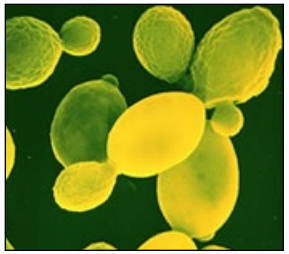
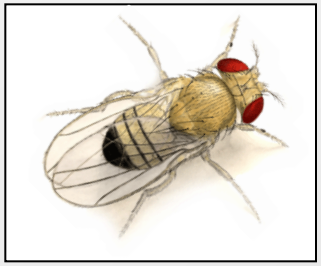


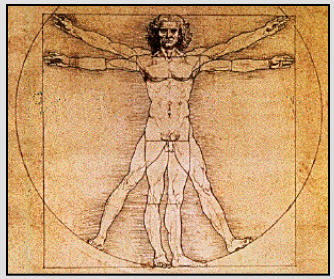


Unweaving the circuitry of complex disease

Manolis Kellis

Broad Institute of MIT and Harvard

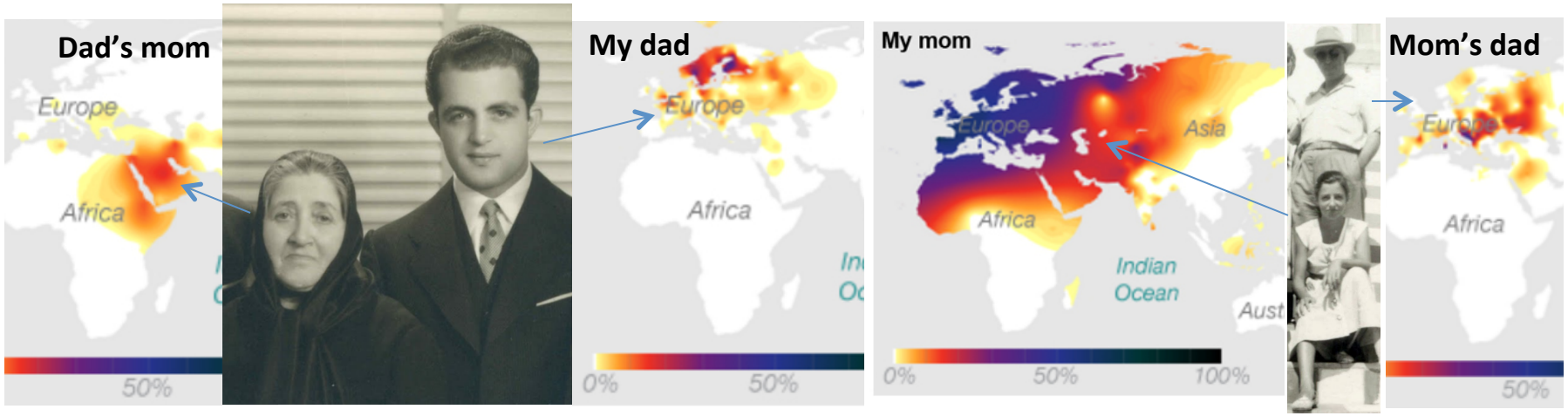
MIT Computer Science & Artificial Intelligence Laboratory



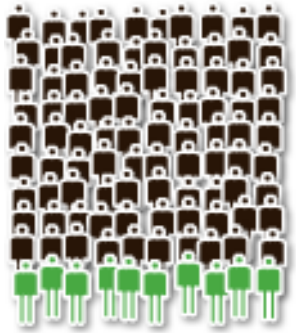
Personal genomics today: 23 and Me

Family Inheritance

Human ancestry



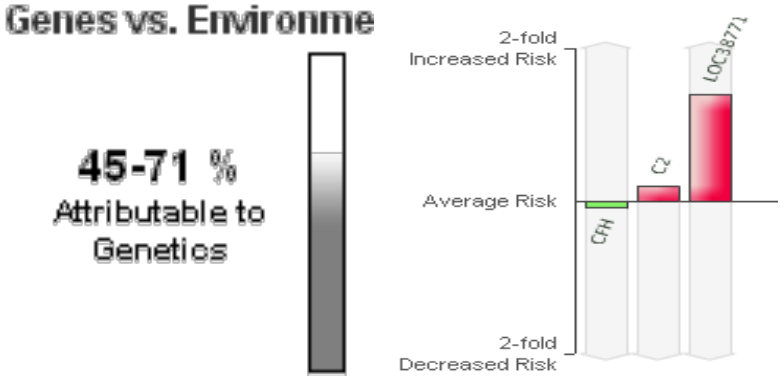
Disease risk



Manolis Kamvysselis

10.5 out of 100

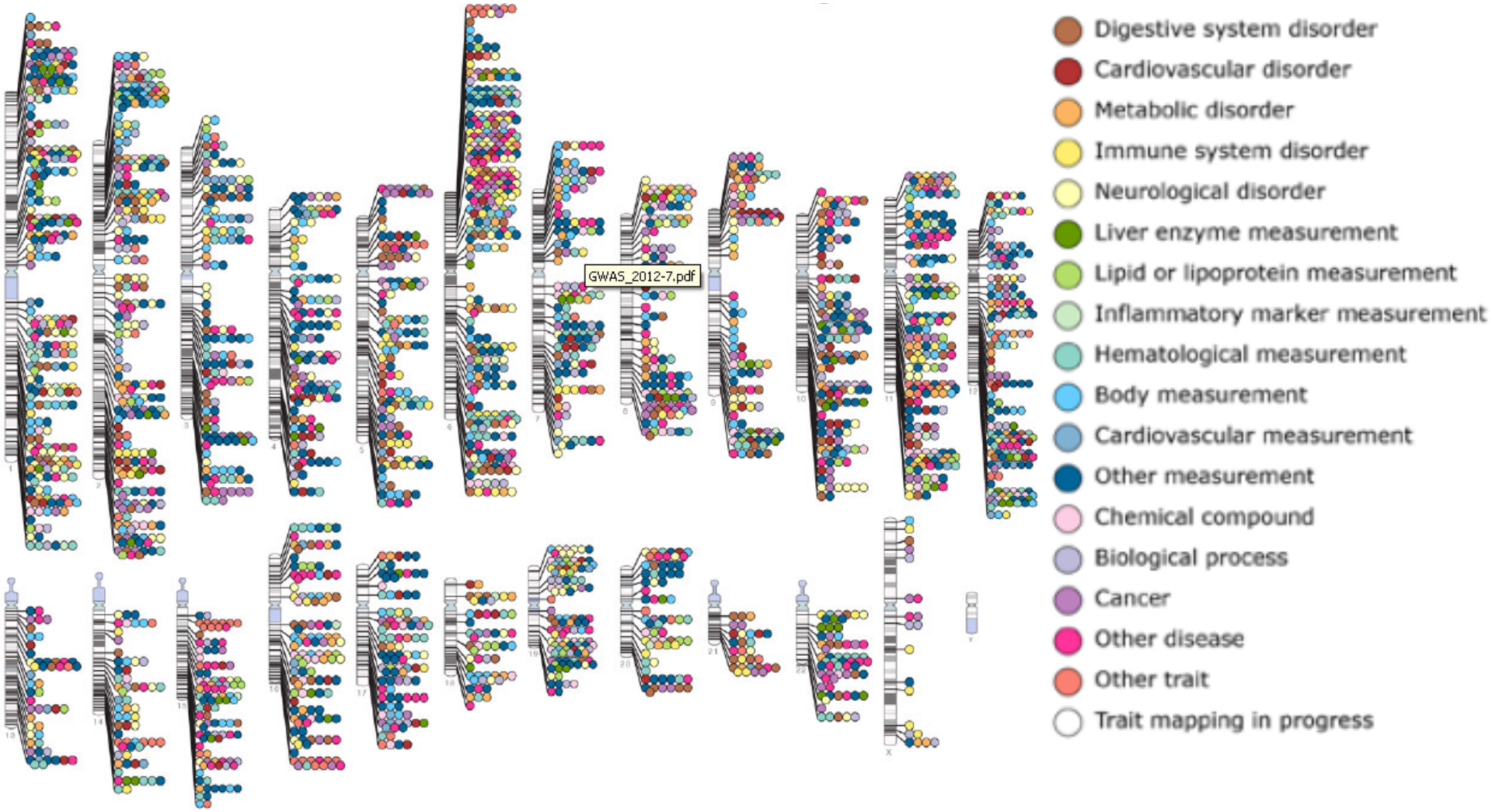
men of European ethnicity who share Manolis Kamvysselis's genotype will develop Age-related Macular Degeneration between the ages of 43 and 79.



