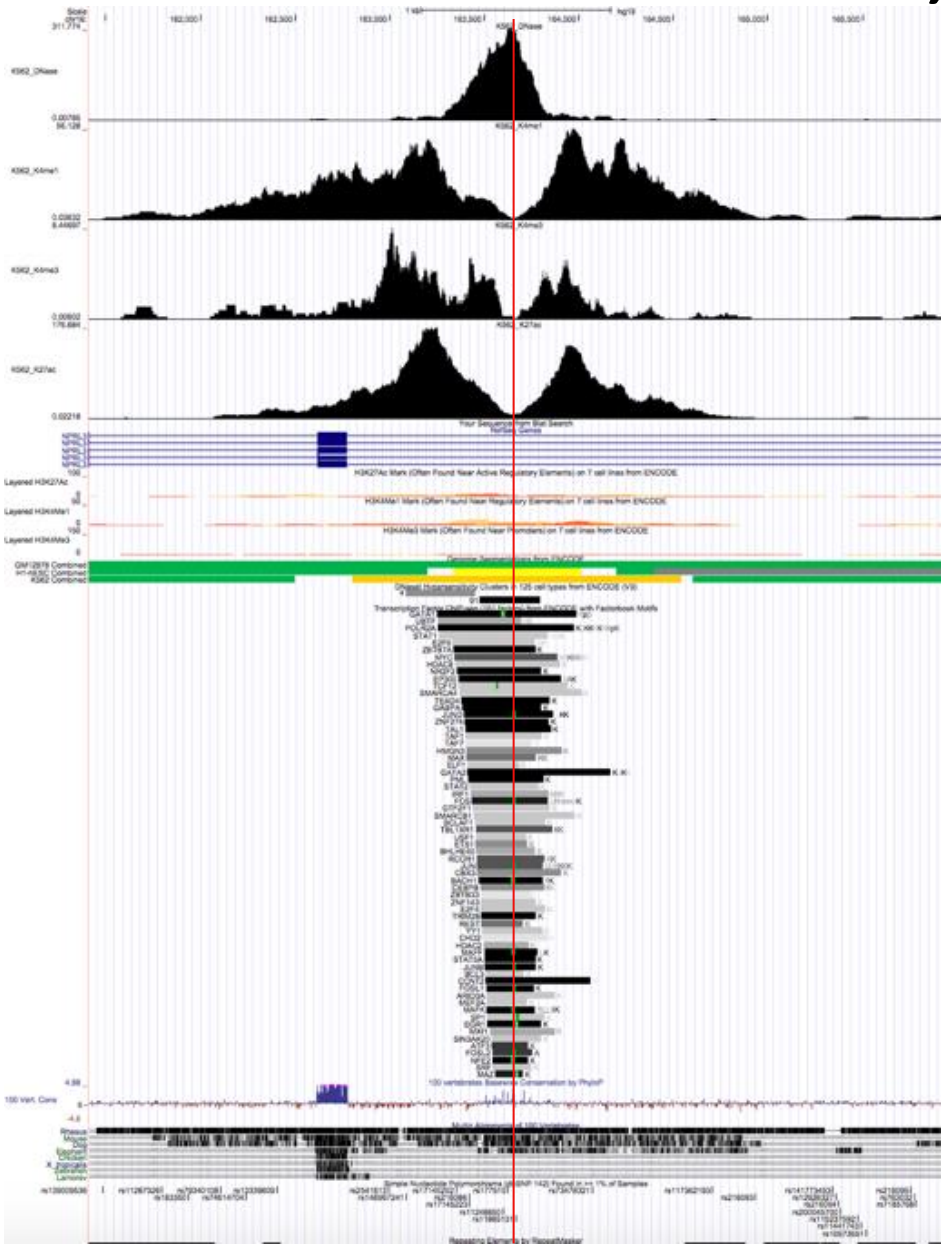
- Functionally-associated SNPs enriched in states, constraint

Allele-specific chromatin marks: cis-vs-trans effects

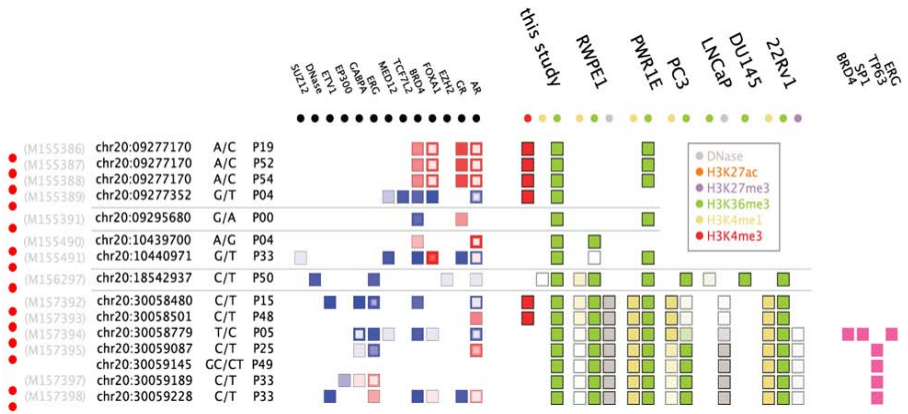


- **Maternal and paternal GM12878 genomes sequenced**
- **Map reads to phased genome, handle SNPs indels**
- **Correlate activity changes with sequence differences**

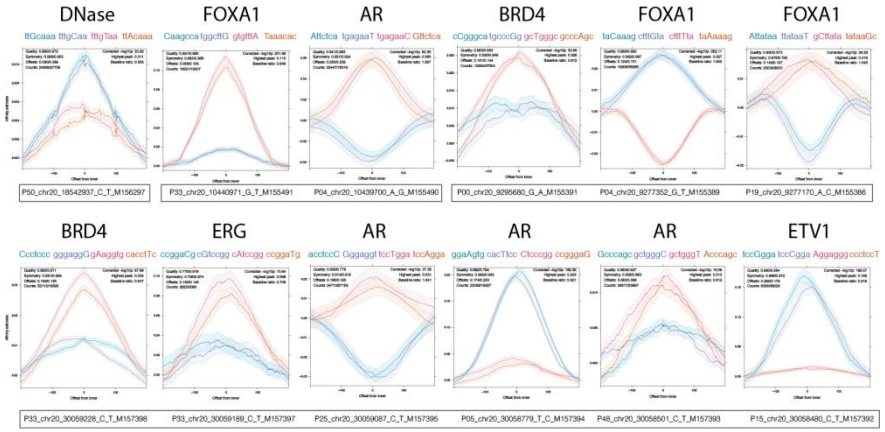
Predict effect of common, rare, somatic mutations



Common: allelic activity in heterozygous lines



All: Regulatory and epigenomic annotations



Rare/somatic: Predict TF binding disruption

Richard Sallari
Xinchen Wang

HaploReg: public resource for dissecting GWAS

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
chr3:12366308	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 organ						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
chr3:12396645	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

Source: Ward, Lucas D. and Manolis Kellis. "HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants." Nucleic Acids Research 40, no. D1 (2012): D930-D934.

- **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**

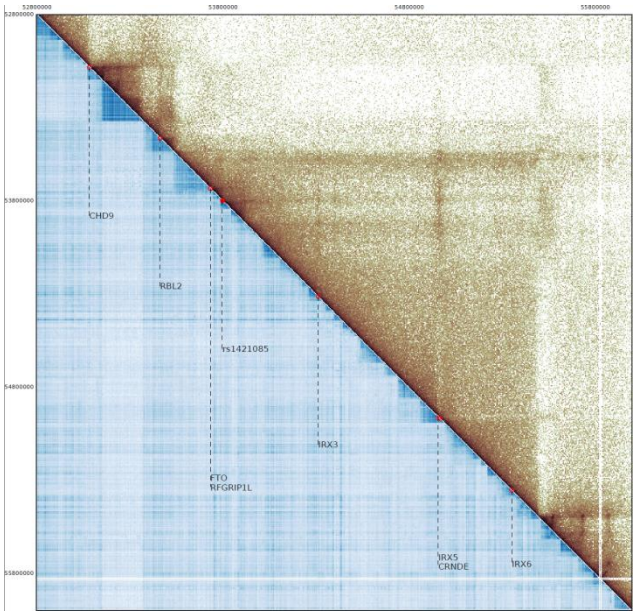
Ward, Kellis NAR 2011

37

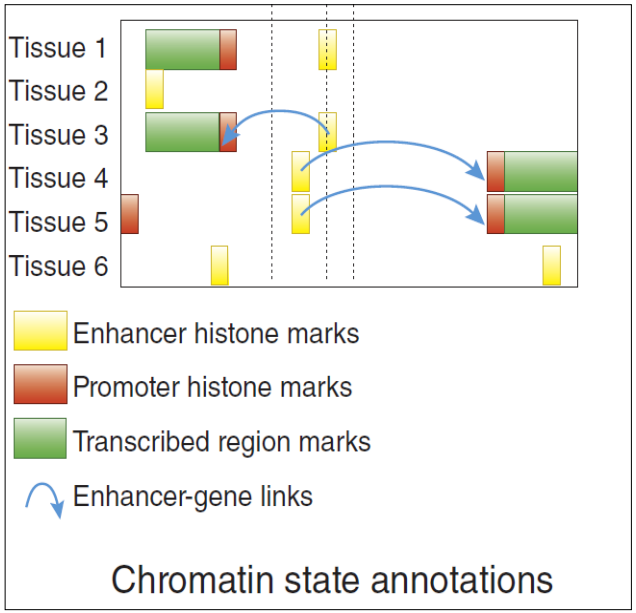
Predicting target genes

Three lines of linking evidence

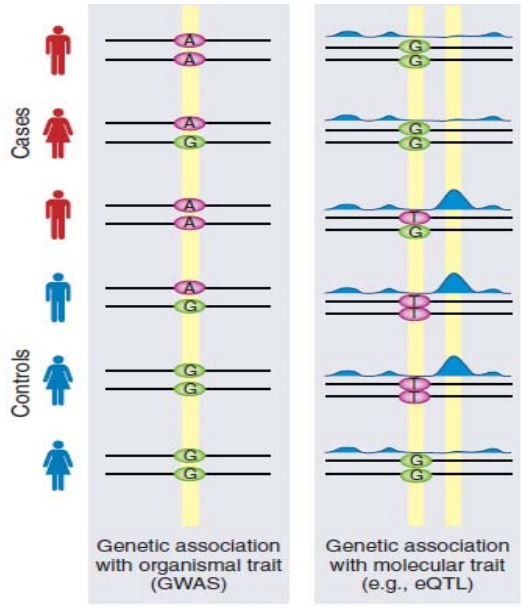
Physical



Functional



Genetic



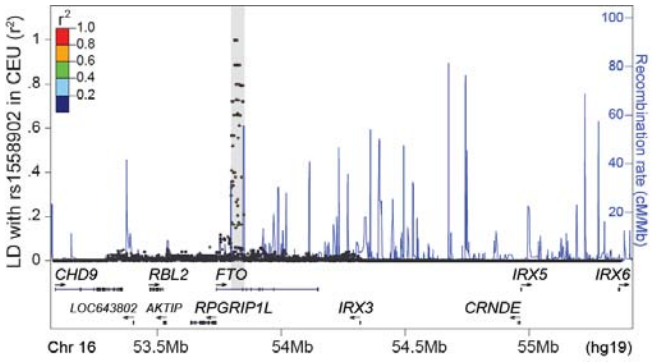
Courtesy of Macmillan Publishers Limited. Used with permission. Ward, L. D., & Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol Nature Biotechnology, 30(11), 1095-1106. doi:10.1038/nbt.2422. Used with permission.

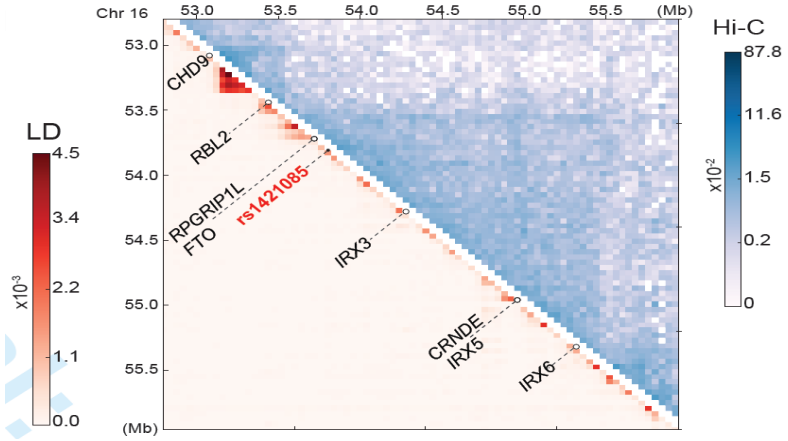
Hi-C: Physical proximity in 3D

Enhancer-gene activity correlation

eQTL evidence: SNP effect on expression

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

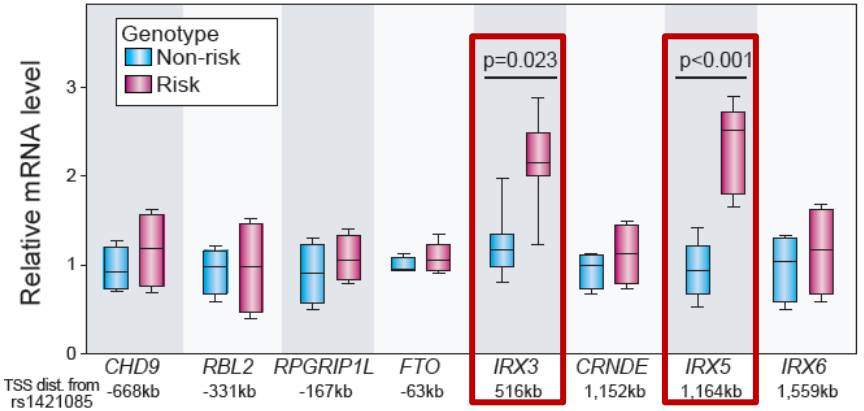




Dixon, Nature 2012

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

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



eQTL targets: IRX3 and IRX5

Risk allele: increased expression (gain-of-function)

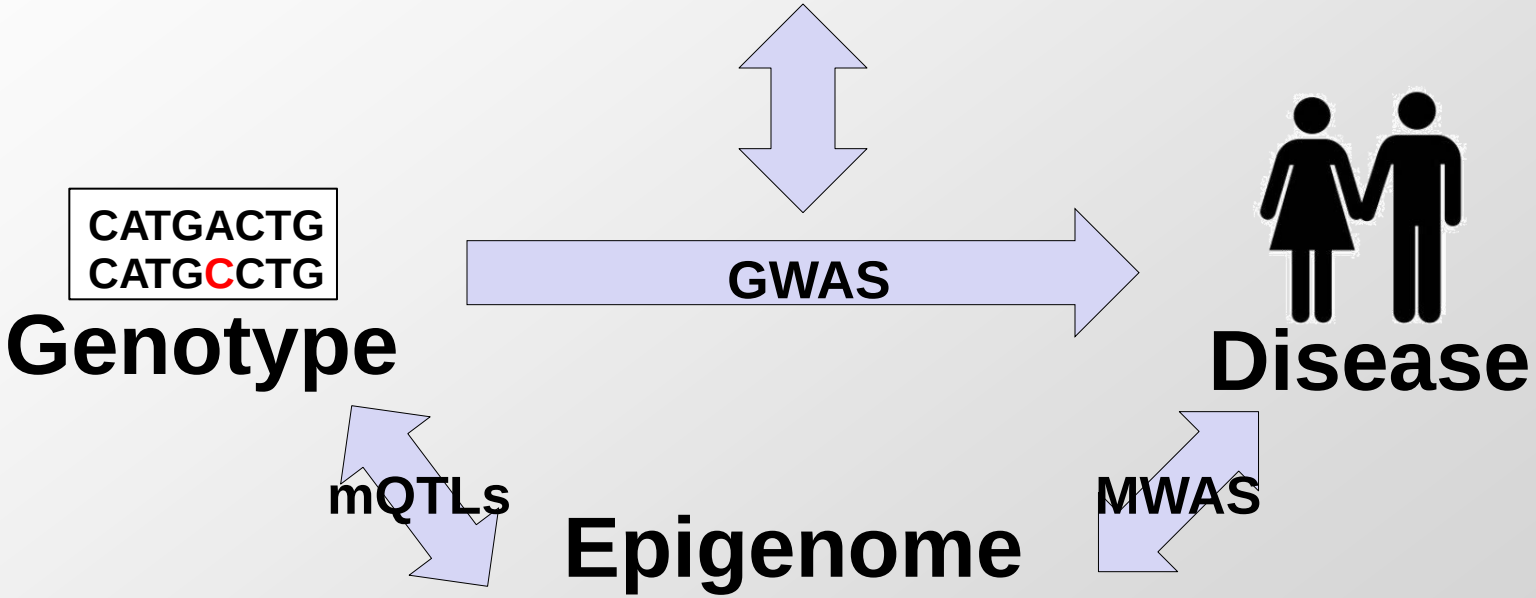
Goal: Personalized and Predictive Medicine

1. Intro to Epidemiology: basis of human disease
2. Genetic Epidemiology:
   - Genetic basis: GWAS and screening
   - Interpreting GWAS with functional genomics
   - Calculating functional enrichments for GWAS loci
3. Molecular epidemiology
   - meQTLs: Genotype-Epigenome association (cis-/trans-)
   - EWAS: Epigenome-Disease association
4. Resolving Causality
   - Statistical: Mendelian Randomization
   - Application to genotype + methylation in AD
5. Systems Genomics and Epigenomics of disease
   - Beyond single loci: polygenic risk prediction models
   - Sub-threshold loci and somatic heterogeneity in cancer

Interpreting disease-association signals

(1) Interpret variants using Epigenomics

- Chromatin states: Enhancers, promoters, motifs
- Enrichment in individual loci, across 1000s of SNPs in T1D

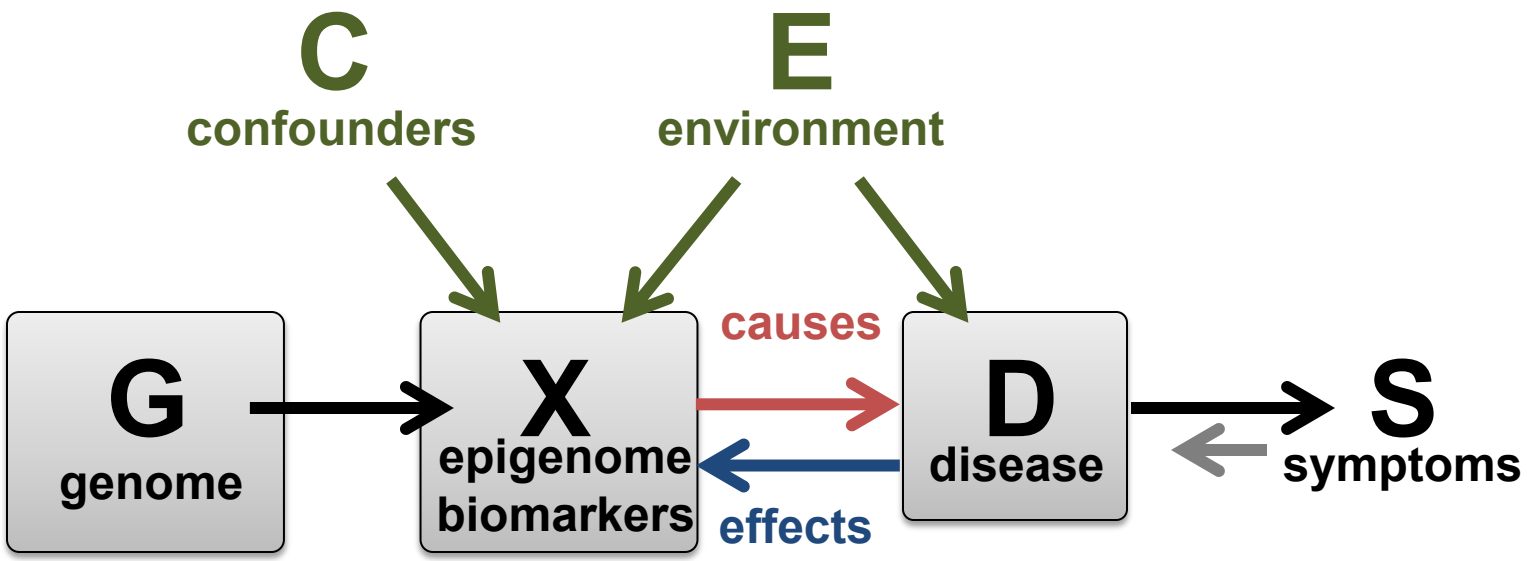


(2) Epigenome changes in disease

- Intermediate molecular phenotypes associated with disease
- Variation in brain methylomes of Alzheimer's patients

Goal: Personalized and Predictive Medicine

1. Intro to Epidemiology: basis of human disease
2. Genetic Epidemiology:
   - Genetic basis: GWAS and screening
   - Interpreting GWAS with functional genomics
   - Calculating functional enrichments for GWAS loci
3. Molecular epidemiology
   - meQTLs: Genotype-Epigenome association (cis-/trans-)
   - EWAS: Epigenome-Disease association
4. Resolving Causality
   - Statistical: Mendelian Randomization
   - Application to genotype + methylation in AD
5. Systems Genomics and Epigenomics of disease
   - Beyond single loci: polygenic risk prediction models
   - Sub-threshold loci and somatic heterogeneity in cancer



Molecular Epidemiology

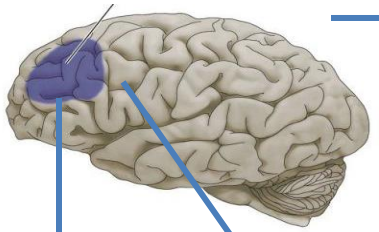
Molecular Biomarkers of disease state:

Gene expression, DNA methylation, chromatin in specific cell types

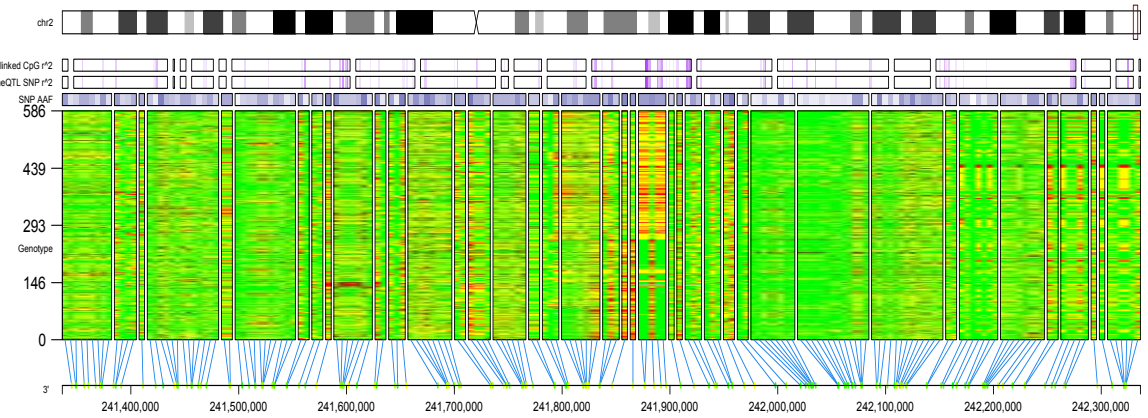
Genetic and epigenetic data in 750 Alzheimer's patients/controls

MAP Memory and Aging Project
+ ROS Religious Order Study

Dorsolateral PFC

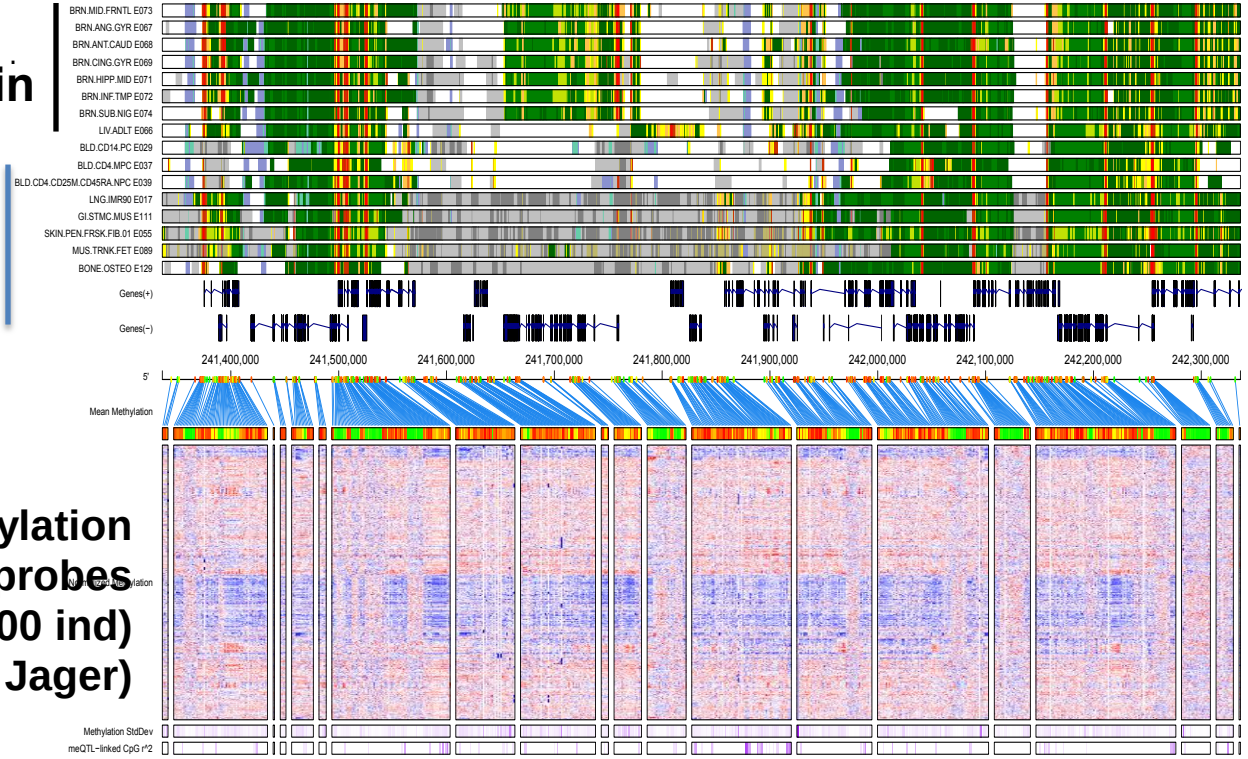


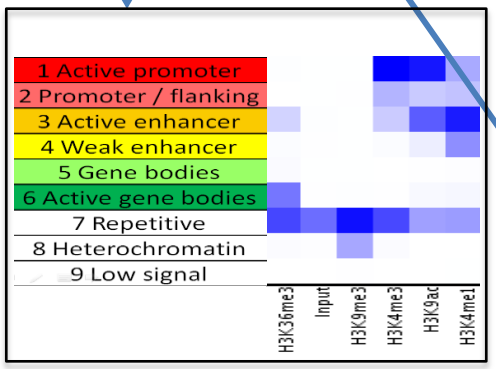
Genotype (1M SNPs x700 ind.) (De Jager)



