Regulatory genomics to interpret complex disease genetics

Luke Ward

Manolis Kellis lab, MIT

July 18, 2013

Regulatory genomics to interpret complex disease genetics

1. Regulatory annotations of the human genome: an overview
2. Using regulatory annotations to interpret GWAS
   a. Locus level
   b. Systems level
3. Beyond GWAS
   a. Molecular variability
   b. Empowering rare-variant and pathway analysis

Regulatory genomics to interpret complex disease genetics

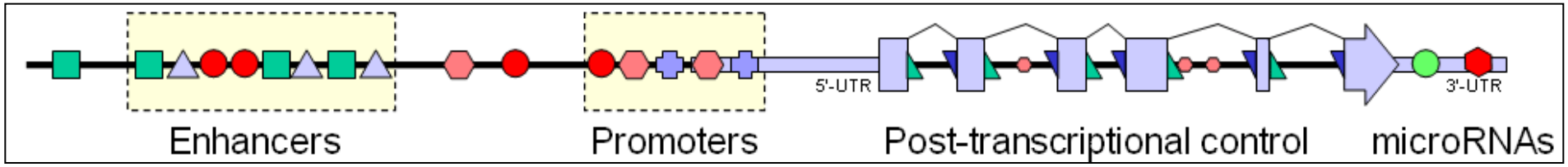
1. Regulatory annotations of the human genome: an overview
2. Using regulatory annotations to interpret GWAS
   a. Locus level
   b. Systems level
3. Beyond GWAS
   a. Molecular variability
   b. Empowering rare-variant and pathway analysis

There are lots of noncoding regulatory mechanisms for disease genetics

Noncoding element disrupted	Molecular function and effect of mutations.	Disease association
Splice-junction and splicing-enhancer	Splicing is constitutive for some transcripts and highly tissue-specific for others, relying on both canonical sequences at the exon-intron junction as well as weakly-specified sequence motifs distributed throughout the transcript. Mutations affecting constitutive splice sites can have an effect similar to nonsense or missense mutations, resulting in aberrantly included introns or skipped exons, sometimes resulting in nonsense-mediated decay (NMD).	Splicing regulatory variants are implicated in several diseases.
		A recent analysis suggests that the majority of disease-causing point mutations in OMIM may exert their effects through splicing.
		Alternative splice site variants in the *WT1* gene are involved in Frasier Syndrome (FS)
		Skipping of exon 7 of the *SMN* gene is involved in spinal muscular atrophy (SMA)
Sequences regulating translation, stability, and localization	Sequences in the 5'-untranslated regions (UTRs) of mRNAs can influence translation regulation, such as upstream ORFs, premature AUG or AUC codons, and palindromic sequences that form inhibitory stem loops. Sequence motifs in the 3'-UTR are recognized by microRNAs and RNA-binding proteins (RBPs).	Loss-of-function mutations in the 5'-UTR of *CDKN2A* predispose individuals to melanoma.
		A rare mutation that creates a binding site for the miRNA hs-miR-189 in the transcript of the gene *SLITRK1* is associated with Tourette's syndrome.
Genes encoding trans-regulatory RNA	Non-coding RNAs participate in a panoply of regulatory functions, ranging from the well-understood transfer and ribosomal RNA to the recently-discovered long non-coding RNAs.	Both rare and common mutations in the gene *RMRP* encoding an RNA component of the mitochondrial RNA processing ribonuclease have been associated with cartilage-hair hypoplasia
		Non-coding RNA mutations can cause many other diseases.
Promoter	Promoter regions are an essential component of transcription initiation and the assembly of RNA polymerase and associated regulators. Mutations can affect binding of activators or repressors, chromatin state, nucleosome positioning, and also looping contacts of promoters with distal regulatory elements. Genes with coding disease mutations can also harbor independently-associated regulatory variants that correlate with expression, are bound by proteins in an allele-specific manner, and disrupt or create regulatory motifs	Mutations in the promoter of the HIV1-progression associated gene *CCR5*, are correlated with expression of the receptor it encodes and bind differentially to at least three transcription factors
		APOE promoter mutations are associated with Alzheimer's disease
		Heme oxygenase-1 (*HO-1*) promoter mutations lead to expression changes and are associated with many diseases
Enhancer	Enhancers are distal regulatory elements that often lie 10,000 to 100,000 nucleotides from the start of their target gene. Mutations within them can disrupt sequence motifs for sequence-specific transcription factors, chromatin regulators, and nucleosome positioning signals. Structural variants including inversions and translocations can disrupt their regulatory activity by moving them away from their targets, disrupting local chromatin conformation, or creating interactions with insulators or repressors that can hinder their action. While it is thought that looping interactions with promoter regions play a role, the rules of enhancer-gene targeting are still poorly understood.	The role of distal enhancers in disease was suggested even before GWAS by many Mendelian disorders for which some patients had translocations or other structural variants far from the promoter
		In one early study, point mutations were mapped in an unlinked locus in the intron of a neighboring gene, a million nucleotides away from the developmental gene *Shh* ; this distal locus acted as an enhancer of *Shh* and recapitulated the polydactyly phenotype in mouse.
		A number of GWAS hits have been validated as functional enhancers; for example, common variants associated with cancer susceptibility map to a gene desert on chromosome 8, with one SNP demonstrated to disrupt a TCF7L2 binding site and to inhibit long-range activation of the oncogene *MYC*.
Synonymous mutations within protein-coding sequences	All of the aforementioned regulatory elements can also be encoded within the protein-coding exons themselves. Thus, synonymous mutations within protein-coding regions may be associated with non-coding functions, acting pre-transcriptionally at the DNA level, or post-transcriptionally at the RNA level.	A synonymous variant in the dopamine receptor gene *DRD2* associated with schizophrenia and alcoholism has been shown to modulate receptor production through differences in mRNA folding and stability.

Ward and Kellis, 2012

What we can model with regulatory annotations



Goal: A systems-level understanding of genomes and gene regulation:

- The regulators: Transcription factors, microRNAs, sequence specificities
- The regions: enhancers, promoters, and their tissue-specificity
- The targets: TFs→targets, regulators→enhancers, enhancers→genes
- The grammars: Interplay of multiple TFs → prediction of gene expression

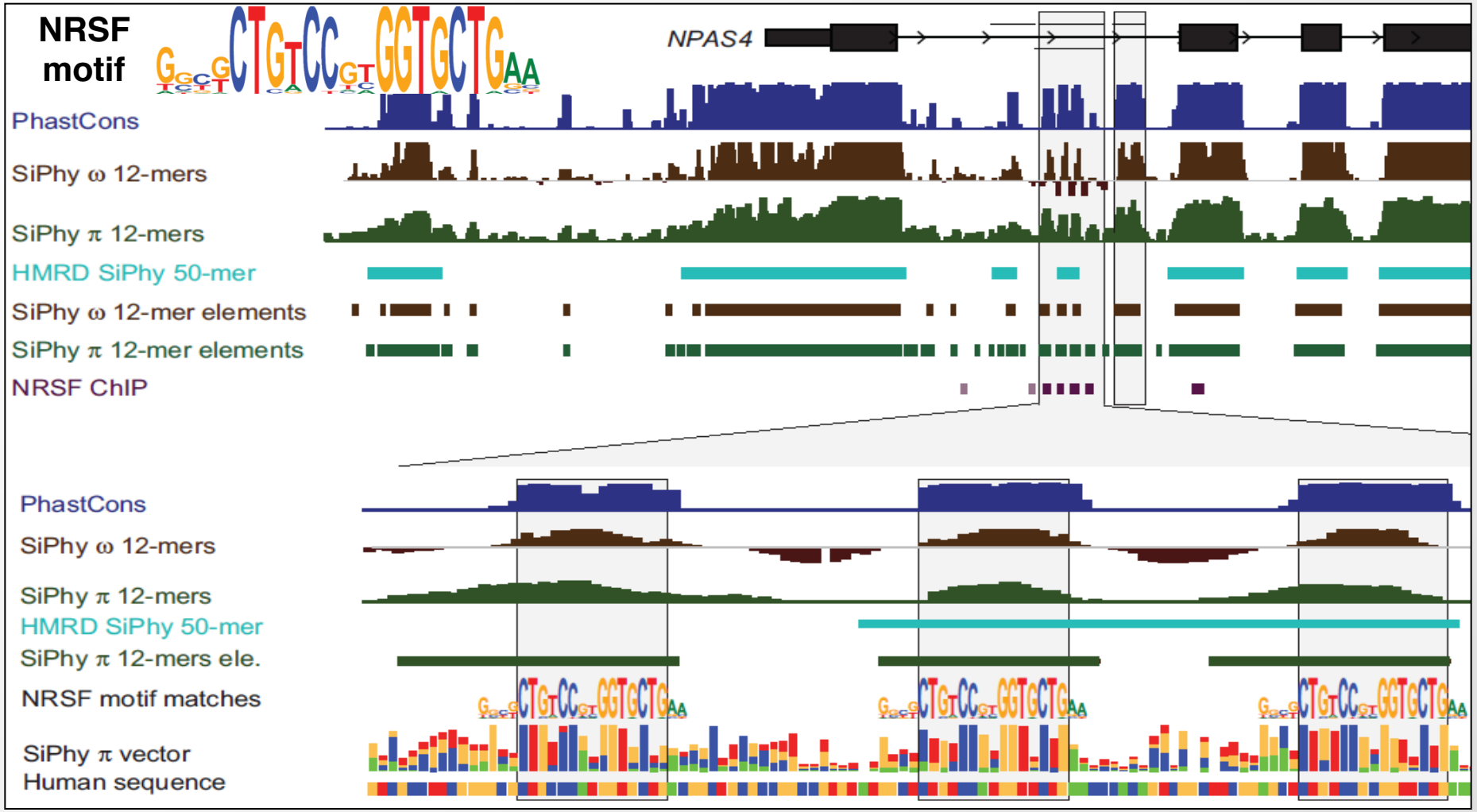
➔ The parts list = Building blocks of gene regulatory networks

Our tools: Comparative genomics & large-scale experimental datasets.

- Evolutionary signatures for coding/non-coding genes, microRNAs, motifs
- Chromatin signatures for regulatory regions and their tissue specificity
- Activity signatures for linking regulators → enhancers → target genes
- Predictive models for gene function, gene expression, chromatin state

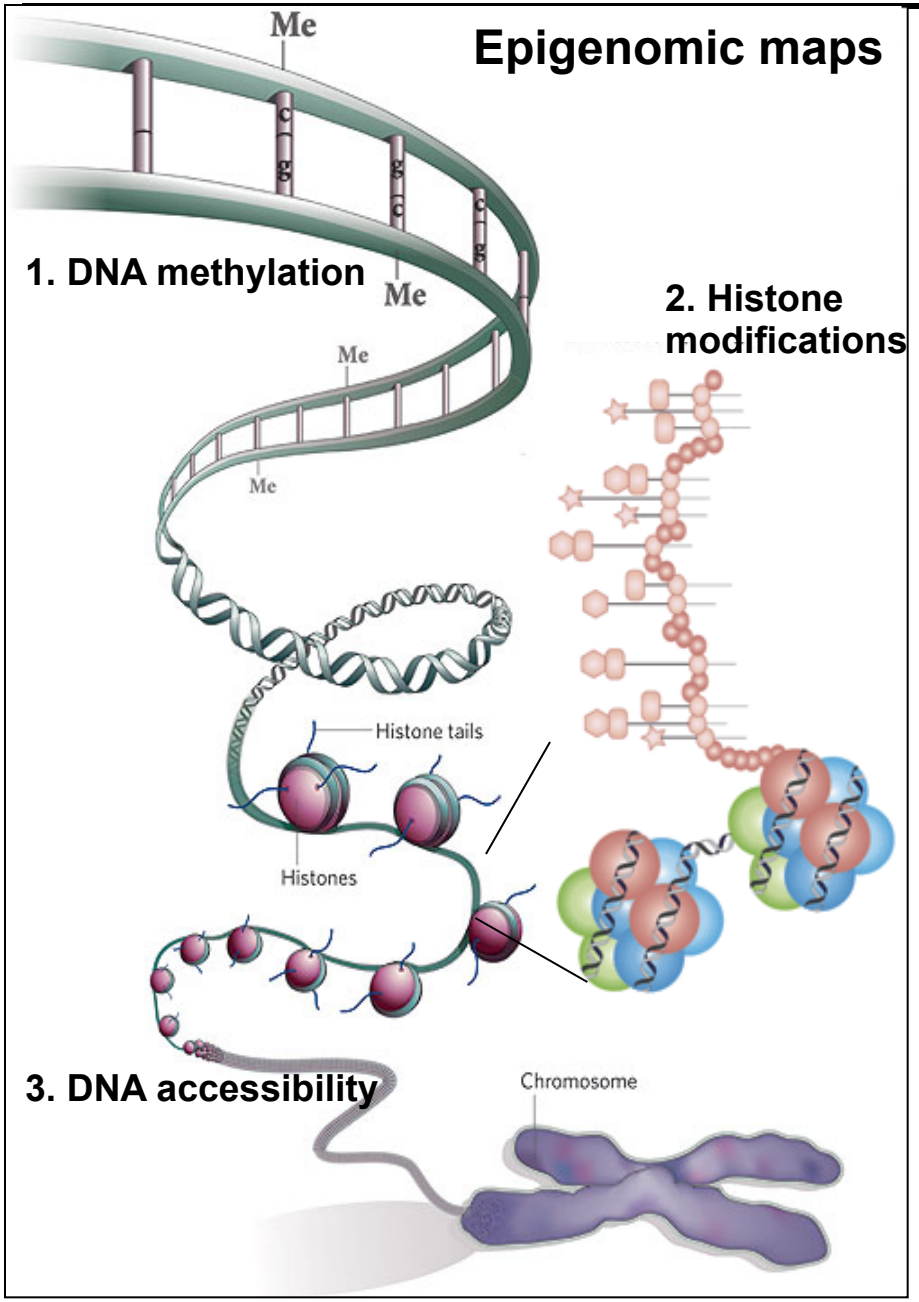
➔Integrative models = Define roles in development, health, disease

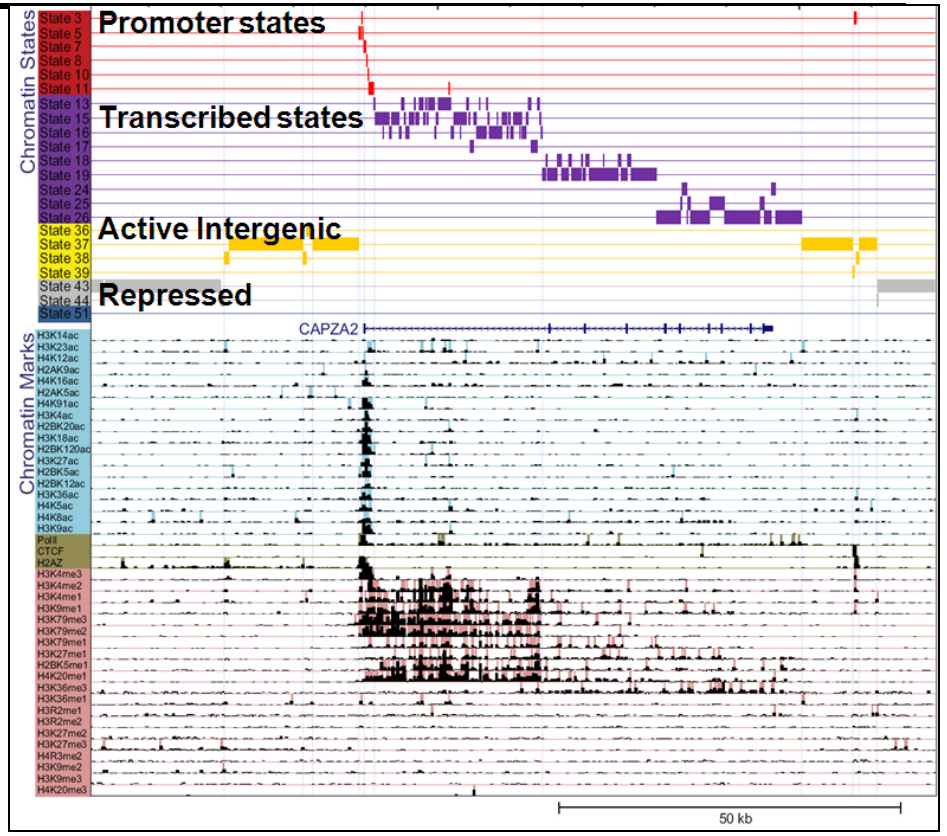
Measuring constraint at individual nucleotides



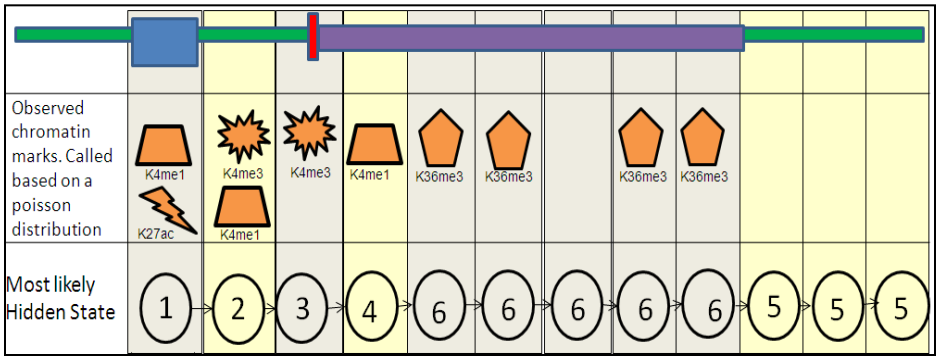
- Reveal individual transcription factor binding sites
- Within motif instances reveal position-specific bias
- More species: motif consensus directly revealed