Genomics: Regions → mechanisms →

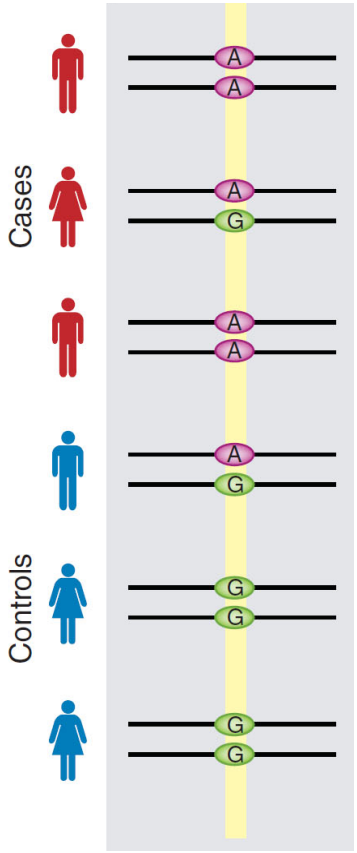
Systems: genes → combinations →

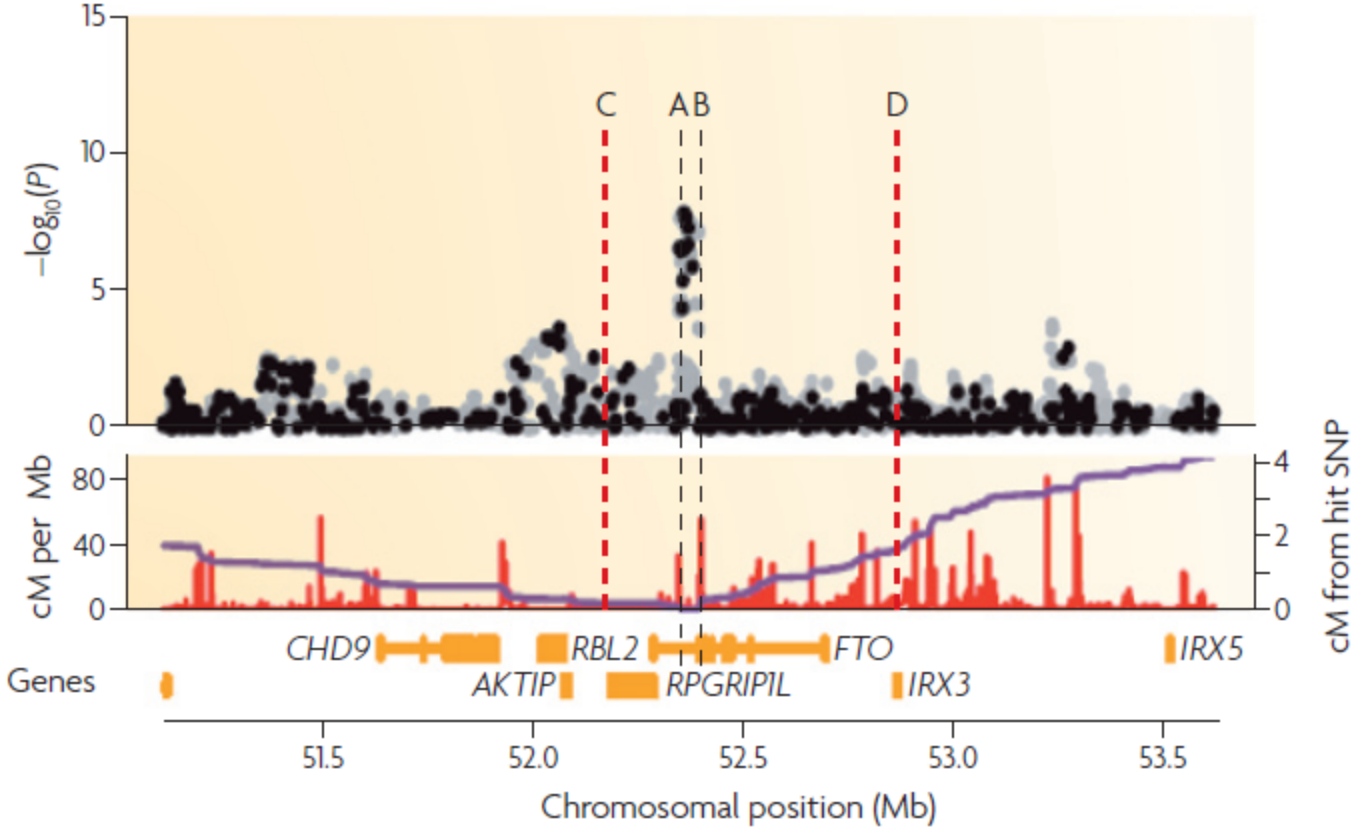
1000s of disease-associated loci from GWAS



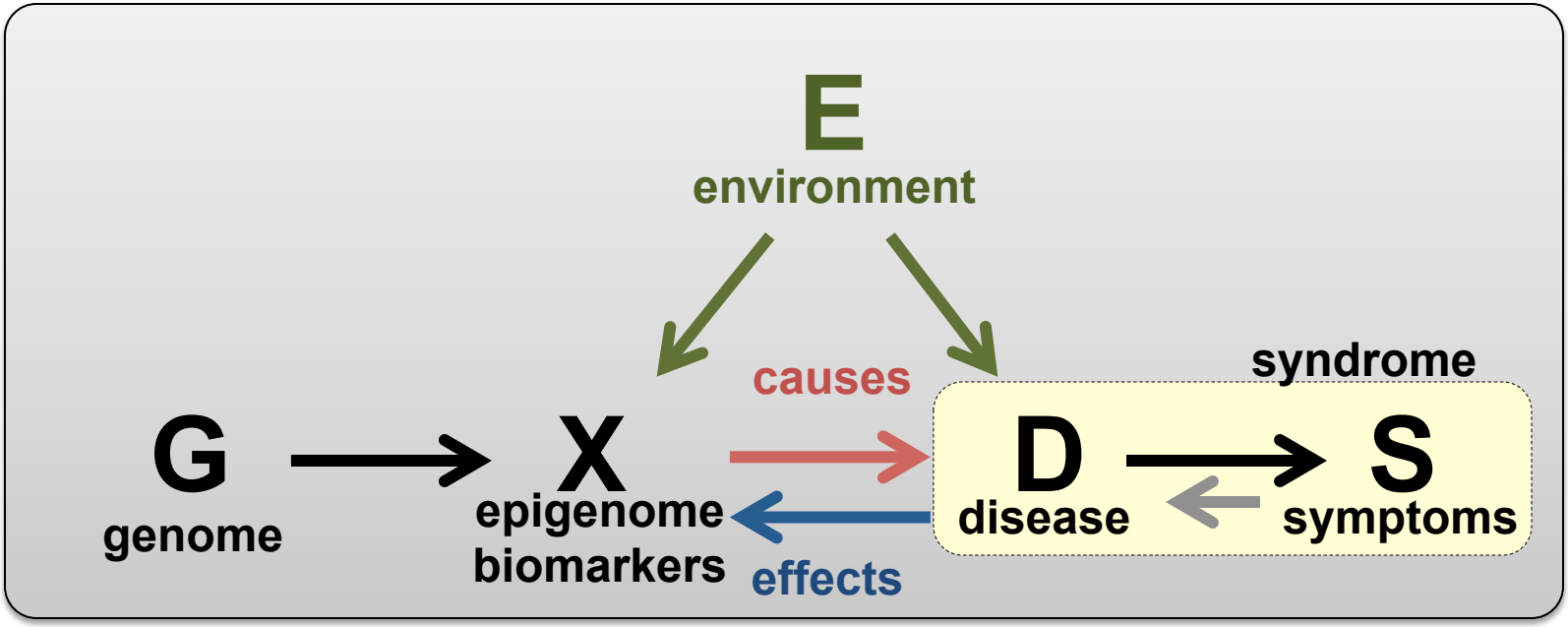
- **Hundreds of studies, each with 1000s of individuals**
  - Power of genetics: find loci, whatever the mechanism may be
  - Challenge: mechanism, cell type, drug target, unexplained heritability