Brain

Liver
Blood
Lung
GI
Skin
Muscle
Bone



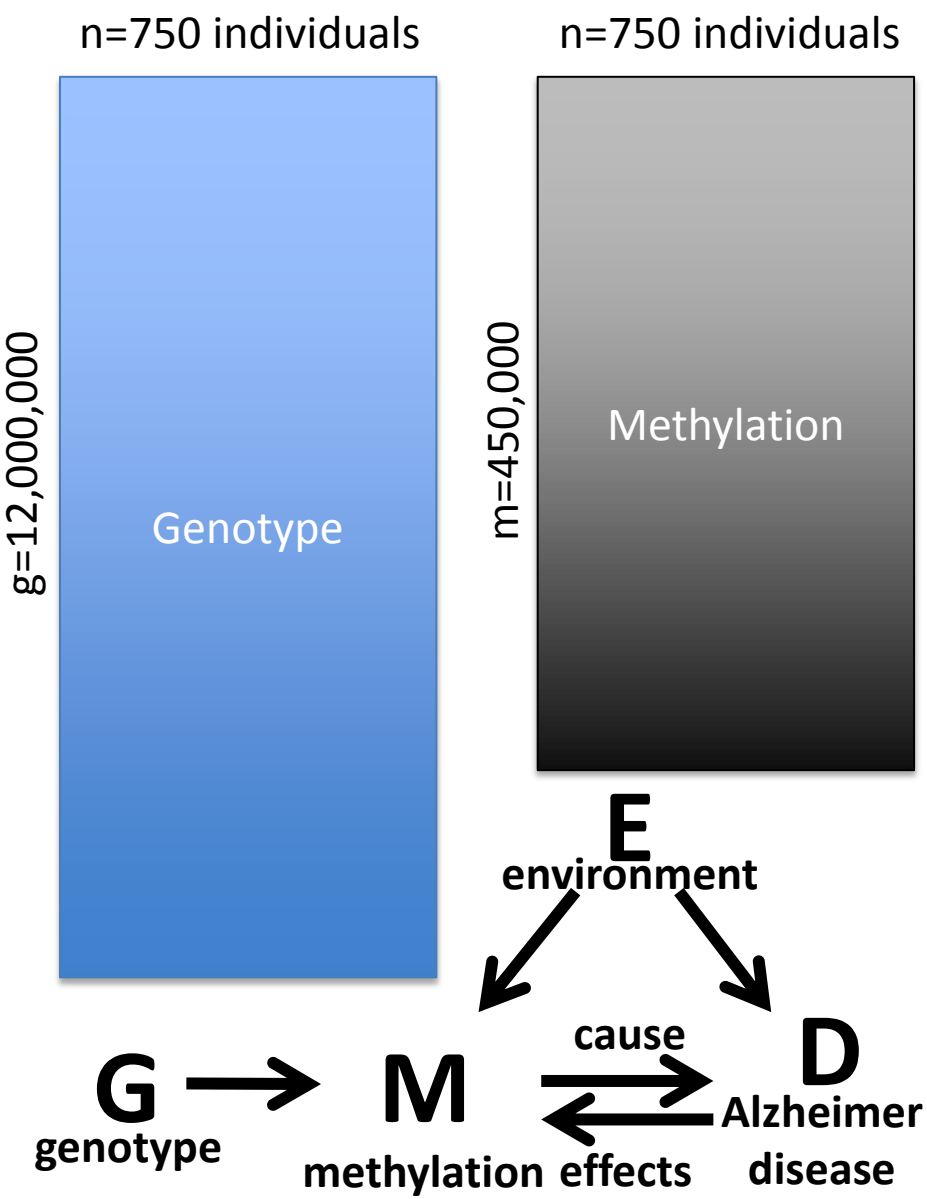


Reference Chromatin states (Bernstein)

Methylation (450k probes x 700 ind) (De Jager)

750 subjects, initially cognitively normal, Alzheimer's diagnosed by pathology. (Bennett)

Data Matrices – An example scenario



M - Illumina Methylation 450k array, 450,000 probes targeting CpGs genome-wide.

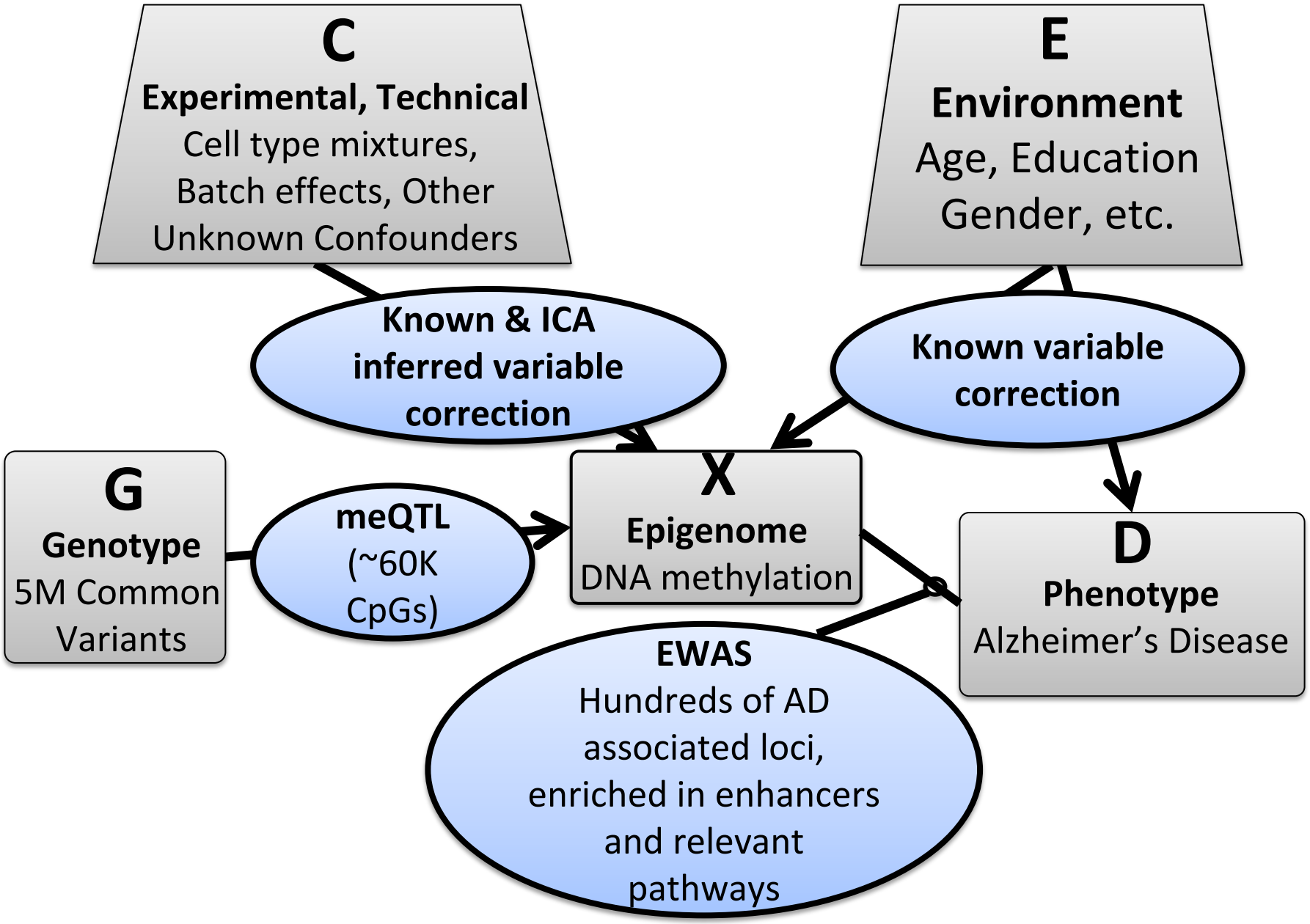
G - Affy SNP arrays, imputed against CEU thousand genomes reference panel, yielding 12m SNPs.

E - Clinical covariates that might mask the variation due to our phenotype, e.g. gender, smoking, age or sample batch.

P - Phenotype of interest, sometimes measured with multiple markers (clinical Alz. diagnosis vs. pathology Alz. diagnosis vs. count of neuritic plaques).

n -> number of individuals in cohort.

EWAS: Capturing variability in the Epigenome attributable to disease



Excluding discovered and known covariates

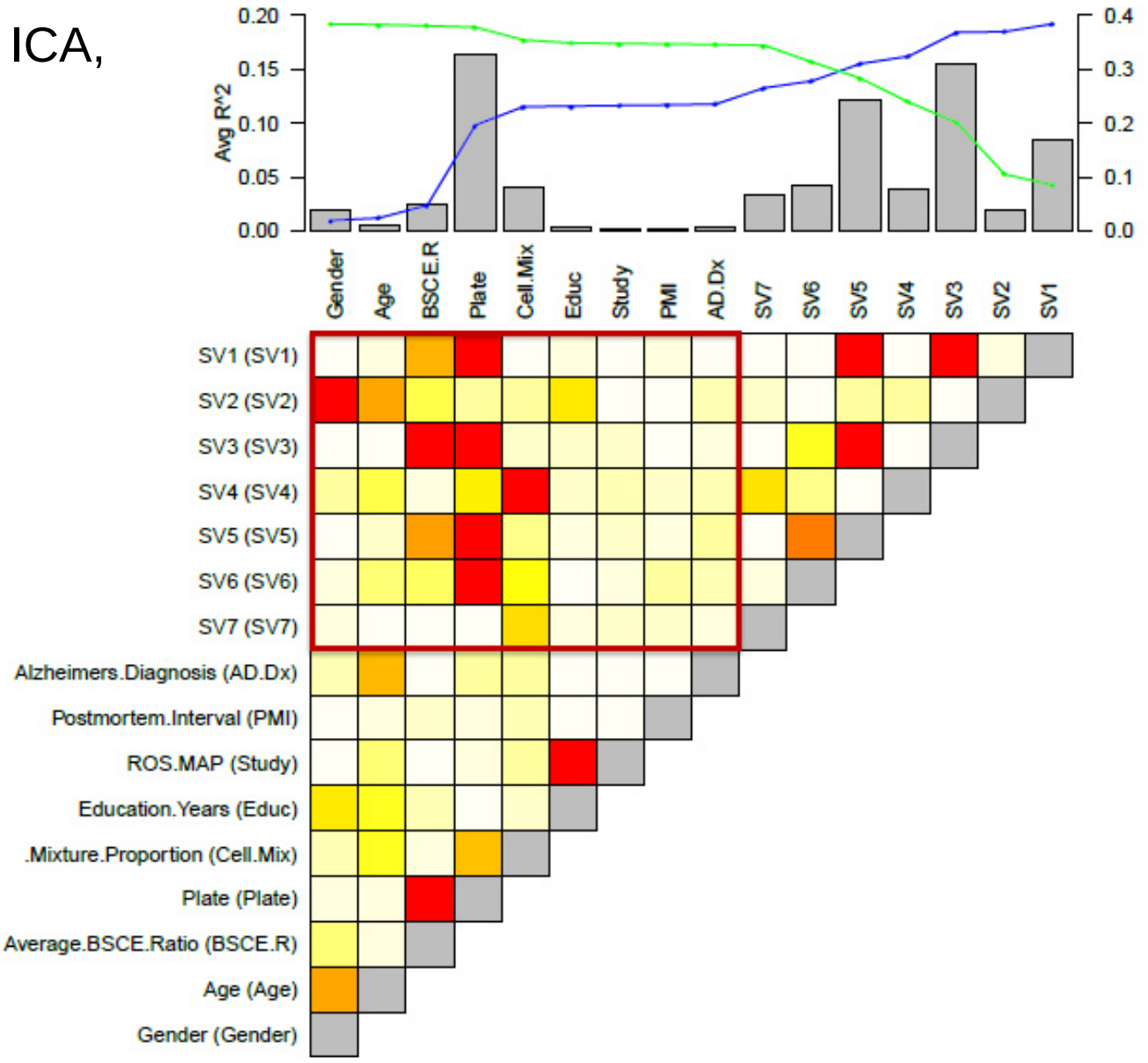
Infer covariates using ICA, compare to known, exclude both.

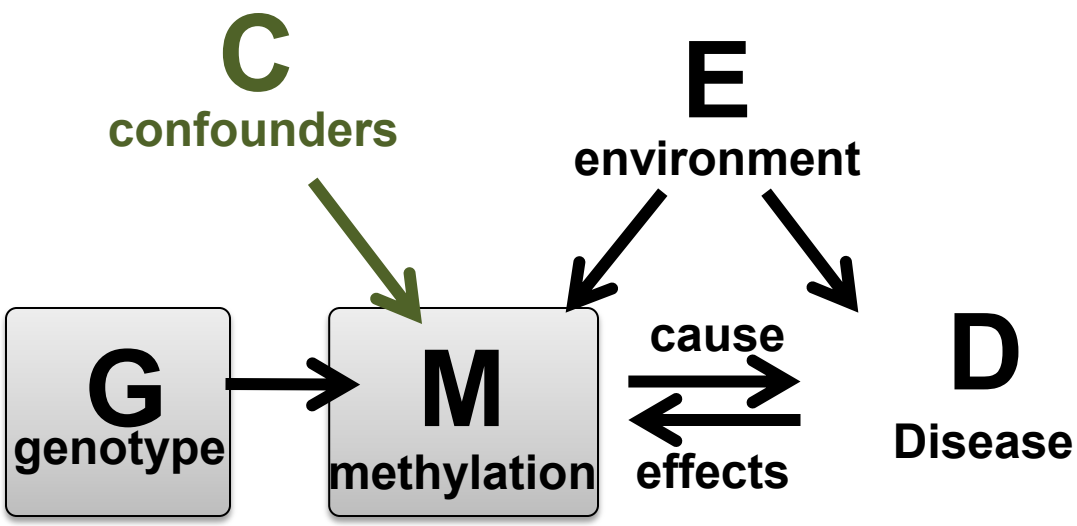
Strongest effects:

- **Plate (batch)**
- **Cell mixture**
- Bisulfite conversion
- Gender
- Age

Variance explained:

- Known: 25%
- Inferred: 35%
- Together: 40%



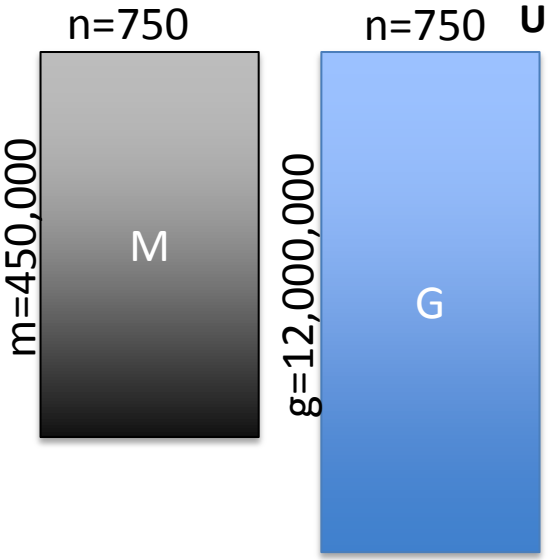


Genotype→Methylation

Discovering mQTLs

Methylation Quantitative Trait Loci

cis-meQTLs



Use linear models to identify *cis*-meQTLs w/in some genomic window.

For methyl mark m_i and SNP g_j:
$m_i = \beta 0 + \beta 1(g_j) + \varepsilon$

- Given several predictors: is additional predictor increasing accuracy more than complexity introduced?
- Likelihood ratio testing paradigm: predict methylation with and without genotype (only works for nested models)

- Null hypothesis H_0: $\beta 1=0$: Additional model complexity doesn't explain a significant portion of variation in response

LM1: $m_i = \beta 0 + \varepsilon$
LM2: $m_i = \beta 0 + \beta 1(g_j) + \varepsilon$

<u>Test using F statistic:</u>

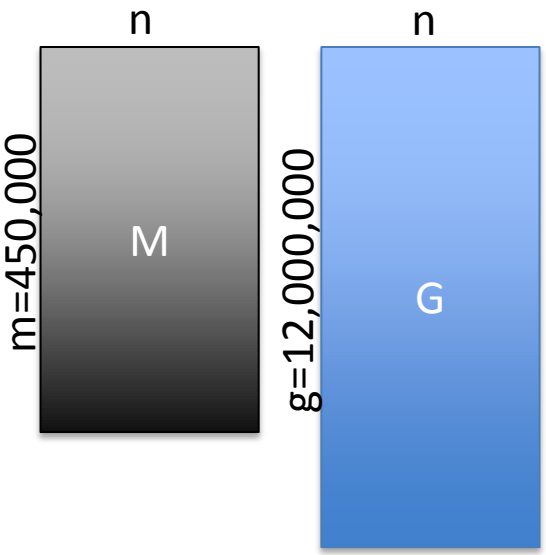
- p is the number of parameters in LM1
- q is the number of parameters in LM2
- n is the sample size
- RSS: Residual sum of squares
- β: parameters to learn. ε: residual error term.

Under null hypothesis: $((RSS_{LM1} - RSS_{LM2}) / (q - p)) / (RSS_{LM2} / (n - q))$
Is distributed as F distribution with (q-p, n-q) degrees of freedom

➔ If F statistic significant: reject null: This p-value is what we report in a meQTL study

➔ Otherwise, no meQTL: i.e. $RSS_{LM1} - RSS_{LM2}$ too small vs. increase in model complexity

cis-meQTLs

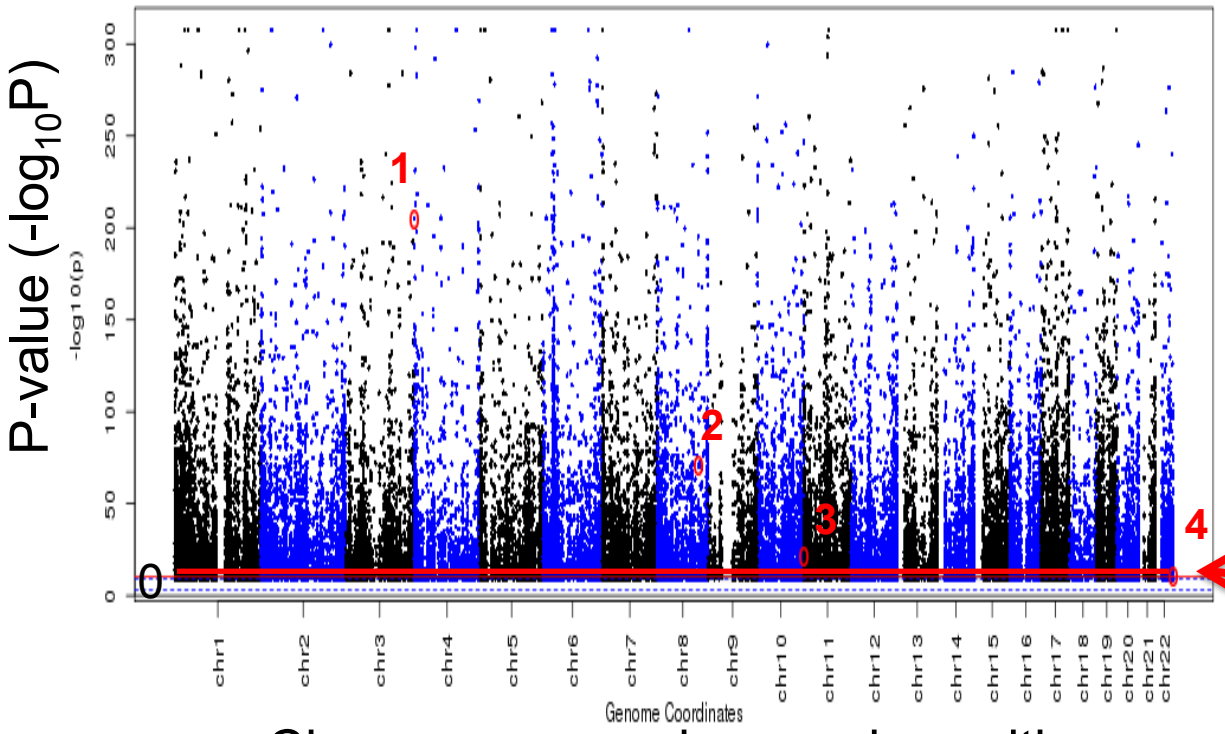


Alternative methods of detection:

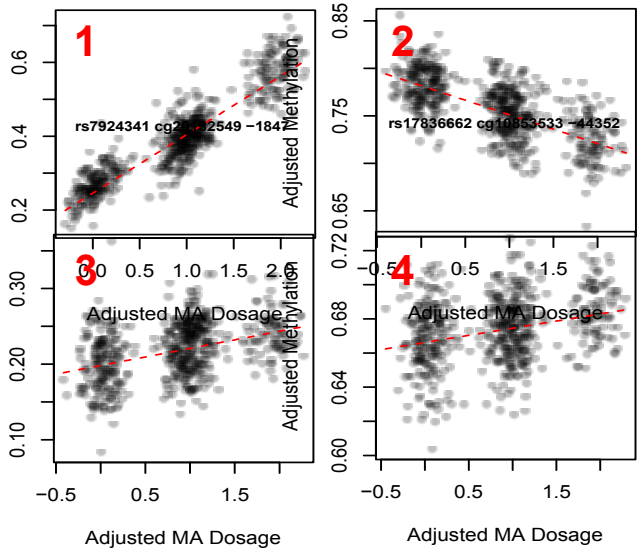
- Permutation:
  - Correlate methylation and genotype.
  - For i in 1 -> nperm:
    - Permute genotypes
    - Correlate methylation and genotype
  - Generate empirical p-value from permuted correlations
- LMM: Linear mixed models.

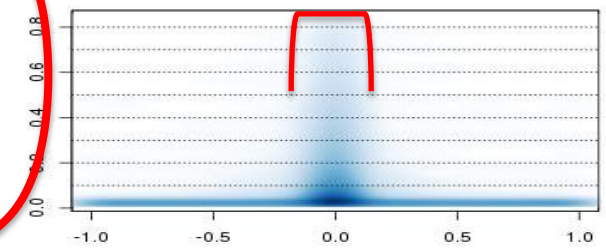
Most epigenomic variability is genotype-driven

Manhattan plot of 450,000 methylation probes



Chromosome and genomic position



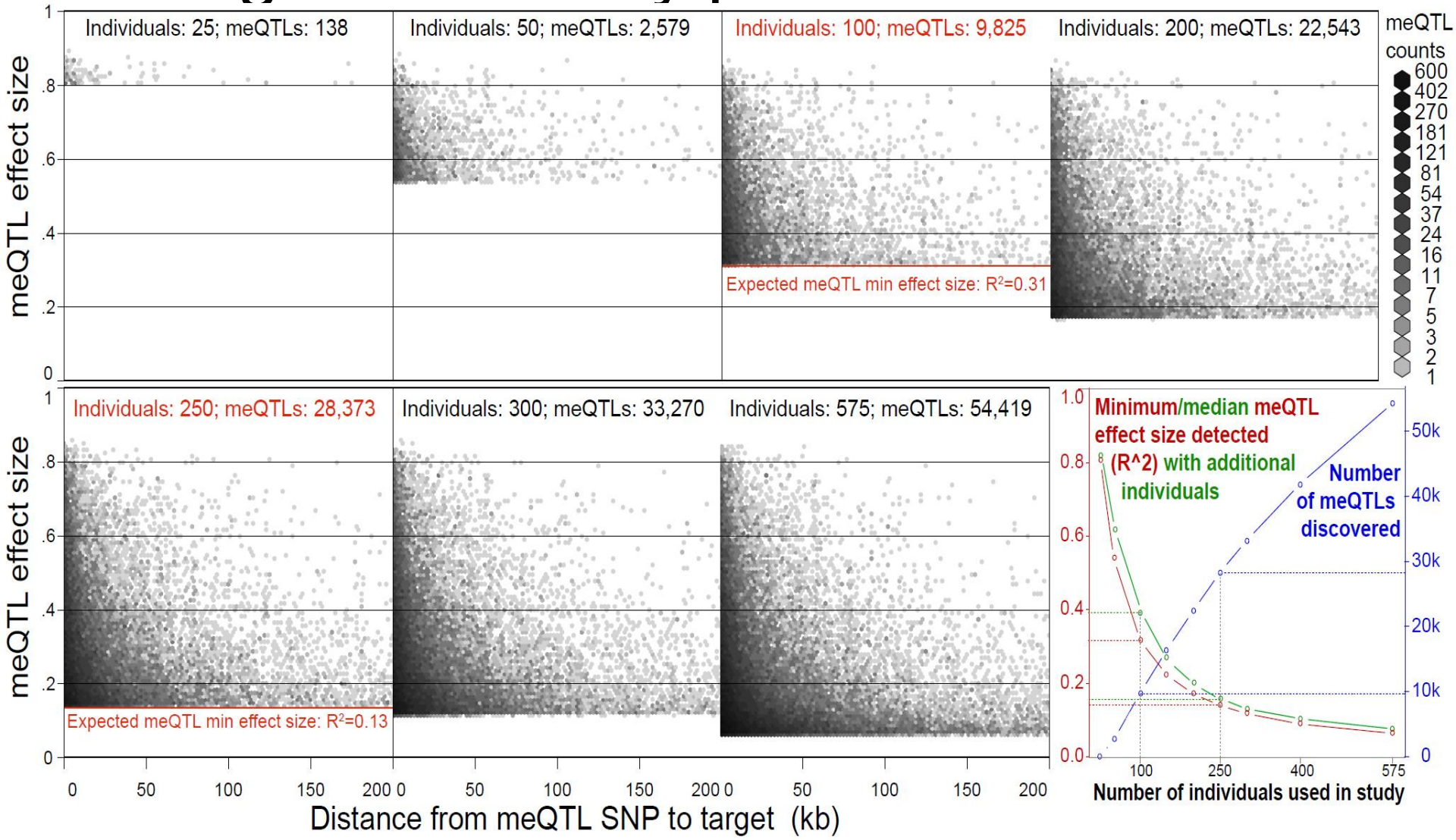


- Genome-wide significance at $p<3x10^{-10}$
- Prune for probes disrupted by SNP.