Chromatin signatures for genome annotation





Ernst et al Nature Biotech 2010

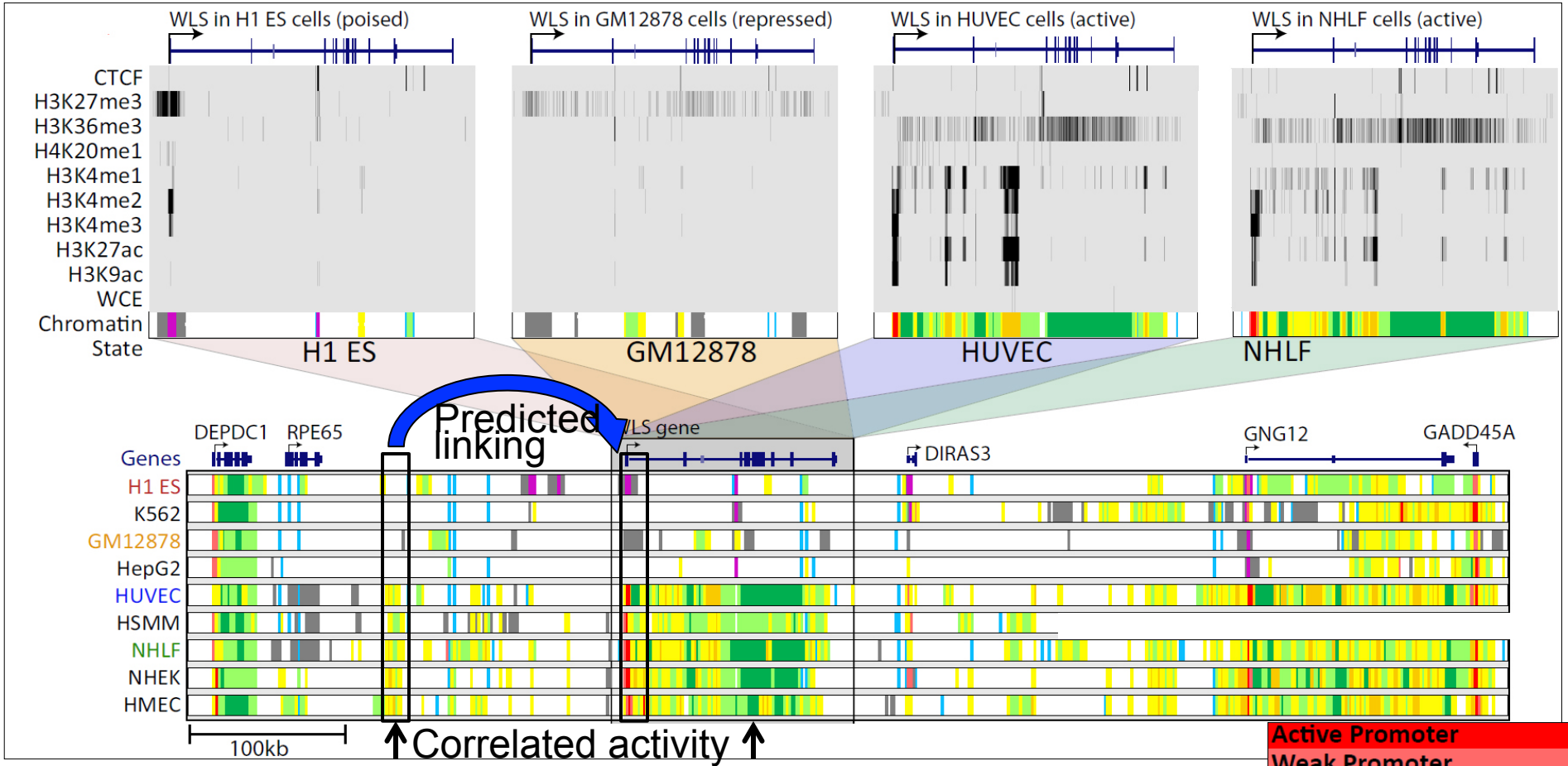


Multi-variate HMM captures combinations

Epigenomics Roadmap: 90 reference epigenomes

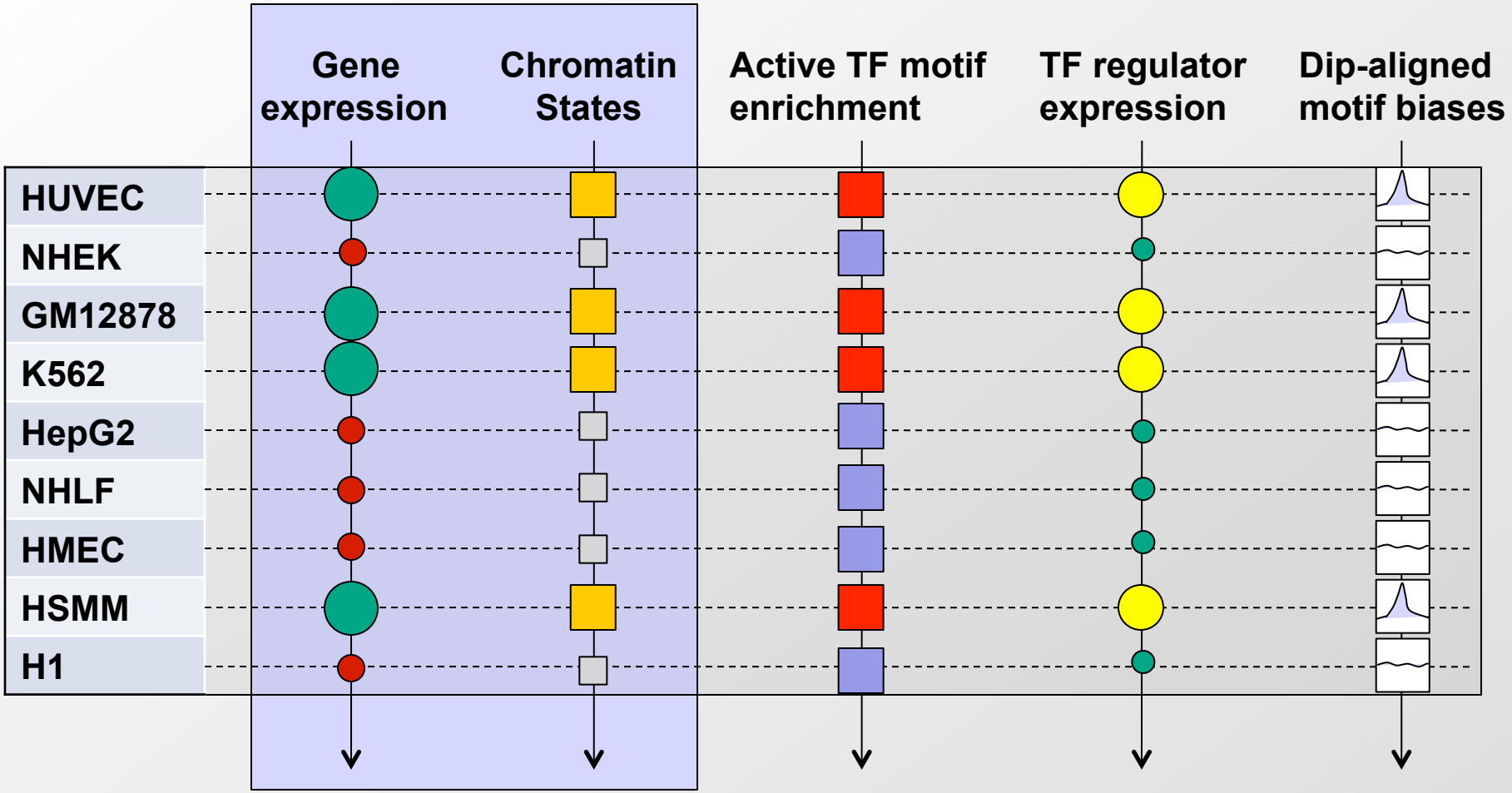
Interpret GWAS, global effects, reveal relevant cell types

Chromatin state dynamics reveal linking/regulators



- Single annotation track for each cell type
- Summarize cell-type activity at a glance
- Study activity pattern across tissues

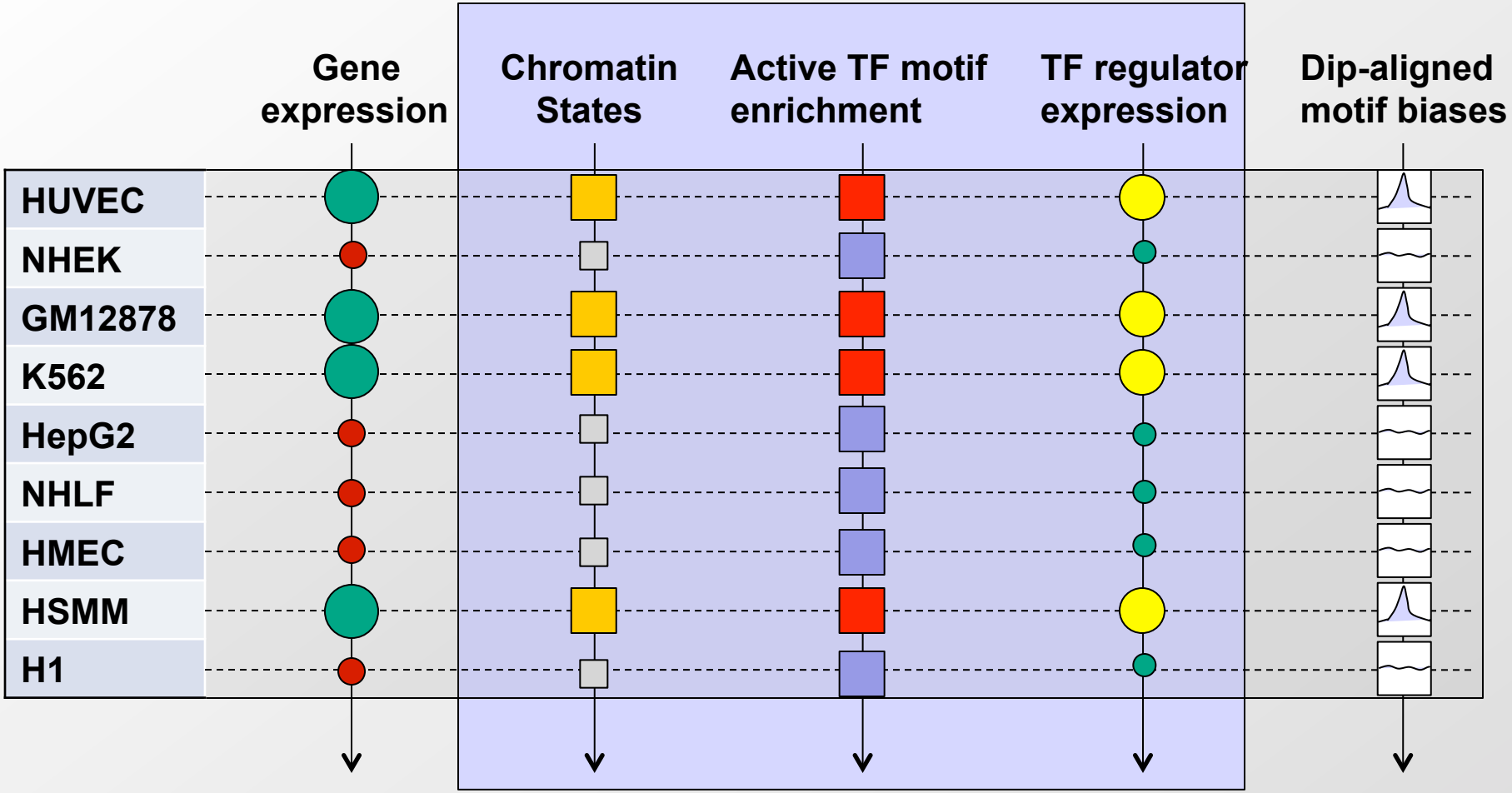
Multi-cell activity profiles connect enhancers



Link enhancers to target genes

ON
OFF
Active enhancer
Repressed
Motif enrichment
Motif depletion
TF On
TF Off
Motif aligned
Flat profile

Multi-cell activity profiles connect enhancers

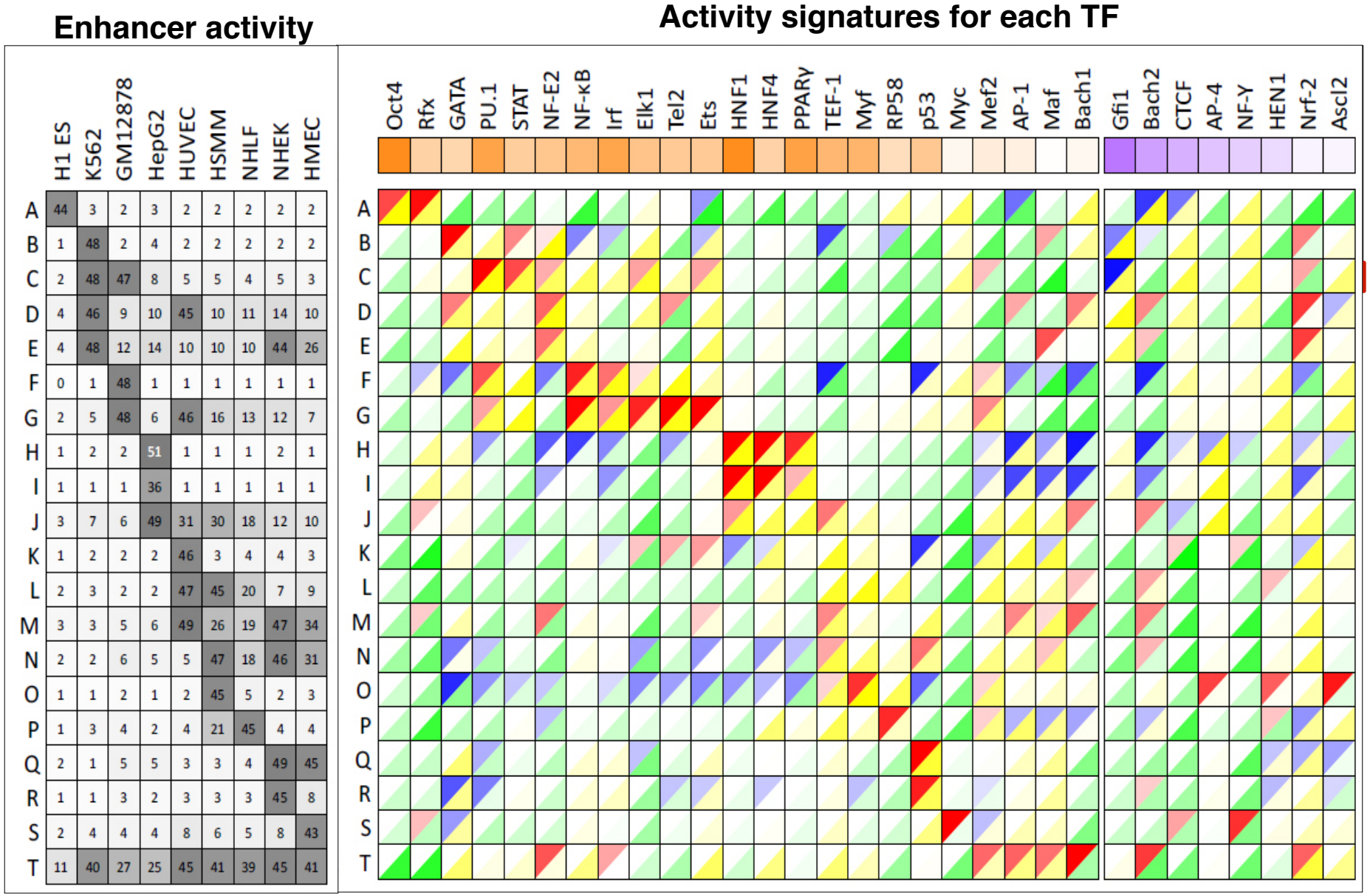


Link TFs to target enhancers

Predict activators vs. repressors

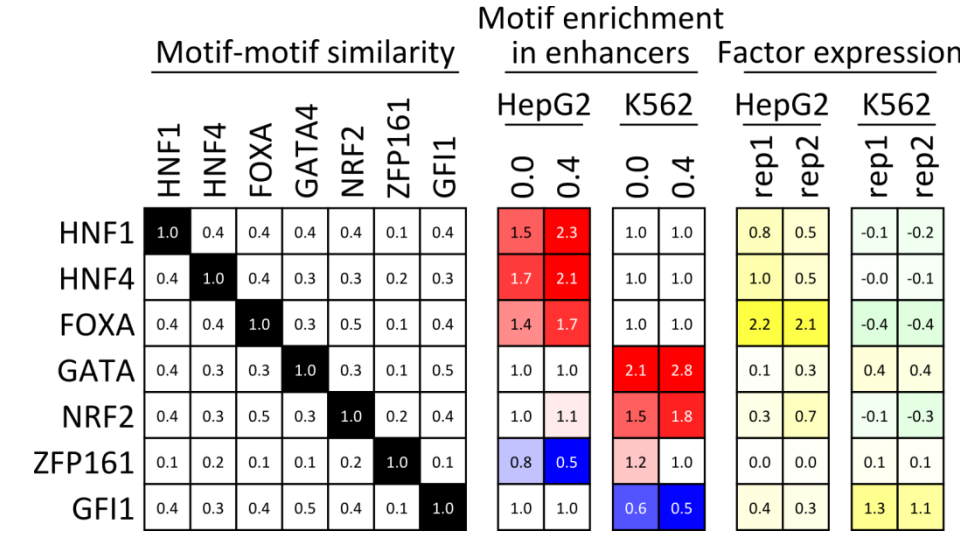
ON
OFF
Active enhancer
Repressed
Motif enrichment
Motif depletion
TF On
TF Off
Motif aligned
Flat profile

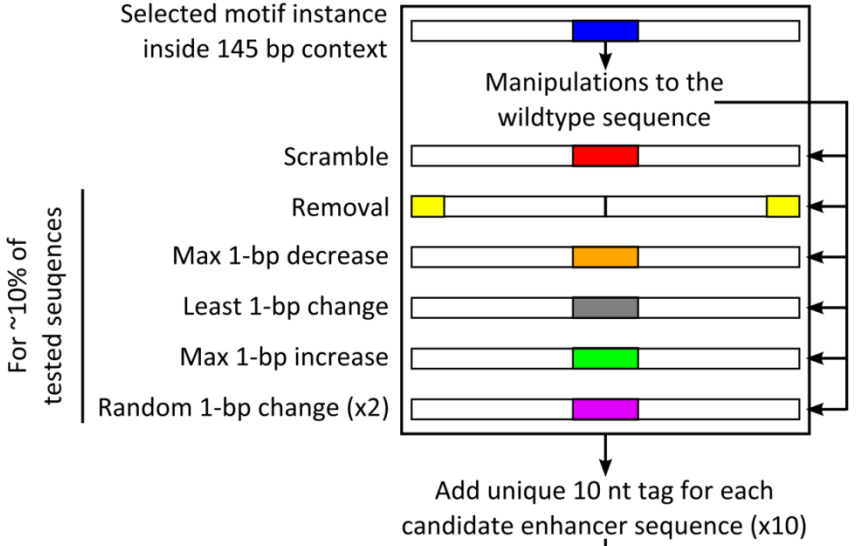
Coordinated activity reveals activators/repressors

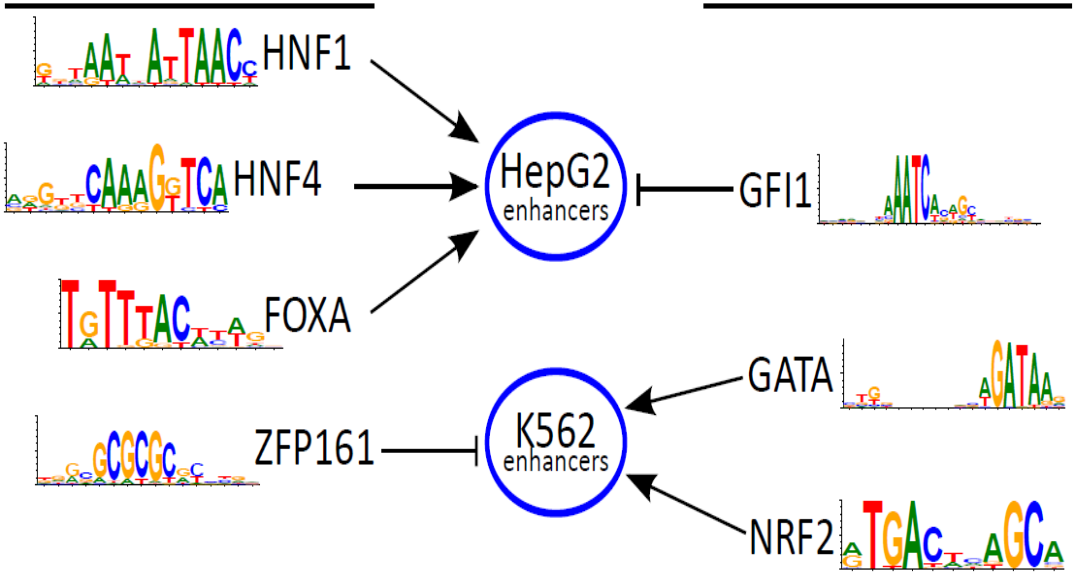


- Enhancer networks: Regulator → enhancer → target gene

Experimental validation of motif activity in tissue-specific enhancers







54000+ measurements (x2 cells, 2x repl)

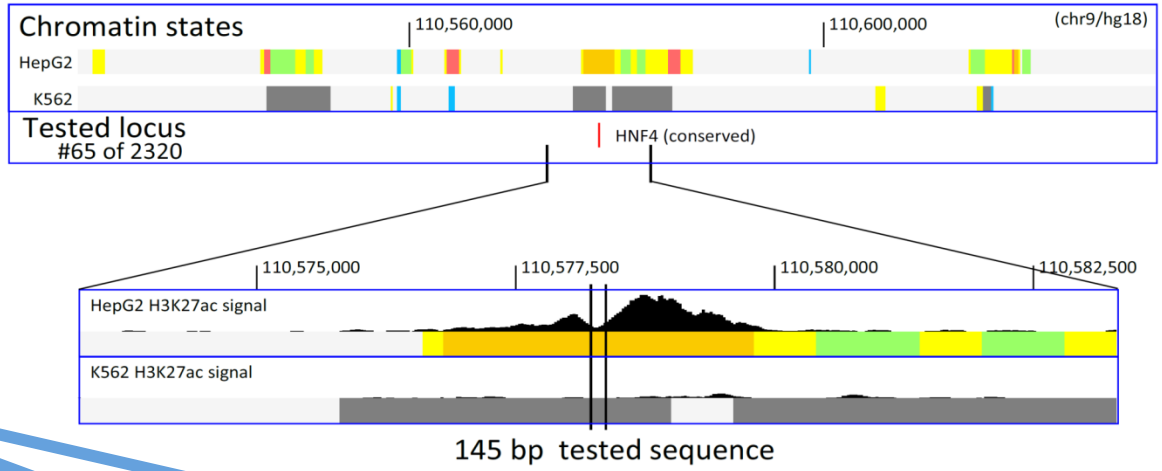
Example activator: conserved HNF4 motif match