Genome-wide association studies (GWAS)



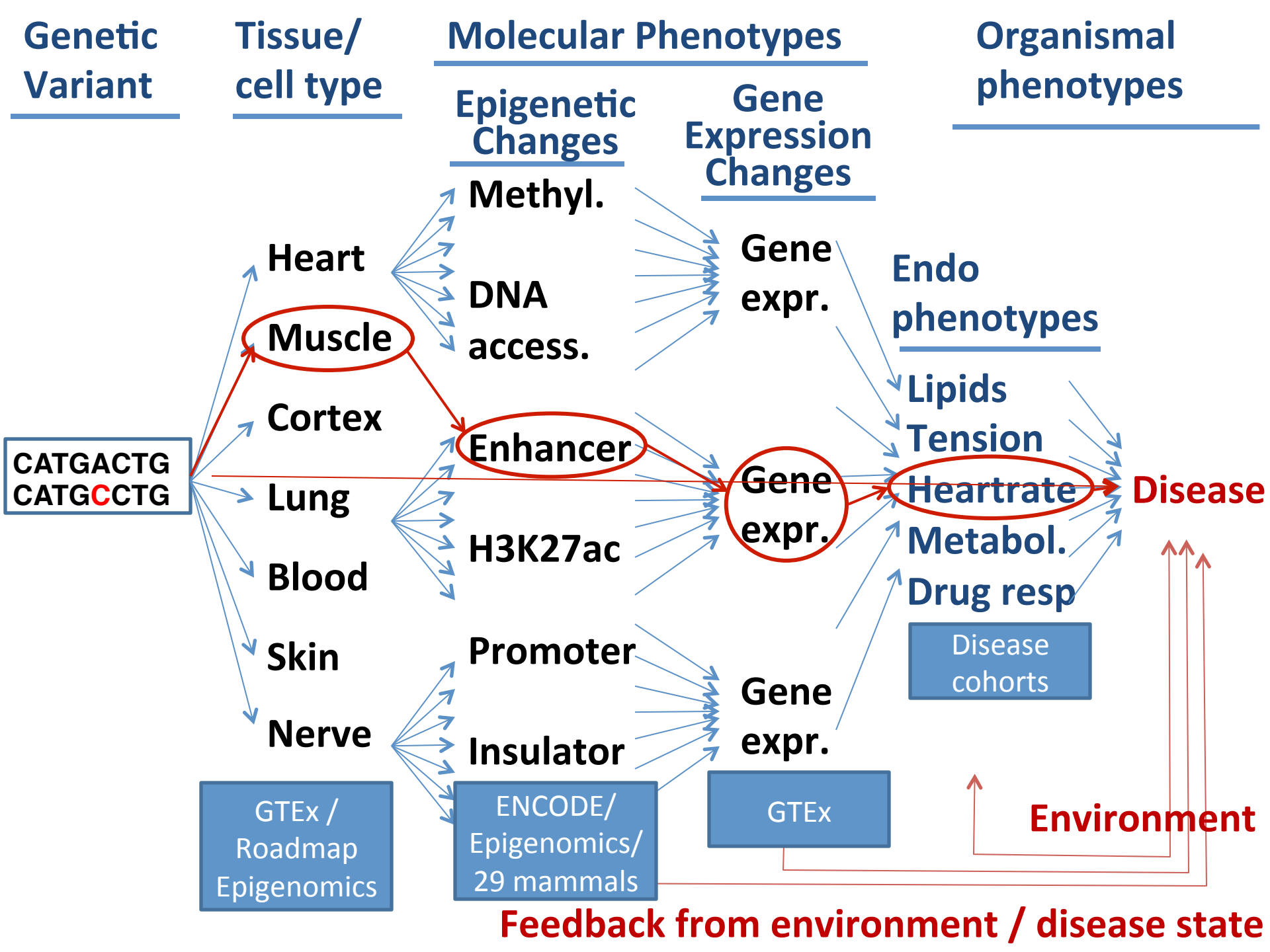


- Identify regions that co-vary with the disease
- Risk allele G more frequent in patients, A in controls
- But: large regions co-inherited ➔ find causal variant
- Genetics does not specify cell type or process

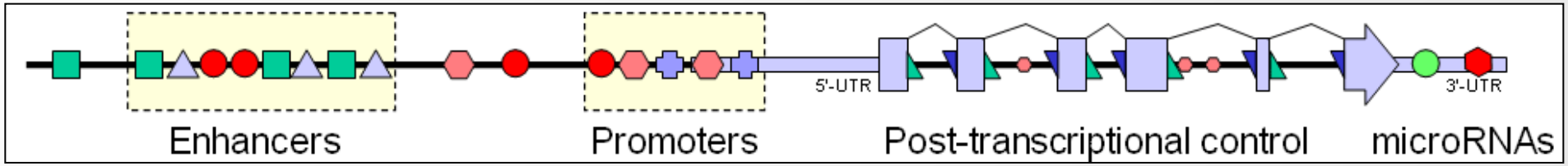


Epidemiology

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


Genetic Variant
Tissue/ cell type
Molecular Phenotypes
Epigenetic Changes
Gene Expression Changes
Organismal phenotypes
Methyl.
DNA access.
Enhancer
H3K27ac
Promoter
Insulator
Heart
Muscle
Cortex
Lung
Blood
Skin
Nerve
CATGACTG
CATGCCTG
Gene expr.
Gene expr.
Gene expr.
Endo phenotypes
Lipids
Tension
Heartrate
Metabol.
Drug resp
Disease
Disease cohorts
GTEx / Roadmap Epigenomics
ENCODE/ Epigenomics/ 29 mammals
GTEx
Environment
Feedback from environment / disease state