➔ 140,000 CpGs associated with genotype at 1% FDR

➔ 55,000 at Bonferroni-corrected P-value of 10^{-2}

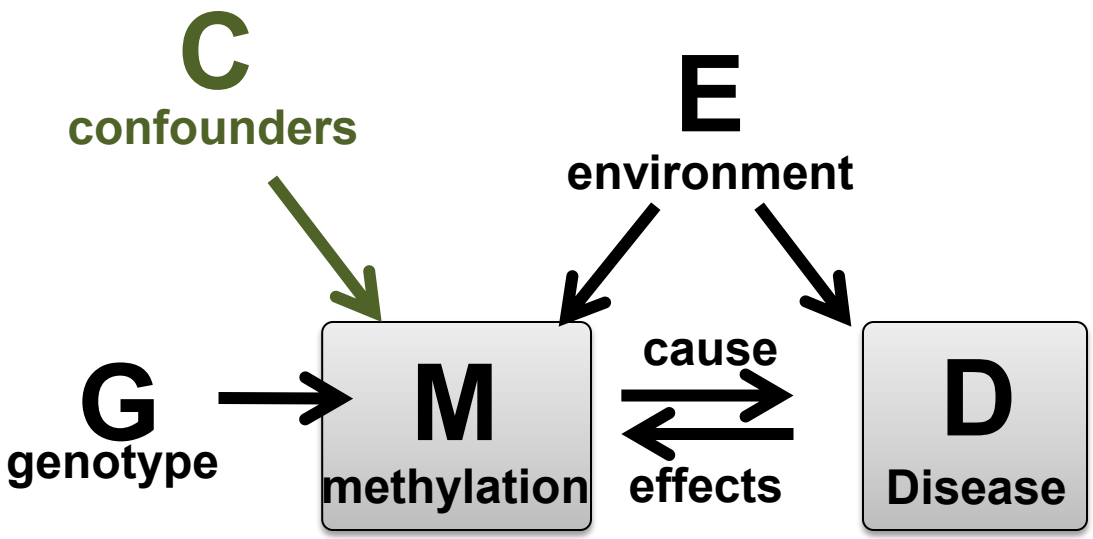
Scaling of discovery power with individuals



- Number of meQTLs continues to increase linearly
- Weak-effect meQTLs: median R^2<0.1 after 400 indiv.

Goal: Personalized and Predictive Medicine

1. Intro to Epidemiology: basis of human disease
2. Genetic Epidemiology:
   - Genetic basis: GWAS and screening
   - Interpreting GWAS with functional genomics
   - Calculating functional enrichments for GWAS loci
3. Molecular epidemiology
   - meQTLs: Genotype-Epigenome association (cis-/trans-)
   - EWAS: Epigenome-Disease association
4. Resolving Causality
   - Statistical: Mendelian Randomization
   - Application to genotype + methylation in AD
5. Systems Genomics and Epigenomics of disease
   - Beyond single loci: polygenic risk prediction models
   - Sub-threshold loci and somatic heterogeneity in cancer

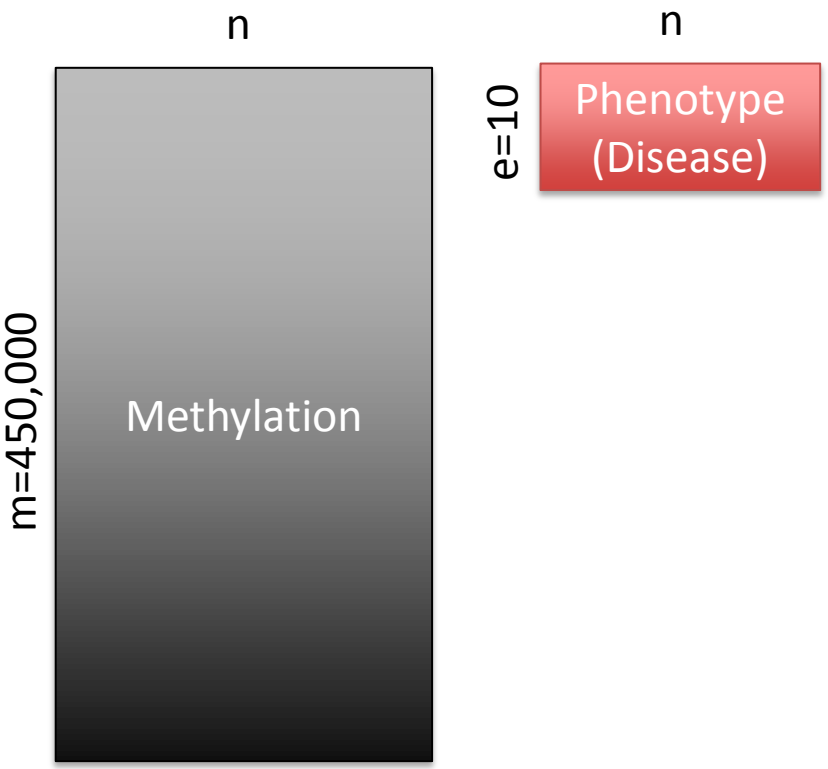


Methylation→Disease

EWAS

Epigenome-wide association study

eWAS

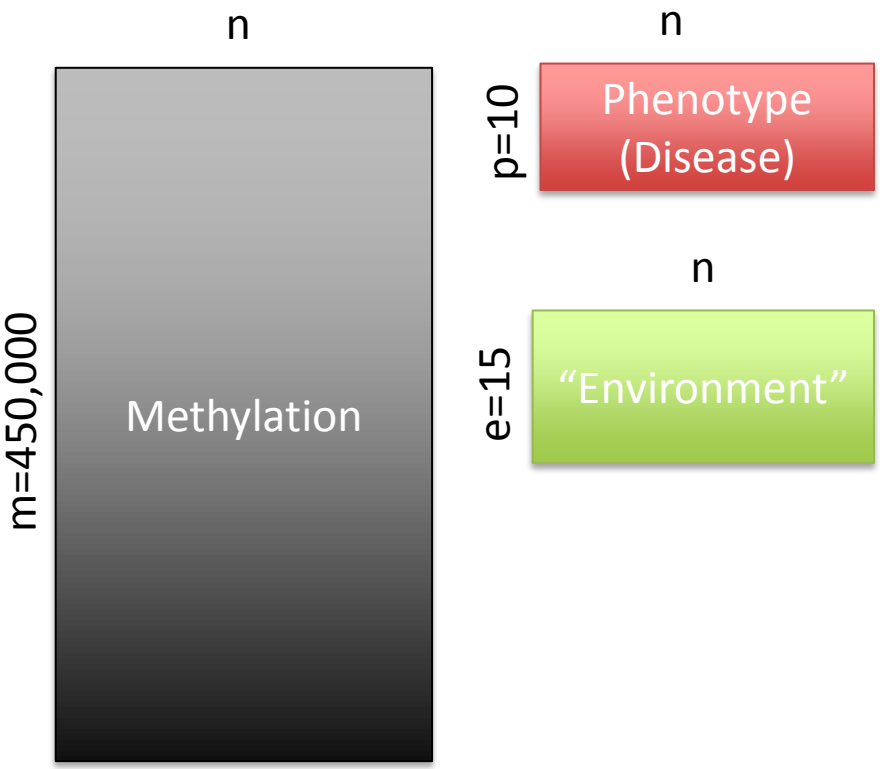


Link methylation⇔phenotype (~cis-eQTLs):

- linear models and hypothesis testing
- Predict phenotype using methylation

LM1: $p_i = \beta 0 + \varepsilon$
LM2: $p_i = \beta 0 + \beta 1(m_j) + \varepsilon$

eWAS



Link methylation⇔phenotype (~cis-eQTLs):
- linear models and hypothesis testing
- Predict phenotype using methylation

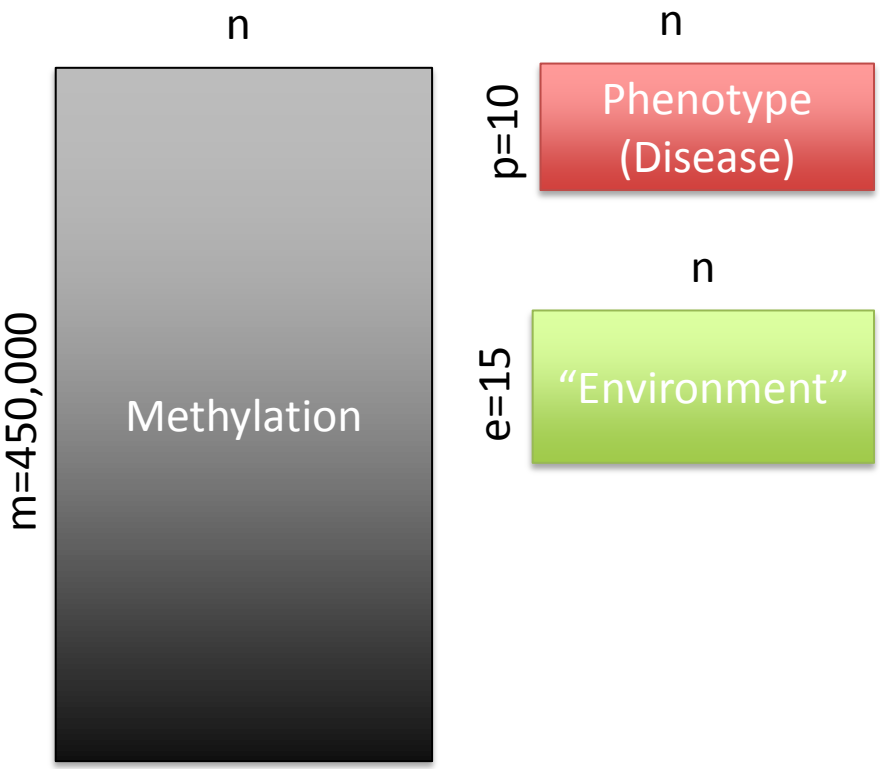
Problem:
variance due to phenotype probably very small (unless your phenotype is cancer)
➔ Needle in a haystack

Control for other sources of variance to make the variance due to the phenotype stand out.

If phenotype is Alzheimer's (AD), gender incorporates more variance into your M matrix than does AD.

LM1: $AD = \beta 0 + \beta 2(gender) + \varepsilon$
LM2: $AD = \beta 0 + \beta 1(m_j) + \beta 2(gender)\ \varepsilon$

eWAS



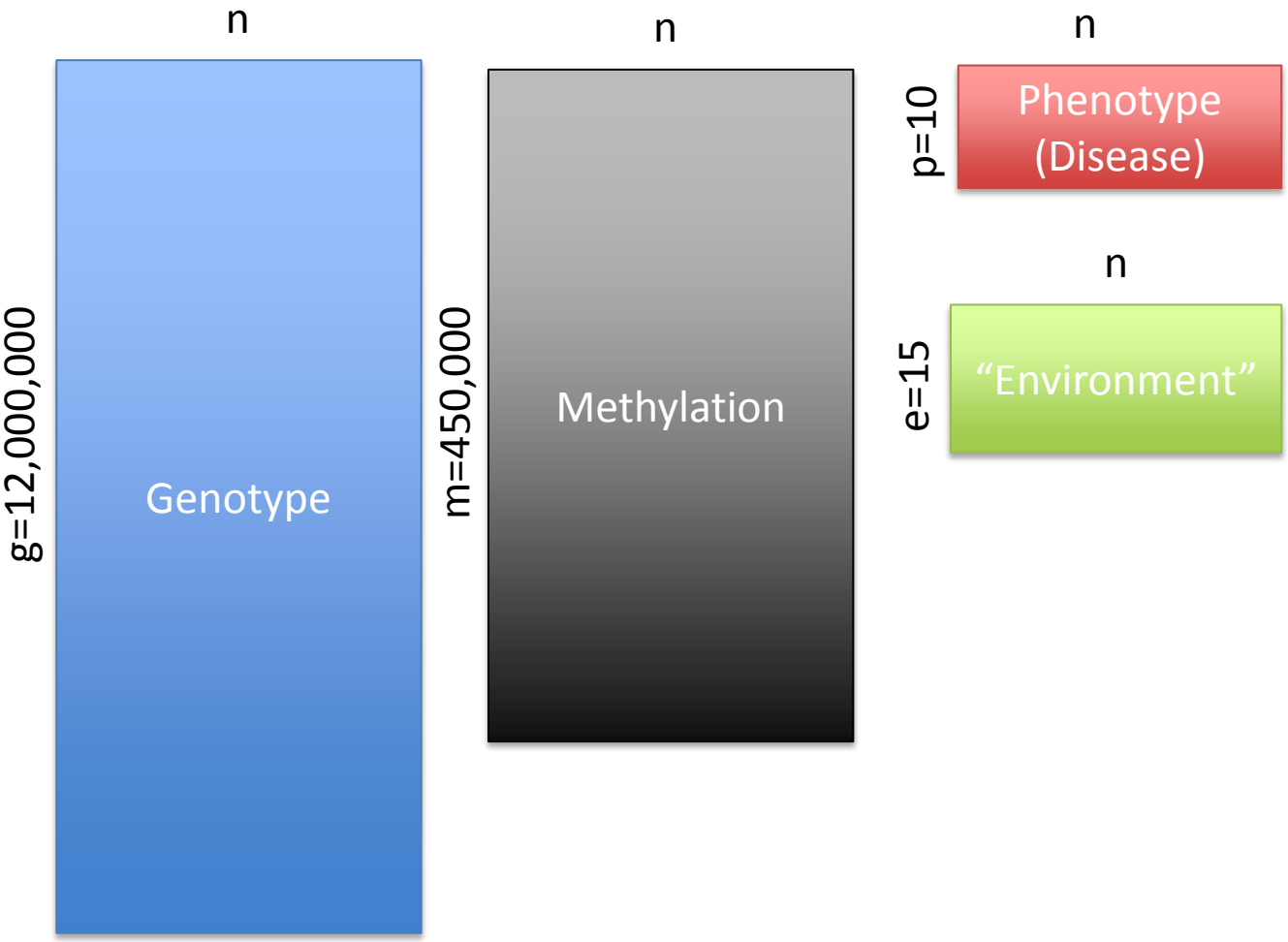
Might have many environmental variables to control for.

LM1: AD = β0 + β2(gender) + β3(age) + β4(education) + … + ε

LM2: AD = β0 + β1(m_j) + β2(gender) + β3(age) + β4(education) + … + ε

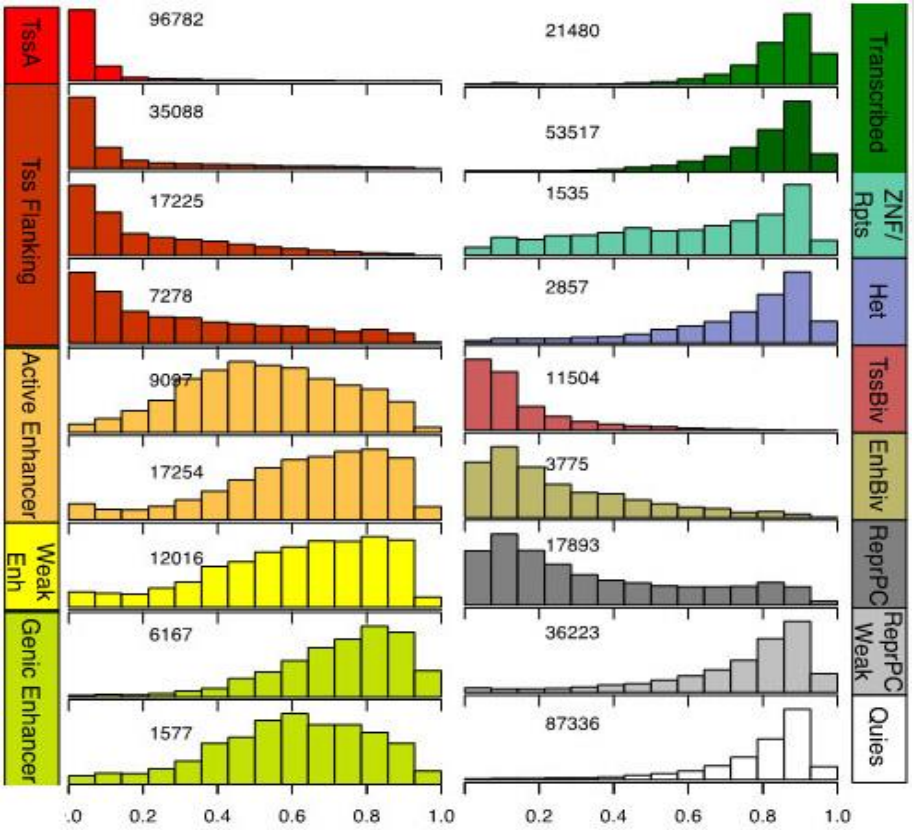
eWAS

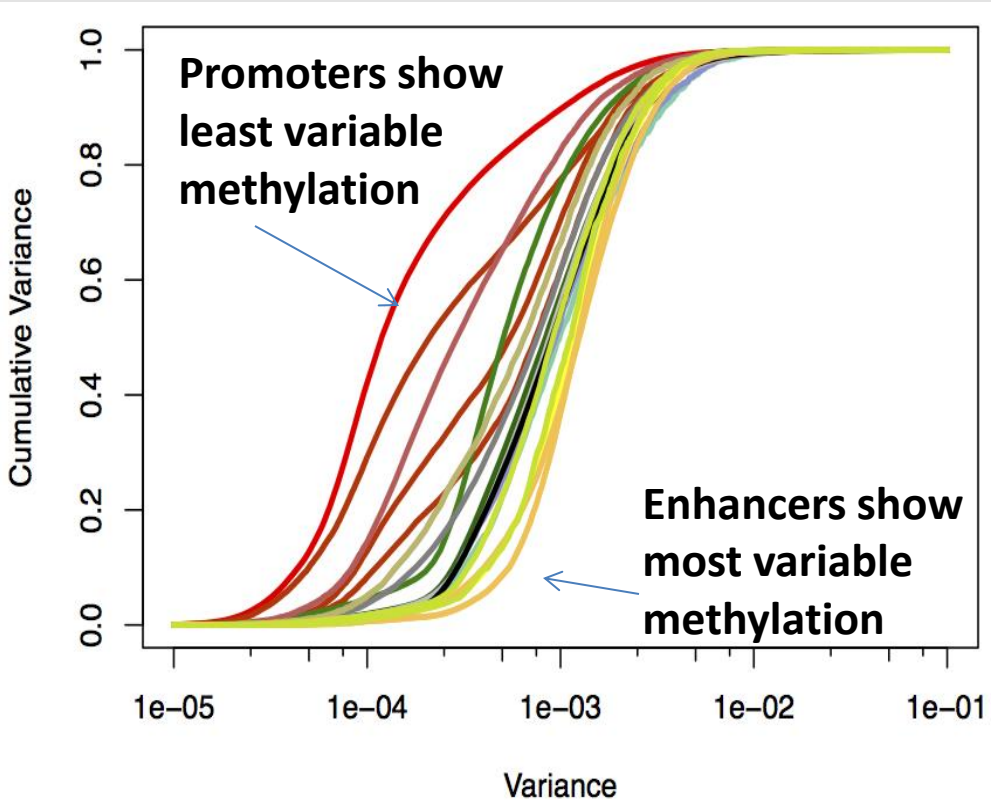
Need to account for variance due to genotype as well.



Role of enhancers vs. promoters in Alzheimer's disease association

Enhancers are hemi-methylated and highly variable

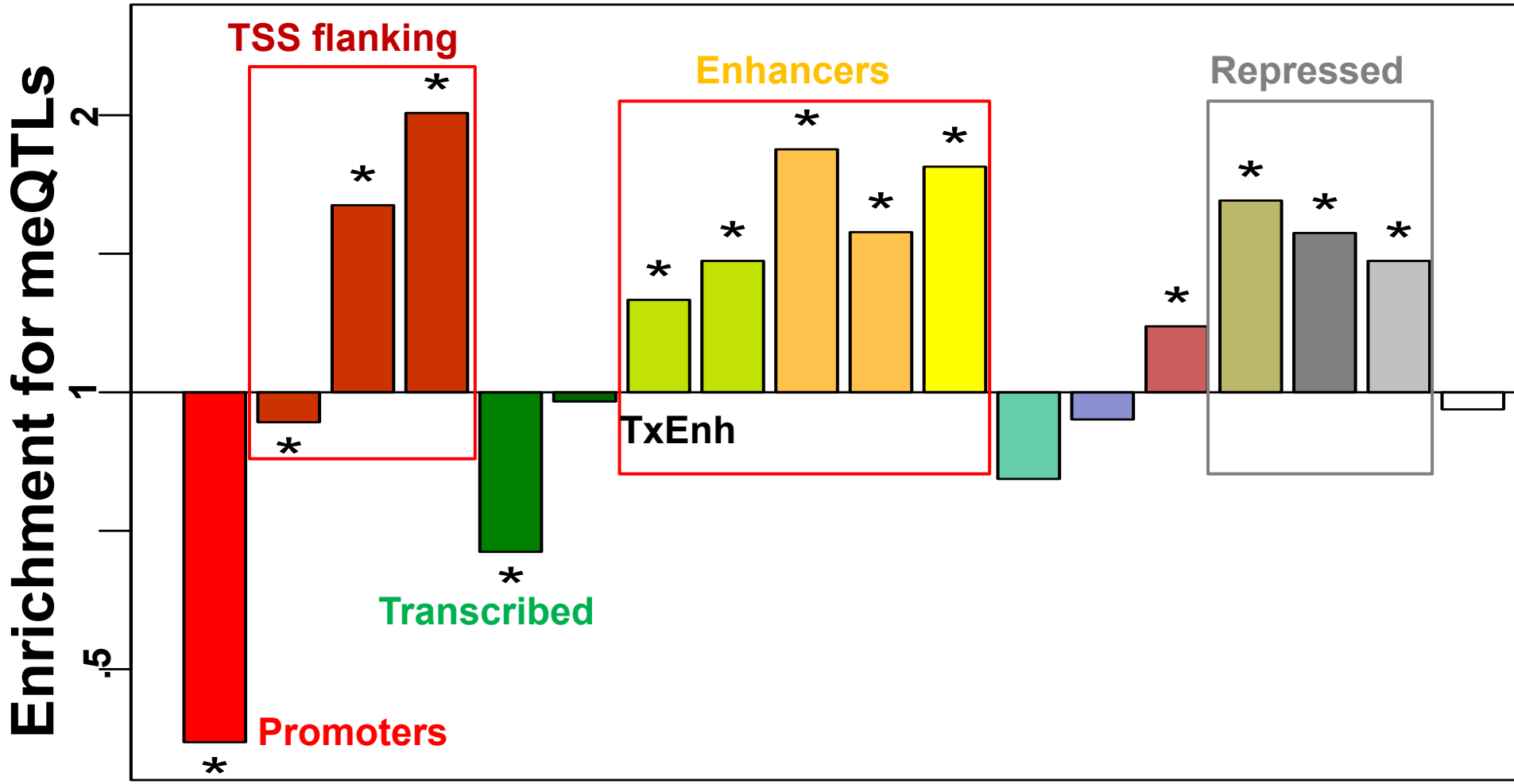




Methylation level Methylation level

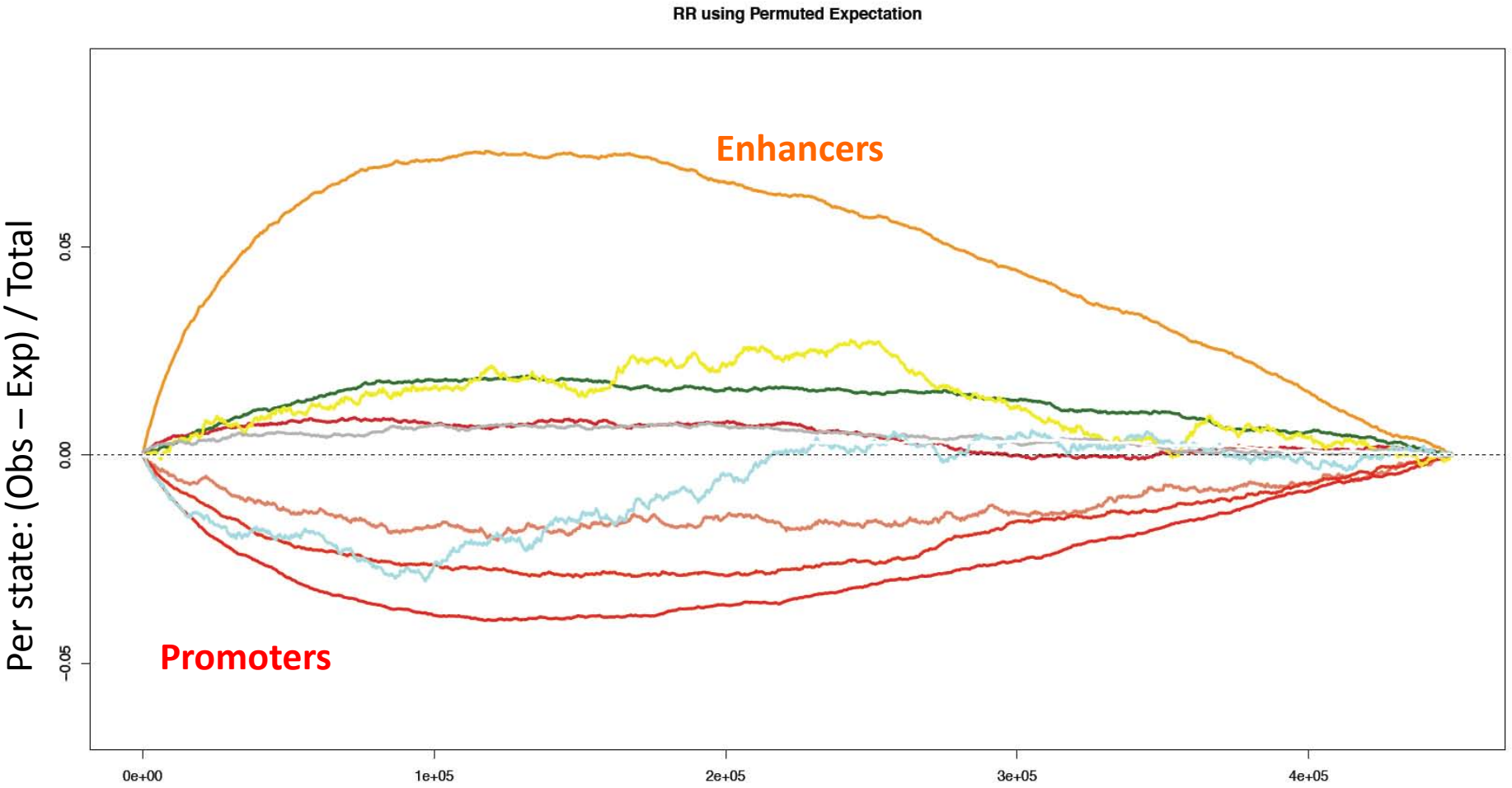
- **Highly distinct signatures for promoters vs. enhancers**
- **Enhancers hemi-methylated in each person (not bimodal)**

SNP-associated CpGs in enhancers, not promoters



- Promoter methylation less affected by genetics
- Enhancer methylation highly genotype-driven
- TSS-flanking and repressed regions also genetic

AD-associated probes in distal enhancers



- After cleaning with known and inferred covariates.
- Distal and transcribed enhancers enriched.
- Proximal regulators (promoters) depleted.