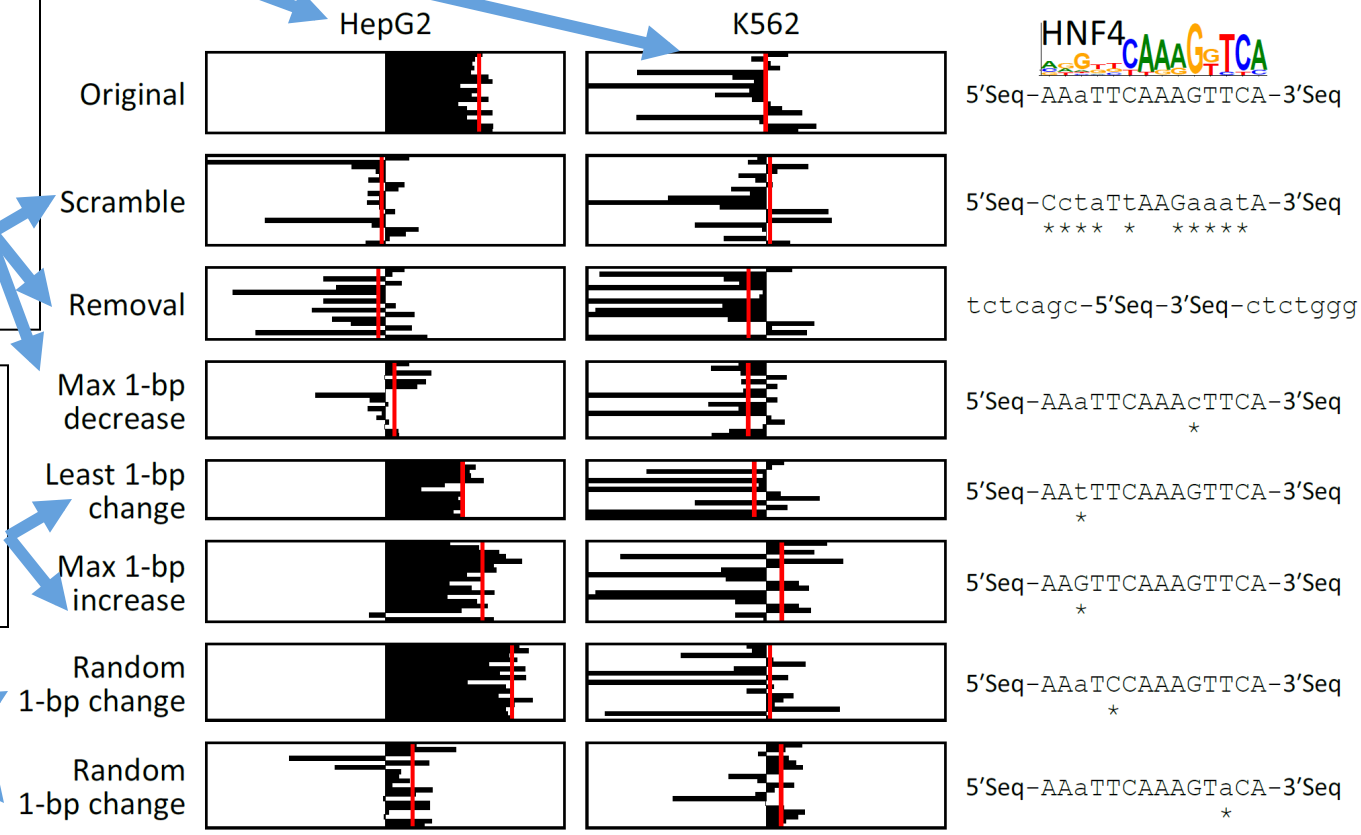
WT expression specific to HepG2

Motif match disruptions reduce expression to background

Non-disruptive changes maintain expression

Random changes depend on effect to motif match





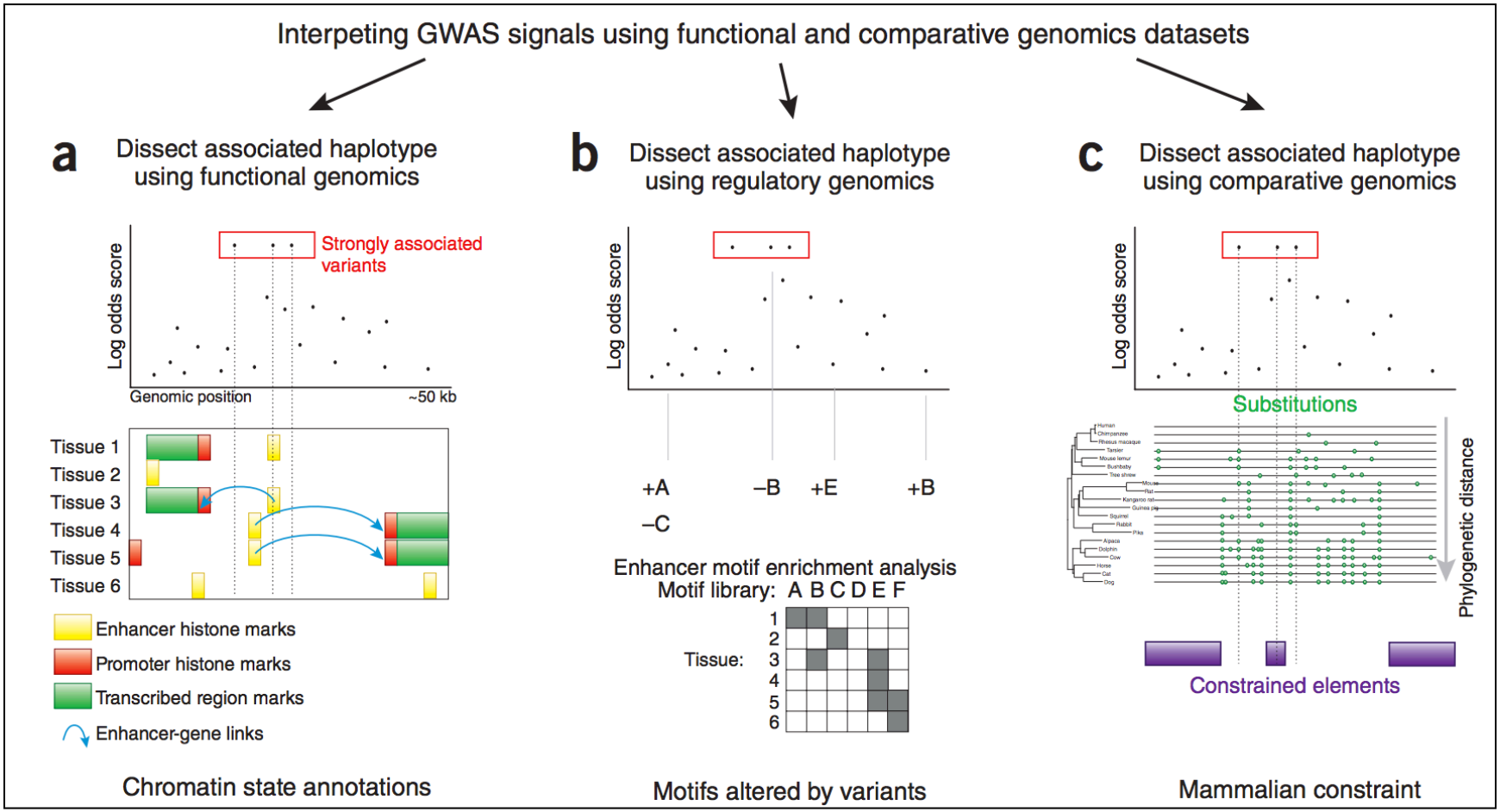
5'Seq = cccacacagccctcccccacctcccagatttgctgggcctgtcatttgaggctaagtcatatagac

3'Seq = ctctgggacatgagcaagggaagacaacaaaacacacattcttctctctgaaatcttttcccagg

Regulatory genomics to interpret complex disease genetics

1. Regulatory annotations of the human genome: an overview
2. Using regulatory annotations to interpret GWAS
   a. Locus level
   b. Systems level
3. Beyond GWAS
   a. Molecular variability
   b. Empowering rare-variant and pathway analysis

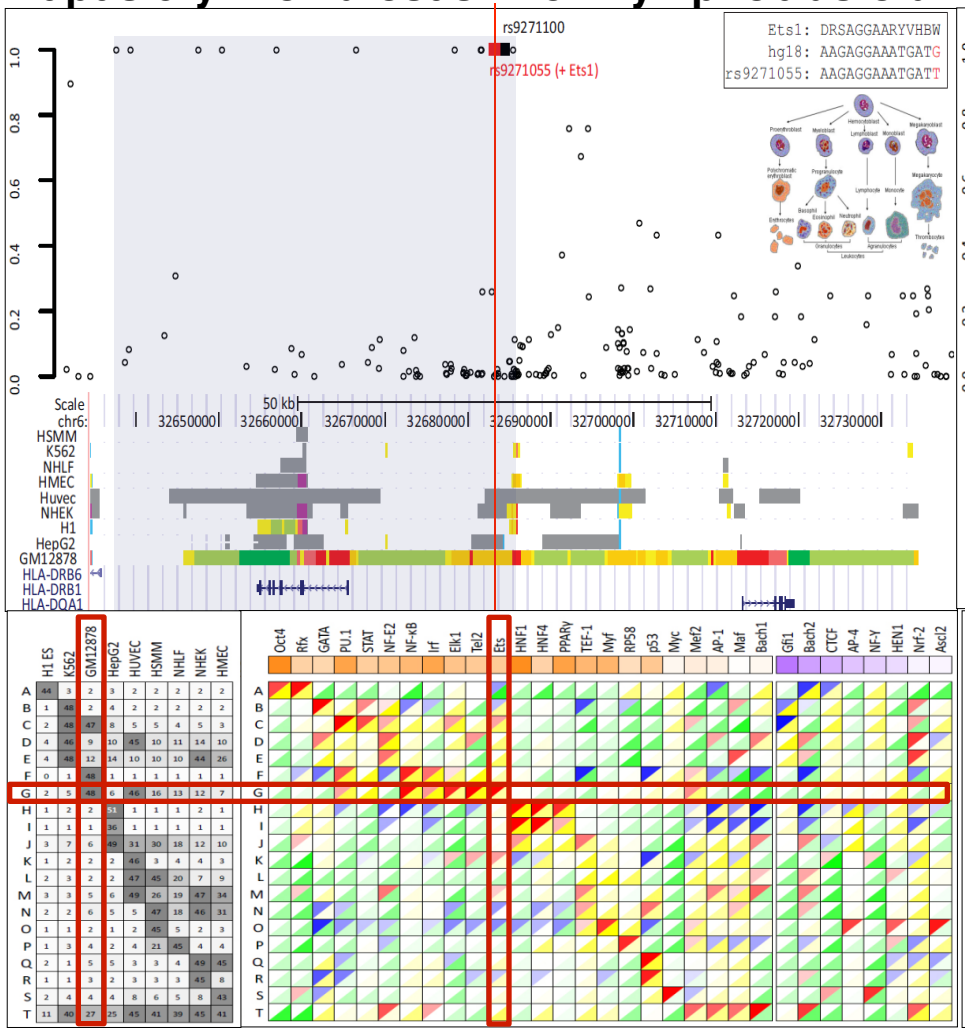
Faced with resolution-limiting LD, annotations can help



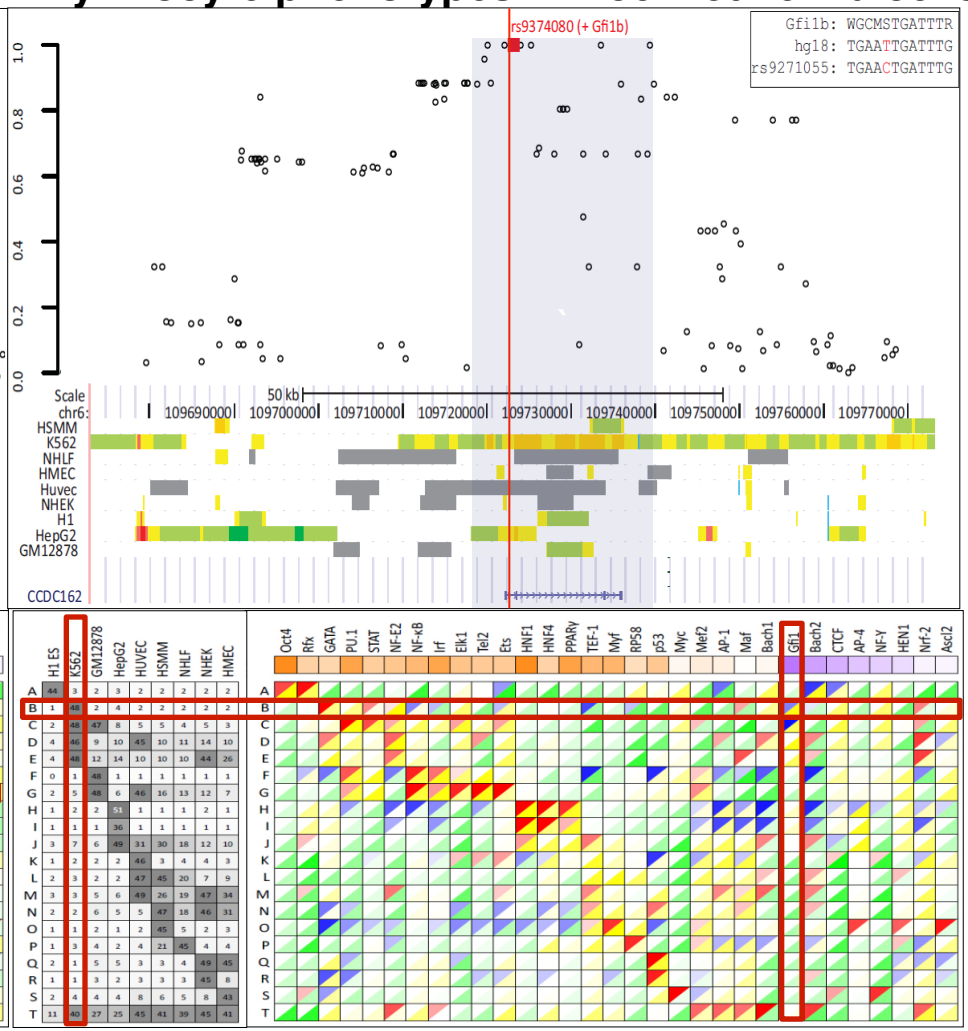
Ward and Kellis, 2012

Mechanistic predictions for top disease-associated SNPs

Lupus erythromatosus in GM lymphoblastoid



Erythrocyte phenotypes in K562 leukemia cells



Disrupt activator Ets-1 motif

➔ Loss of GM-specific activation

➔ Loss of enhancer function

➔ Loss of HLA-DRB1 expression

Creation of repressor Gfi1 motif

➔ Gain K562-specific repression

➔ Loss of enhancer function

➔ Loss of CCDC162 expression

HaploReg: systematic regulatory mining of variants (compbio.mit.edu/HaploReg)

Query SNP: rs17145713 and variants with r^2 >= 0.95

chr	pos (hg19)	LD	variant	Ref	Alt	ASN freq	CEU freq	YRI freq	GERP cons	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	GENCODE genes	RefSeq genes	dbSNP func annot
7	72842724	1	7:72480660	GAC	G	0	0.15	0					PANC-1			5.4kb 5' of FZD9	5.4kb 5' of FZD9	
7	72856430	1	rs1178979	T	C	0.13	0.18	0.3					CLL		GATA	BAZ1B	BAZ1B	intronic
7	72857049	1	rs1178977	A	G	0.14	0.18	0.3							AREB6,DEC	BAZ1B	BAZ1B	intronic
7	72857713	1	rs34604283	CA	C	0.13	0.1	0.2					8 cell types		Sox	BAZ1B	BAZ1B	intronic
7	72868522	1	rs1306476	A	G	0.12	0.18	0.36								BAZ1B	BAZ1B	intronic
7	72883106	1	rs62465144	T	C	0.14	0.18	0.29								BAZ1B	BAZ1B	intronic
7	72885810	1	rs6976930	G	A	0.14	0.18	0.39								BAZ1B	BAZ1B	intronic
7	72904810	1	rs17145713	C	T	0.14	0.18	0.3							ATF3	BAZ1B	BAZ1B	intronic
7	72939244	1	rs11983997	G	C	0.13	0.18	0.26				GM12878, K562	GM12864,GM12878,K562			2.6kb 5' of BAZ1B	2.6kb 3' of BAZ1B	
7	72977249	1	rs34594435	C	T	0.12	0.18	0.03				K562	CMK	KAP1,SETDB1		4.9kb 5' of BCL7B	5.2kb 3' of BCL7B	
7	72988069	1	rs35659126	C	T	0.13	0.18	0.08								TBL2	TBL2	intronic
7	72989141	1	rs34550818	C	CA	0.11	0.12	0						POL2		TBL2	TBL2	intronic
7	72989390	1	rs11974409	A	G	0.13	0.18	0.14								TBL2	TBL2	intronic
7	72998952	1	rs9638180	A	G	0.12	0.18	0.08							Zbtb3	5.8kb 5' of TBL2	5.9kb 3' of TBL2	
7	72999105	1	rs9638182	T	G	0.12	0.18	0.14								6kb 5' of TBL2	6.1kb 3' of TBL2	
7	73007943	1	rs1051921	G	A	0.12	0.18	0.08					4 cell types	POL2		MLXIPL	MLXIPL	3'-UTR

HMEC,iPS,A549,H1-hESC,H7-hESC,HA-h,HNPCEpiC,HRE

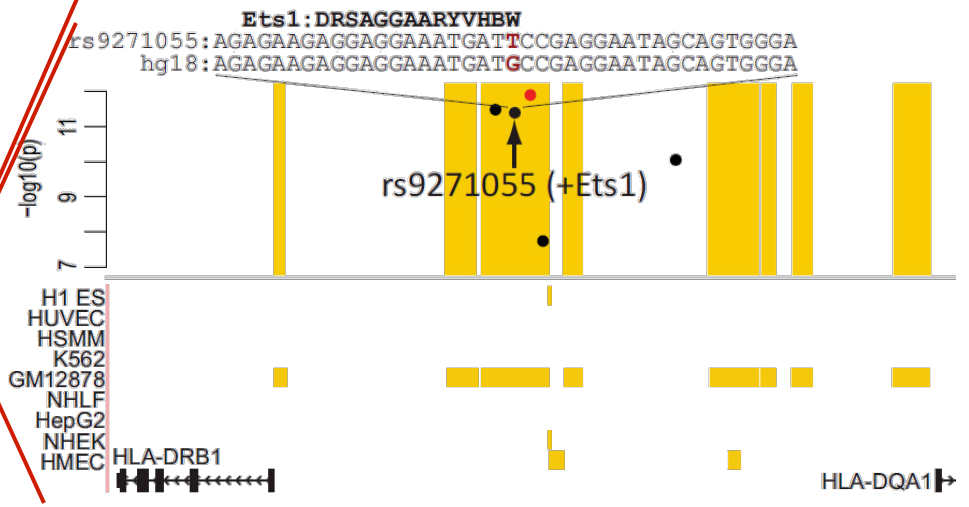
- Start with any list of SNPs or select a GWA study
  - Mine publically available ENCODE+Roadmap data for hit LD blocks
  - Hundreds of assays, dozens of cells, conservation, motifs
  - Integration with published eQTL from GTEx portal (future: ASE)
  - Report systems-level enrichments (future: R tools)

Regulatory genomics to interpret complex disease genetics

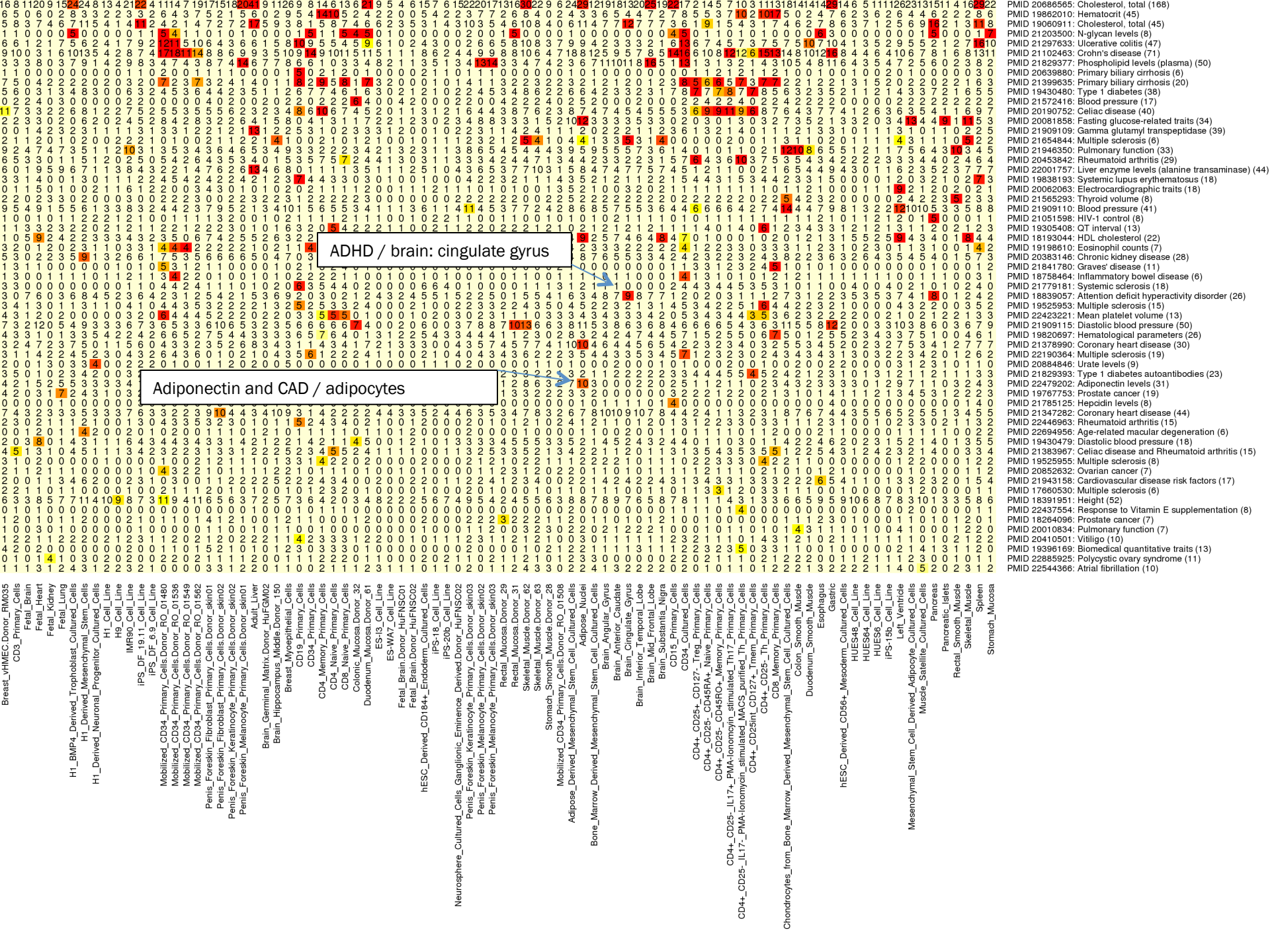
1. Regulatory annotations of the human genome: an overview
2. Using regulatory annotations to interpret GWAS
   a. Locus level
   b. Systems level
3. Beyond GWAS
   a. Molecular variability
   b. Empowering rare-variant and pathway analysis

Phenotype	Top Cell Type	Total #SNPs from Study	#SNPs in enh. States 4 and 5	p-value	FDR	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC
Erythrocyte phenotypes (Ref. 38)	K562	35	9	$<10^{-7}$	0.02	9	17	4	0	0	1	2	1	1
Blood lipids (Ref. 39)	HepG2	101	13	$<10^{-7}$	0.02	3	5	0	11	2	3	3	4	3
Rheumatoid arthritis (Ref. 40)	GM12878	29	7	2.0×10^{-7}	0.03	0	0	15	0	2	0	0	2	3
Primary billary cirrhosis (Ref. 41)	GM12878	6	4	6.0×10^{-7}	0.03	0	11	41	0	0	0	0	8	8
Systemic lupus erythromatosus (Ref. 42)	GM12878	18	6	9.0×10^{-7}	0.03	0	4	21	0	5	8	0	3	5
Lipoprotein cholesterol/triglycerides (Ref. 43)	HepG2	18	5	1.2×10^{-6}	0.03	17	8	0	24	3	6	4	3	3
Hematological traits (Ref. 44)	K562	39	7	1.7×10^{-6}	0.03	0	12	10	2	1	0	0	1	0
Hematological parameters (Ref. 45)	K562	28	6	2.2×10^{-6}	0.03	0	15	7	0	5	7	7	3	2
Colorectal cancer (Ref. 46)	HepG2	4	3	3.8×10^{-6}	0.03	0	0	0	66	0	12	0	12	12
Blood pressure (Ref. 47)	K562	9	4	5.0×10^{-6}	0.04	0	30	14	0	10	6	7	5	11