Building systems-level views of genome regulation



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

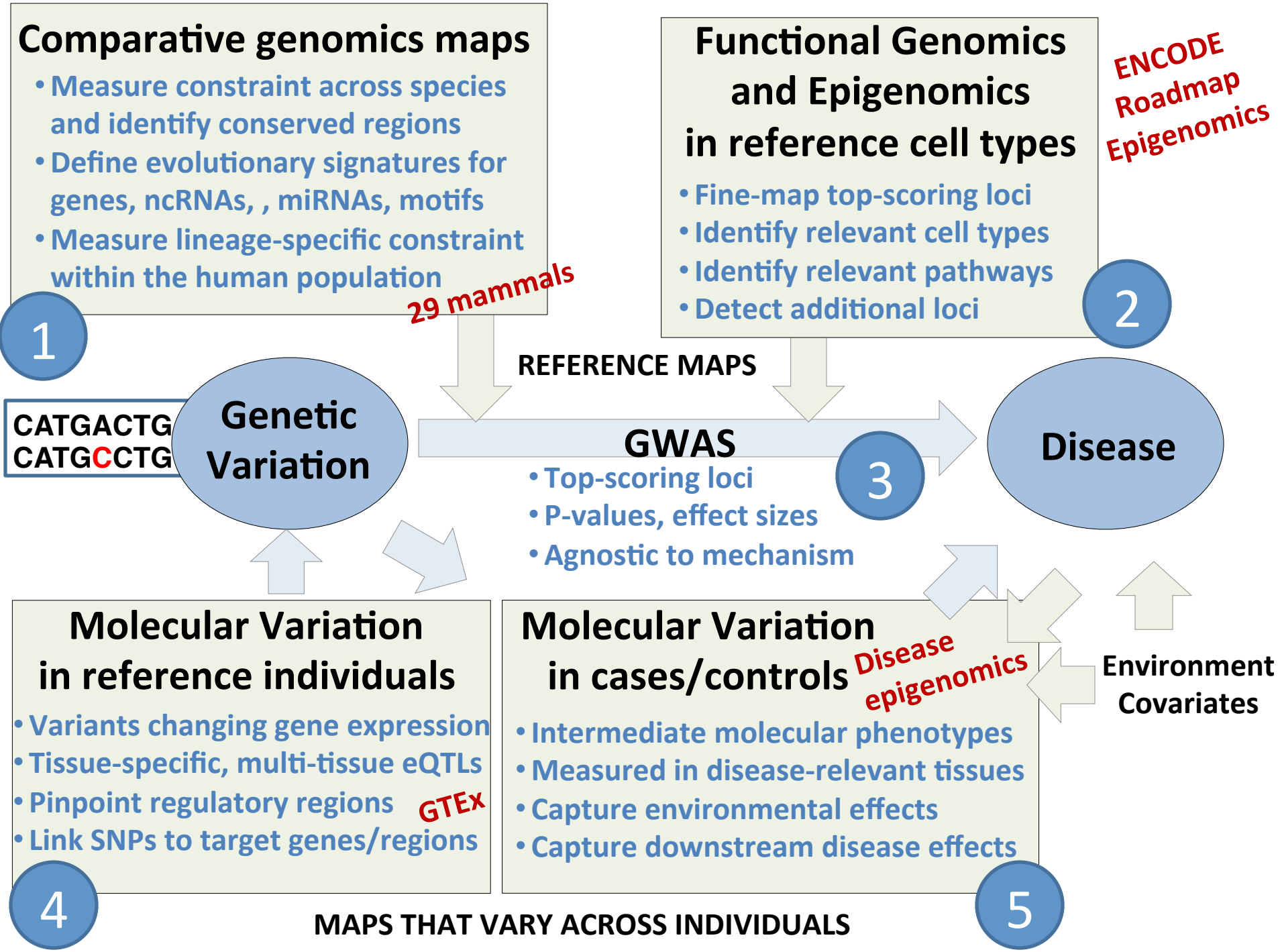
- The regulators: Transcription factors, microRNAs, sequence specificities
- The regions: enhancers, promoters, and their tissue-specificity
- The targets: TFs→targets, regulators→enhancers, enhancers→genes
- The grammars: Interplay of multiple TFs → prediction of gene expression

➔ 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

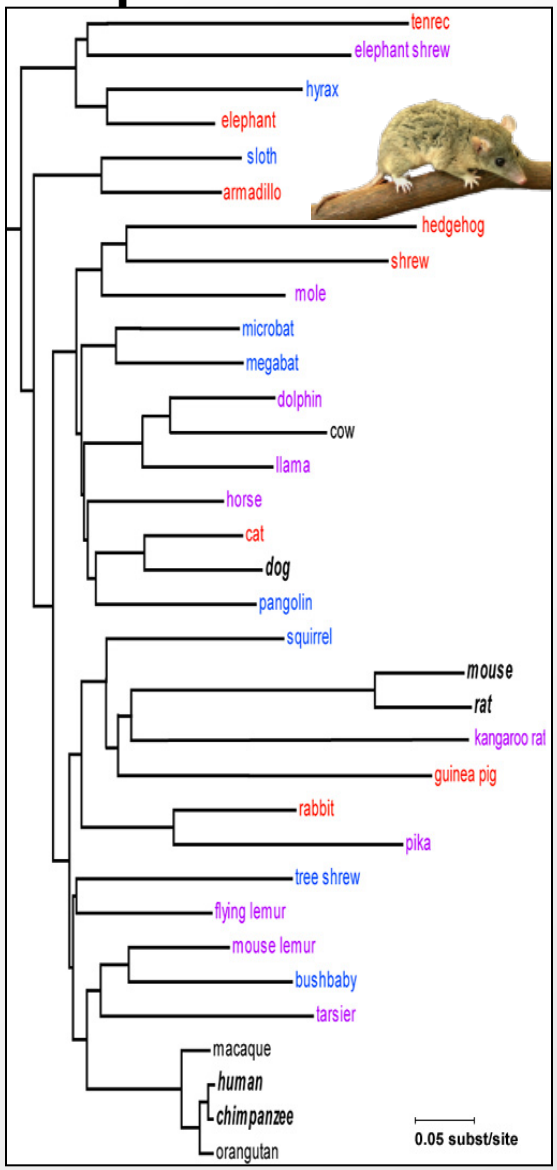

Comparative genomics maps
• Measure constraint across species and identify conserved regions
• Define evolutionary signatures for genes, ncRNAs, , miRNAs, motifs
• Measure lineage-specific constraint within the human population
1
29 mammals
Functional Genomics and Epigenomics in reference cell types
• Fine-map top-scoring loci
• Identify relevant cell types
• Identify relevant pathways
• Detect additional loci
2
ENCODE
Roadmap
Epigenomics
REFERENCE MAPS
CATGACTG
CATGCCTG
Genetic Variation
GWAS
• Top-scoring loci
• P-values, effect sizes
• Agnostic to mechanism
3
Disease
Molecular Variation in reference individuals
• Variants changing gene expression
• Tissue-specific, multi-tissue eQTLs
• Pinpoint regulatory regions
• Link SNPs to target genes/regions
GTEx
4
Molecular Variation in cases/controls
Disease epigenomics
• Intermediate molecular phenotypes
• Measured in disease-relevant tissues
• Capture environmental effects
• Capture downstream disease effects
5
Environment
Covariates
MAPS THAT VARY ACROSS INDIVIDUALS

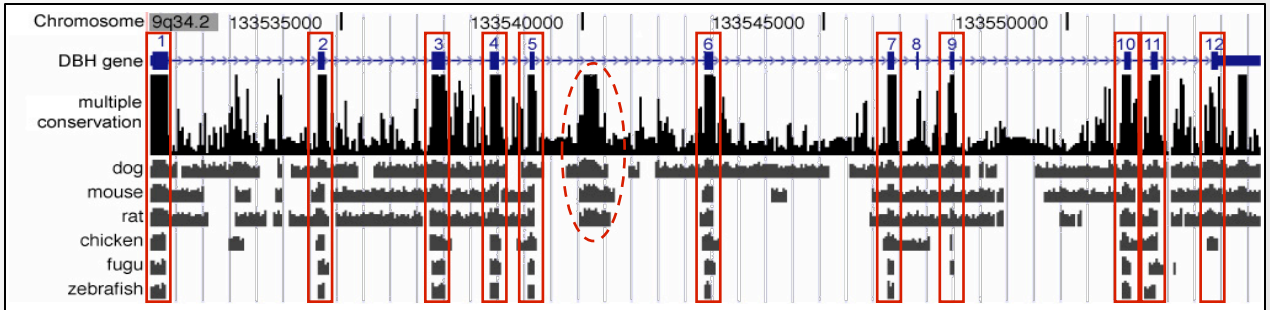
Genomic/Epigenomic tools for disease SNPs

- **Evolutionary signatures → gene/genome annotation**
  - High-resolution annotation: genes, RNAs, motif instances
  - Measuring selection within the human population
- **Chromatin states → regulatory region annotation**
  - Classes of promoter/enhancer/transcribed/repressed/etc
  - Prioritizing chromatin experiments, mark/state imputation
- **Activity signatures → linking enhancer networks**
  - Activity-based linking of TFs → enhancers → targets
  - Testing activators/repressors in 2000+ human enhancers
- **Interpreting disease-associated sequence variants**
  - Mechanistic predictions for individual top-scoring SNPs
  - Functional roles of 1000s of disease-associated SNPs
- **Personal (epi)genomes: geno-/phenotype, cis/trans**
  - Allele-specific activity. Alzheimer's and brain methylation