ICA covariate correction cleans up enhancer signal

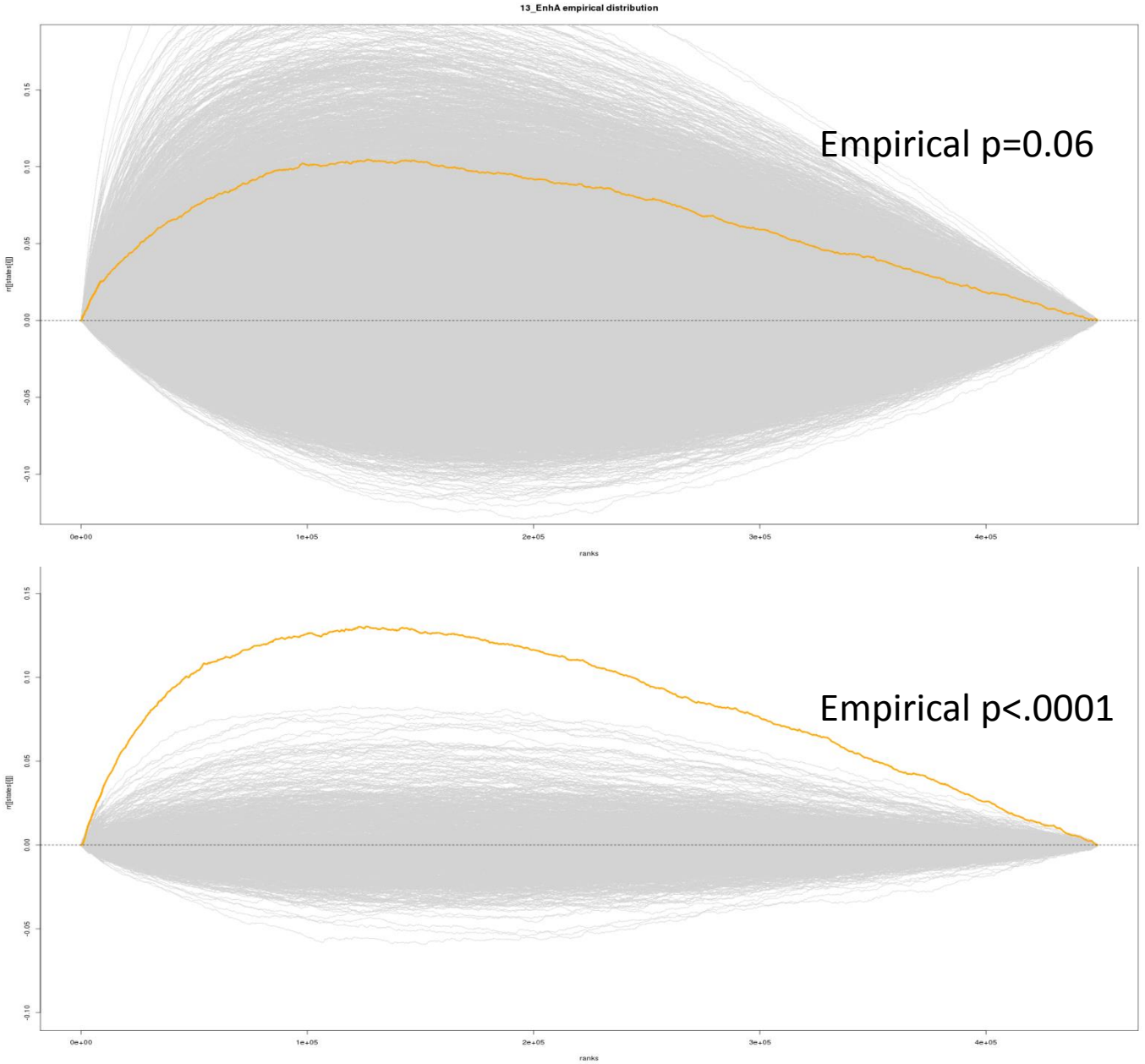
Before:

Orange: Enrichment of enhancer probes for association with the real phenotype.

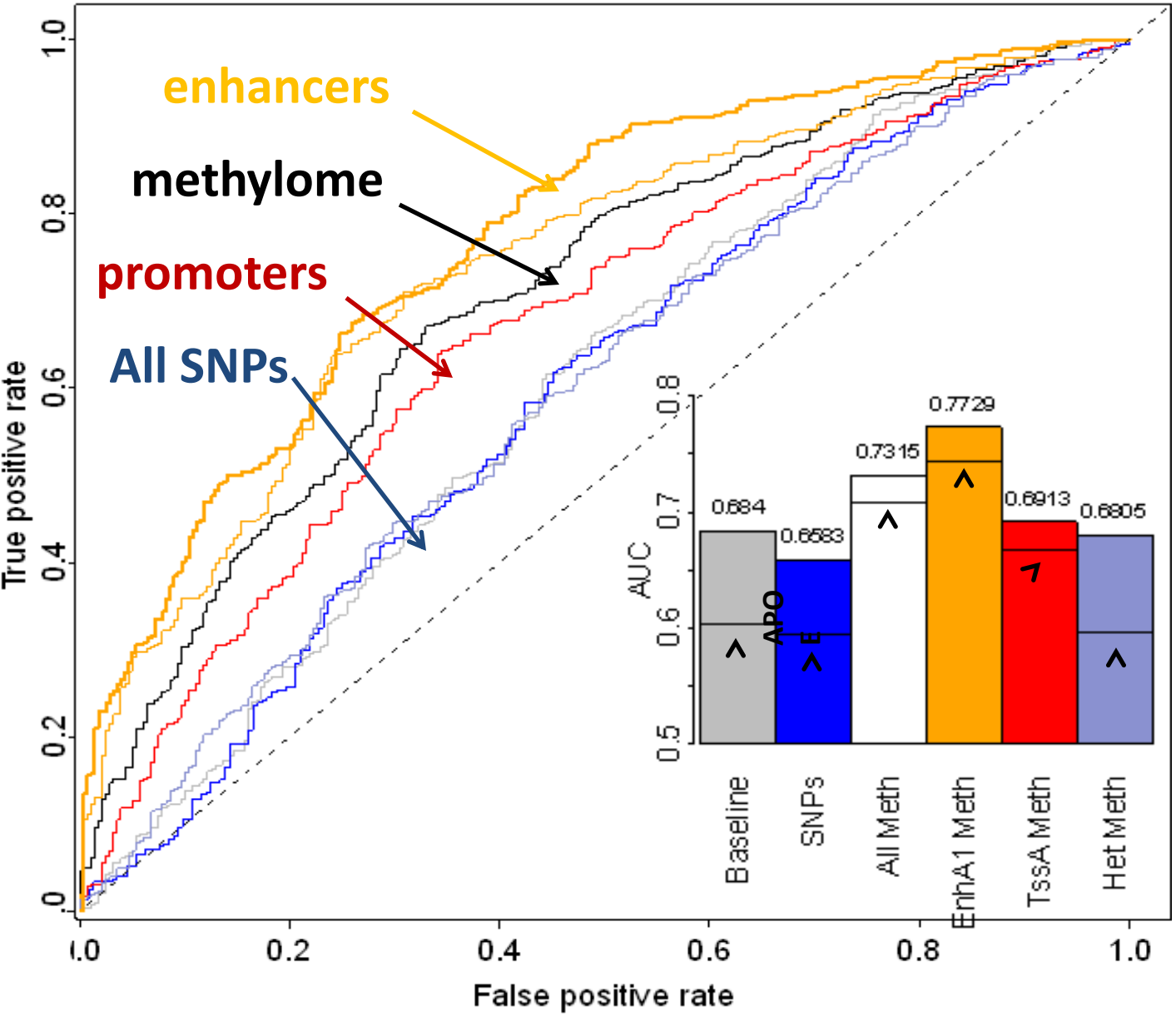
Grey: Enrichment of enhancer probes for a scrambled phenotype.

After:

(**After** conditioning on 7 surrogate variables discovered with ICA.)



AD predictive power highest in enhancers

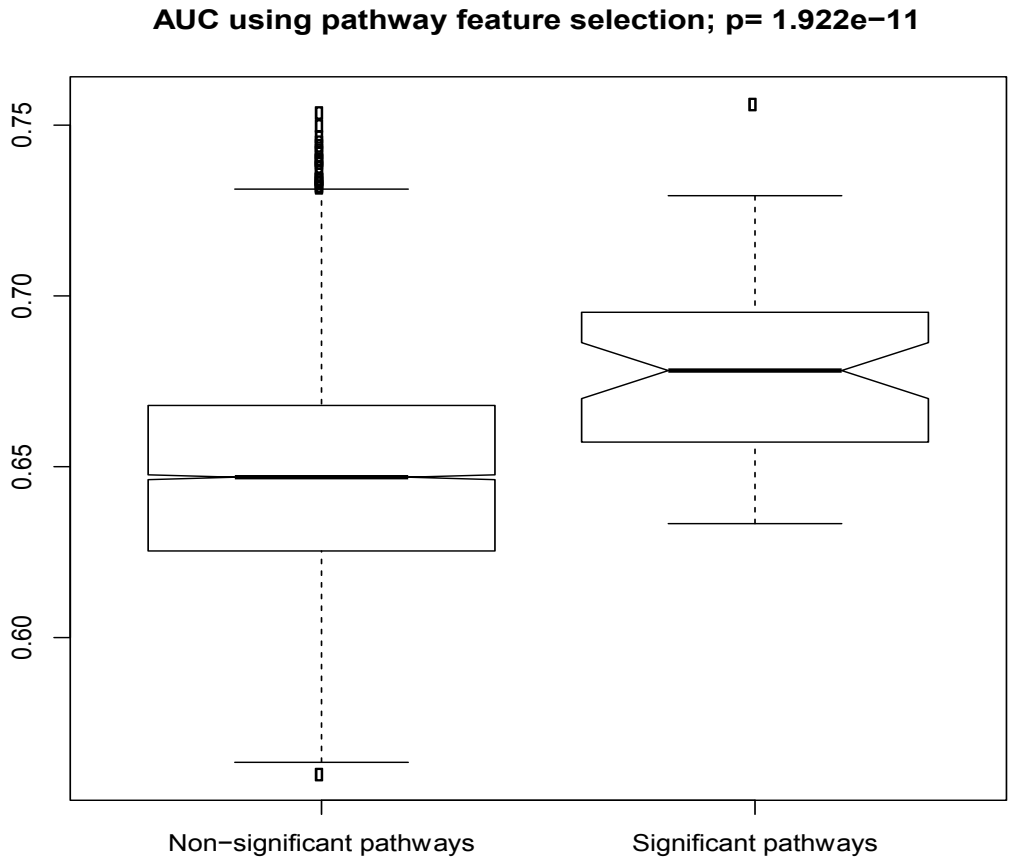


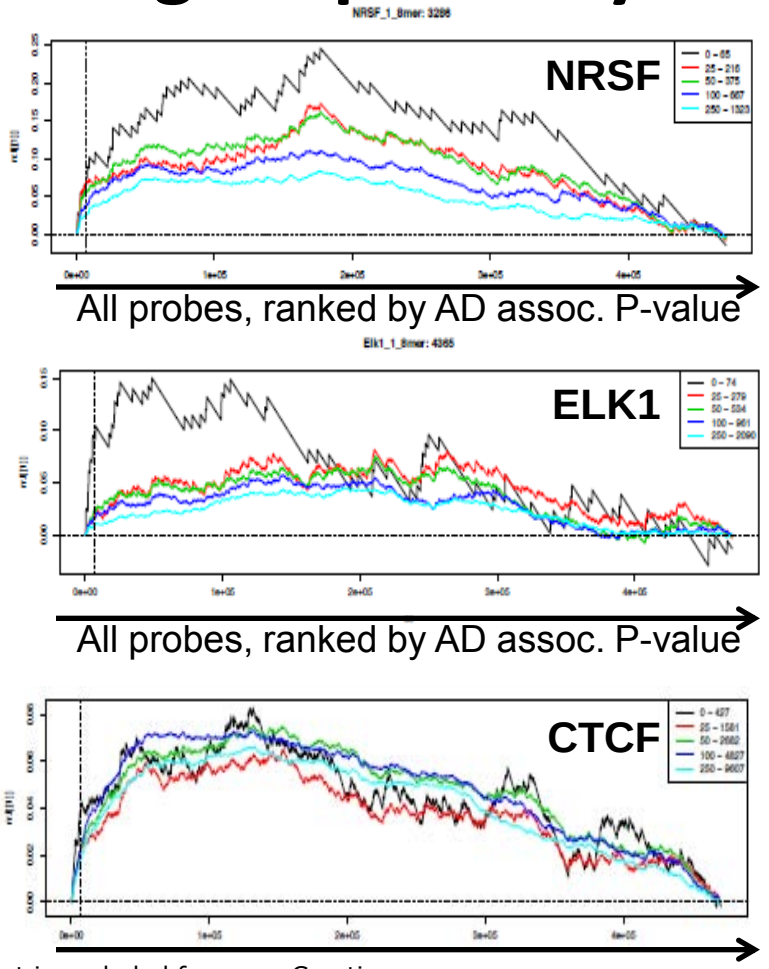
Top predictive features are:

- Enhancer methylation
- All methyl.
- TSS, Het
- Genetics (incl. APOE)

- Causality?
- Common pathways?

AD prediction reveals likely biological pathways





Enriched regulatory motifs suggest potential pathways

HEB/Tcf12: proliferating neural and progenitor cells

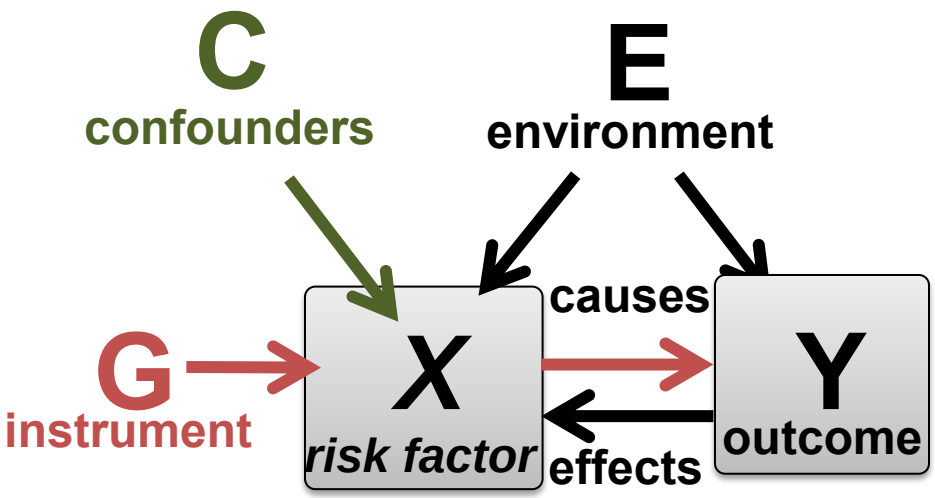
GATA: cell growth, blood, cell development

TLX1/NFIC: Neuronal cell fates ➔ **Mouse AD models**

Goal: Personalized and Predictive Medicine

1. Intro to Epidemiology: basis of human disease
2. Genetic Epidemiology:
   - Genetic basis: GWAS and screening
   - Interpreting GWAS with functional genomics
   - Calculating functional enrichments for GWAS loci
3. Molecular epidemiology
   - meQTLs: Genotype-Epigenome association (cis-/trans-)
   - EWAS: Epigenome-Disease association
4. Resolving Causality
   - Statistical: Mendelian Randomization
   - Application to genotype + methylation in AD
5. Systems Genomics and Epigenomics of disease
   - Beyond single loci: polygenic risk prediction models
   - Sub-threshold loci and somatic heterogeneity in cancer

Risk factor causality w/ instrumental variables



If X⇔Y are correlated, possible scenarios are:

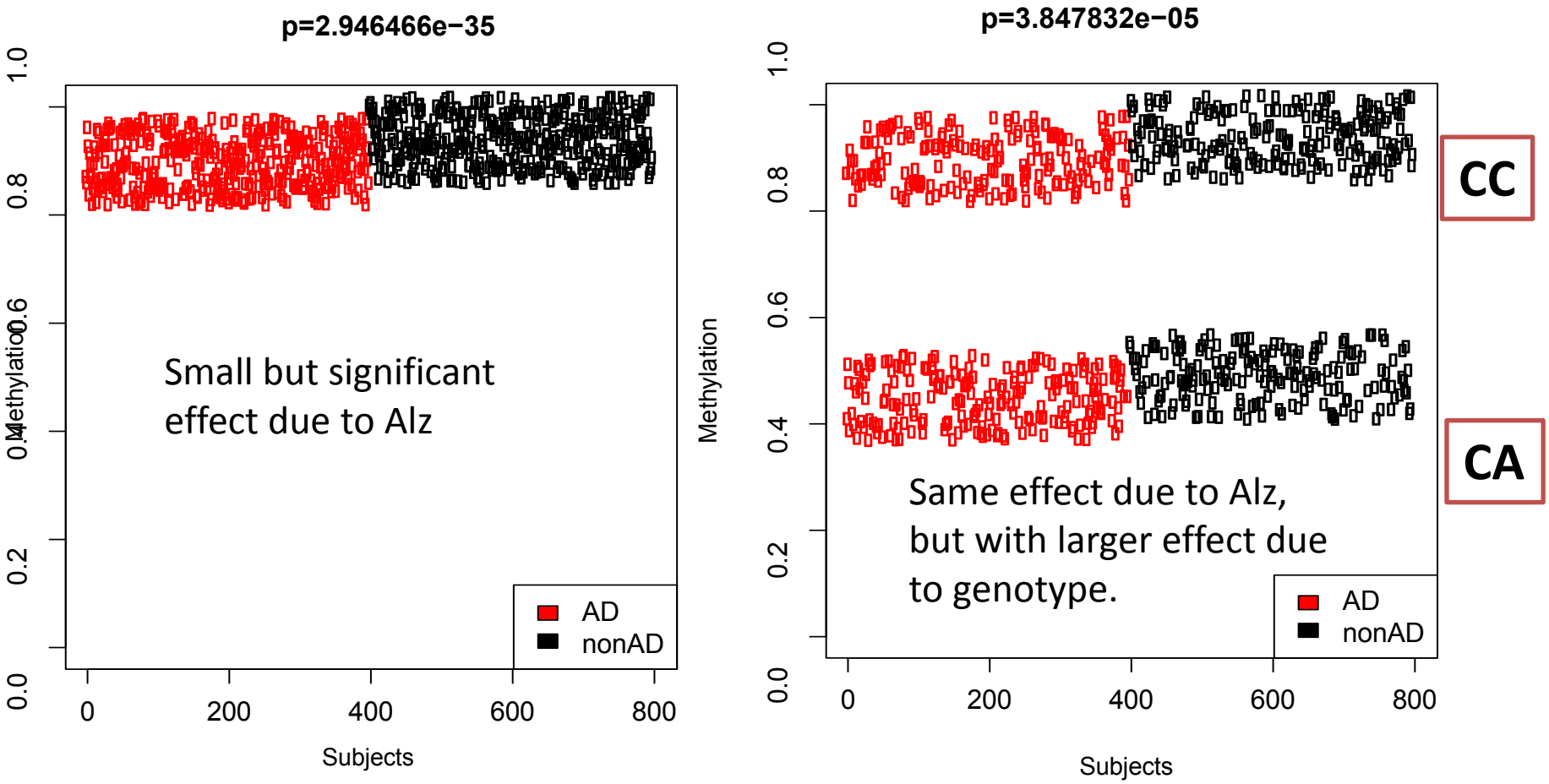
- X→Y
- X←Y
- X←U→Y

To distinguish, need controlled random experiment

- Is risk factor X causing disease Y (or a consequence)?
  - E.g. alcohol addiction, smoking, blood cholesterol, fever, stress

  ➔ Randomized experiment, with and without X: feasibility? ethics?
- **G ⇔ randomized experiment** (e.g. random Mendelian inheritance), as only some subjects have genotype
- G (**i.v.**)must be correlated with Y ***but only through X*** i.e. if X known, G gives no additional information about Y

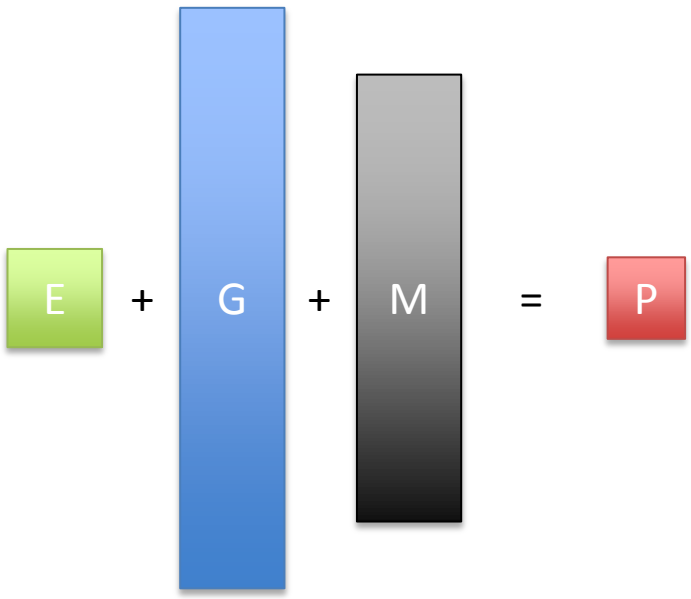
In silico thought experiment



Hemi-methylation associated with meQTL yields a p-value that's 30 orders of magnitude lower for the AD phenotype.