SNP	H1 ES	K562	GM	HepG2	Huvec	HSMM	NHLF	NHEK	HMEC	Chrom. Band	Gene	Link Scc	Distanc
rs13385731										2p22			
rs10036748										5q33			
rs1385374										12q24	MGC16384	-	1
rs2230926										6q23	TNFAIP3	3.7	7
rs4728142										7q32	IRF5	-	4
rs9271100										6p21	HLA-DRB1	4.5	19
rs4917014										7p12	IKZF1	2.2	38
rs7812879										8p23	BLK	2.9	11
rs2205960										1q25			
rs548234										6q21			



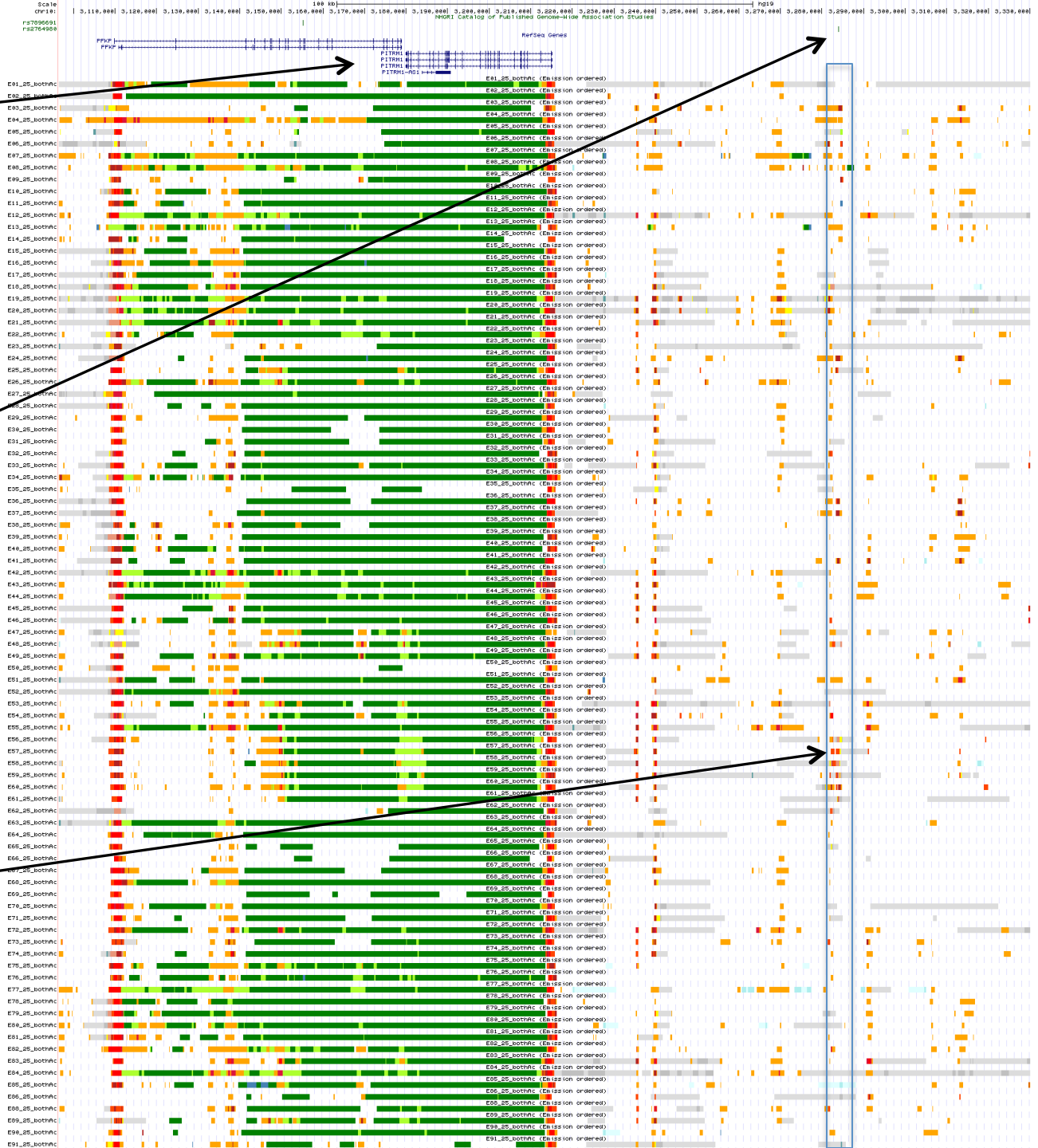
- Disease-associated SNPs enriched for enhancers in relevant cell types
- E.g. lupus SNP in GM enhancer disrupts Ets1 predicted activator


ADHD / brain: cingulate gyrus
Adiponectin and CAD / adipocytes
PMID 20686565: Cholesterol, total (168)
PMID 19862010: Hematocrit (45)
PMID 19060911: Cholesterol, total (45)
PMID 21203500: N-glycan levels (8)
PMID 21297633: Ulcerative colitis (47)
PMID 21102463: Crohn's disease (71)
PMID 21829377: Phospholipid levels (plasma) (50)
PMID 20639880: Primary biliary cirrhosis (6)
PMID 21399635: Primary biliary cirrhosis (20)
PMID 19430480: Type 1 diabetes (38)
PMID 21572416: Blood pressure (17)
PMID 20190752: Celiac disease (40)
PMID 20081856: Fasting glucose-related traits (34)
PMID 21909109: Gamma glutamyl transpeptidase (39)
PMID 21654844: Multiple sclerosis (6)
PMID 21946350: Pulmonary function (33)
PMID 20453842: Rheumatoid arthritis (29)
PMID 22001757: Liver enzyme levels (alanine transaminase) (44)
PMID 19838193: Systemic lupus erythematosus (18)
PMID 20062063: Electrocardiographic traits (18)
PMID 21565293: Thyroid volume (8)
PMID 21909110: Blood pressure (41)
PMID 21051598: HIV-1 control (8)
PMID 19305408: QT interval (13)
PMID 18193044: HDL cholesterol (22)
PMID 19198610: Eosinophil counts (7)
PMID 20383146: Chronic kidney disease (28)
PMID 21841780: Graves' disease (11)
PMID 18758464: Inflammatory bowel disease (6)
PMID 21779181: Systemic sclerosis (18)
PMID 18839057: Attention deficit hyperactivity disorder (26)
PMID 19525953: Multiple sclerosis (15)
PMID 22423221: Mean platelet volume (13)
PMID 21909115: Diastolic blood pressure (50)
PMID 19820697: Hematological parameters (26)
PMID 21378990: Coronary heart disease (30)
PMID 22190364: Multiple sclerosis (19)
PMID 20884846: Urate levels (9)
PMID 21829393: Type 1 diabetes autoantibodies (23)
PMID 22479202: Adiponectin levels (31)
PMID 19767753: Prostate cancer (19)
PMID 21785125: Hepcidin levels (8)
PMID 21347282: Coronary heart disease (44)
PMID 22446963: Rheumatoid arthritis (15)
PMID 22694956: Age-related macular degeneration (6)
PMID 19430479: Diastolic blood pressure (18)
PMID 21383967: Celiac disease and Rheumatoid arthritis (15)
PMID 19525955: Multiple sclerosis (8)
PMID 20852632: Ovarian cancer (7)
PMID 21943158: Cardiovascular disease risk factors (17)
PMID 17660530: Multiple sclerosis (6)
PMID 18391951: Height (52)
PMID 22437554: Response to Vitamin E supplementation (8)
PMID 18264096: Prostate cancer (7)
PMID 20010834: Pulmonary function (7)
PMID 20410501: Vitiligo (10)
PMID 19396169: Biomedical quantitative traits (13)
PMID 22885925: Polycystic ovary syndrome (11)
PMID 22544366: Atrial fibrillation (10)
Breast_vHMEC.Donor_RM035
CD3_Primary_Cells
Fetal_Brain
Fetal_Heart
Fetal_Kidney
Fetal_Lung
H1_BMP4_Derived_Trophoblast_Cultured_Cells
H1_Derived_Mesenchymal_Stem_Cells
H1_Derived_Neuronal_Progenitor_Cultured_Cells
H1_Cell_Line
H9_Cell_Line
IMR90_Cell_Line
iPS_DF_19.11_Cell_Line
iPS_DF_6.9_Cell_Line
Mobilized_CD34_Primary_Cells.Donor_RO_01480
Mobilized_CD34_Primary_Cells.Donor_RO_01536
Mobilized_CD34_Primary_Cells.Donor_RO_01549
Mobilized_CD34_Primary_Cells.Donor_RO_01562
Penis_Foreskin_Fibroblast_Primary_Cells.Donor_skin01
Penis_Foreskin_Fibroblast_Primary_Cells.Donor_skin02
Penis_Foreskin_Keratinocyte_Primary_Cells.Donor_skin02
Penis_Foreskin_Melanocyte_Primary_Cells.Donor_skin01
Adult_Liver
Brain_Germinal_Matrix.Donor_HuFGM02
Brain_Hippocampus_Middle.Donor_150
Breast_Myoepithelial_Cells
CD19_Primary_Cells
CD34_Primary_Cells
CD4_Memory_Primary_Cells
CD4_Naive_Primary_Cells
CD8_Naive_Primary_Cells
Colonic_Mucosa.Donor_32
Duodenum_Mucosa.Donor_61
ES-I3_Cell_Line
ES-WA7_Cell_Line
Fetal_Brain.Donor_HuFNSC01
Fetal_Brain.Donor_HuFNSC02
hESC_Derived_CD184+_Endoderm_Cultured_Cells
iPS-18_Cell_Line
iPS-20b_Cell_Line
Neurosphere_Cultured_Cells_Ganglionic_Eminence_Derived.Donor_HuFNSC02
Penis_Foreskin_Keratinocyte_Primary_Cells.Donor_skin03
Penis_Foreskin_Melanocyte_Primary_Cells.Donor_skin02
Penis_Foreskin_Melanocyte_Primary_Cells.Donor_skin03
Rectal_Mucosa.Donor_29
Rectal_Mucosa.Donor_31
Skeletal_Muscle.Donor_62
Skeletal_Muscle.Donor_63
Stomach_Smooth_Muscle.Donor_28
Mobilized_CD34_Primary_Cells.Donor_RO_01508
Adipose_Derived_Mesenchymal_Stem_Cell_Cultured_Cells
Adipose_Nuclei
Bone_Marrow_Derived_Mesenchymal_Stem_Cell_Cultured_Cells
Brain_Angular_Gyrus
Brain_Anterior_Caudate
Brain_Cingulate_Gyrus
Brain_Inferior_Temporal_Lobe
Brain_Mid_Frontal_Lobe
Brain_Substantia_Nigra
CD15_Primary_Cells
CD34_Cultured_Cells
CD4+_CD25+_CD127-_Treg_Primary_Cells
CD4+_CD25-_CD45RA+_Naive_Primary_Cells
CD4+_CD25-_CD45RO+_Memory_Primary_Cells
CD4+_CD25-_IL17+_PMA-Ionomycin_stimulated_Th17_Primary_Cells
CD4+_CD25-_IL17-_PMA-Ionomycin_stimulated_MACS_purified_Th_Primary_Cells
CD4+_CD25int_CD127+_Tmem_Primary_Cells
CD4+_CD25-_Th_Primary_Cells
CD8_Memory_Primary_Cells
Chondrocytes_from_Bone_Marrow_Derived_Mesenchymal_Stem_Cell_Cultured_Cells
Colon_Smooth_Muscle
Duodenum_Smooth_Muscle
Esophagus
Gastric
hESC_Derived_CD56+_Mesoderm_Cultured_Cells
HUES48_Cell_Line
HUES64_Cell_Line
HUES6_Cell_Line
iPS-15b_Cell_Line
Left_Ventricle
Mesenchymal_Stem_Cell_Derived_Adipocyte_Cultured_Cells
Muscle_Satellite_Cultured_Cells
Pancreas
Pancreatic_Islets
Rectal_Smooth_Muscle
Skeletal_Muscle
Spleen
Stomach_Mucosa

PITRM1 gene

ADHD SNP rs2764980

Intergenic brain-specific promoter



HaploReg view of ADHD SNP rs17658378

Query SNP: rs17658378 and variants with r^2 >= 0.8

chr	pos (hg19)	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	GENCODE genes	dbSNP func annot
8	116394075	1	1	rs17658378	A	G	0.02	0.13	0.00	0.09		BN.SN, BN.ITL, BN.AC	9 cell types				4 altered motifs	27kb 3' of TRPS1	

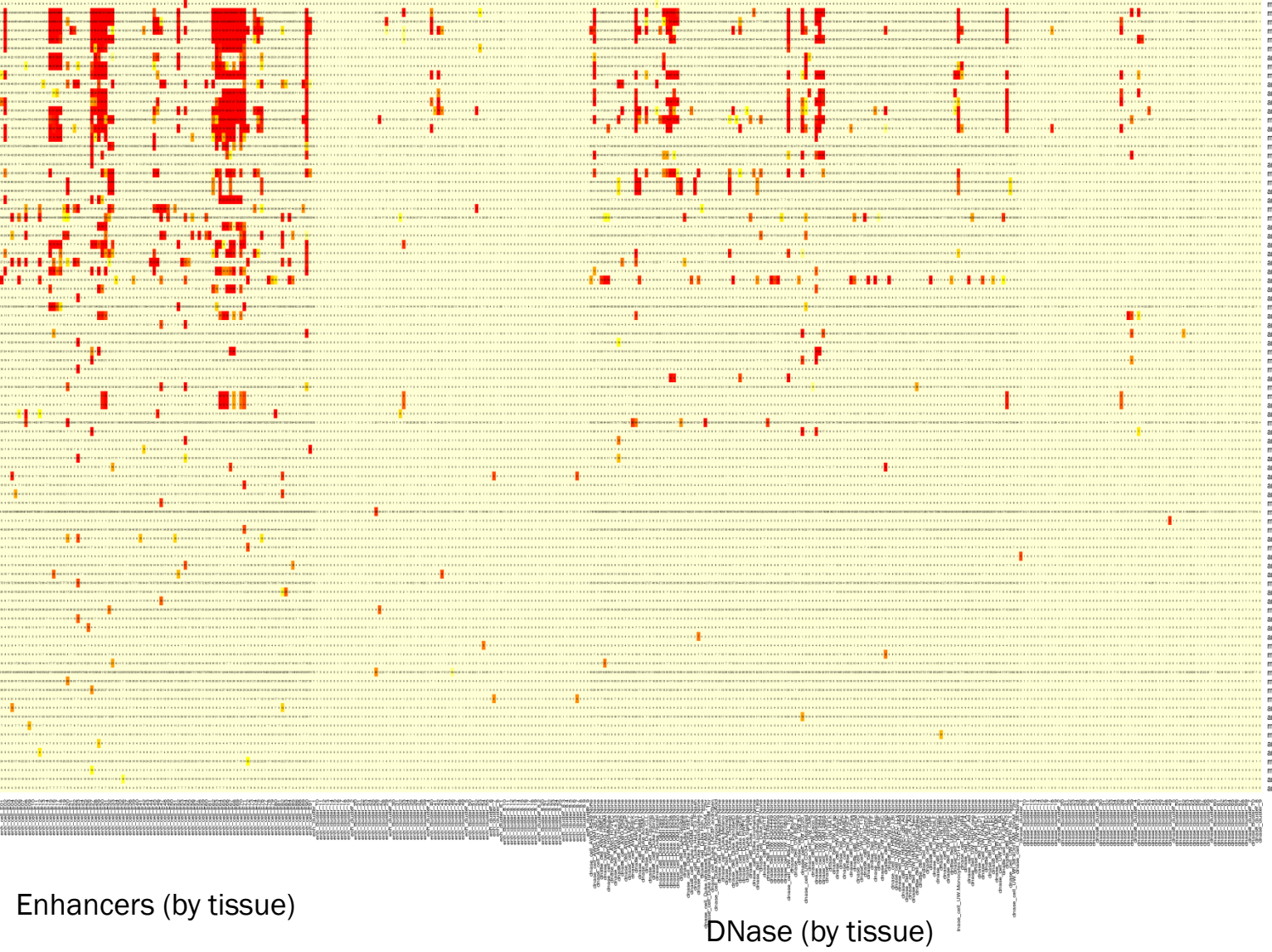
Regulatory chromatin states (Roadmap)

Cell ID	Cell description	State (25-state HMM)
BN.SN	Brain Substantia Nigra	6_TssD2
BN.ITL	Brain Inferior Temporal Lobe	6_TssD2
BN.AC	Brain Anterior Caudate	6_TssD2
BR.H35	Breast vHMEC.Donor RM035	11_EnhWk1
IPS.18	iPS-18 Cell Line	12_EnhWk2
H1.BMP4DM	H1 BMP4 Derived Mesendoderm Cultured Cells	12_EnhWk2
BN.CC	Brain Cingulate Gyrus	12_EnhWk2
PFK.2	Penis Foreskin Keratinocyte Primary Cells.Donor skin02	12_EnhWk2
HUES6	HUES6 Cell Line	12_EnhWk2
BN.HM150	Brain Hippocampus Middle.Donor 150	12_EnhWk2
ESO	Esophagus	12_EnhWk2
BN.AG	Brain Angular Gyrus	12_EnhWk2

Regulatory motifs altered

PWM	Strand	Ref	Alt	Match on:
				Ref: CTTGTTCTCTTTCCCAGGCCATAGCGGCTATCAGGAACTTGTAGCCATCTGGGGGTCAG Alt: CTTGTTCTCTTTCCCAGGCCATAGCGGCTGTCAGGAACTTGTAGCCATCTGGGGGTCAG
PU.1_disc1	-	-33.3	-21.6	AWGRGGAAGT
PU.1_known3	-	11.7	9.8	NHASTTCCBYHWHN
Pbx3_disc3	-	6.6	11.2	TTGGYYVNNNBNSCYGYCMVT
SETDB1_disc1	+	2.3	9.9	CRNDGMHYBMYGGRARWKGTAGTYY
Znf143_disc3	+	9.9	11.2	GSVBBSBGGGVVNBGBRGB

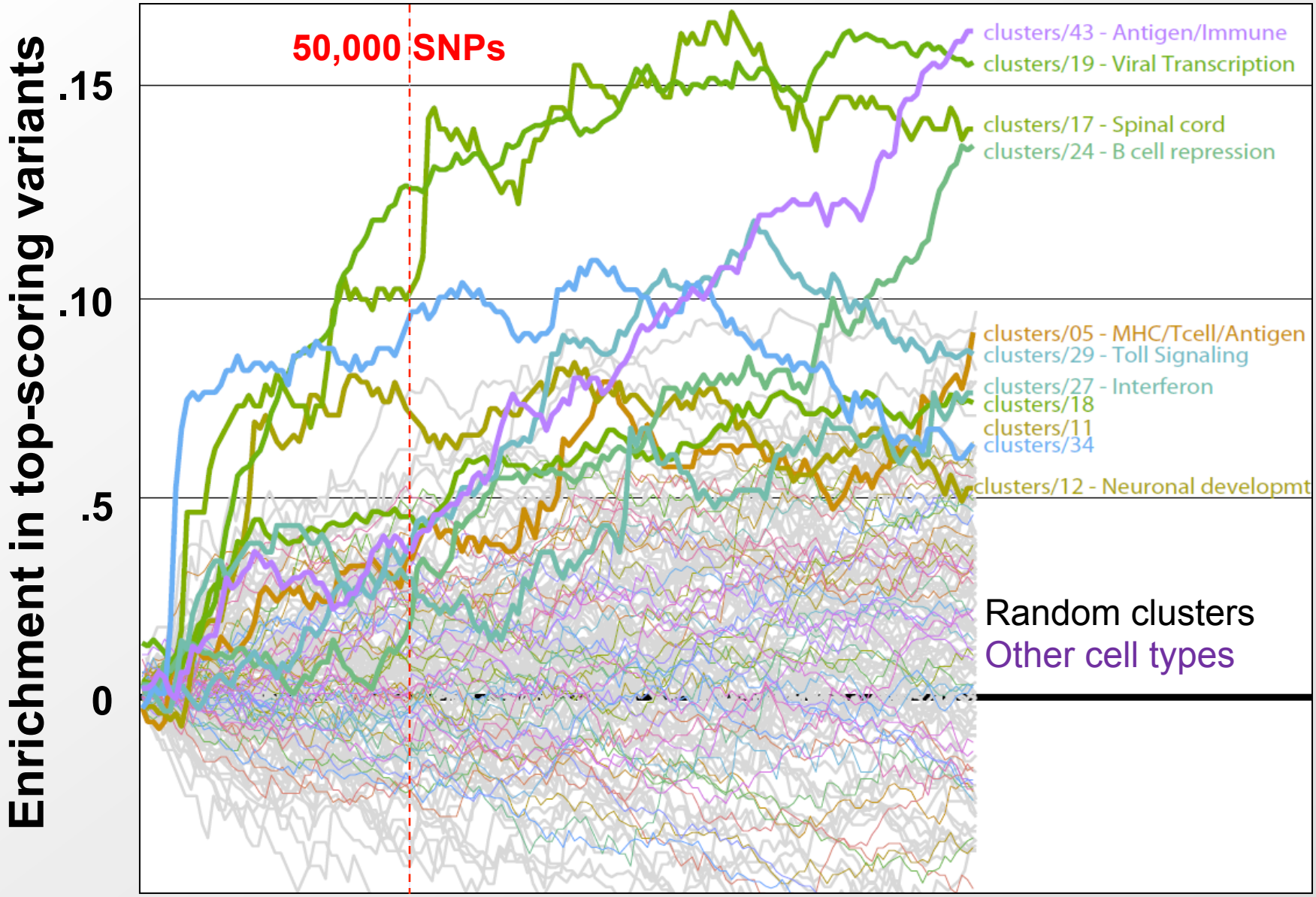
rs17658378 / TRPS1
From Lasky-Su et al (PMID 2008), who do not consider it one of the interesting loci
TRPS1 is a TF; locus has been associated by GWAS with weight fluctuation (PMID 22911880) and major depressive disorder (PMID 22472876)