Evolutionary signatures reveal genes, RNAs, motifs

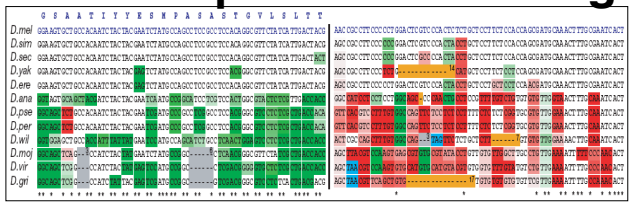
Compare 29 mammals

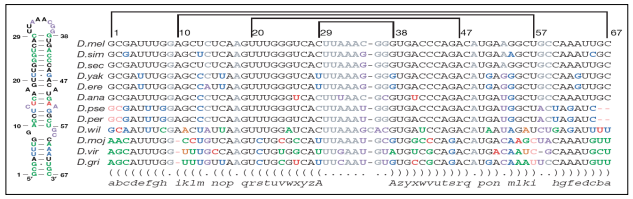


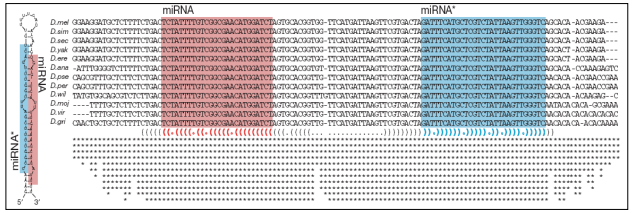


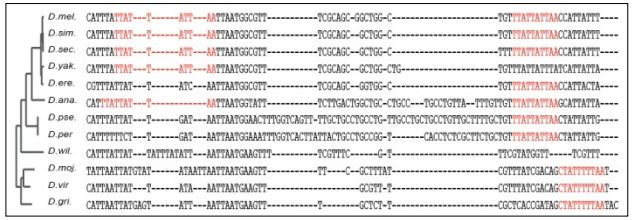
Increased conservation pinpoints functional regions

Distinct patterns of change distinguish diff. functions









Protein-coding genes
- Codon Substitution Frequencies
- Reading Frame Conservation

RNA structures
- Compensatory changes
- Silent G-U substitutions

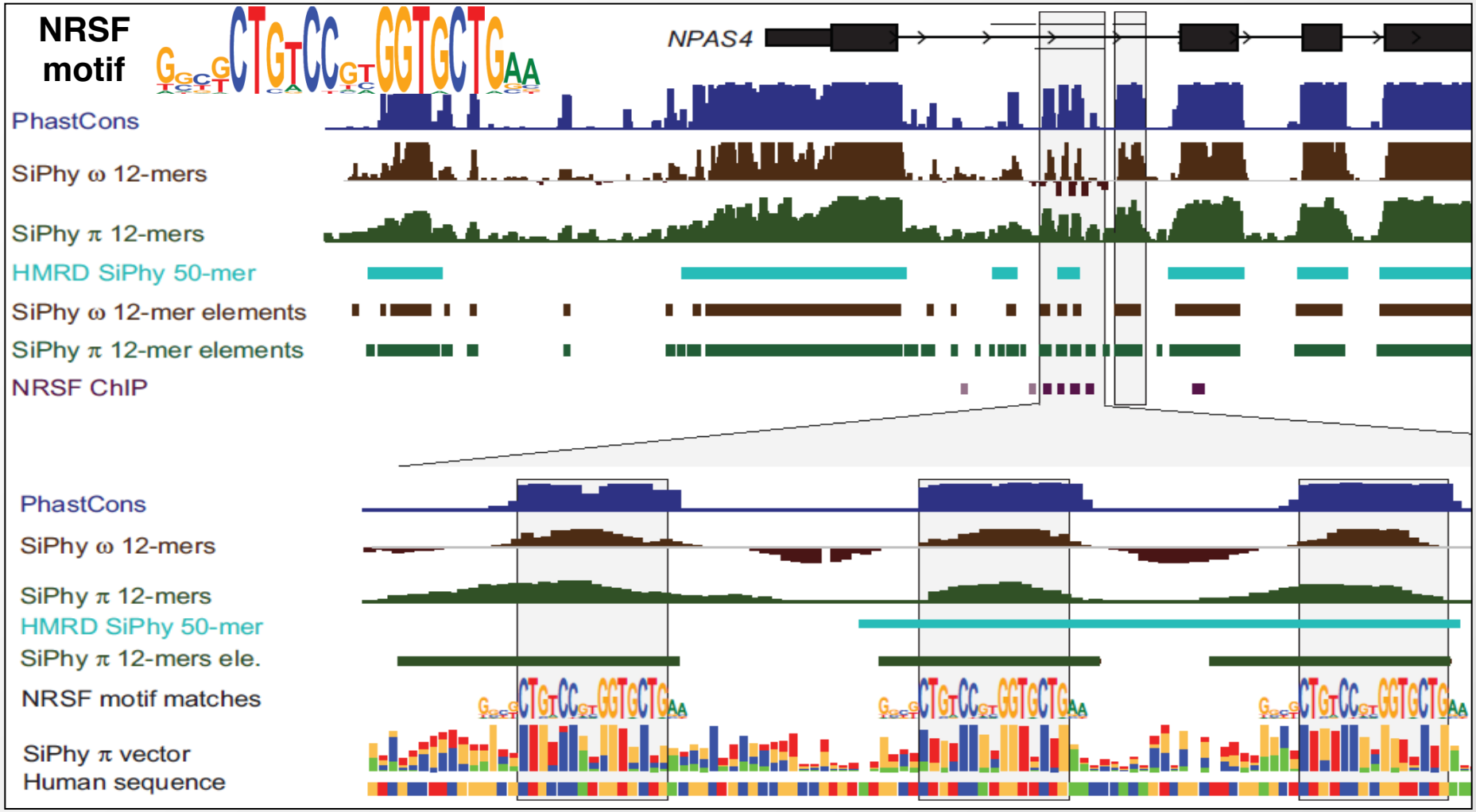
microRNAs
- Shape of conservation profile
- Structural features: loops, pairs
- Relationship with 3'UTR motifs

Regulatory motifs
- Mutations preserve consensus
- Increased Branch Length Score
- Genome-wide conservation