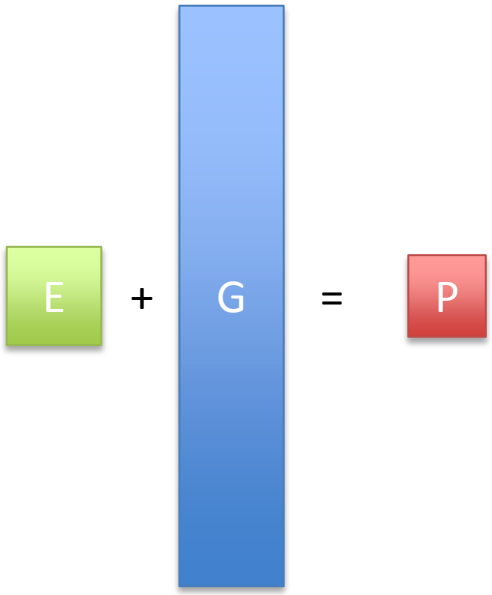
Mendelian randomization approach

Account for variance due to genotype, how much does methylation add?



From G, include probe-specific terms for cis-meQTLs, as well as including trans-meQTLs in all comparisons.

VS



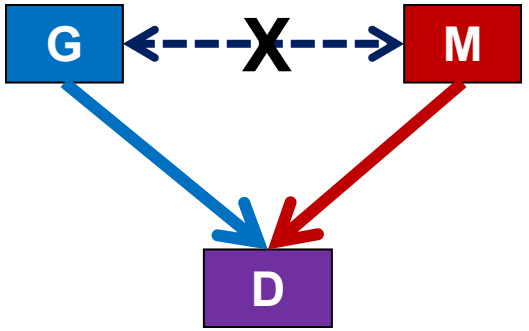
With variability due to genotype and environmental covariates removed, the effect due to phenotype should become more prevalent.

Causality testing

Modeling complex Human diseases

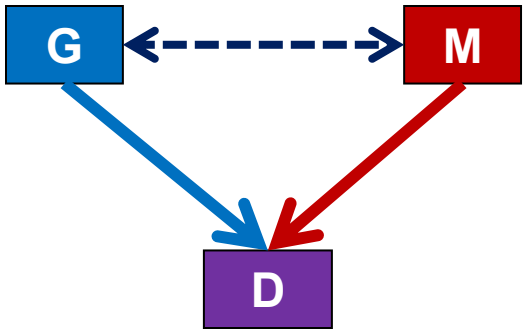
- Three possible models:

1. Independent Associations



2. Causal Pathway Model

3. Interaction Model

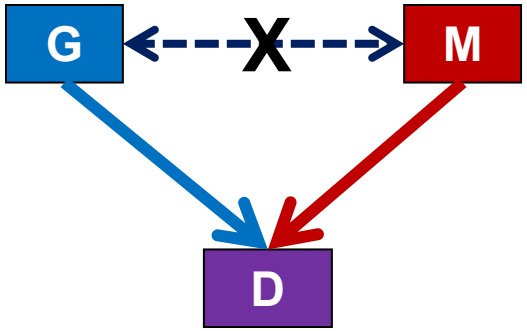


G Genotype

M Methylation

D Disease

(1) Independent Associations



- Association between Factor A and Disease
- Association between Factor B and Disease
- No association between Factor A and Factor B

- Example: 2 independent risk genes

G Factor A

X Factor B

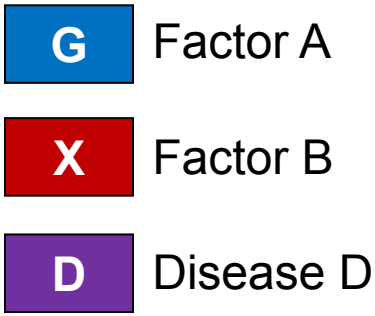
Y Disease D

(2) Causal Pathway Models

- Is the a direct link between risk factor (A) and disease (D)?

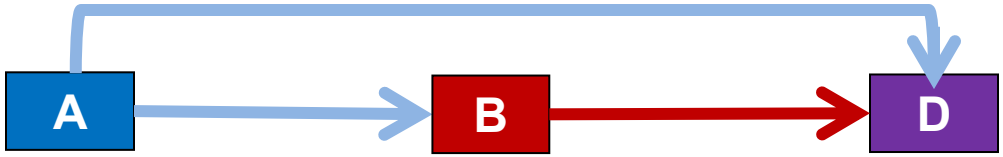
- Does the risk factor's (A) effect on disease (D) depend on an intermediate step (B)?

- To test:
  - A is associated with B and D
  - B is associated with D
  - A is not associated with D when controlling for B
  - Note: A **MUST** come before B temporally

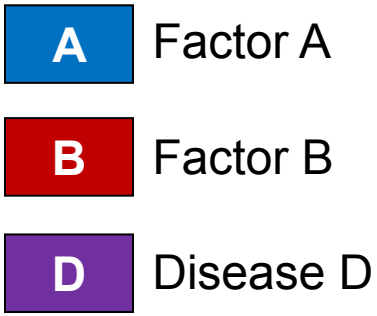


(2) Causal Pathway Models

- In reality its a little of both. A's affect on D is partially *mediated* through B

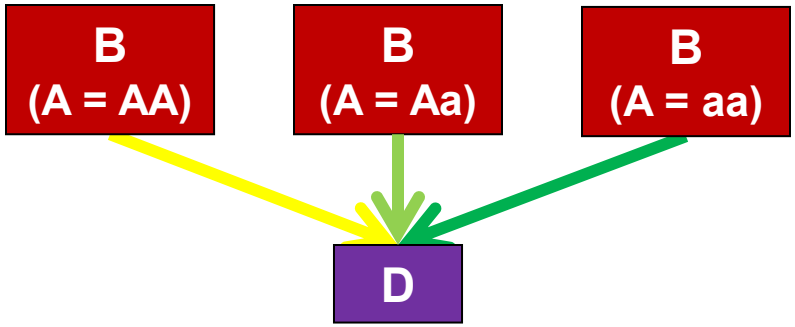


- To test:
  - A is associated with B and D
  - B is associated with D
  - The effect size of A on D is decreased when controlling for B
  - Note: A **MUST** come before B temporally
- Example: *CR1* effect on cognitive decline



(3) Interaction Models

- Factor B's effect on D is different depending on value for factor A



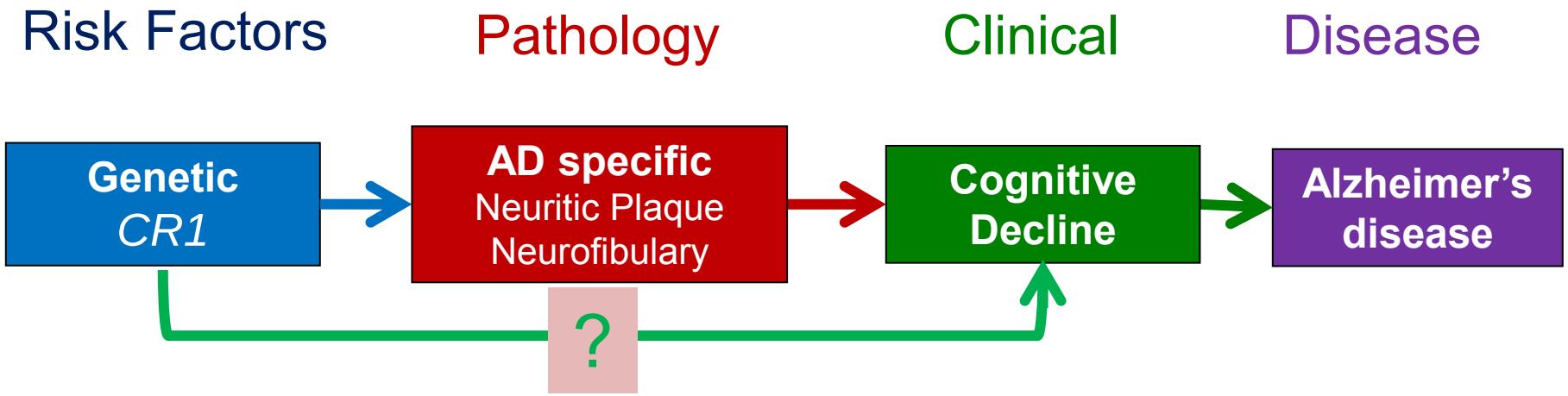
- To test:
  - A + B + A*B → D, if estimate for A*B is significant then
  - Stratify by levels of A

- Example:
  - A drug's effect is different depending on genotype
  - More to come…

Application to 12 AD GWAS loci

Gene	locus	reference	Published AD	AD	NP
ABCA7	rs3764650	Hollingsworth 2010	5.0×10^{-21}	0.747	0.187
APOE	Any ε4			1.2×10^{-13}	1.8×10^{-23}
BIN1	rs744373	Seshadri 2010	1.6×10^{-11}	0.204	0.480
CD2AP	rs9349407	Naj 2011/Hollingsworth 2011	8.6×10^{-9}	0.445	0.221
CD33	rs3865444	Naj 2011/Hollingsworth 2012	1.6×10^{-9}	0.133	0.123
CLU	rs11136000	Lambert 2009/Harold 2009	7.5×10^{-9}	0.762	0.649
CR1	rs6656401	Lambert 2009	3.7×10^{-9}	0.0009	0.057
EPHA1	rs11767557	Naj 2011/Hollingsworth 2011	6.0×10^{-10}	0.562	0.391
MS4A4A	rs4938933	Naj 2011	1.7×10^{-9}	0.792	0.567
MS4A6A	rs610932	Hollingsworth 2010	1.2×10^{-16}	0.534	0.820
MTHFD1L	rs11754661	Naj 2010	1.9×10^{-10}	0.126	0.934
PICALM	rs3851179	Harold 2009	1.9×10^{-8}	0.382	0.171

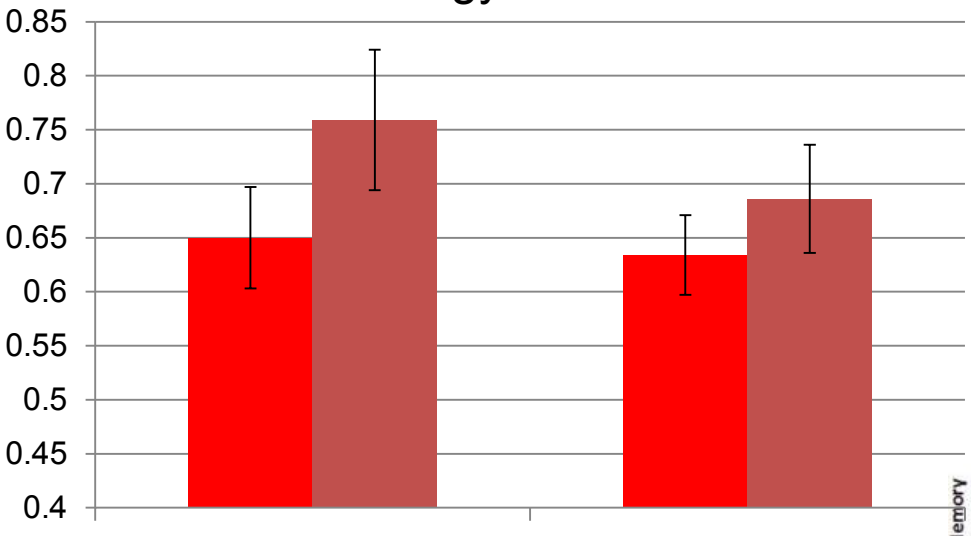
CR1: Causal pathway model



- *CR1* first associated with AD in 2009
- Original associated variant is in an intron, no clear function
- Unclear how *CR1* locus influences AD susceptibility mechanistically
- Questions:
  - Is the effect only on AD?
  - Is there a broader effect on cognitive decline?
  - Is there an association with AD pathology?
  - Does it go through pathology to have an effect of cognitive decline?

CR1 (rs6656401)

CR1 → Pathology



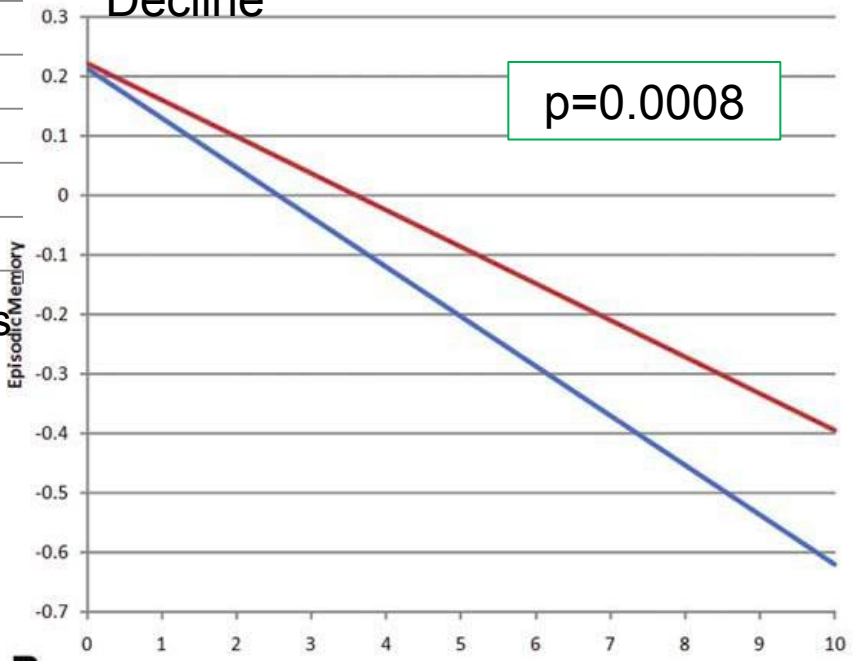
p=0.008

p=0.10

TT

AT/AA (risk allele)

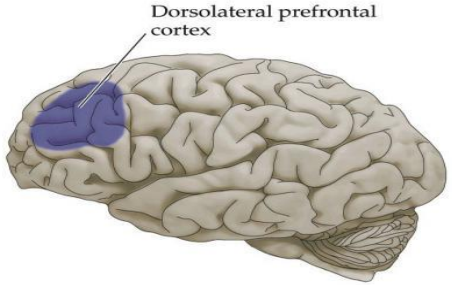
CR1 → Global Cognitive Decline

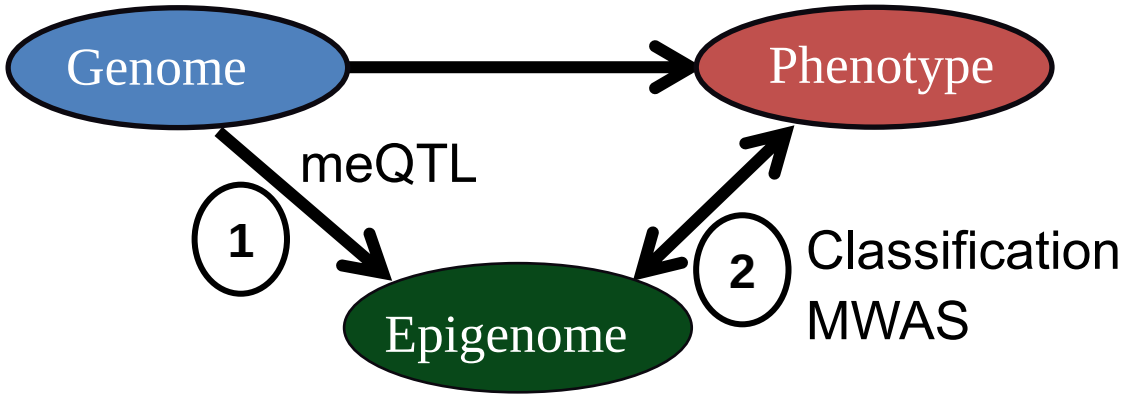


Pathology → Global Cognitive Decline

p < 0.0001

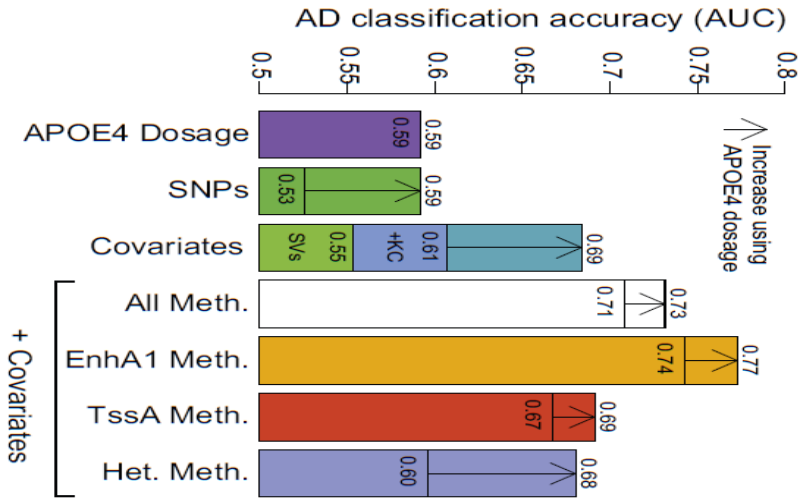
Genetic + Epigenetic variation in Alzheimer's



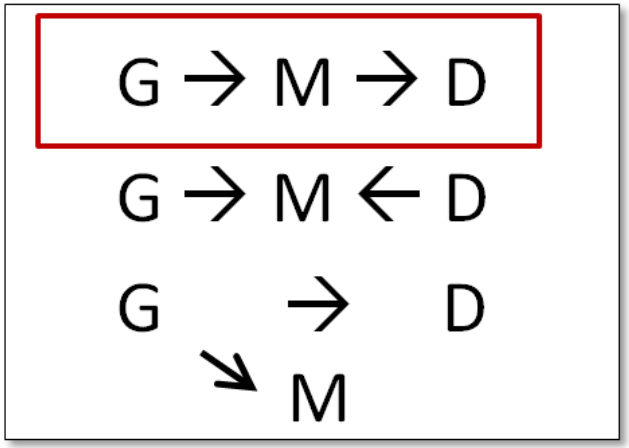


Methylation variation in 723 AD patients & controls

Relate to genotype and AD variation



Methylation >> SNPs
Enhancers >> promoters



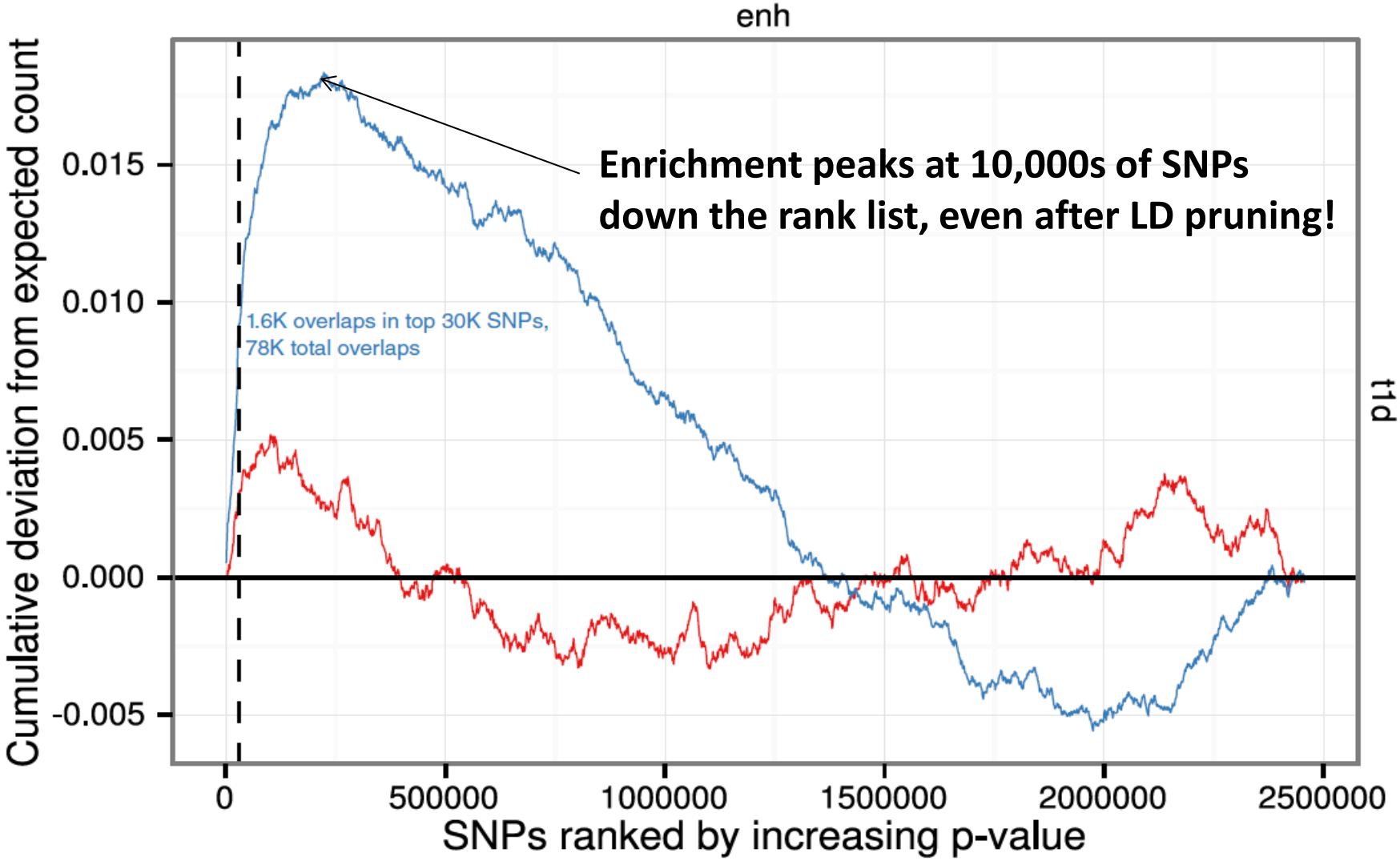
Estimate causal M roles: regression of meQTL effects reduces M⇔D

Goal: Personalized and Predictive Medicine

1. Intro to Epidemiology: basis of human disease
2. Genetic Epidemiology:
   - Genetic basis: GWAS and screening
   - Interpreting GWAS with functional genomics
   - Calculating functional enrichments for GWAS loci
3. Molecular epidemiology
   - meQTLs: Genotype-Epigenome association (cis-/trans-)
   - EWAS: Epigenome-Disease association
4. Resolving Causality
   - Statistical: Mendelian Randomization
   - Application to genotype + methylation in AD
5. Systems Genomics and Epigenomics of disease
   - Beyond single loci: polygenic risk prediction models
   - Sub-threshold loci and somatic heterogeneity in cancer

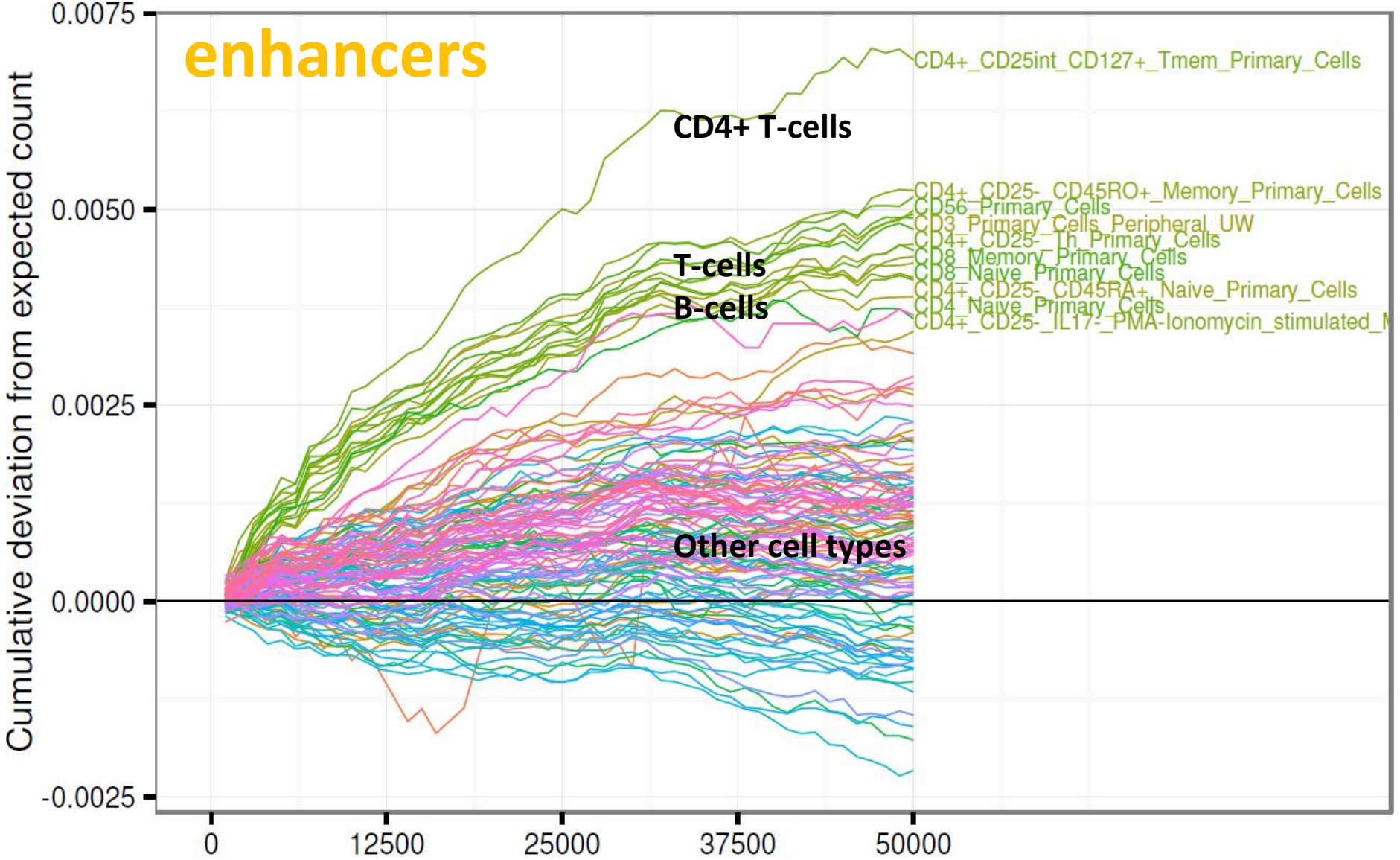
Beyond top-scoring hits: 1000s of variants of weak effect cluster in cell type specific enhancers