Examples where the clustered enhancers give a previously undetectable enrichment

(Cardiac traits <-> enhancer cluster with fluid shear GO)

T1D enrichment relative to randomized cell types



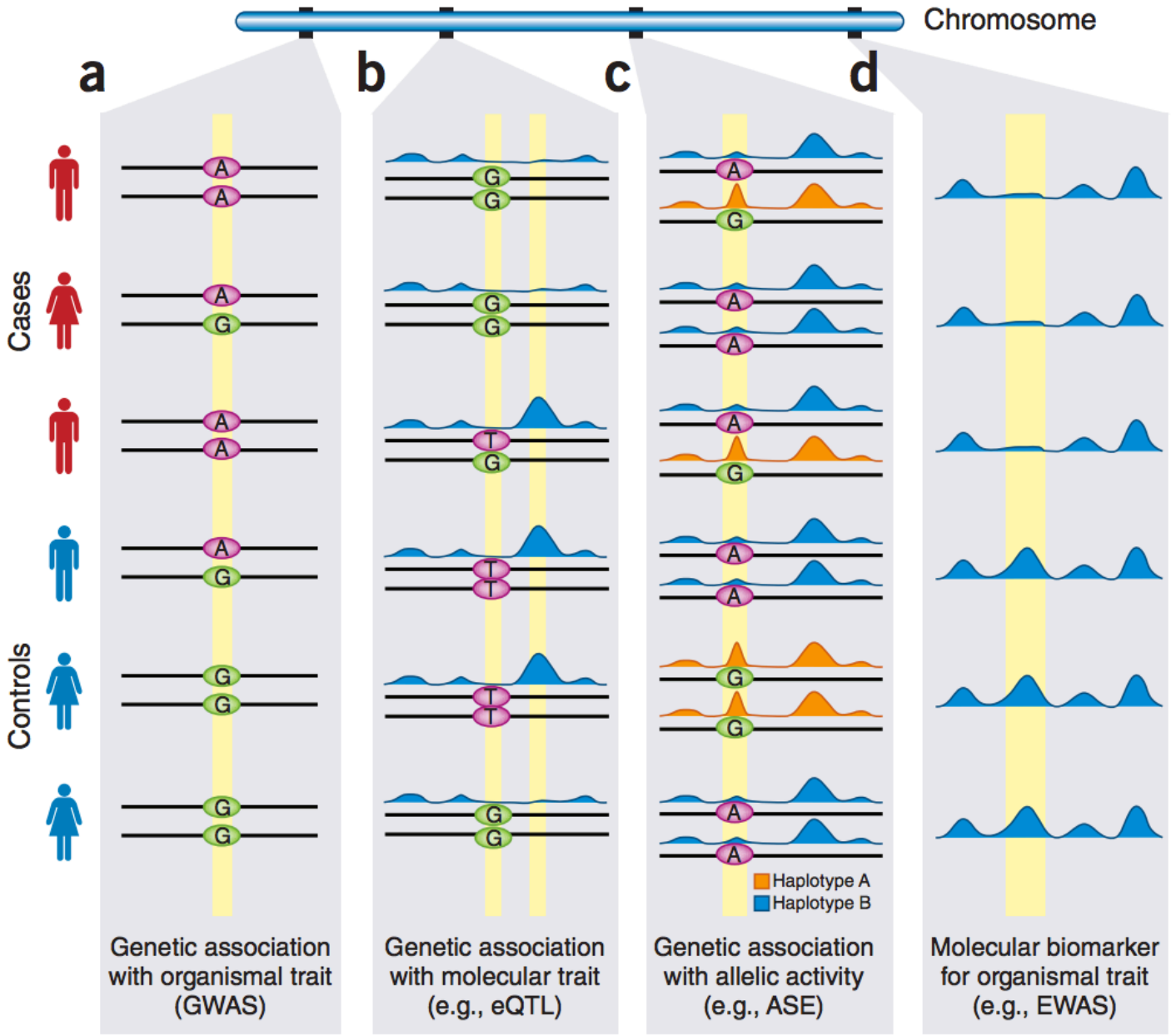
T1D/RA-enriched enhancers spread across genome

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

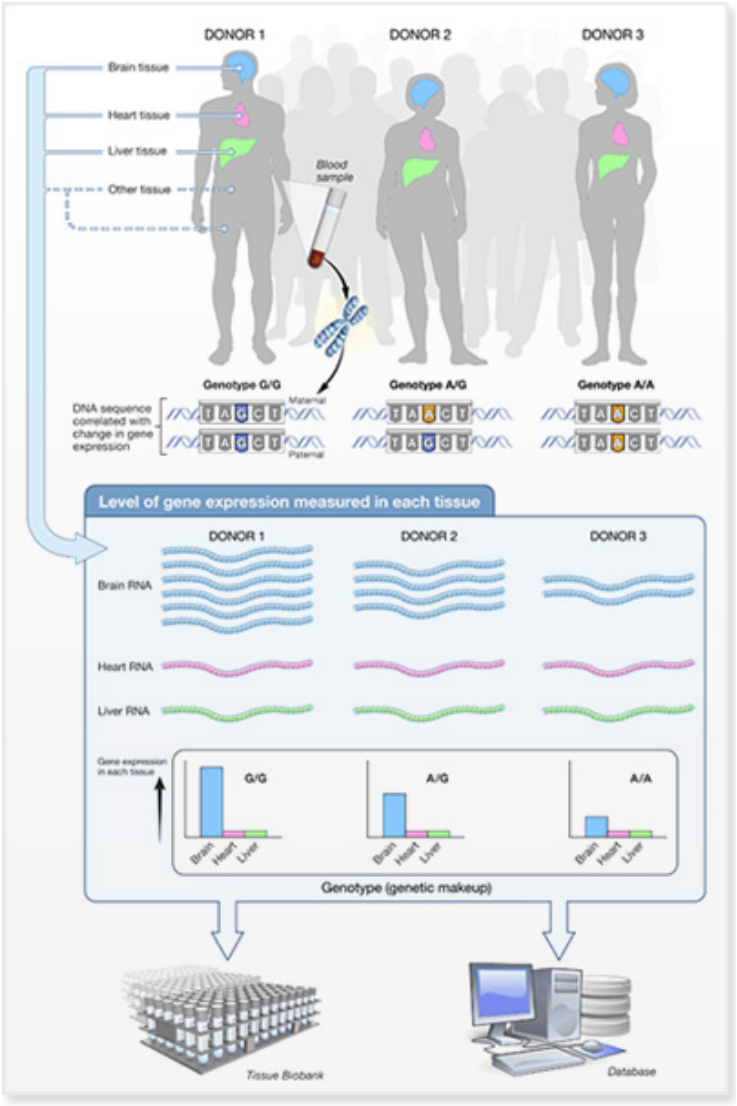
Regulatory genomics to interpret complex disease genetics

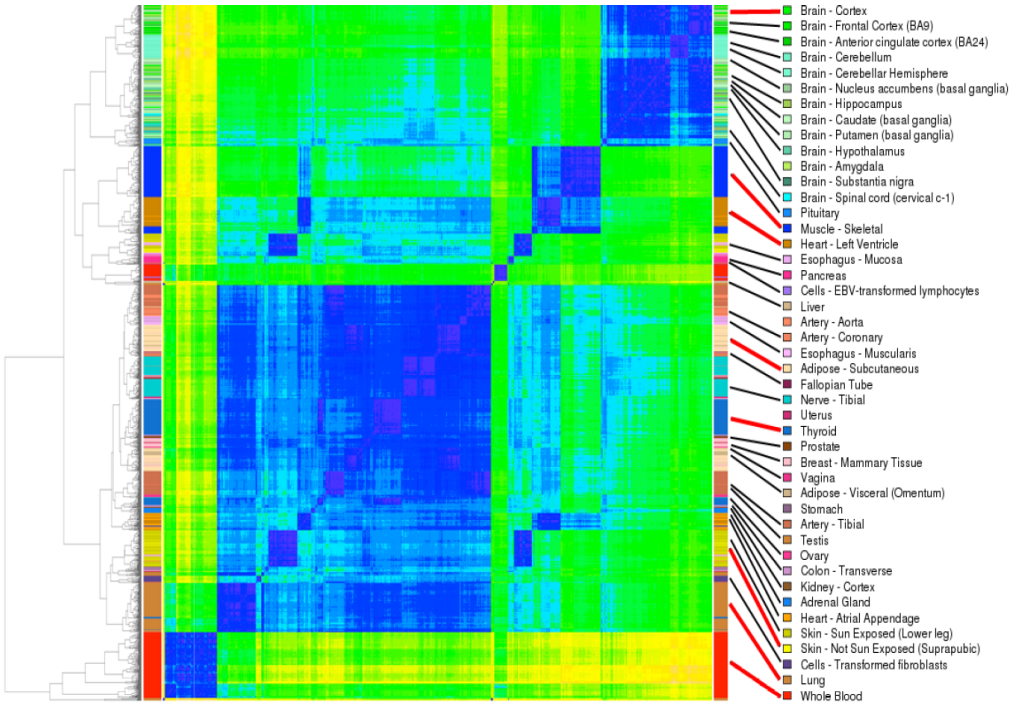
1. Regulatory annotations of the human genome: an overview
2. Using regulatory annotations to interpret GWAS
   a. Locus level
   b. Systems level
3. Beyond GWAS
   a. Molecular variability
   b. Empowering rare-variant and pathway analysis

Beyond GWAS: Molecular variability



eQTLs: The GTEx Project



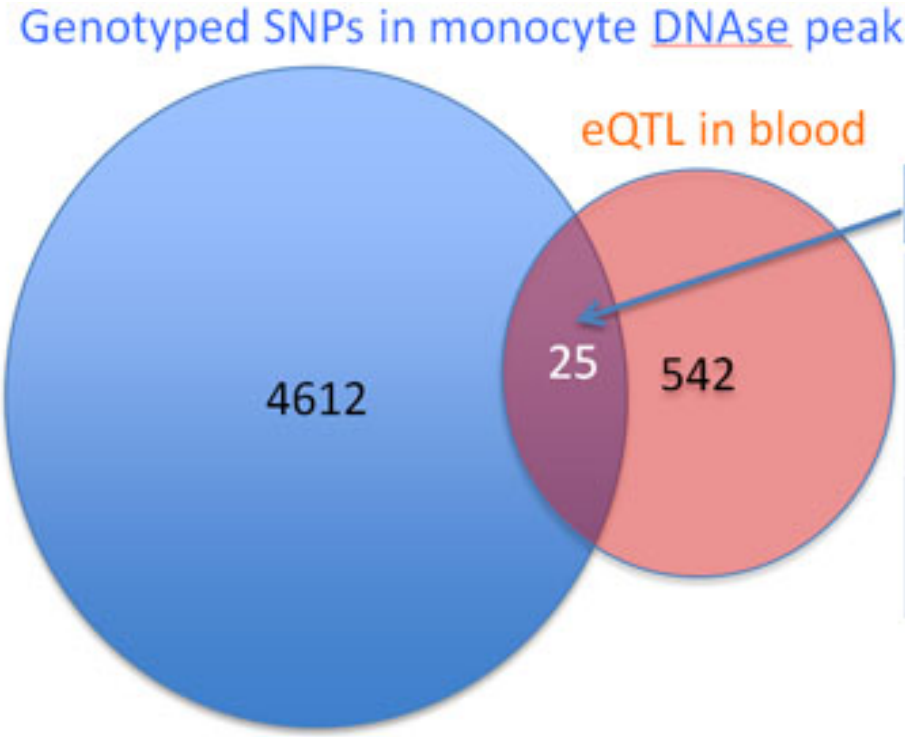


17 GTEx tissues w/ close Roadmap match

	GTEX tissues	Current Roadmap tissues
1	Adipose	Adipose nuclei (E54)
2	Blood	Peripheral blood mononuclear primary cells (E87)
3	Brain - Midbrain - Substantia nigra	Brain - midbrain - substantia nigra (E61)
4	Brain - Cortex - Frontal Cortex (BA9)	Brain - cerebral cortex - mid frontal lobe (E60)
5	Brain - Cerebellar Hemisphere Brain - Cerebellum Brain - Cortex - Anterior cingulate cortex (BA24) Brain - Cortex	Brain - cerebral cortex - inferior temporal lobe (E59) Brain - cerebral cortex - angular gyrus (E56) Brain - cerebral cortex - cingulate gyrus (E58)
6	Brain - Cerebrum - Subcortical - Hippocampus Brain - Cerebrum - Subcortical - Caudate (basal ganglia) Brain - Cerebrum - Subcortical - Basal ganglia - Putamen Brain - Cerebrum - Subcortical - Basal ganglia - Nucleus accumbens	Brain - cerebrum - basal ganglia - anterior caudate (E57) Brain - cerebrum - hippocampus (E25)
7	Heart	Heart - left ventricle (E82) Heart - fetal heart (E04)
8	Lung	Lung - fetal lung (E06)
9	Muscle - Skeletal	Muscle - skeletal muscle (E47, E48, E89)
10	Pancreas	Pancreas (E85) Pancreatic islets (E86)
11	Skin	Penis foreskin fibroblast (E19, E20) Penis foreskin keratinocyte (E21, E42) Penis foreskin melanocyte (E22, E33, E44)

- Expected to increase with additional coverage
- Expression correlation metric for unbiased matching

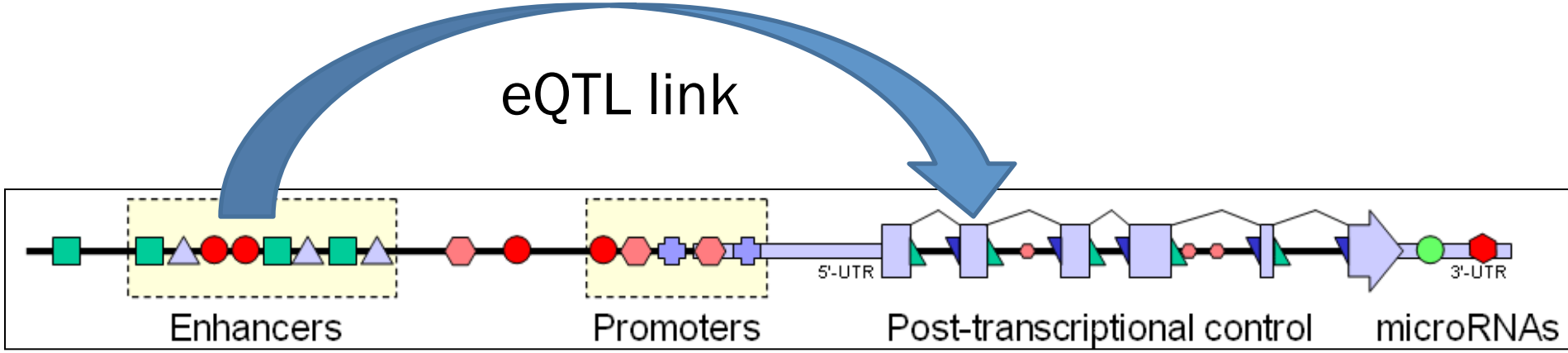
Example of functional overlap



Background model	Expected	P-value
All genes	5	4.7×10^{-12}
Genes expressed in tissue	7.3	4.4×10^{-8}
Genes expressed & eQTL TSS distance	9.4	5.2×10^{-6}

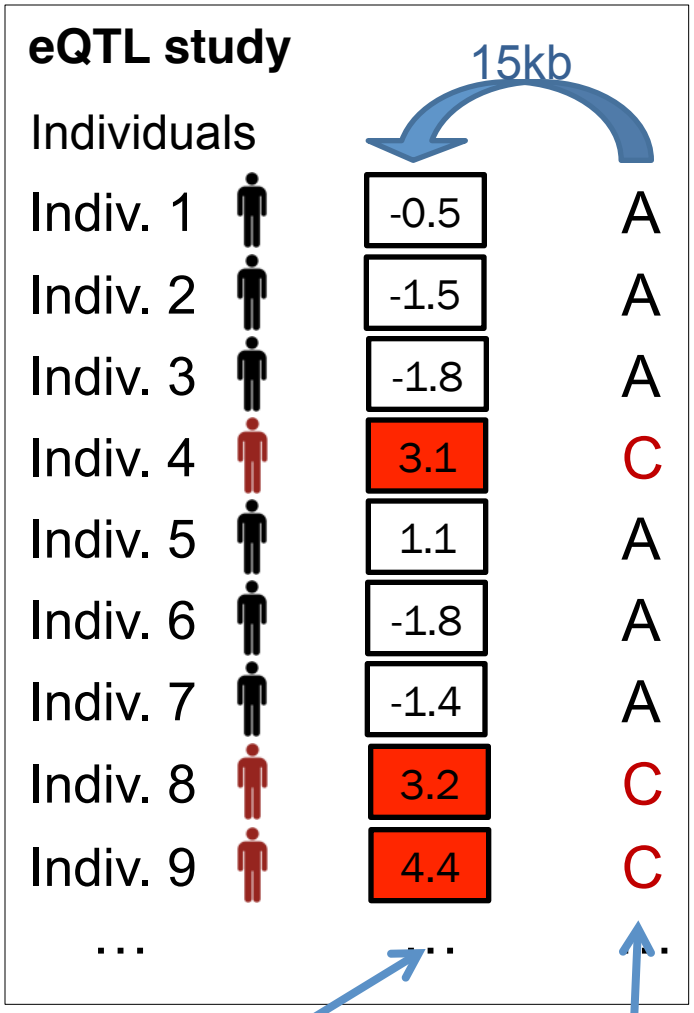
- 25 eQTLs in blood from GTEx
  lie within monocyte DNase peaks from ENCODE
- Significantly more than expected by chance
- Even after correcting for TSS distance and tissue
  ➔ eQTLs identify likely regulatory elements

Functional characterization of eQTLs



1. **Functional roles:** specific region, motif, link, ASE
2. Exploit GTEx matrix for **systems biology** studies
3. **Disease roles:** modules, tissues, genome-wide

Enhancer-gene links supported by eQTL-gene links



Validation rationale:

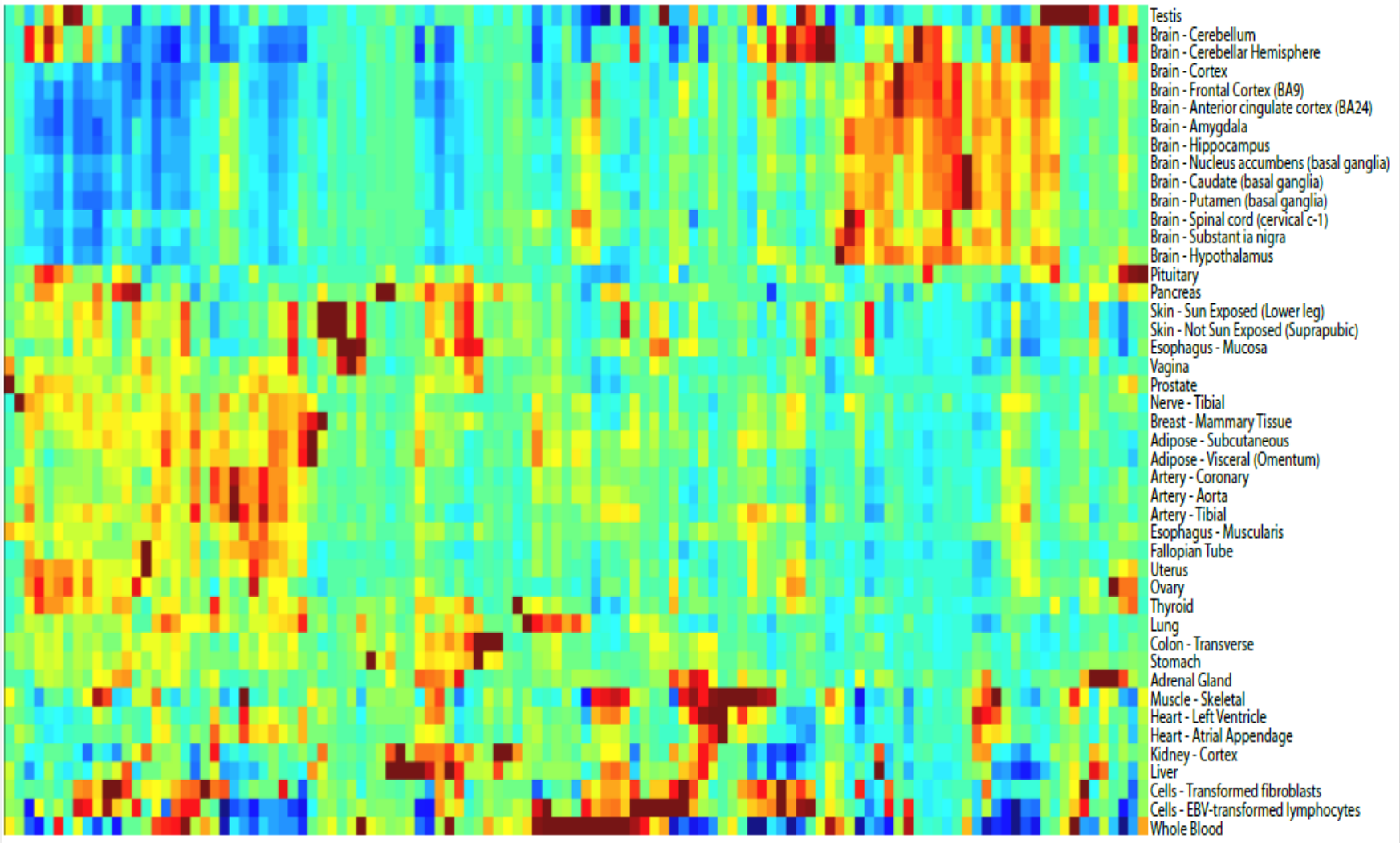
- Expression Quantitative Trait Loci (eQTLs) provide **independent SNP-to-gene** links
- Do they agree with activity-based links?