Lindblad-Toh Nature 2011

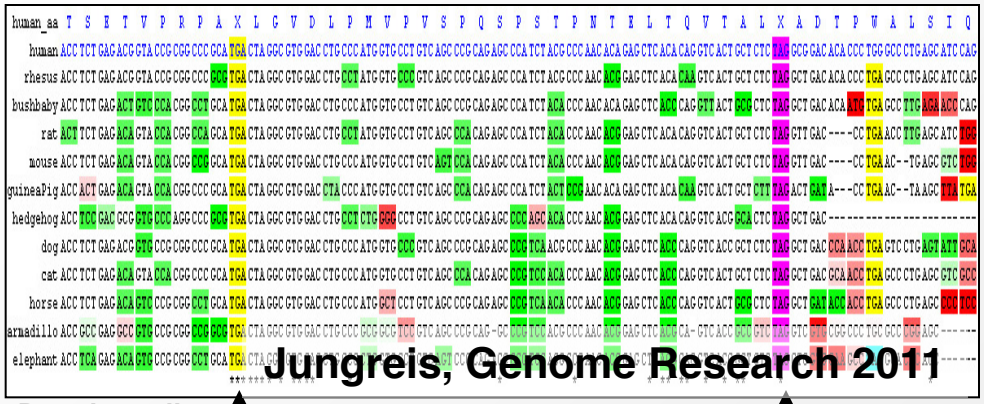
Stark Nature 2007

Measuring constraint at individual nucleotides



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

Translational read-through in human & fly



Protein-coding conservation

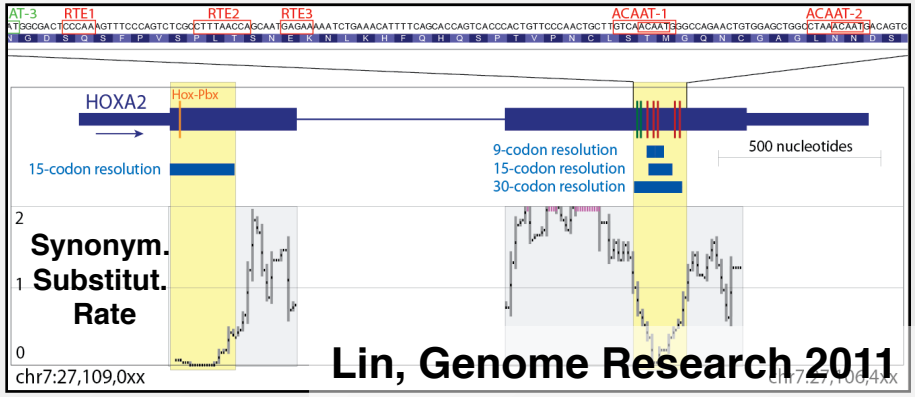
Stop codon read through

Continued protein-coding conservation

2nd stop codon

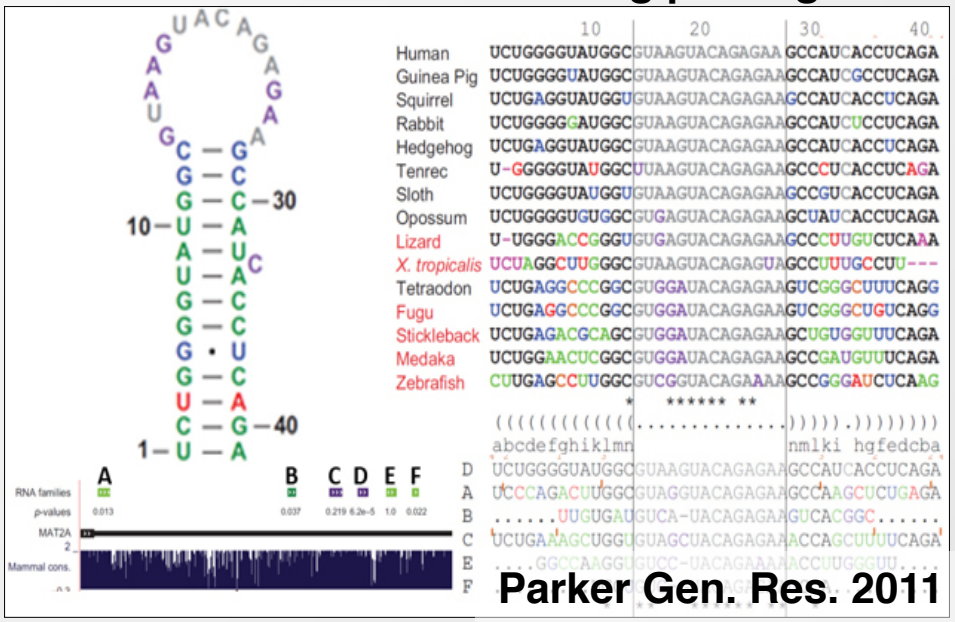
No more conserv

Overlapping selection in human exons

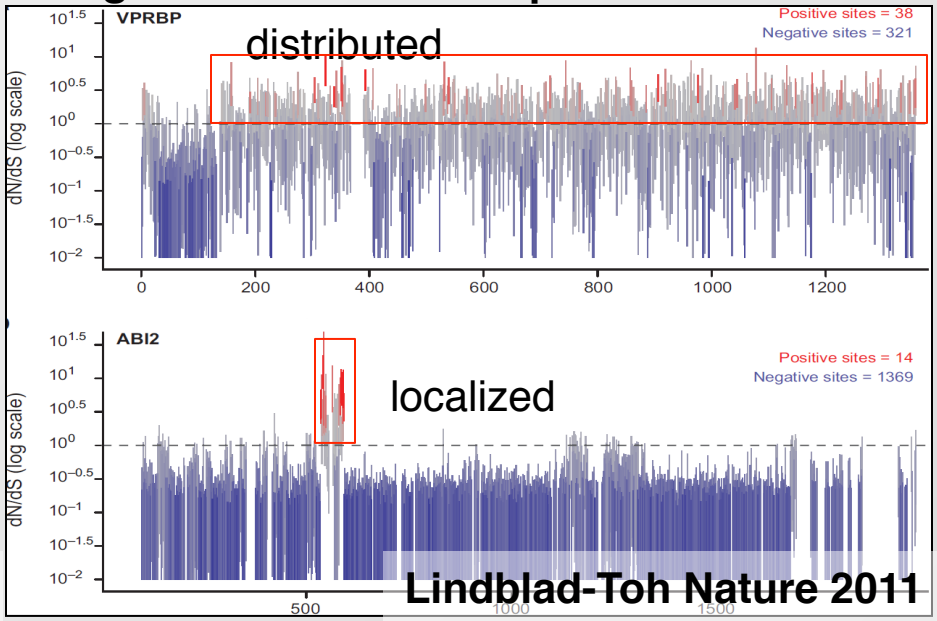


Reveal splicing signals, RNA structures, enhancer motifs, dual-coding genes

RNA structure families: ortholog/paralog cons



Regions of codon-level positive selection



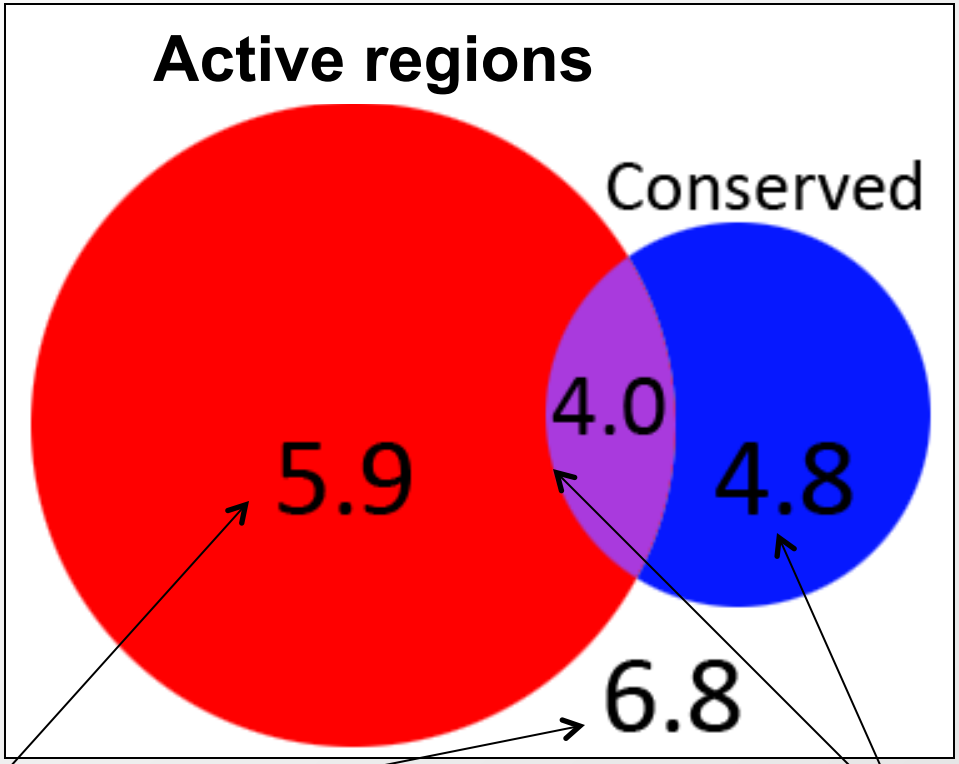
Ex:MAT2A S-adeosyl-methionic level detection