Rank-based functional testing of weak associations



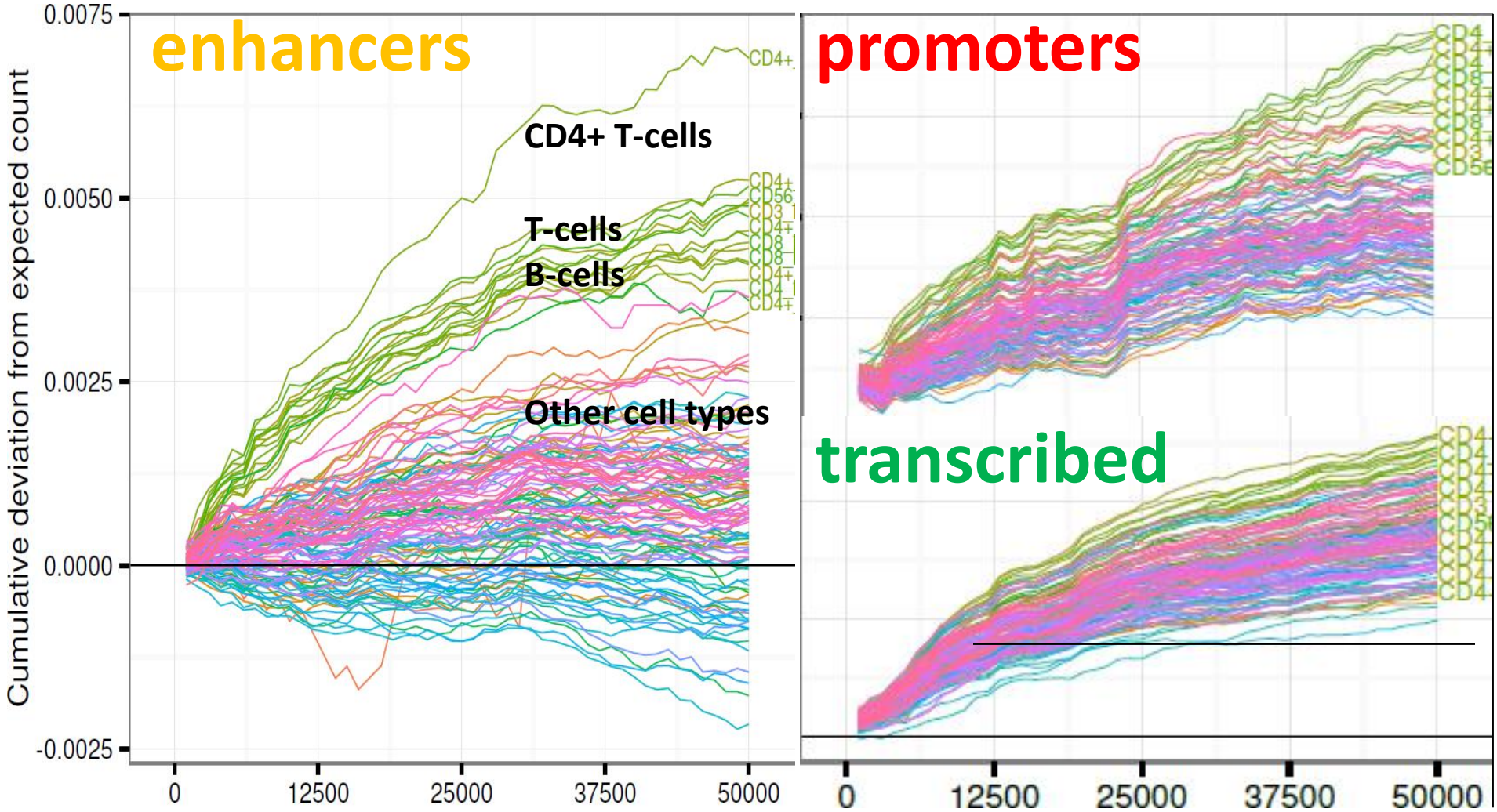
- **Rank all SNPs based on GWAS signal strength**
- **Functional enrichment for cell types and states**

Weak-effect T1D hits in 50k T-cell enhancers



- LD-pruning (CEU r^2>.2): 50k ➔ 41k independ. loci

Cell type specificity stronger for enhancers



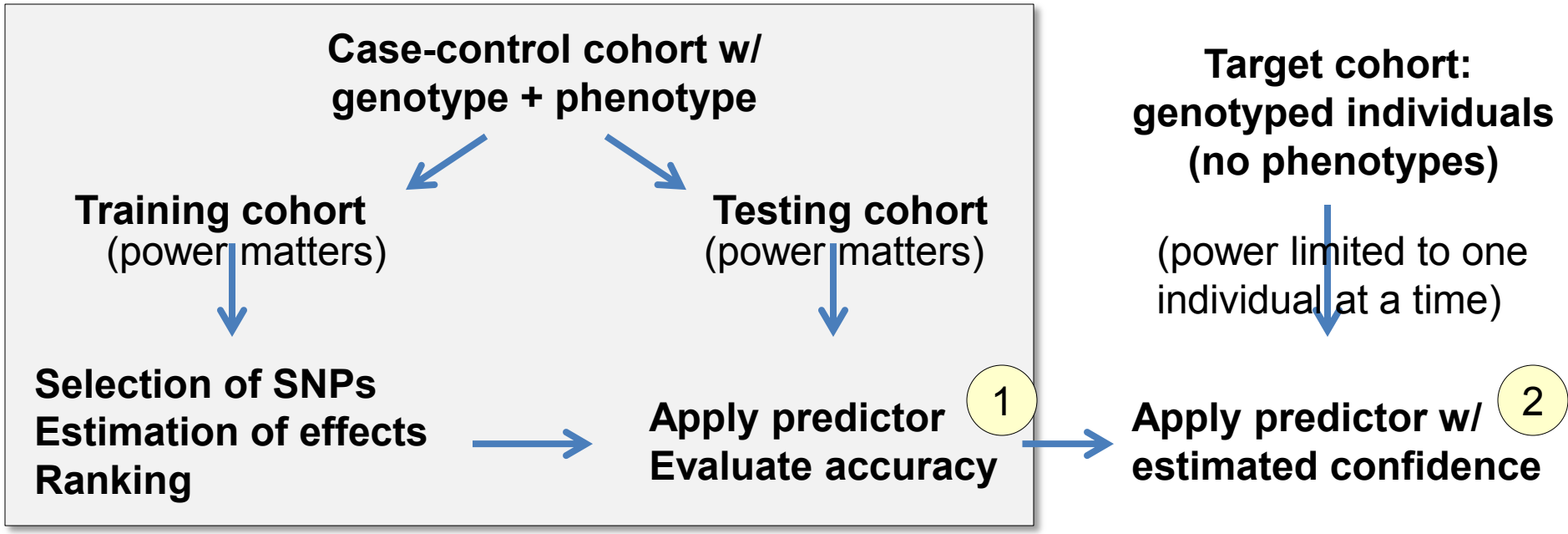
- T/B-cells also enriched for promoters, transcribed
- Enhancer enrichment much more cell type specific

T1D/RA-enriched enhancers spread across genome

- **High concentration of loci in MHC, high overlap**
- **Yet: many distinct regions, 1000s of distinct loci**

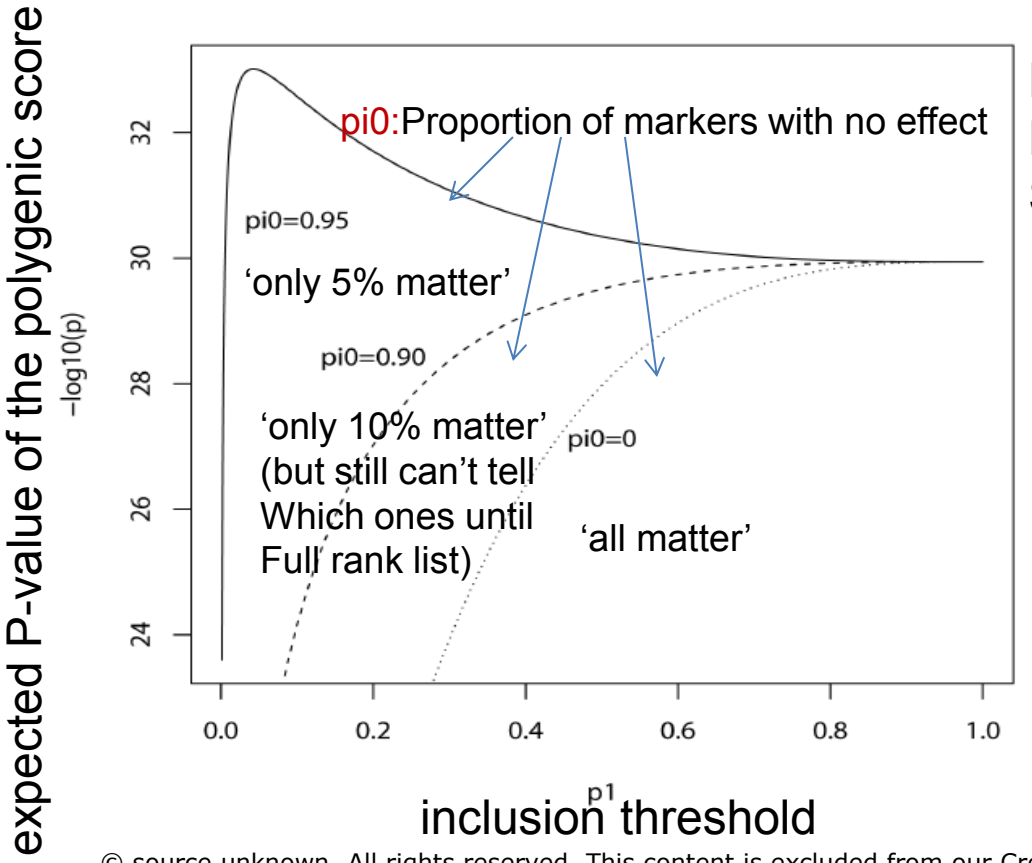
Implications for genetic predisposition: polygenic models for risk prediction

Basic setup of polygenic risk prediction studies



- Applications 1 (testing cohort)
  - Understand total heritability captured in common variants
  - Understand disease "architecture": number of SNPs
  - Recognize functional classes associated with weak genetic associations
- Applications 2 (new individuals)
  - Provide health recommendations at the individual level
  - Prioritize high-risk individuals for subsequent testing at population level

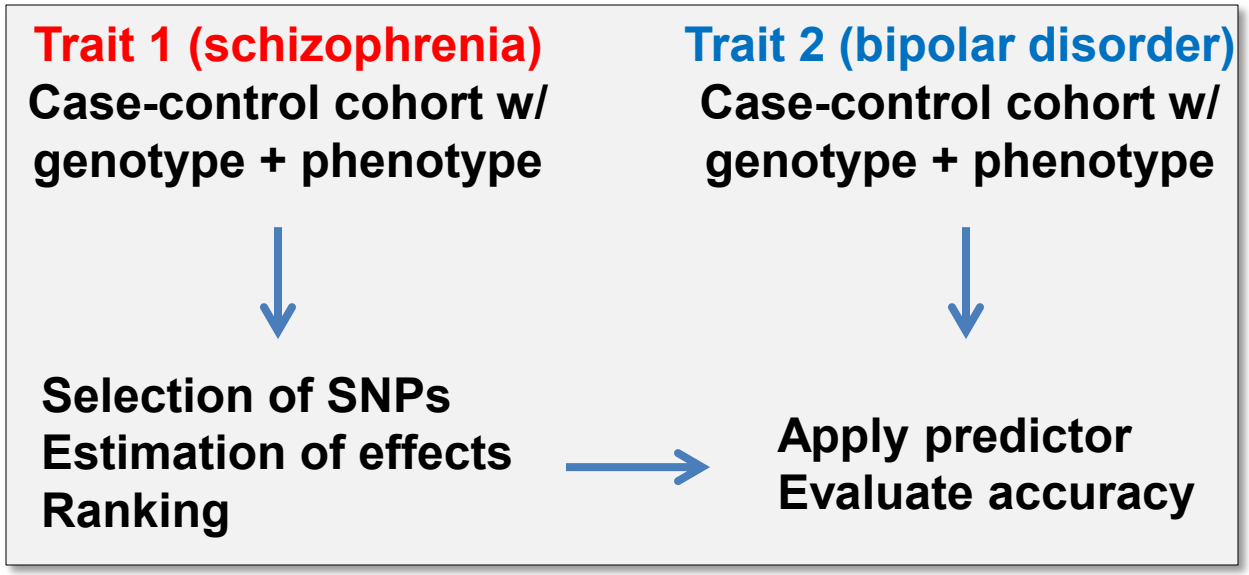
How many SNPs to include in model?



Dudridge PLoS Genetics 2013
Purcell Nature 2009
Schizophrenia risk prediction

- It depends on:
  - Architecture: Fraction of SNPs that are estimated to be functional
  - Power: Number of individuals in cohort, i.e. ability to rank correctly
- It only peaks at 5% (≈1-pi0) when sufficient power to rank
  - Large fraction of associated markers are hidden within non-significant SNPs
- For pi0=0.90, still need to include all SNPs to maximize predictive power

Application to pleiotropy and common risk



- Ability to assess common genetic risk
  - Are the highly-ranked SNPs for one study relevant to a different study?
  - Is there a shared genetic architecture between seemingly unrelated traits?
- First use showed schizophrenia and bipolar disorder common risk
  - Schizophrenia-ranked SNPs in one cohort…
  - … are predictive of bipolar disorder diagnosis
  - … but not predictive of unrelated (cardiovascular) traits

Important points/caveats for risk prediction

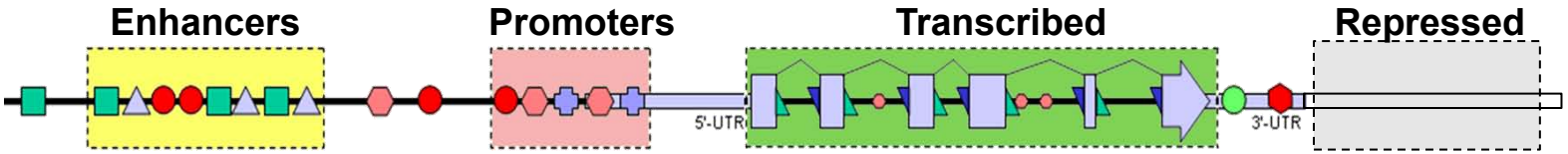
- Always limited by genetic component
  - Environment, random effects play big role for most traits
- Mendelian=deterministic vs. common variants=prob.ic
  - Only a first screen for individuals at risk
- Limited by discovery power
  - Cohort size limits discriminative power and ranking ability
- Limited by genotyped SNPvs vs. all SNPs
  - Selection pushes fitness-reducing variants to lower freq
  - Genotyped SNPs selected to be common
- Even if SNPs are correctly identified, their effects are not
  - Winner's curse: over-estimate above-threshold true effect
- Training and testing cohort non-independence
  - Relatives, cryptic relatedness, population stratification inflate est.

Goal: Personalized and Predictive Medicine

1. Intro to Epidemiology: basis of human disease
2. Genetic Epidemiology:
   - Genetic basis: GWAS and screening
   - Interpreting GWAS with functional genomics
   - Calculating functional enrichments for GWAS loci
3. Molecular epidemiology
   - meQTLs: Genotype-Epigenome association (cis-/trans-)
   - EWAS: Epigenome-Disease association
4. Resolving Causality
   - Statistical: Mendelian Randomization
   - Application to genotype + methylation in AD
5. Systems Genomics and Epigenomics of disease
   - Beyond single loci: polygenic risk prediction models
   - Sub-threshold loci and somatic heterogeneity in cancer

This talk: From loci to mechanisms

Building a reference map of the regulatory genome



Regions: Enhancers, promoters, transcribed, repressed
Cell types: Predict tissues and cell types of epigenomic activity
Target genes: Link variants to their target genes using eQTLs, activity, Hi-C
Nucleotides: Regulatory consequence of mutation: Conservation, PWMs
Regulators: Upstream regulators whose activity is disrupted by mutation

Application to GWAS, hidden heritability, and Cancer

GWAS hits

CATGCCTG
CGTGTCTA

- 93% top hits non-coding → Mechanism? Cell type?
- Lie in haplotype blocks → Causal variant(s)?

'Hidden' heritability

CATGCCTG
CGTGTCTA

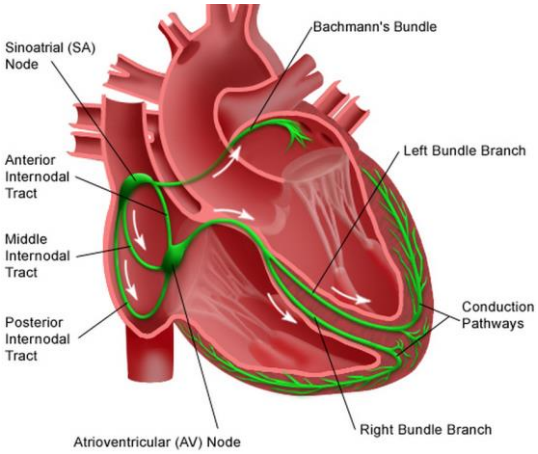
- Many variants, small effects → Pathway-level burden/load
- Many false positives → Prioritize w/ regulatory annotations

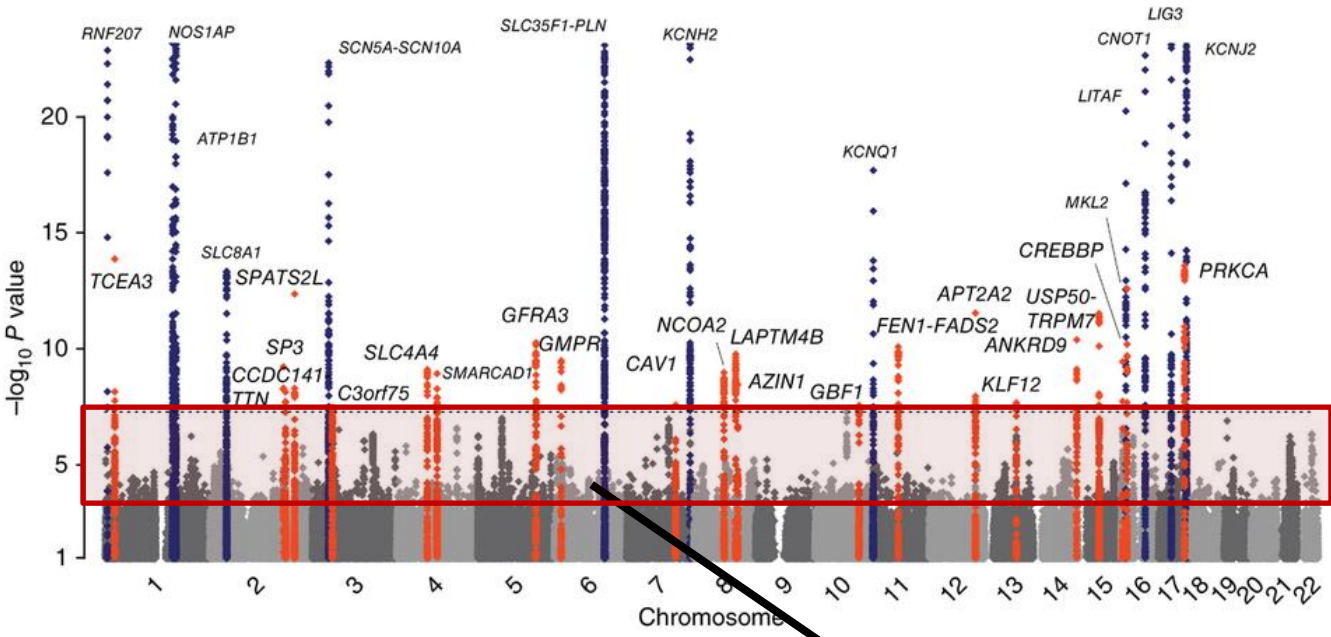
Cancer mutations

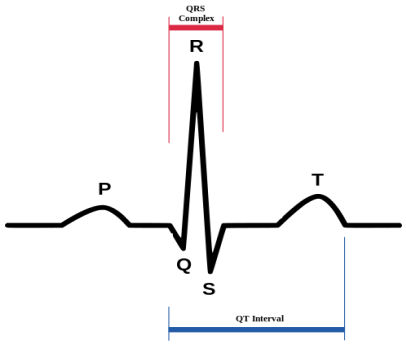
CATGCCTG
CATCCCTG

- Loss of function → Protein-coding variants, convergence
- Gain of function → Regulatory variants, heterogeneity

Characterizing sub-threshold variants in heart arrhythmia







Focus on sub-threshold variants (e.g. rs1743292 $P=10^{-4.2}$)

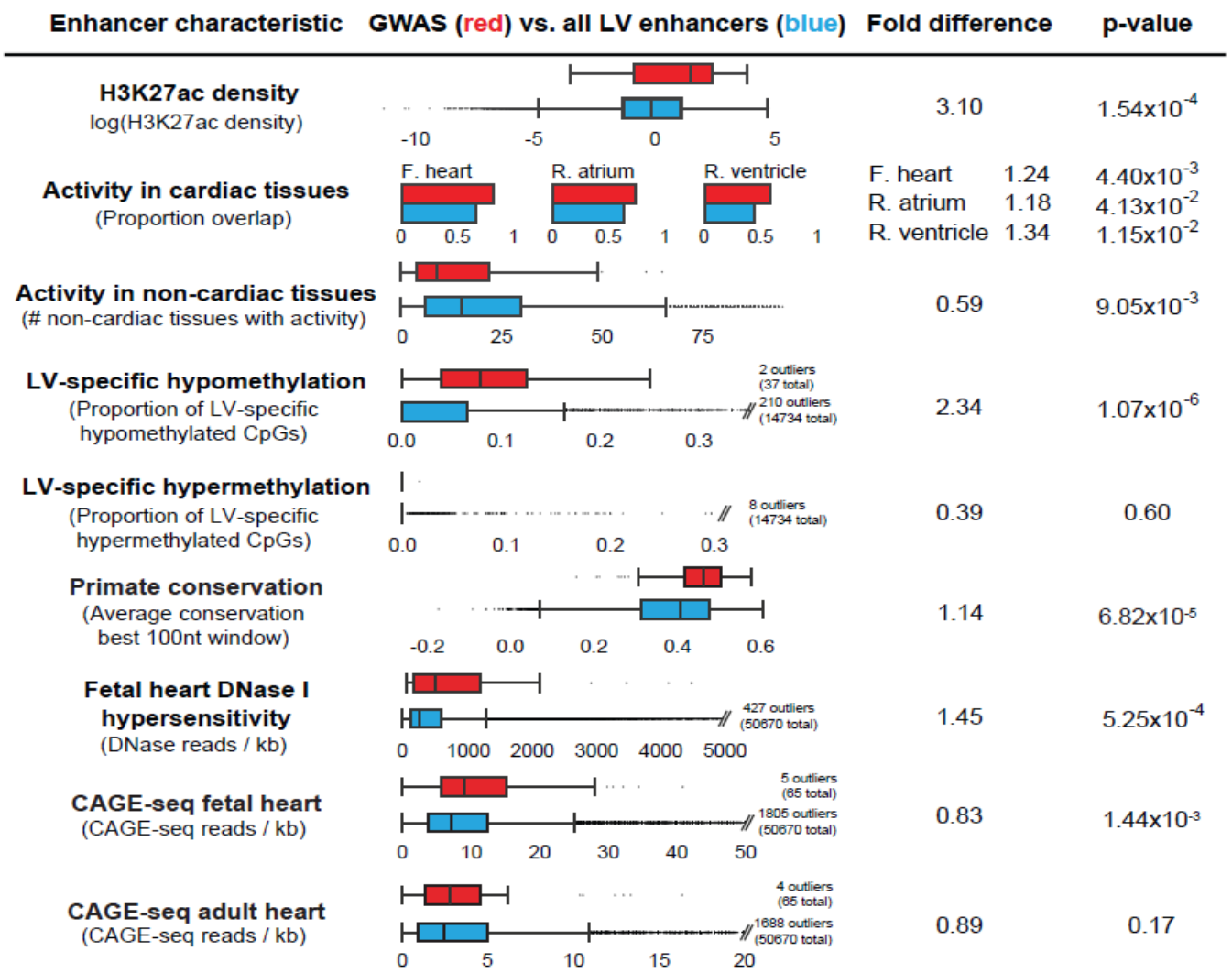
From Arking, D. E., Pulit, S. L., Crotti, L., Harst, P. V., Munroe, P. B., Koopmann, T. T., . . . Newton-Cheh, C. (2014). Genetic association study of QT interval highlights role for calcium myocardial repolarization. Nature Genetics Nat Genet, 46(8), 826-836. Used with permission.