Example: Lymphoblastoid (GM) cells study

- **Expression/genotype** across 60 individuals (Montgomery *et al*, Nature 2010)
- **120** eQTLs are eligible for enhancer-gene linking based on our datasets
- **51** actually linked (43%) using predictions
  ➔ **4-fold enrichment** (10% exp. by chance)

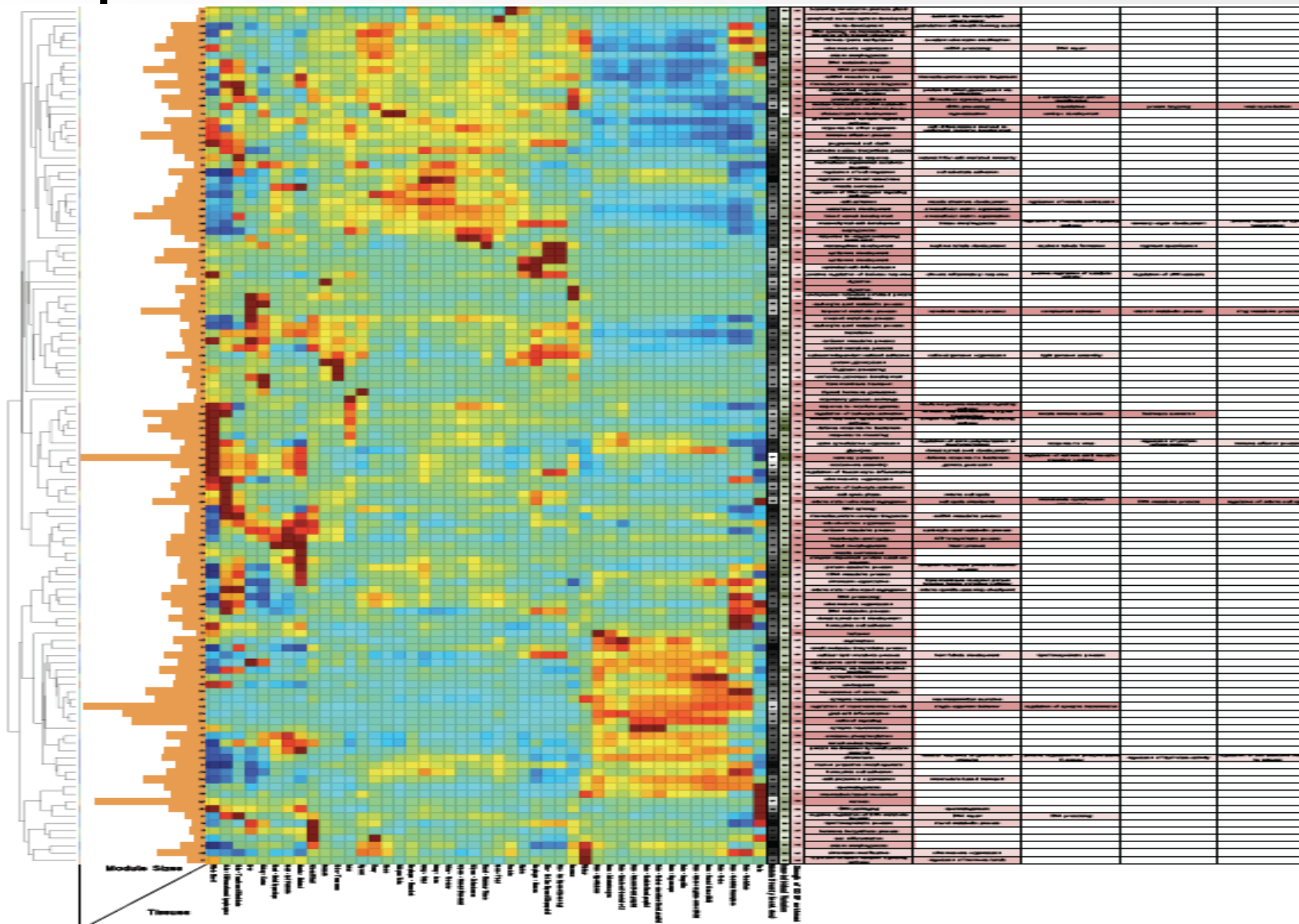
- Independent validation of links.
- Relevance to disease datasets.

GTEx eQTLs enable us to define multi-tissue expression modules in individuals

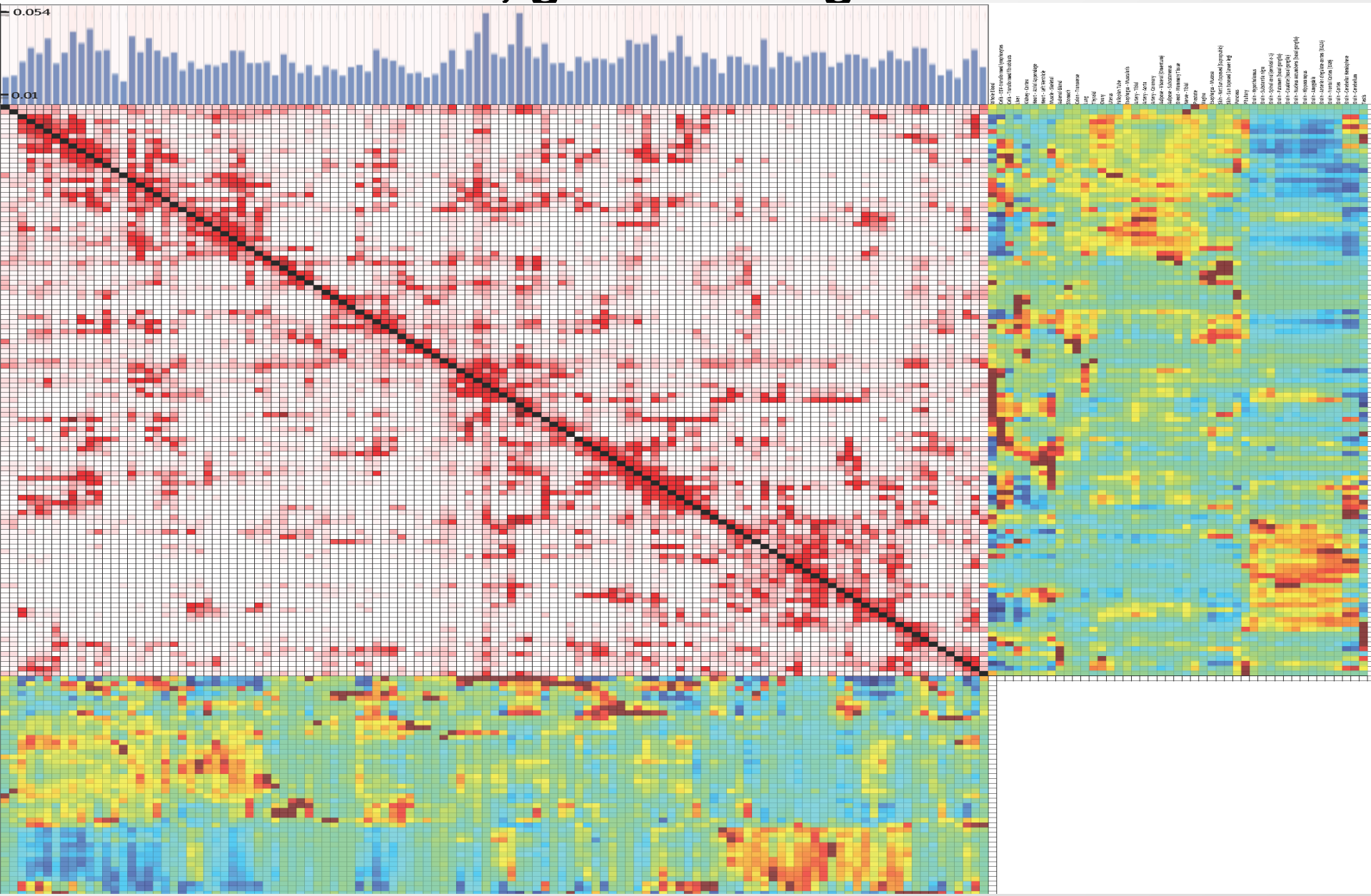


- Regulators and motifs affect gene expression patterns
- Exploit multi-tissue multi-individual nature of dataset

Expression modules enriched in relevant GO terms

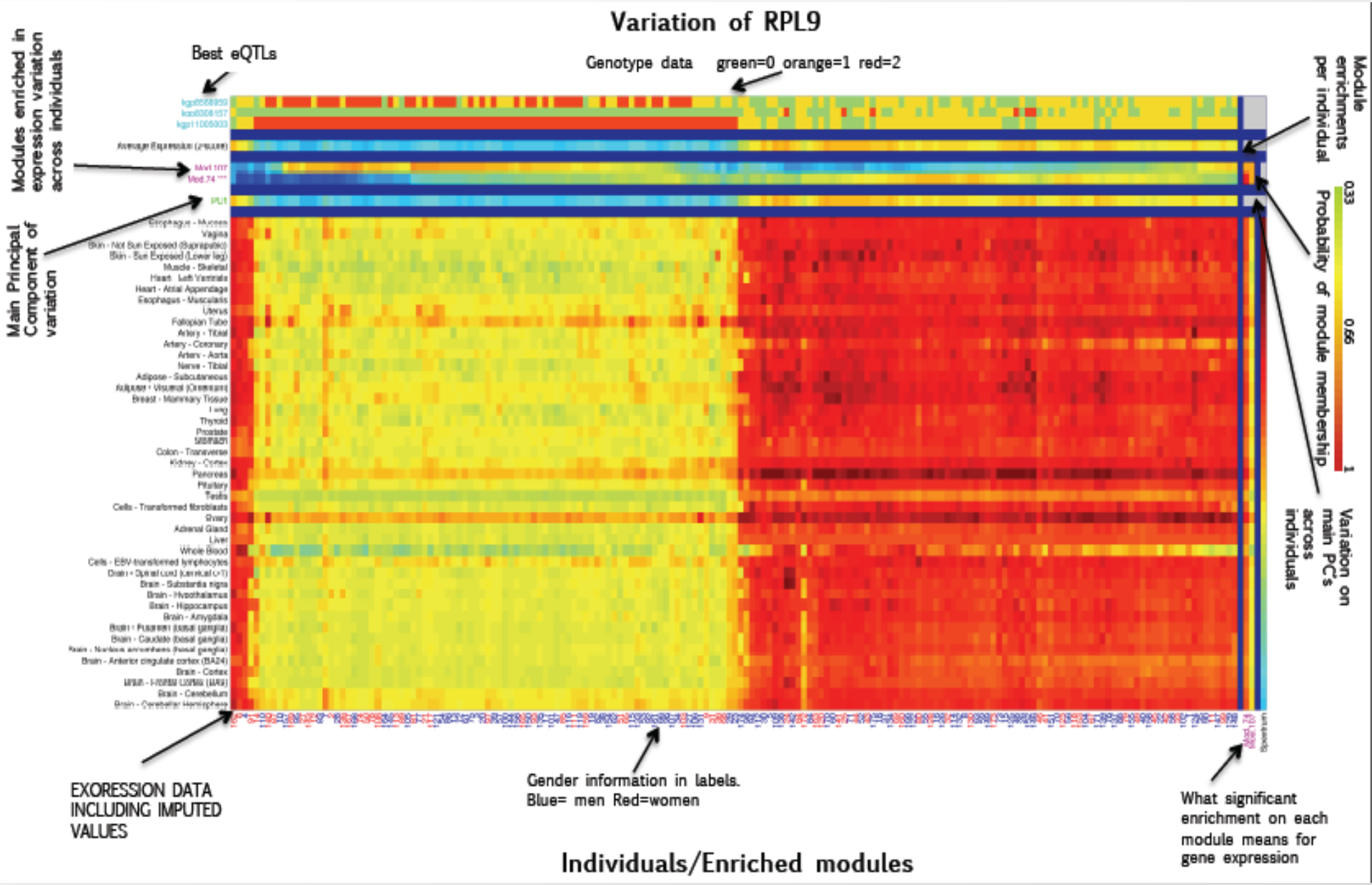


Across individuals, genes change between modules



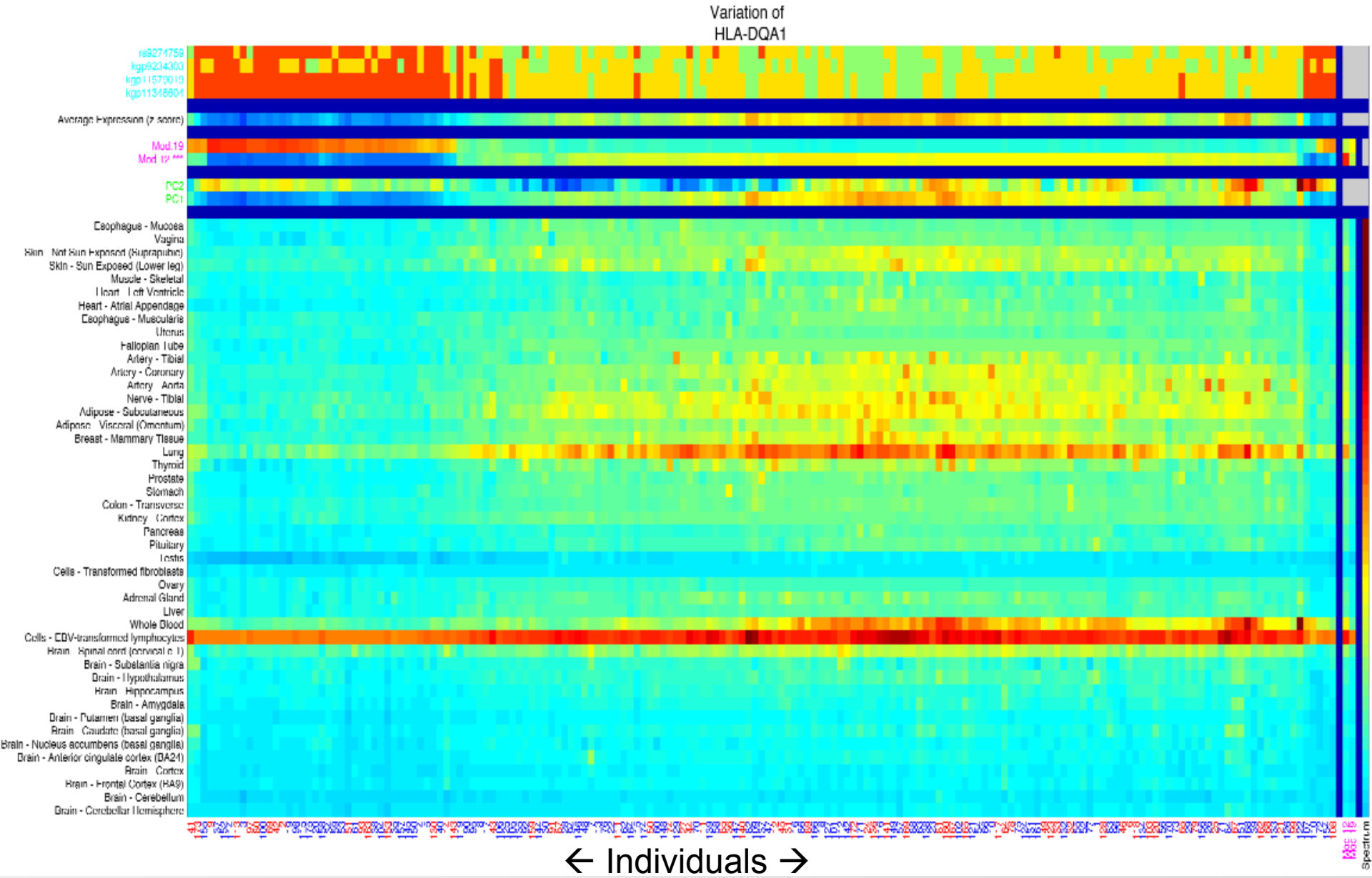
➔ Use module membership prob. as quantitative traits

Identify SNPs underlying module changes: netQTLs



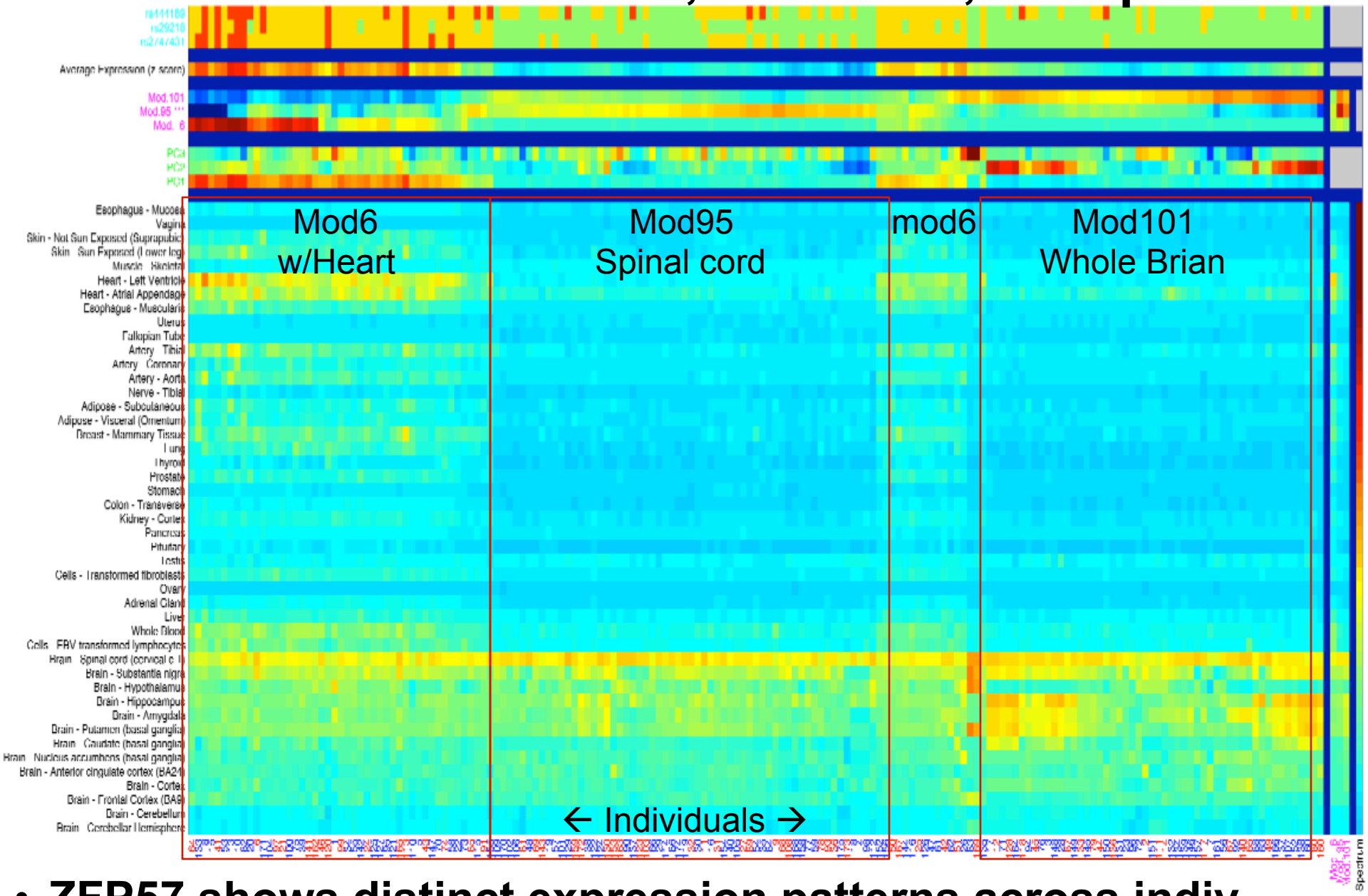
- **Single-tissue expression patterns only a partial picture**

Ex2: Mod19(lymph only)⇔Mod12(mesdoerm-wide)



- **Multiple SNPs associated with different PCs of variation**

Ex3: ZFP57 three modules, three PCs, multiple SNPs

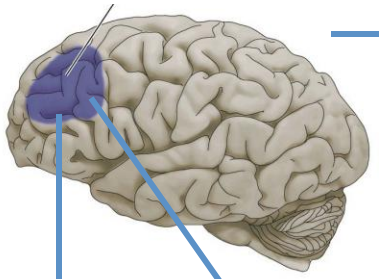


- **ZFP57 shows distinct expression patterns across indiv.**
- **ZNF-KRAB regulator, transient neonatal diabetes mellitus**

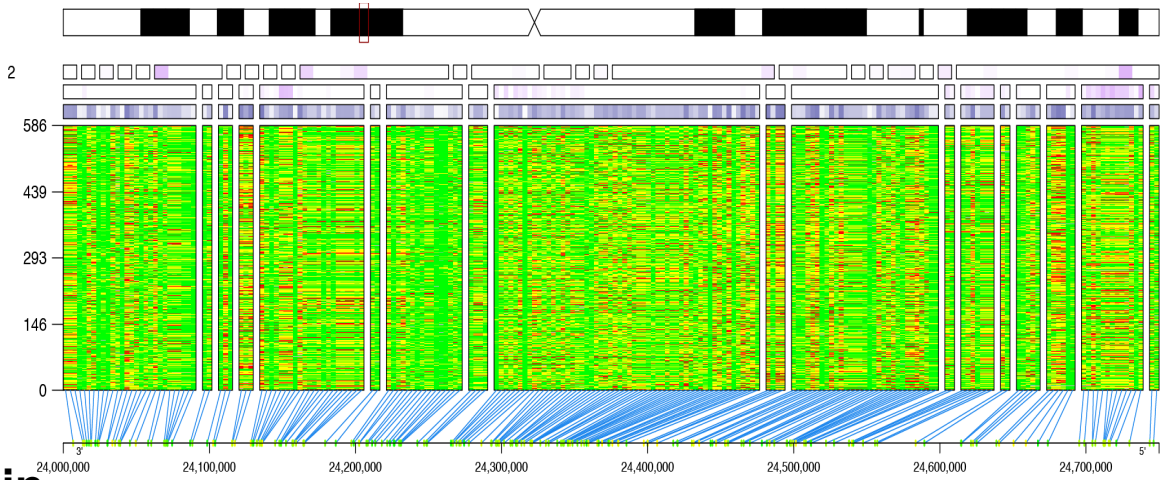
EWAS: Global association of brain enhancers with AD