Structure/loop sequence deep conservation

Distributed vs. localized positive selection

Immunity/taste vs. retinal/bone/secretion

Human constraint outside conserved regions



Average diversity (heterozygosity)

Aggregate over the genome

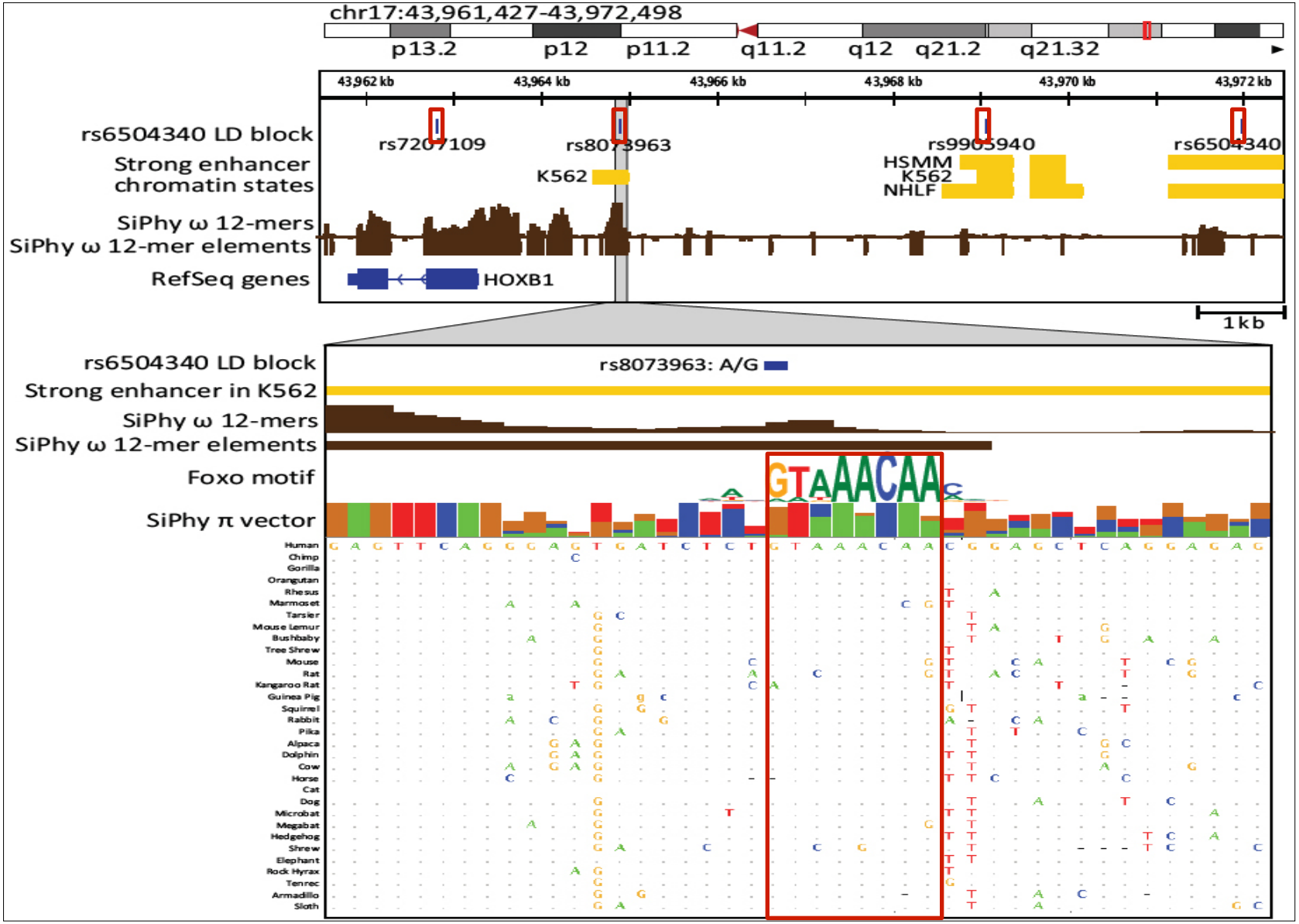
- **Non-conserved regions:**
  - ENCODE-active regions show reduced diversity

➔ Lineage-specific constraint in biochemically-active regions

- **Conserved regions:**
  - Non-ENCODE regions show increased diversity

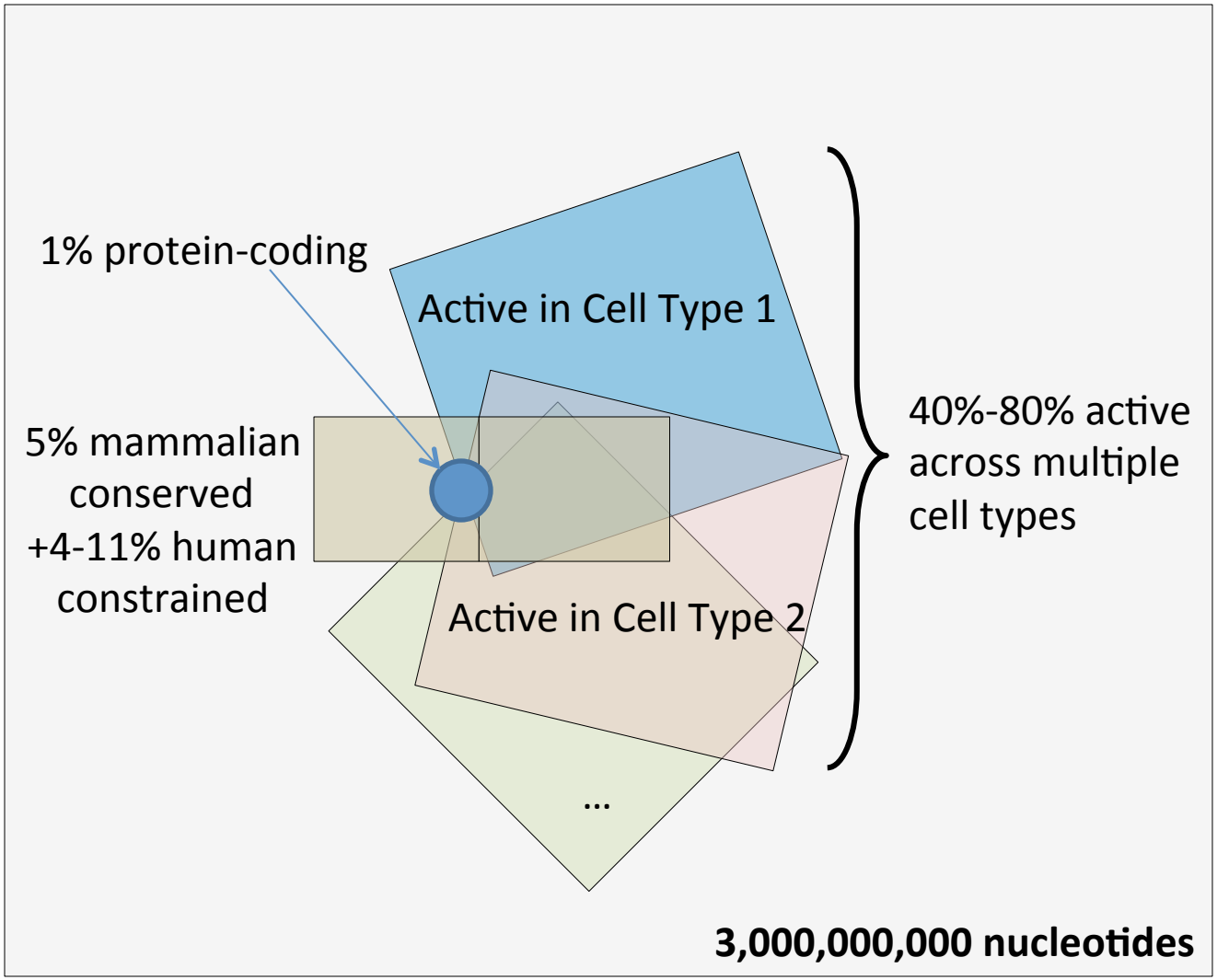
➔ Loss of constraint in human when biochemically-inactive