Trait: QRS/QT interval

(1) Large cohorts, (2) many known hits

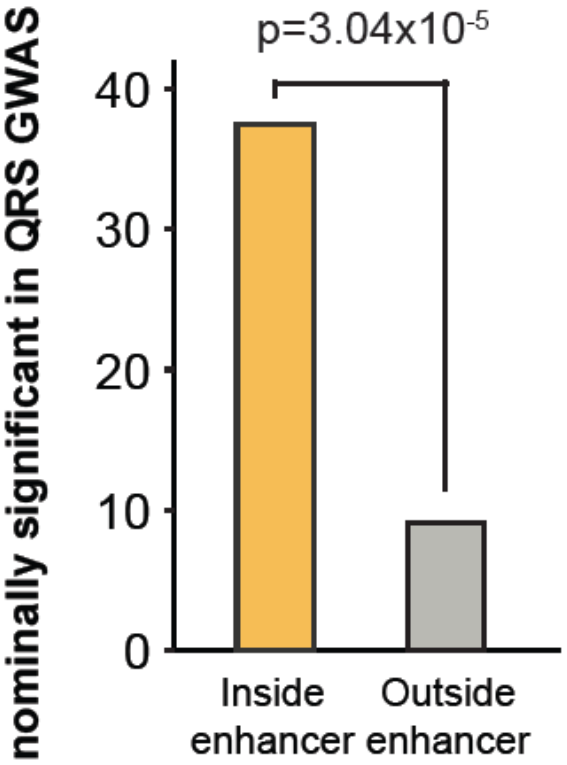
(3) well-characterized tissue drivers

Enhancers overlapping GWAS loci share functional properties

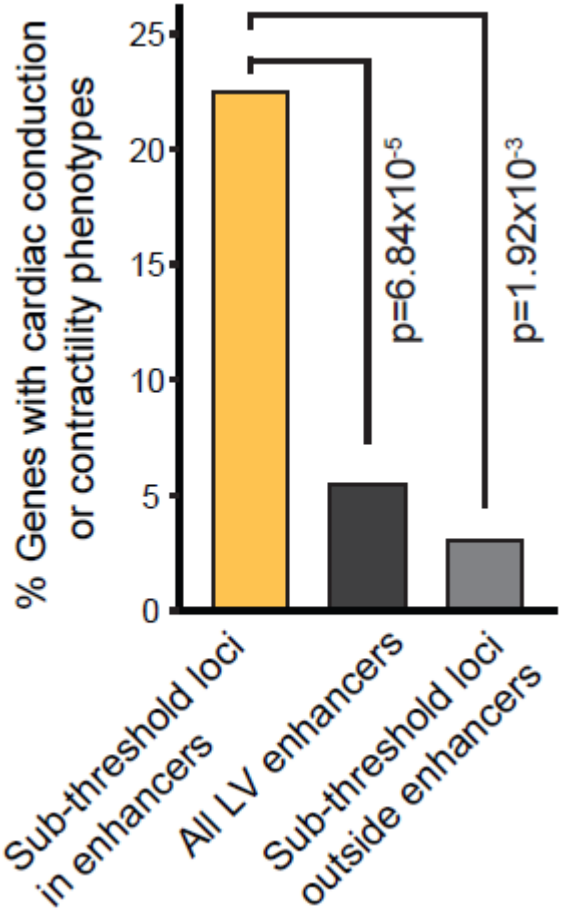


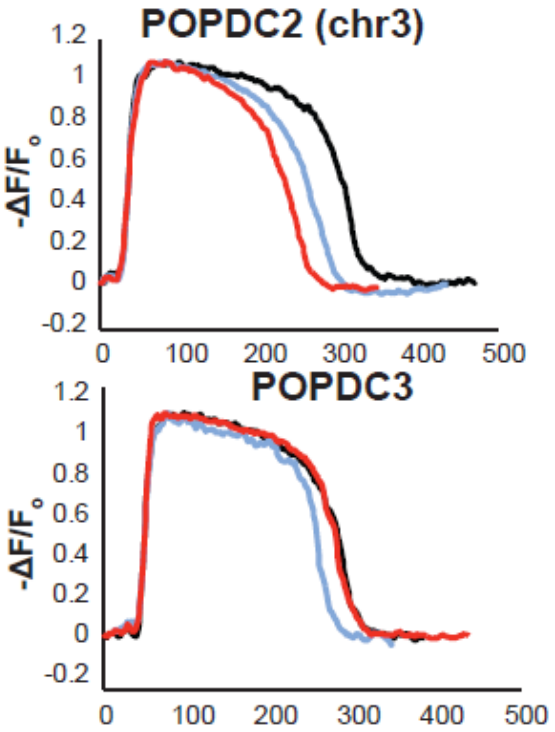
Train machine learning model to prioritize sub-threshold loci

Functional evidence for sub-threshold target genes



Human genetics





Zebrafish phenotypes

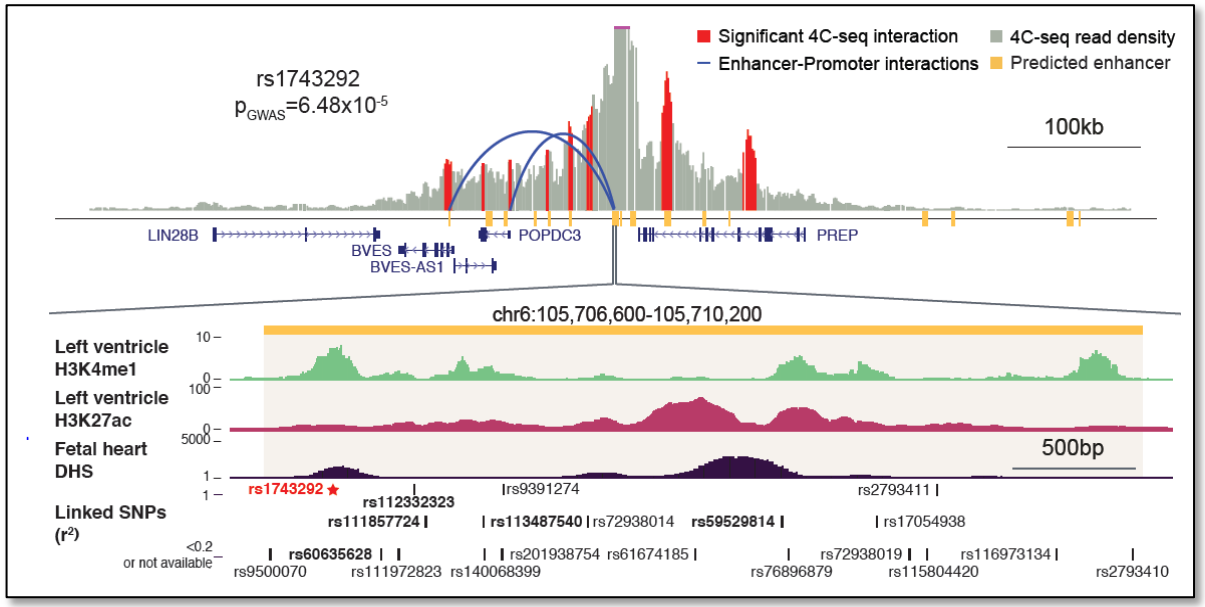
Mouse phenotypes

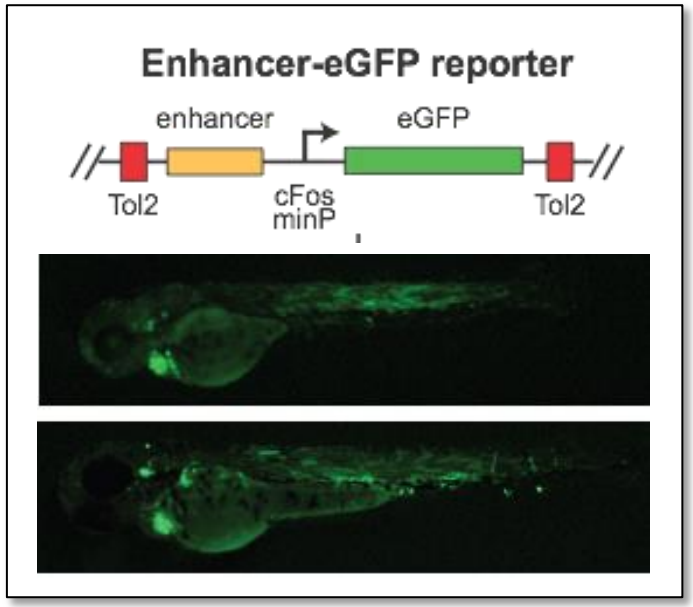
Experimental validation of 11 sub-threshold loci

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

9 of 11 tested loci show allelic activity, chromatin interactions

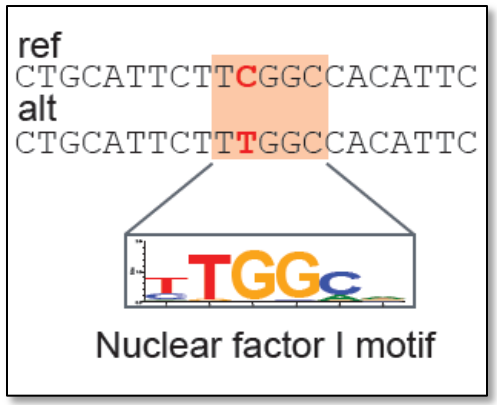
Functional evidence for rs1743292 causality ($P=10^{-4.2}$)

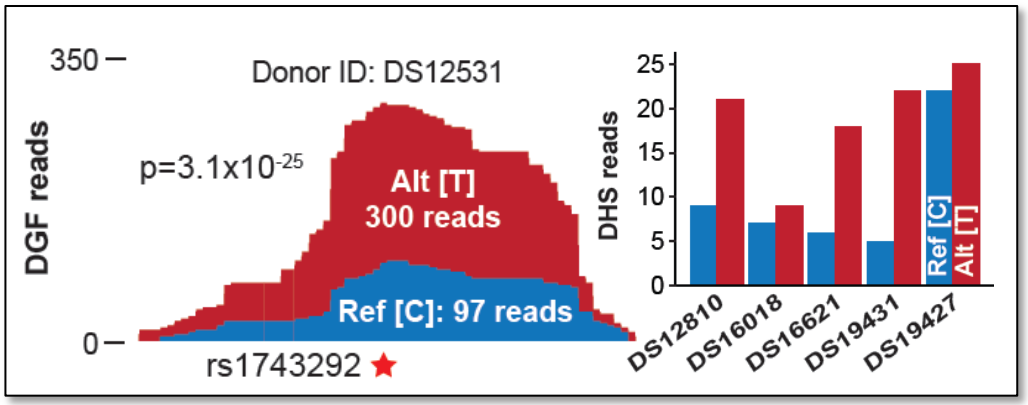


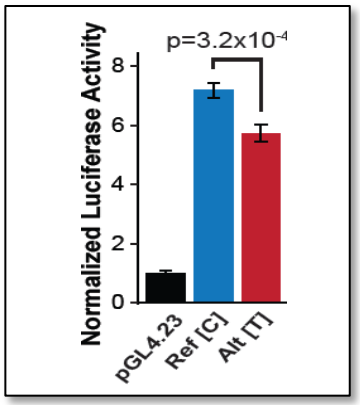


Enhancer 4C links to target gene promoters

Heart enhancer activity







Motif disruption

Allelic DNase in multiple individuals

Allelic enhancer activity

Target gene impact on heart conduction

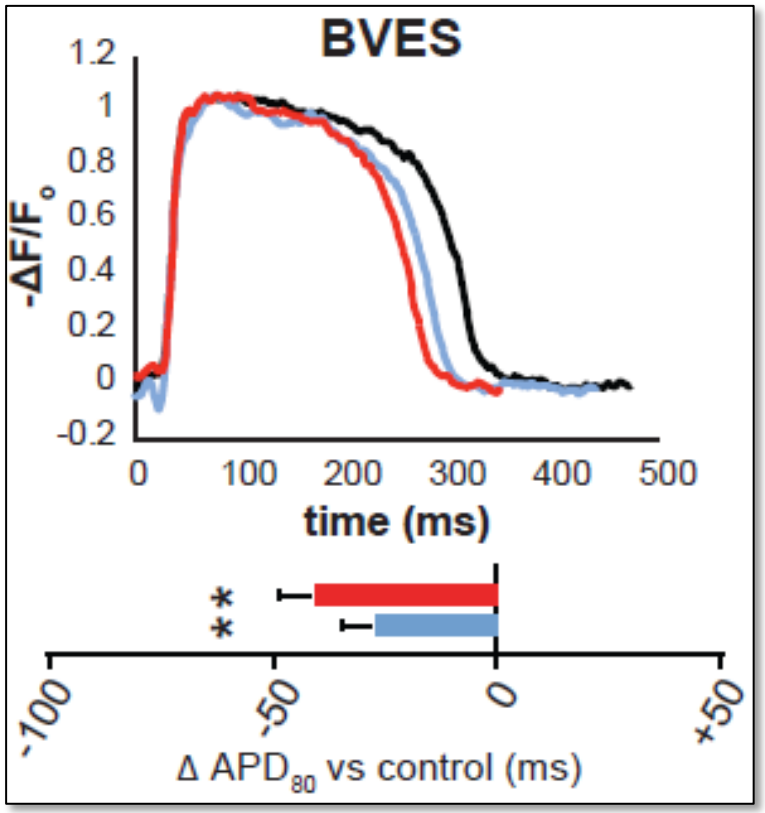
Optical voltage mapping

ventricle

atrium

transmembrane voltage
(ventricle)

zebrafish embryo hearts
voltage-sensitive fluorescent dye



Detection and validation of a new cardiac locus

What would we need to discover rs1743292 without epigenomics?

rs1743292
Minor allele frequency: 0.134
Effect size: -0.5773 +/- 0.17 msec
With 68,900 individuals: 12.8% power to discover at $p<5x10^{-8}$

- rs1743292 has similar effect sizes as many genome-wide significant variants
- Many GWAS variants discovered due to **winner's curse**: often only have 5-20% power to discover
- Combining epigenomics and GWAS can:
  1. Confirm existing GWAS loci are real
  2. Discover new sub-threshold loci with weak effect sizes, low power

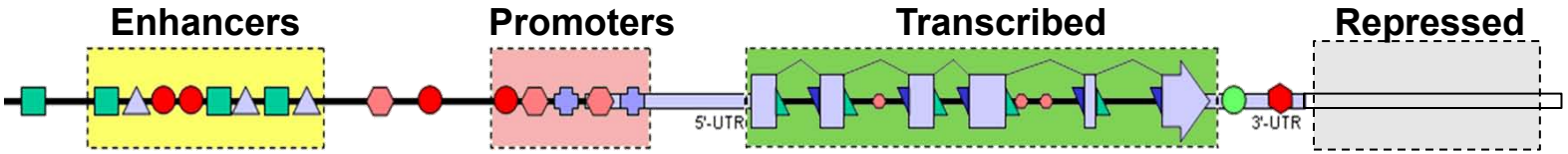
To reach 80% power to discover rs1743292 at $p<5x10^{-8}$,
we need **146,700** individuals!

Goal: Personalized and Predictive Medicine

1. Intro to Epidemiology: basis of human disease
2. Genetic Epidemiology:
   - Genetic basis: GWAS and screening
   - Interpreting GWAS with functional genomics
   - Calculating functional enrichments for GWAS loci
3. Molecular epidemiology
   - meQTLs: Genotype-Epigenome association (cis-/trans-)
   - EWAS: Epigenome-Disease association
4. Resolving Causality
   - Statistical: Mendelian Randomization
   - Application to genotype + methylation in AD
5. Systems Genomics and Epigenomics of disease
   - Beyond single loci: polygenic risk prediction models
   - Sub-threshold loci and somatic heterogeneity in cancer

This talk: From loci to mechanisms

Building a reference map of the regulatory genome



Regions: Enhancers, promoters, transcribed, repressed
Cell types: Predict tissues and cell types of epigenomic activity
Target genes: Link variants to their target genes using eQTLs, activity, Hi-C
Nucleotides: Regulatory consequence of mutation: Conservation, PWMs
Regulators: Upstream regulators whose activity is disrupted by mutation

Application to GWAS, hidden heritability, and Cancer

GWAS hits

CATGCCTG
CGTGTCTA

- 93% top hits non-coding → Mechanism? Cell type?
- Lie in haplotype blocks → Causal variant(s)?

'Hidden' heritability

CATGCCTG
CGTGTCTA

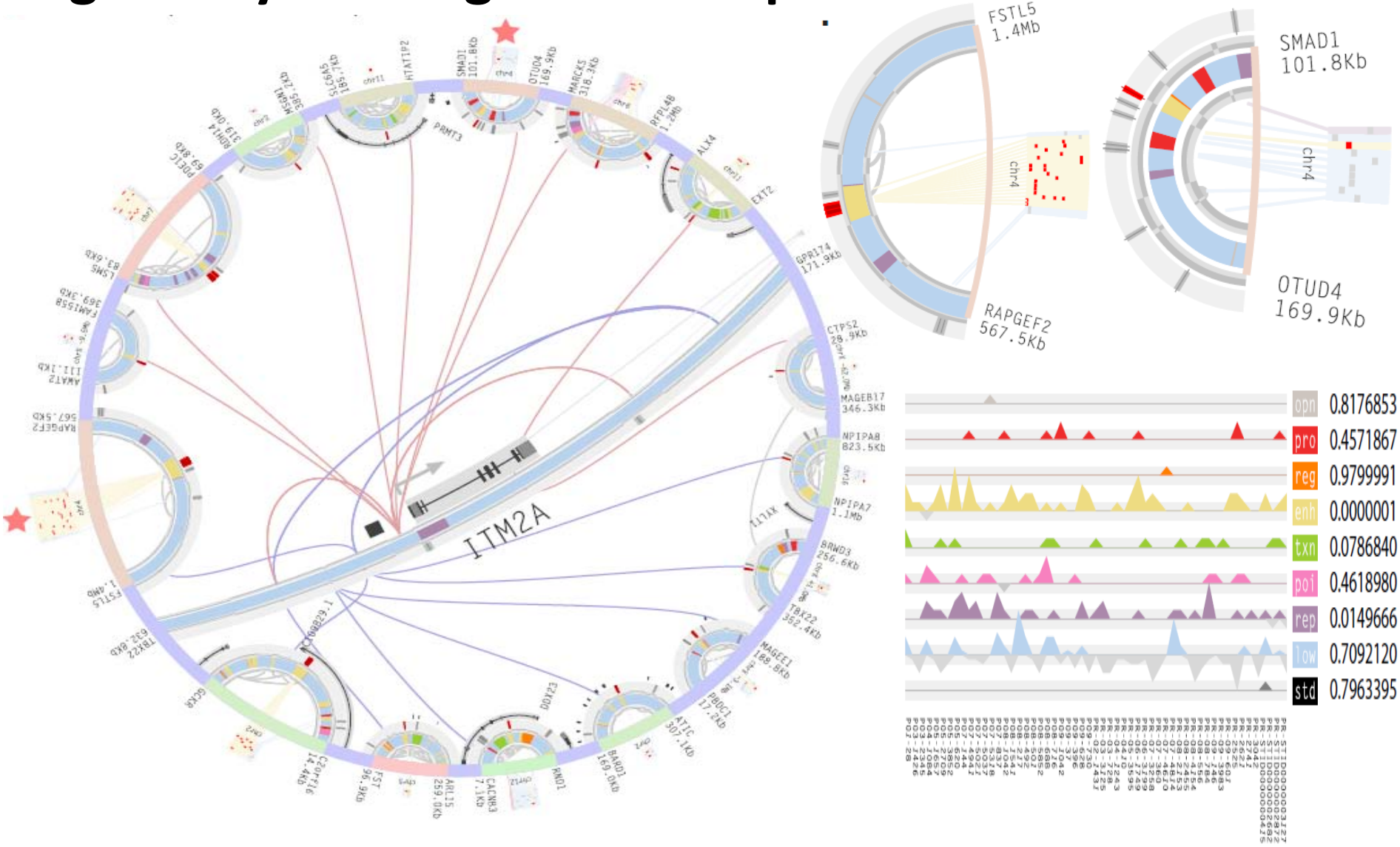
- Many variants, small effects → Pathway-level burden/load
- Many false positives → Prioritize w/ regulatory annotations

Cancer mutations

CATGCCTG
CATCCCTG

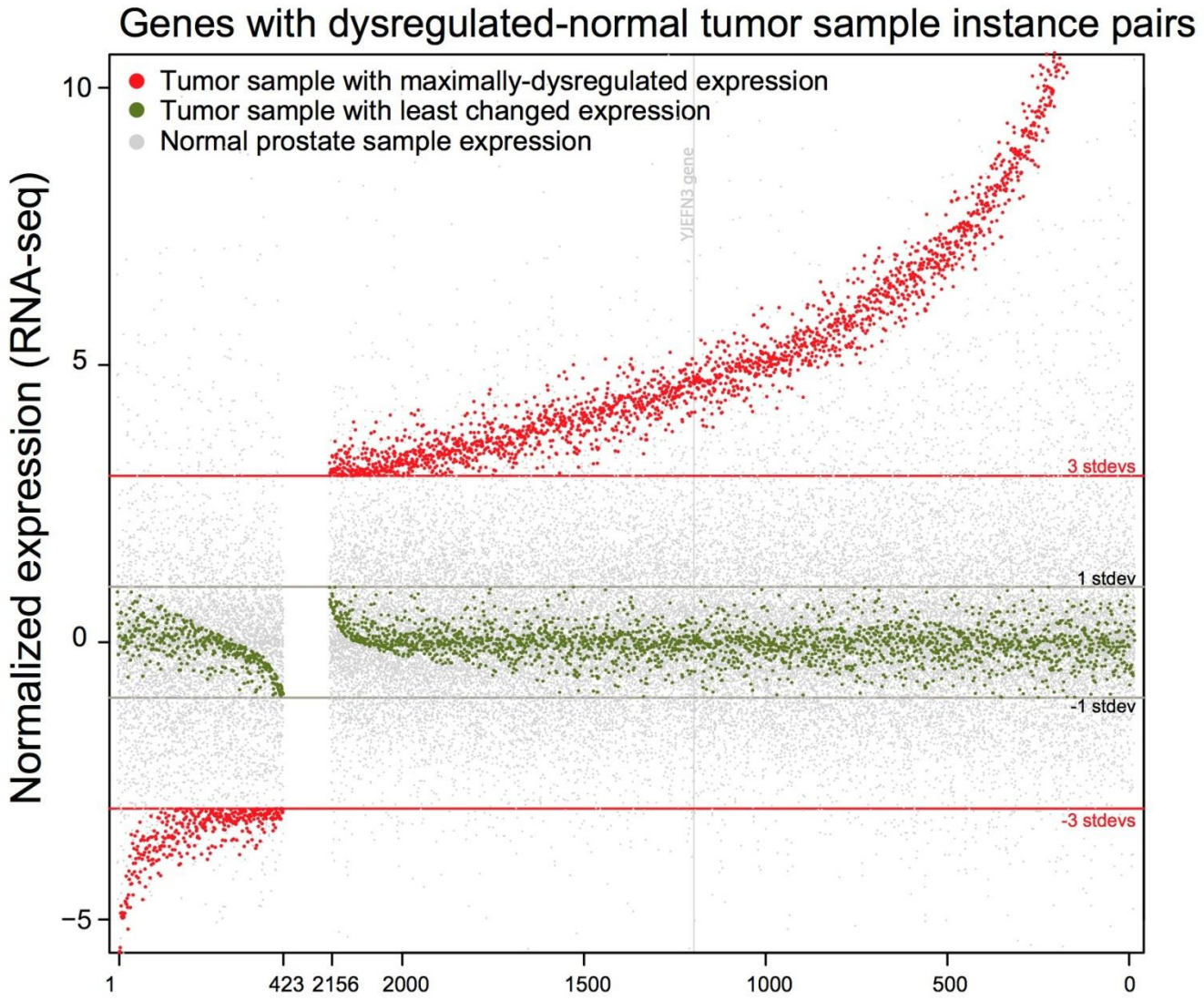
- Loss of function → Protein-coding variants, convergence
- Gain of function → Regulatory variants, heterogeneity