MAP Memory and Aging Project
+ ROS Religious Order Study

Dorsolateral PFC

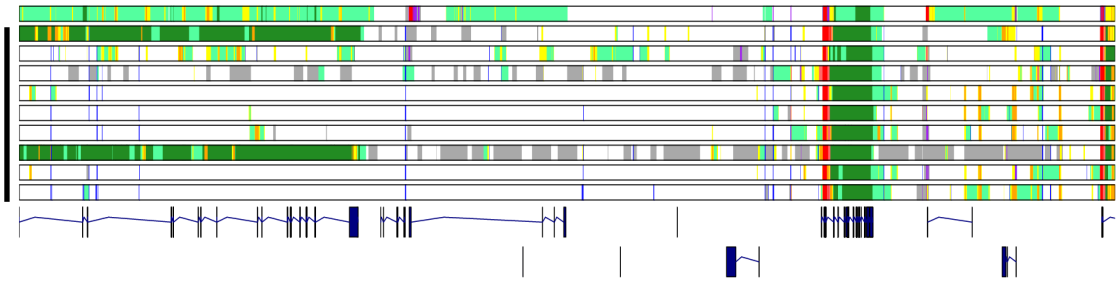


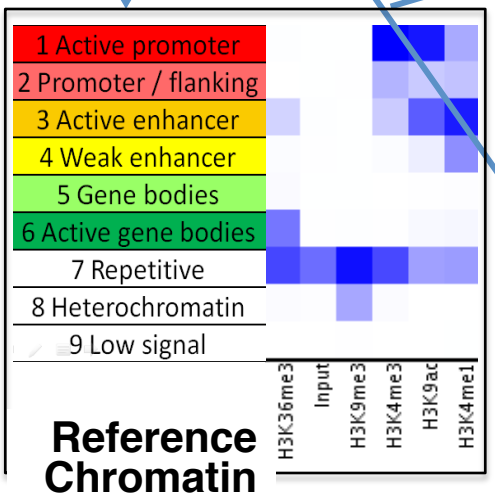
Genotype (1M SNPs x700 ind.) (De Jager)



Brain

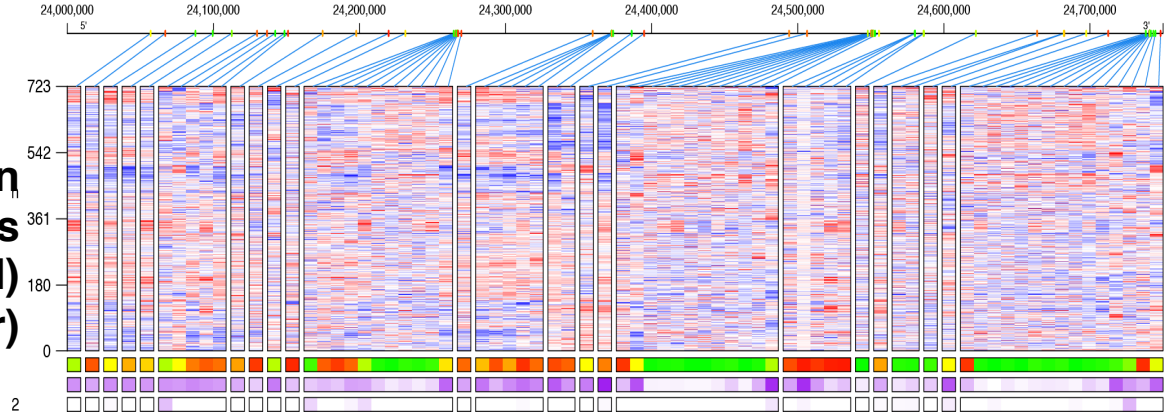
ENCODE





Reference Chromatin states (Bernstein)

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



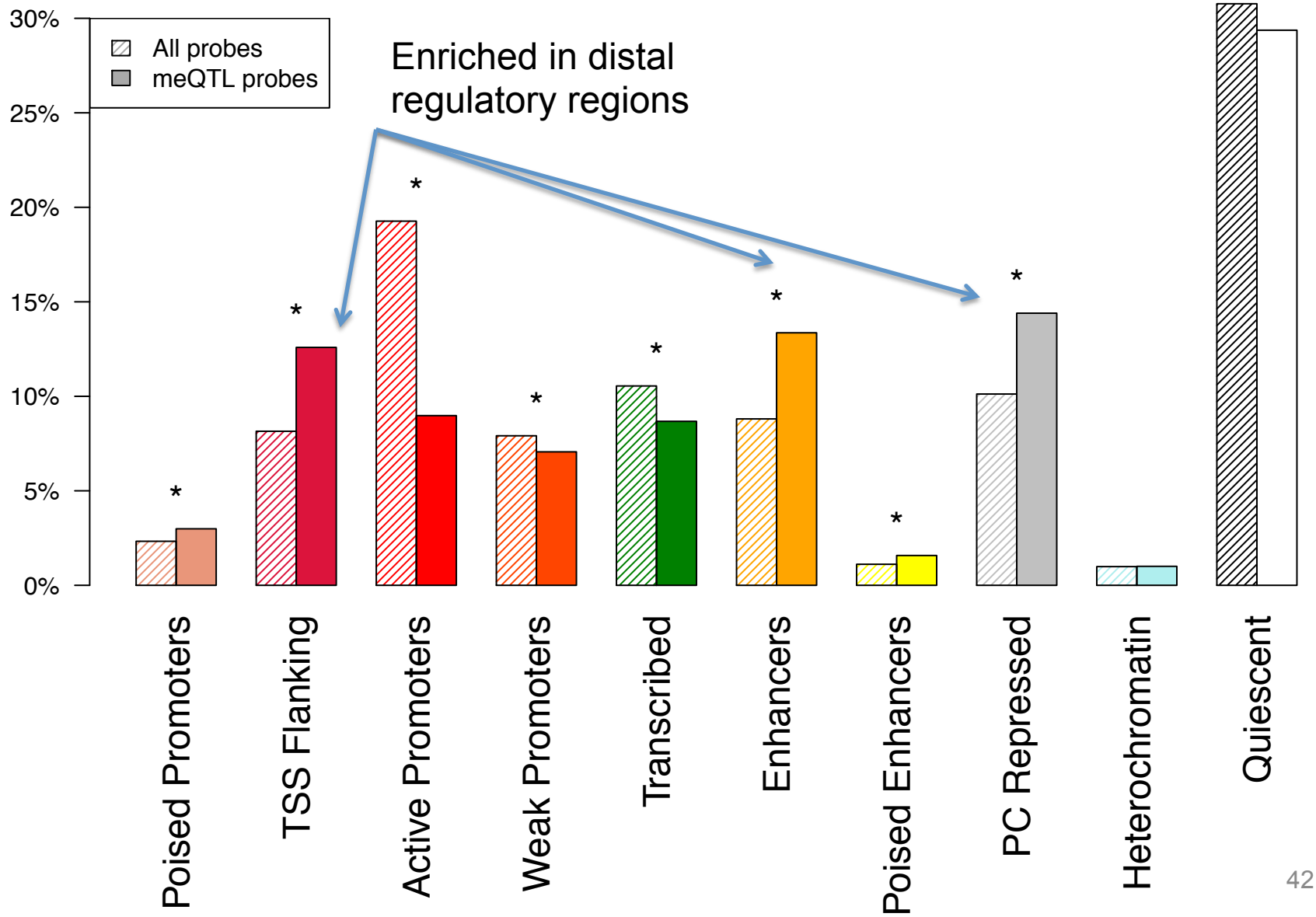
750 subjects, initially cognitively normal, Alzheimer's diagnosed by pathology. (Bennett)

Majority of AD-associated GWAS SNPs are meQTLs

Rank	ad.rsid	Gene	Description	meQTL P-value	meQTL SNP	meQTL Gene	SNP state
1	rs11767557	EPHA1	Ephrene A receptor 1	1.12E-13	rs12703526	cg18997129	24_Quies3
2	rs1532278	CLU	Clusterin	6.68E-125	rs17057441	cg18814083	22_Quies1
3	rs3865444	CD33	Myeloid transmembrane receptor	1.13E-10	rs12971624	cg11581627	22_Quies1
4	rs561655	PICALM	Phosphatidylinositol binding clathrin assembly protein	7.11E-77	rs17817919	cg24166175	22_Quies1
5	rs610932	MS4A2	Immunoglobulin receptor subunit	1.37E-33	rs562028	cg16954525	22_Quies1
6	rs6701713	CR1	Complement Receptor 1	1.39E-21	rs3849266	cg19373649	22_Quies1
7	rs7561528	BIN1	Bridging Integrator Nucleocytoplasmic adaptor protein	3.73E-176	rs4663104	cg02887598	2_TssF
8	rs9349407	CD2AP	Actin Cytoskeleton Regulating Scaffold	7.12E-63	rs2275446	cg16361253	22_Quies1

- Importance of mapping intermediate phenotypes
- Genetic ⇔ Molecular ⇔ Cellular ⇔ Neural ⇔ Disease

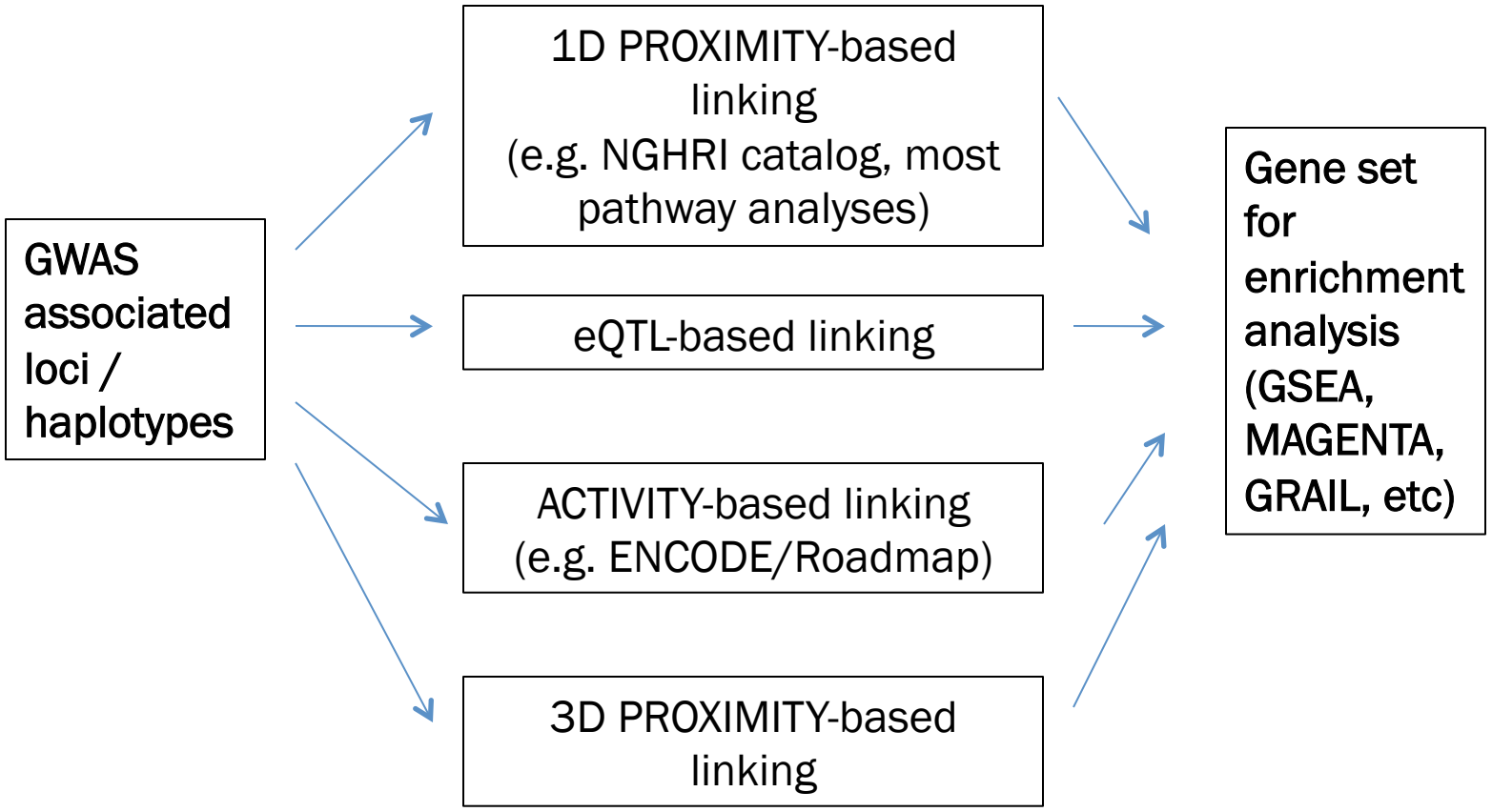
SNP-associated CpGs depleted in promoters, enriched in enhancers



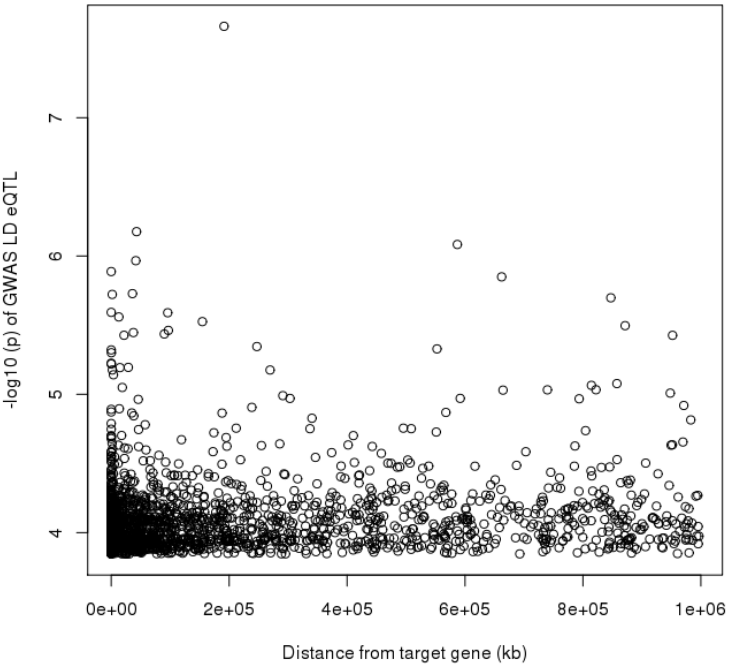
Regulatory genomics to interpret complex disease genetics

1. Regulatory annotations of the human genome: an overview
2. Using regulatory annotations to interpret GWAS
   a. Locus level
   b. Systems level
3. Beyond GWAS
   a. Molecular variability
   b. Empowering rare-variant and pathway analysis

Pathway analysis of GWAS relies on accurately linking cis-regulatory regions to their targets

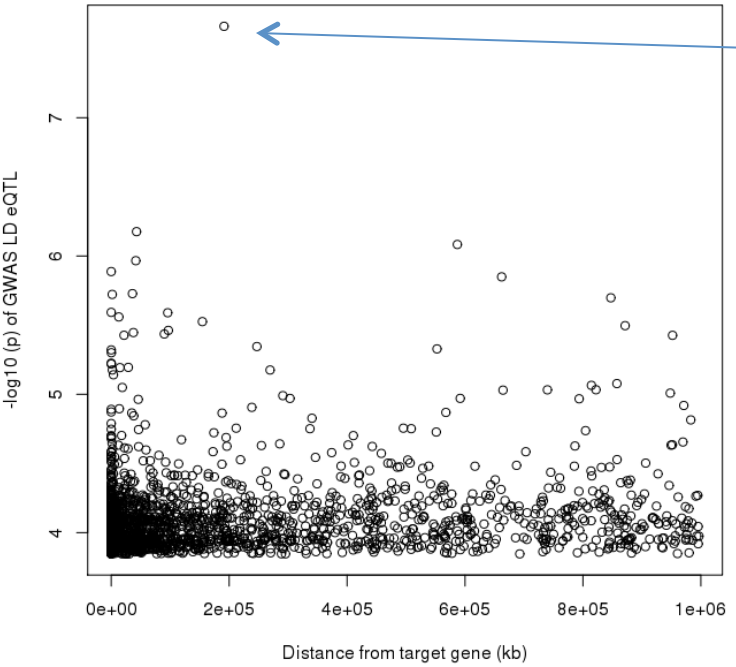


eQTLs reveal that closest isn't always best



- 1877 GWAS SNPs are within r2>=0.8 of the best GTEx eQTL for a gene/tissue combination
- Only 345/1810 (19%) of phenotype-associated GTEx eQTLs show agreement between their strongest eQTL-linked gene(s) and their physically closest or overlapping gene(s)

eQTLs for improved target gene prediction of regulatory GWAS SNPs

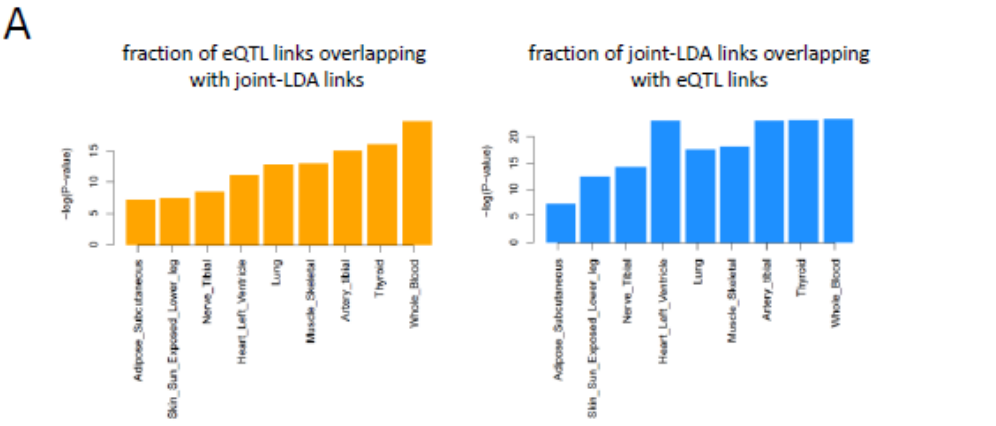


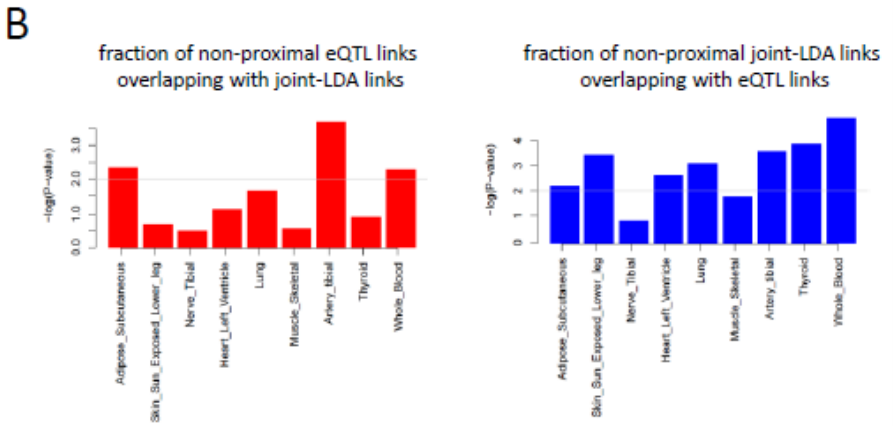
- rs919129 (myocardial infarction) is 25 kb from proximity-based target FBN. Strongest eQTL target is a novel lincRNA 191 kb away, RP11-506F22.2
- Alters a TBX5 motif instance (role in heart development)

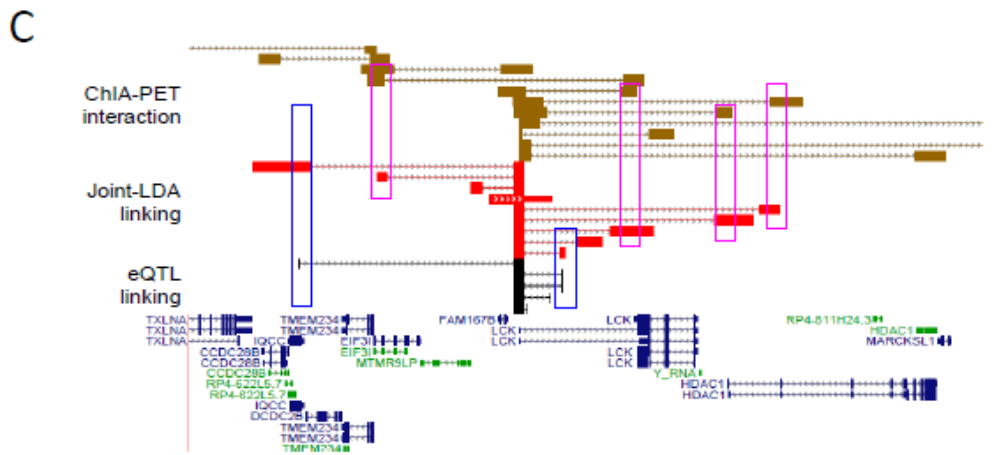
Regulatory motifs altered

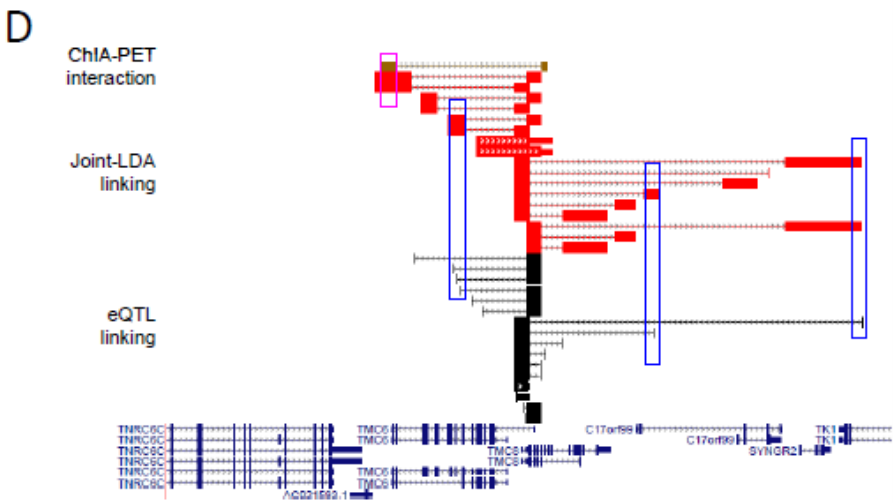
PWM	Strand	Ref	Alt	Match on: Ref: AAGTGCAAAAAGGAGGTATGGGAGAATGTGTGAGT Alt: AAGTGCAAAAAGGAGGTATGGGAGAATGTATGAGT
TBX5_3	+	6.7	-5.3	MAGGTGTGAR

Activity, eQTL, and conformation methods agree on links

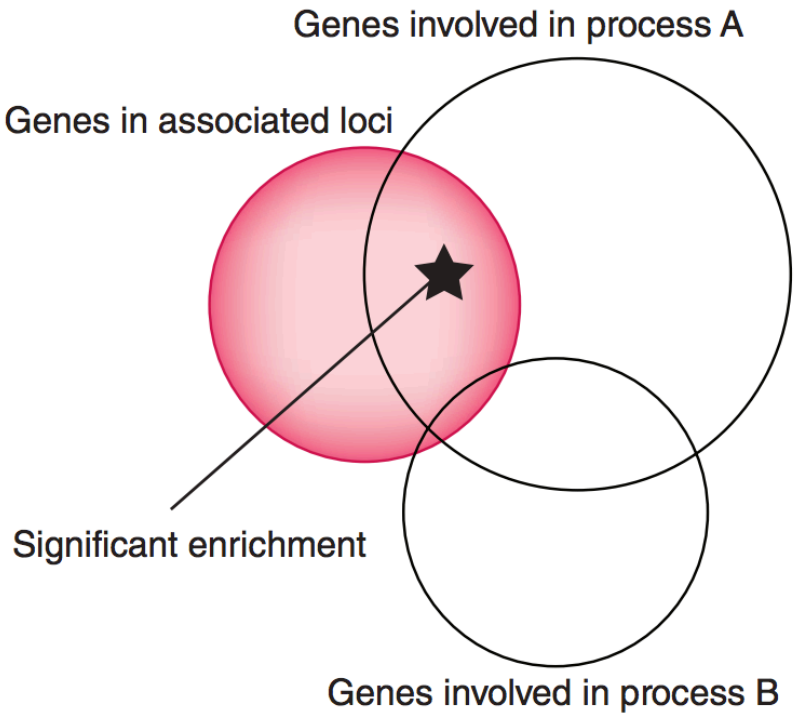


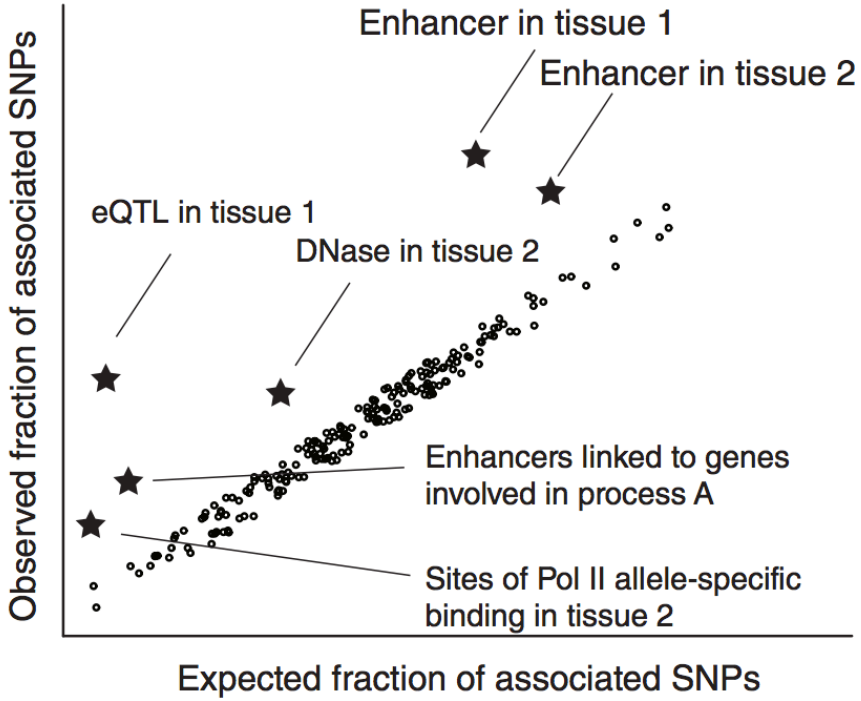






Linking as another way to improve GWAS analysis

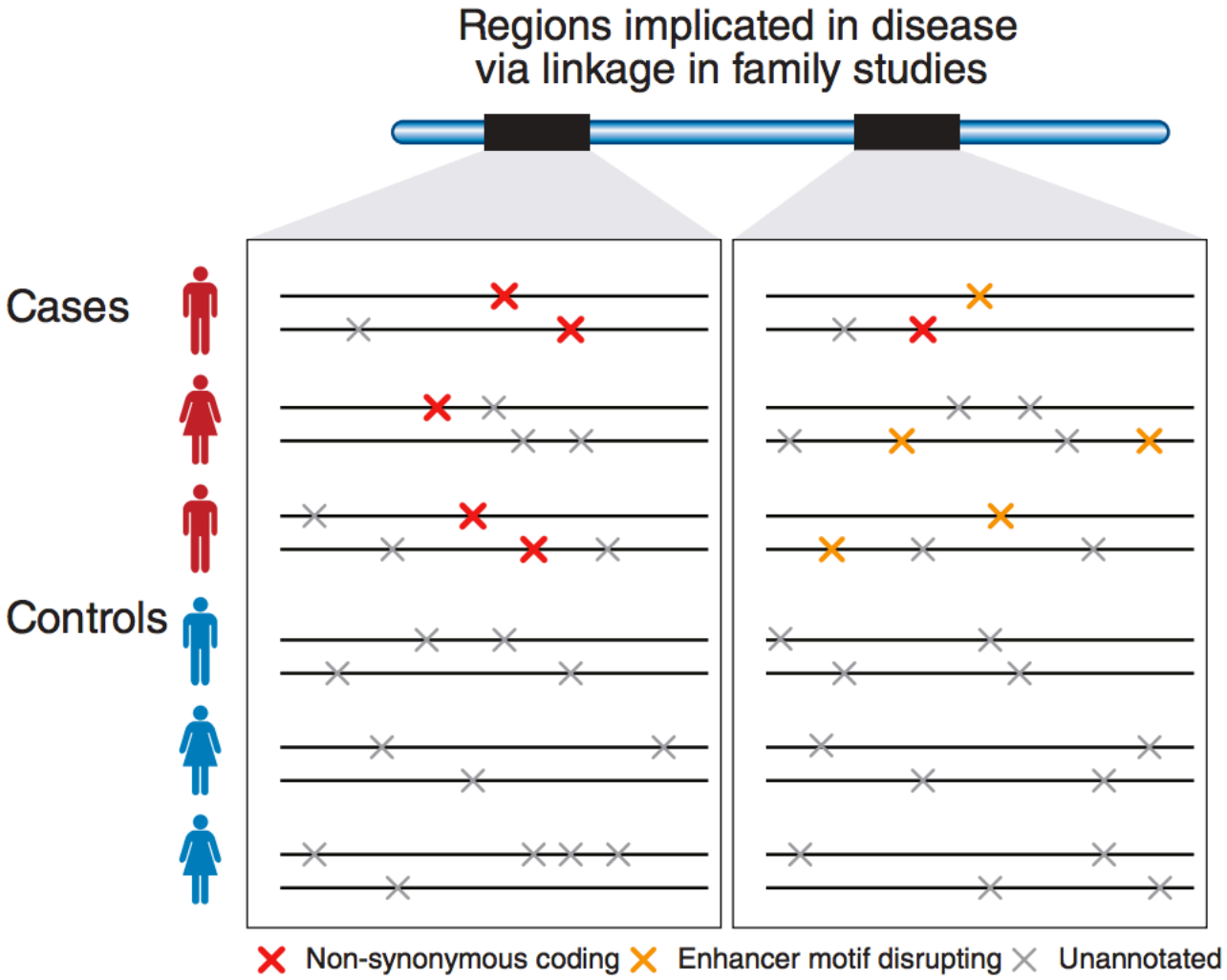




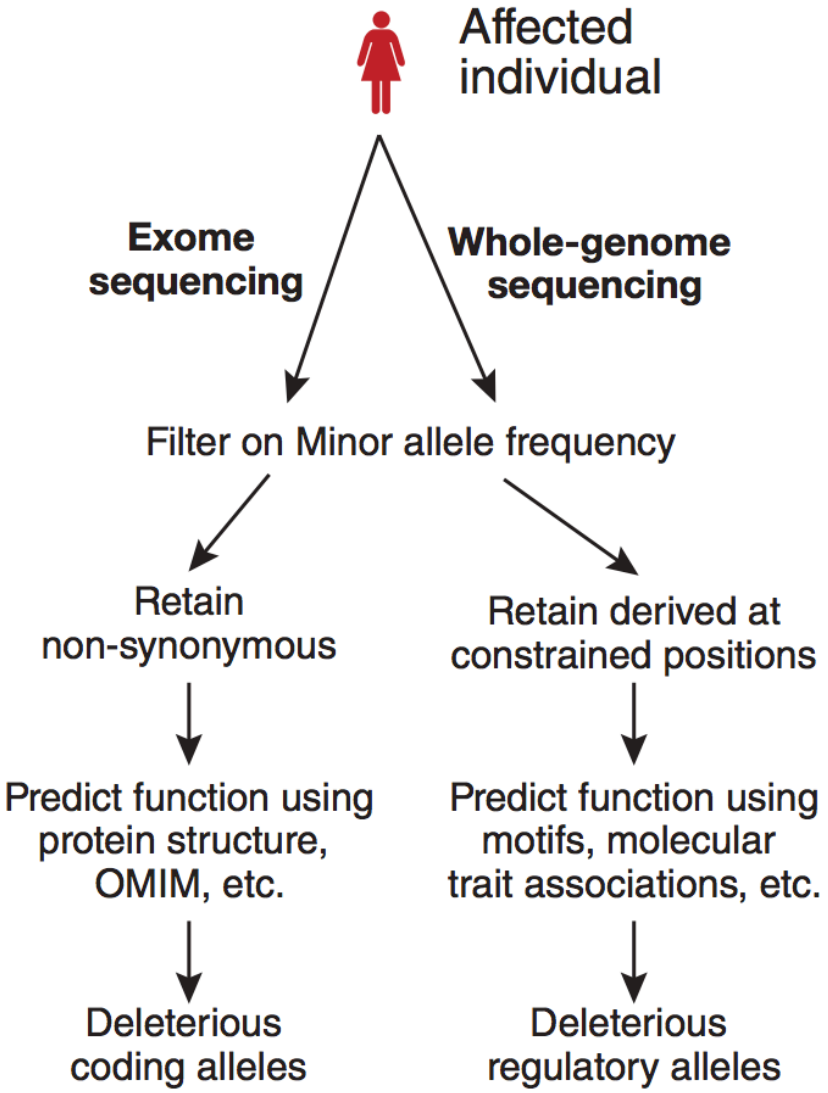
Current methods

Regulatory genomics informed methods

Annotating rare variants: for burden tests on WGS data



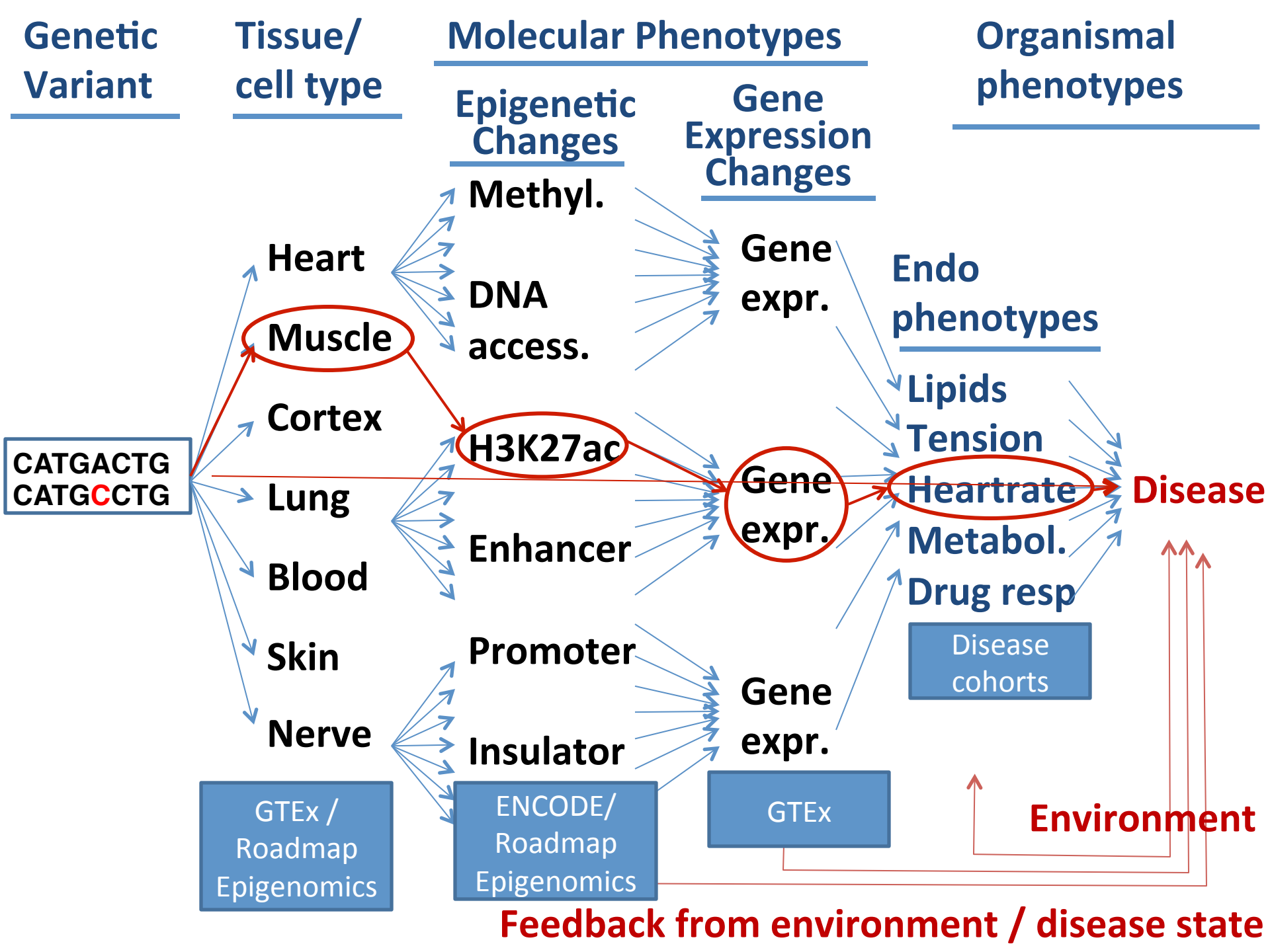
Annotating rare variants: for clinical interpretation pipelines

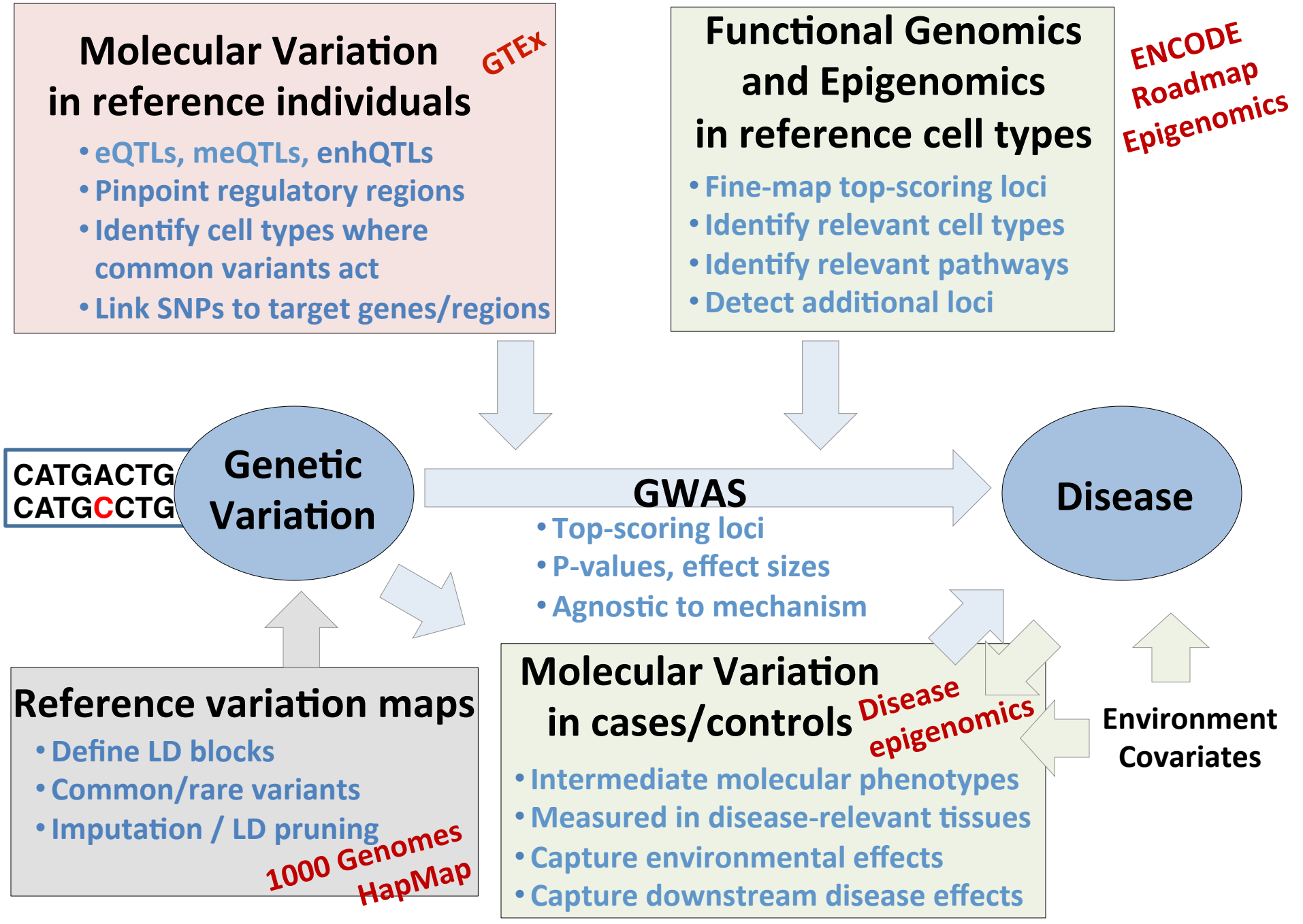


Regulatory genomics to interpret complex disease genetics

1. Regulatory annotations of the human genome: an overview
2. Using regulatory annotations to interpret GWAS
   a. Locus level
   b. Systems level
3. Beyond GWAS
   a. Molecular variability
   b. Empowering rare-variant and pathway analysis

The (very) big picture


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.
H3K27ac
Enhancer
Promoter
Insulator
Gene expr.
Gene expr.
Gene expr.
Endo phenotypes
Lipids
Tension
Heartrate
Metabol.
Drug resp
Disease
Disease cohorts
GTEx / Roadmap Epigenomics
ENCODE/ Roadmap Epigenomics
GTEx
Environment
Feedback from environment / disease state


Molecular Variation in reference individuals
GTEx
• eQTLs, meQTLs, enhQTLs
• Pinpoint regulatory regions
• Identify cell types where common variants act
• Link SNPs to target genes/regions
Functional Genomics and Epigenomics in reference cell types
ENCODE
Roadmap
Epigenomics
• Fine-map top-scoring loci
• Identify relevant cell types
• Identify relevant pathways
• Detect additional loci
CATGACTG
CATGCCTG
Genetic Variation
GWAS
• Top-scoring loci
• P-values, effect sizes
• Agnostic to mechanism
Disease
Reference variation maps
• Define LD blocks
• Common/rare variants
• Imputation / LD pruning
1000 Genomes
HapMap
Molecular Variation in cases/controls
Disease epigenomics
• Intermediate molecular phenotypes
• Measured in disease-relevant tissues
• Capture environmental effects
• Capture downstream disease effects
Environment
Covariates

Acknowledgements

Kellis lab:

– Manolis Kellis
– Pouya Kheradpour
– Jason Ernst
– Benjamin Iriarte
– Abhishek Sarkar
– Matthew Eaton
– Wouter Meuleman
– Anshul Kundaje
– Jianrong Wang

ENCODE Project Consortium

Roadmap Epigenome Mapping Consortium

GTEx Project Consortium

Regulatory genomics to interpret complex disease genetics

Luke Ward

Manolis Kellis lab, MIT

July 18, 2013