Regulatory convergence of dispersed driver mutations

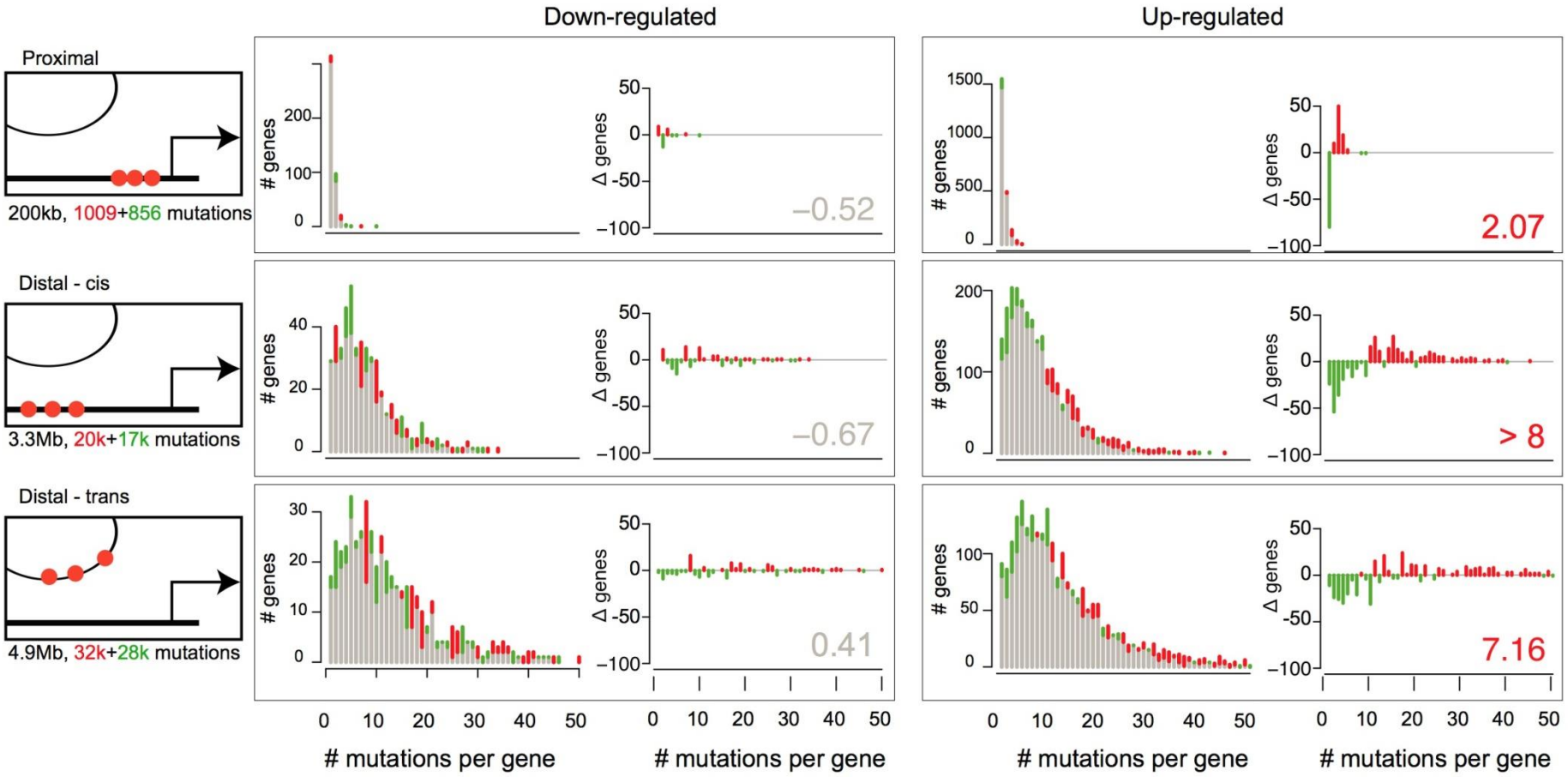


Common mutations in regulatory plexus of each gene

Cancer genes are more likely to be up-regulated

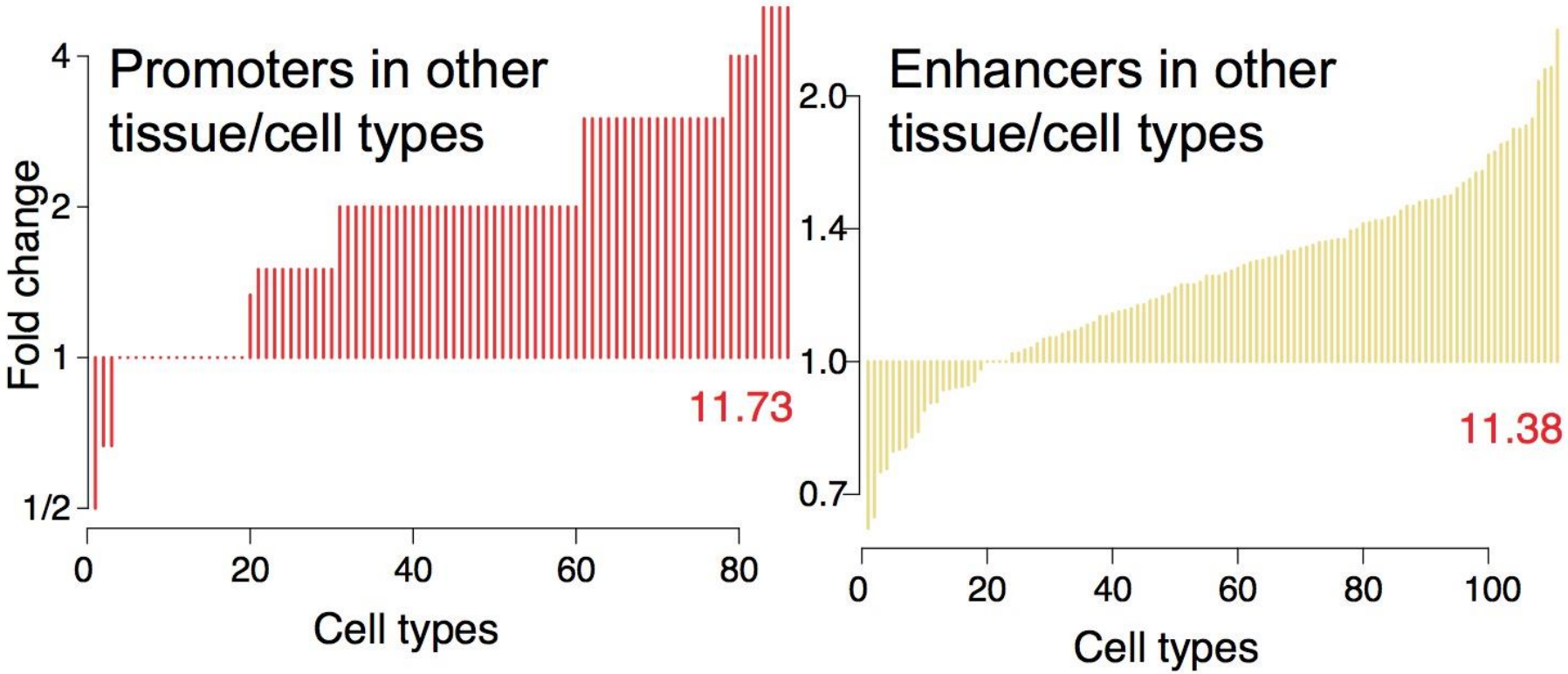


Dysregulated genes show dispersed non-coding mutations



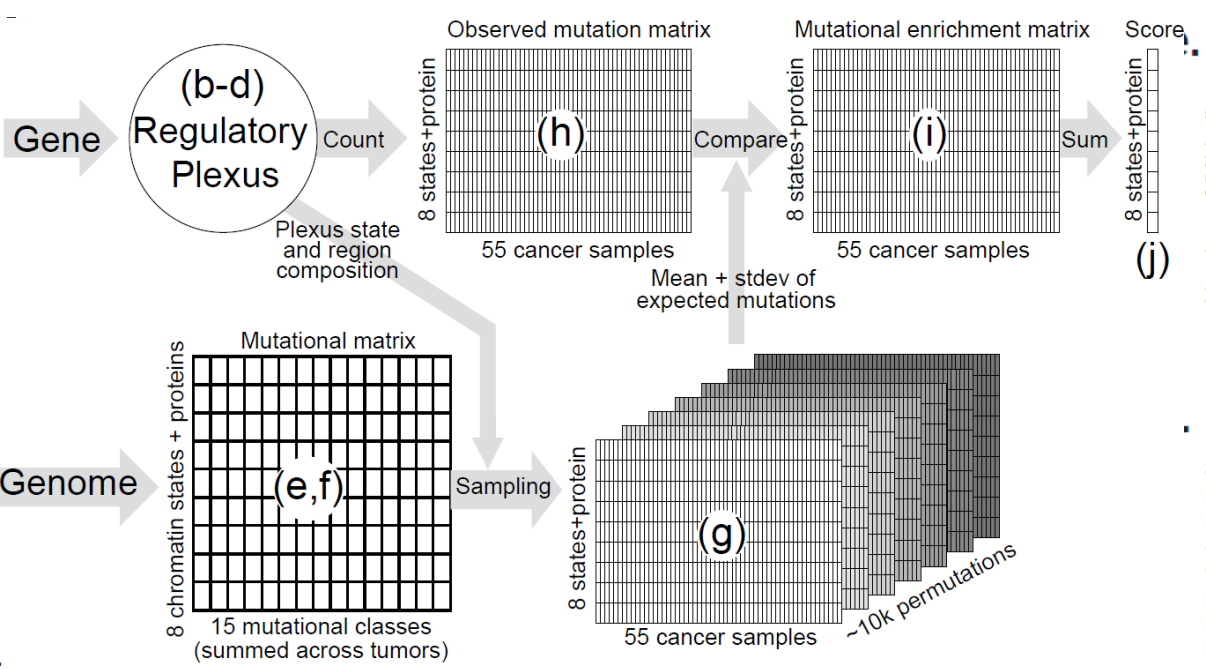
Dysregulated genes are enriched for plexus mutations at all distances.

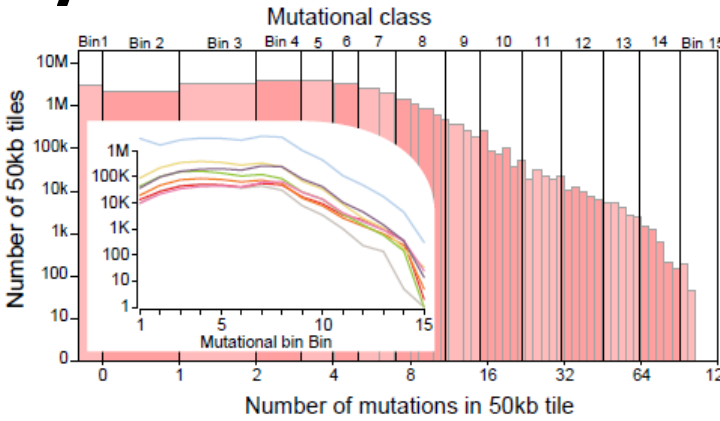
Non-coding mutations enriched in promoters / enhancers active in other cell types



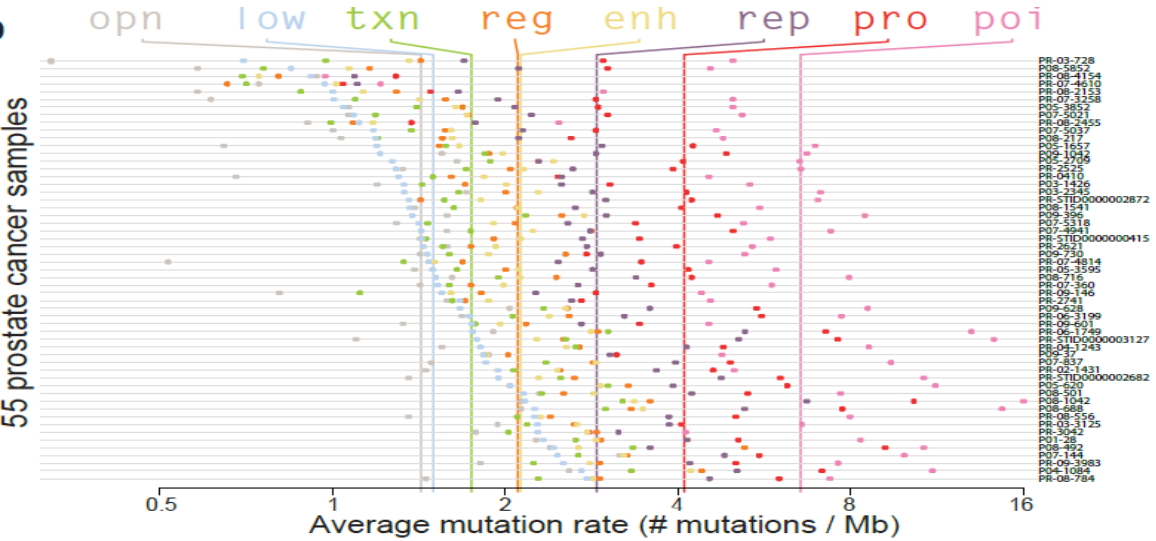
Disruptive mutations in 'low' elements are enriched in enhancers and promoters in other tissues

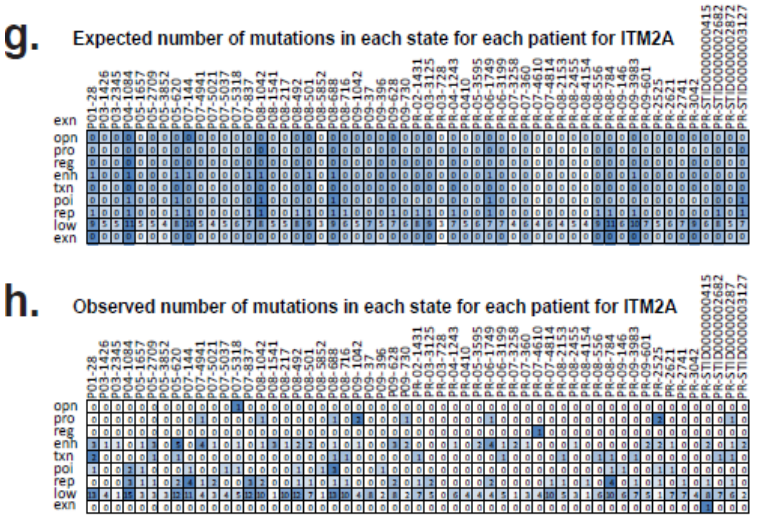
Statistical model for excess of rare/somatic variants





	Bin1	Bin2	Bin3	Bin4	Bin5	Bin6	Bin7	Bin8	Bin9	Bin10	Bin11	Bin12	Bin13	Bin14	Bin15
opn	12	32	37	28	52	33	13	23	8	0	2	0	0	0	0
pro	6	61	78	68	73	55	78	75	17	1	0	0	0	0	0
reg	16	38	115	41	54	90	34	34	9	3	0	0	0	0	0
enh	95	227	274	263	239	186	162	238	85	23	13	0	0	0	0
txn	82	160	339	236	177	105	36	65	32	1	0	0	0	0	0
poi	6	66	59	19	36	37	76	69	65	7	8	0	0	0	0
rep	12	52	64	127	173	219	299	306	90	52	1	0	0	0	0
low	1281	3710	5255	5693	5635	4460	6060	5891	1166	338	233	0	0	0	0

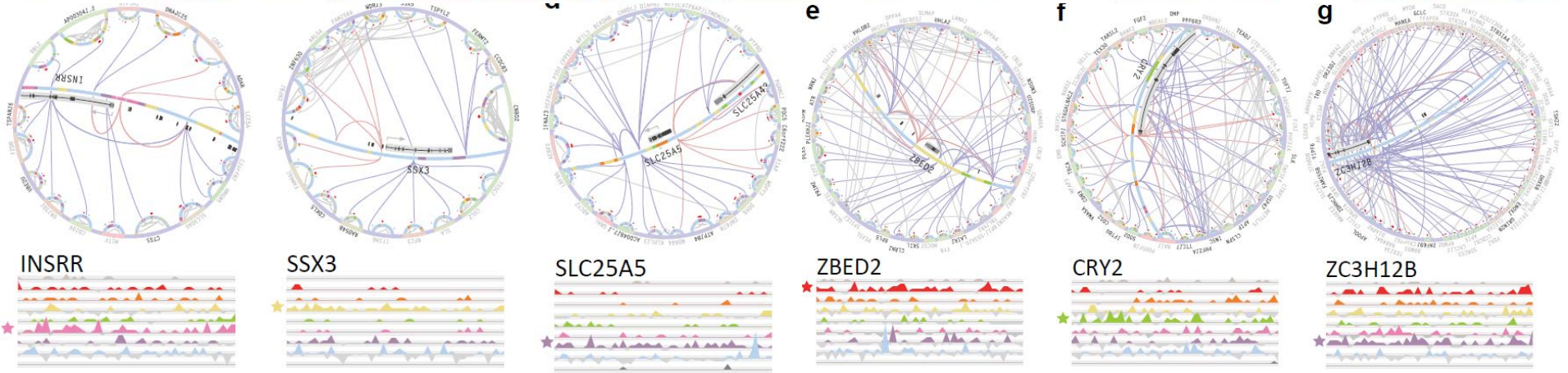




- **Correct for region-, state-, tumor-specific rate variation**

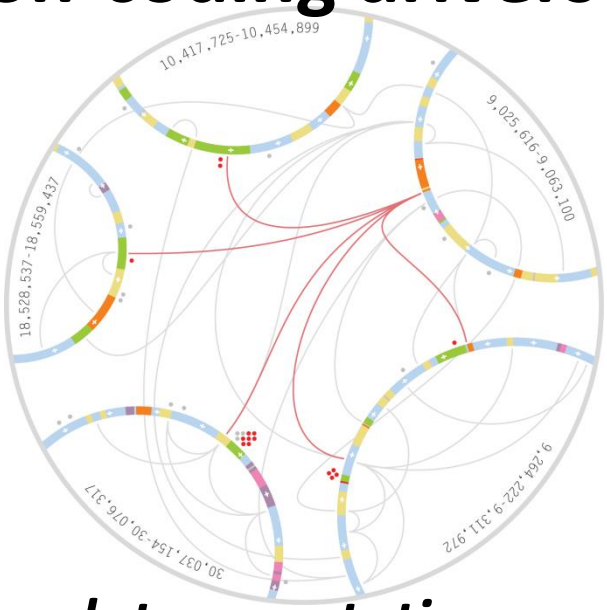
Convergence in immune, signaling, mitoch. functions

Gene	Enriched chromatin state: State	P-val.	Conv	N elmts	kb	Nchr	Nmuts	All regulatory chromatin states: Conv	N elmts	kb	Nchr	Nmuts	Panel	Gene function	ConvGrp
ITM2A	enh	1E-07	56%	245	181	20	58	87%	793	483	20	129		immun-evas	(2)
INSRR	poi	2E-07	62%	142	99	18	76	100%	2727	1756	23	267	(b)	ins-androgen	(1)
ZCCHC16	txn	3E-07	60%	193	251	22	63	98%	2138	1347	23	275		unknown	N/A
ZBED2	pro	9E-07	71%	331	176	21	72	100%	4173	2771	23	421	(e)	immun-evas	(2)
SPANXN3	pro	1E-06	35%	76	37	14	24	85%	1047	710	22	124		spermatogen	(1)
PLCB4	txn	2E-06	67%	341	419	23	79	98%	3576	2497	23	362		ins-androgen	(1)
COQ3	pro	2E-06	64%	247	125	20	58	100%	2913	1957	23	326		mitochondr	(3)
EDNRA	txn	2E-06	75%	452	586	19	102	100%	4238	2734	23	392		blood flow	N/A
CRY2	txn	3E-06	65%	358	455	21	83	100%	3569	2376	23	289	(f)	ins-androgen	(1)
ZC3H12B	rep	3E-06	89%	576	505	22	150	100%	2661	1754	23	415	(g)	immun-evas	(2)
C14orf180	txn	4E-06	51%	133	161	19	49	93%	1700	1243	23	142		secreted	N/A
IDO2	rep	4E-06	87%	378	368	20	98	98%	1525	1058	21	189		immun-evas	(2)
RRAD	poi	4E-06	51%	113	72	20	47	95%	2091	1430	22	180		ins-androgen	(1)
SLC25A5	rep	4E-06	67%	196	189	18	68	91%	1332	853	21	143	(d)	mitochondr	(3)
SSX3	enh	4E-06	65%	230	196	20	53	78%	682	468	21	104	(c)	spermatogen	(1)

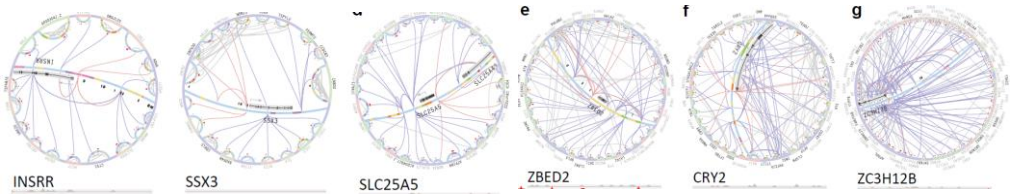


- **Pathway-level convergence, hierarchical model**

Non-coding drivers of prostate cancer dysregulation

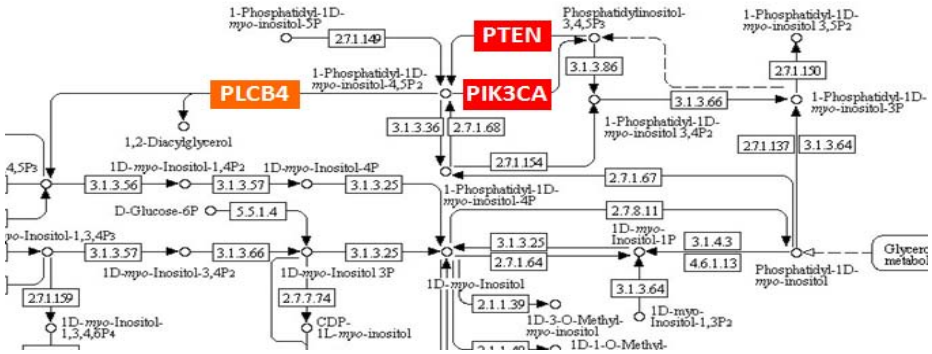


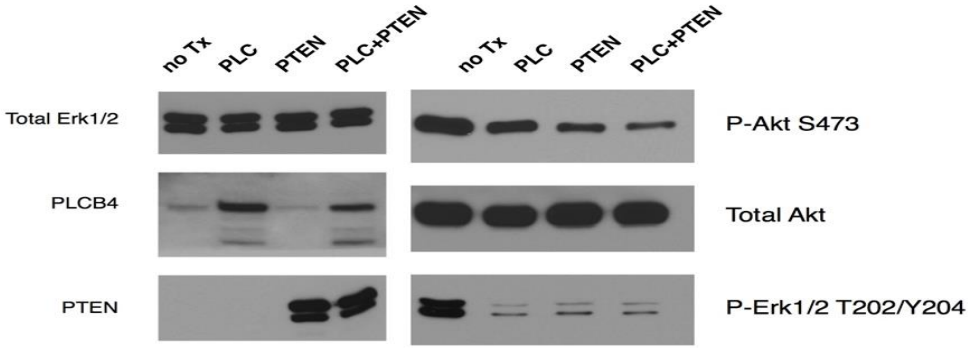
Gene	State	P-val.	Conv	N elmts	kb	Nchr	Nmuts	Conv	N elmts	kb	Nchr	Nmuts	Panel	ConvGrp Gene function	
		Enriched chromatin state						All regulatory chromatin states							
ITM2A	enh	1E-07	56%	245	181	20	58	87%	793	483	20	129		immun-evas	(2)
INSRR	poi	2E-07	62%	142	99	18	76	100%	2727	1756	23	267	(b)	ins-androgen	(1)
ZCCHC16	txn	3E-07	60%	193	251	22	63	98%	2138	1347	23	275		unknown	N/A
ZBED2	pro	9E-07	71%	331	176	21	72	100%	4173	2771	23	421	(e)	immun-evas	(2)
SPANXN3	pro	1E-06	35%	76	37	14	24	85%	1047	710	22	124		spermatogen	(1)
PLCB4	txn	2E-06	67%	341	419	23	79	98%	3576	2497	23	362		ins-androgen	(1)
COQ3	pro	2E-06	64%	247	125	20	58	100%	2913	1957	23	326		mitochondr	(3)
EDNRA	txn	2E-06	75%	452	586	19	102	100%	4238	2734	23	392		blood flow	N/A
CRY2	txn	3E-06	65%	358	455	21	83	100%	3569	2376	23	289	(f)	ins-androgen	(1)
ZC3H12B	rep	3E-06	89%	576	505	22	150	100%	2661	1754	23	415	(g)	immun-evas	(2)
C14orf180	txn	4E-06	51%	133	161	19	49	93%	1700	1243	23	142		secreted	N/A
IDO2	rep	4E-06	87%	378	368	20	98	98%	1525	1058	21	189		immun-evas	(2)
RRAD	poi	4E-06	51%	113	72	20	47	95%	2091	1430	22	180		ins-androgen	(1)
SLC25A5	rep	4E-06	67%	196	189	18	68	91%	1332	853	21	143	(d)	mitochondr	(3)
SSX3	enh	4E-06	65%	230	196	20	53	78%	682	468	21	104	(c)	spermatogen	(1)



Regulatory mutations reveal new cancer driver genes

Convergence in immune, signaling, mitochondrial functions





Convergence in inositol phosphate metabolism adjacent to PTEN, PIK3CA, known cancer genes

PLCB4 overexpression in PC3 prostate cancer reduces Erk/Akt activity, synergistic with PTEN

Personal genomics tomorrow: Already 100,000s of complete genomes

- Health, disease, quantitative traits:
  - Genomics regions ➔ disease mechanism, drug targets
  - Protein-coding ➔ cracking regulatory code, variation
  - Single genes ➔ systems, gene interactions, pathways
- Human ancestry:
  - Resolve all of human ancestral relationships
  - Complete history of all migrations, selective events
  - Resolve common inheritance vs. trait association
- What's missing is the computation
  - New algorithms, machine learning, dimensionality reduction
  - Individualized treatment from 1000s genes, genome
  - Understand missing heritability
  - Reveal co-evolution between genes/elements
  - Correct for modulating effects in GWAS

Challenge ahead: From research to clinic

1. Systematic medical genotyping / sequencing
   – Currently a curiosity, future: medical practice
2. Systematic medical molecular profiling
   – Functional genomics in relevant cell types
3. Systematic perturbation studies for validation
   – 1000s of regulatory predictions x 100s cell types
4. Systematic repurposing of approved drugs
   – Systems-biology view of drug response
5. Genomics of drug response in clinical trials
   – Personalized drug prescription and combinations
6. Partnerships: academia, industry, hospitals
   – Interdisciplinary training in each of the instituttions

Summary: Personalized & Predictive Medicine

1. Intro to Epidemiology: basis of human disease
2. Genetic Epidemiology:
   - Genetic basis: GWAS and screening
   - Interpreting GWAS with functional genomics
   - Calculating functional enrichments for GWAS loci
3. Molecular epidemiology
   - meQTLs: Genotype-Epigenome association (cis-/trans-)
   - EWAS: Epigenome-Disease association
4. Resolving Causality
   - Statistical: Mendelian Randomization
   - Application to genotype + methylation in AD
5. Systems Genomics and Epigenomics of disease
   - Beyond single loci: polygenic risk prediction models
   - Sub-threshold loci and somatic heterogeneity in cancer

6.047 / 6.878 / HST.507 Computational Biology
Fall 2015