Detect SNPs that disrupt conserved regulatory motifs



- Functionally-associated SNPs enriched in states, constraint
- Prioritize candidates, increase resolution, disrupted motifs

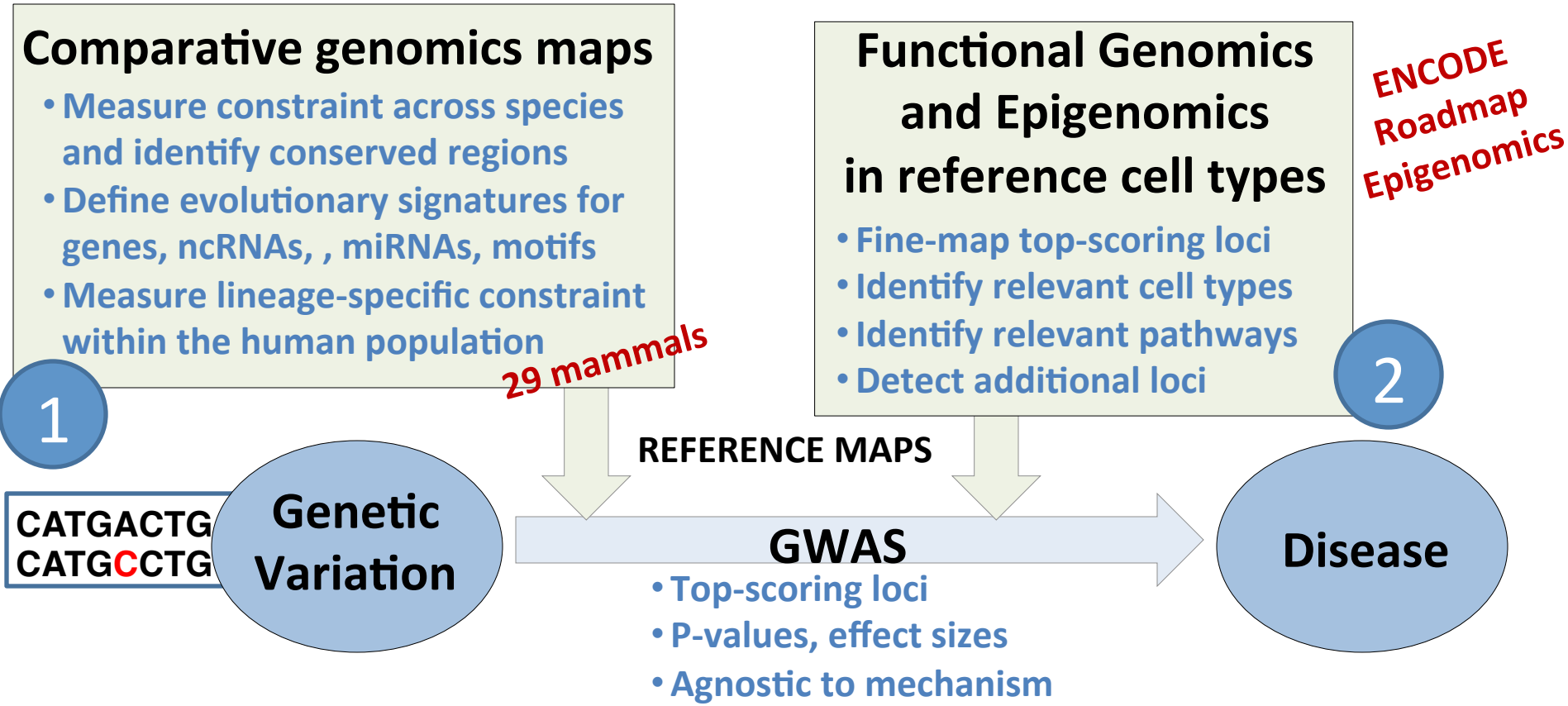
What fraction of the genome matters?



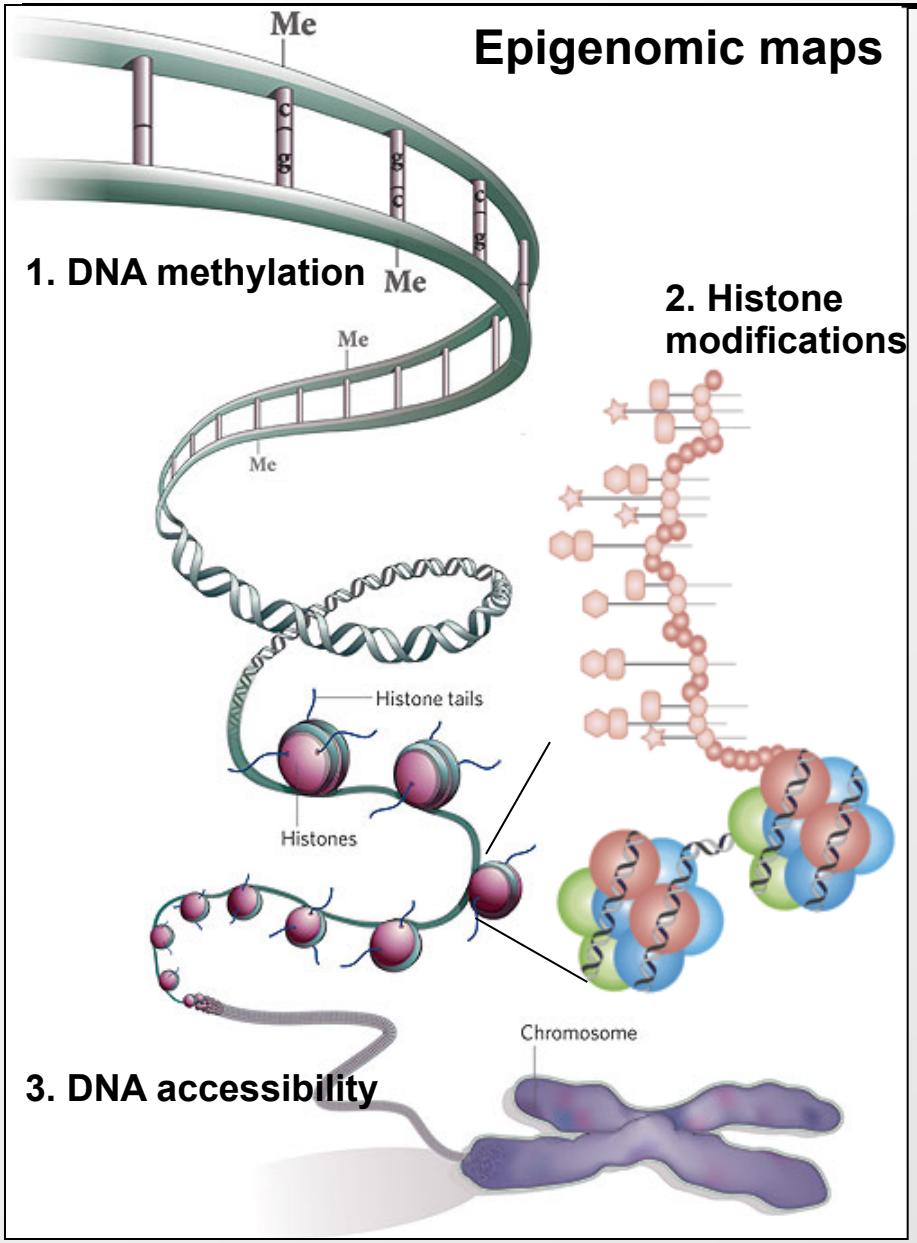
- **Evidence of lineage-specific constraint in human for non-conserved elements**
- **Mutations in non-conserved regions can have late-onset disease phenotypes**
- **Genetic disease requires understanding the whole human genome**

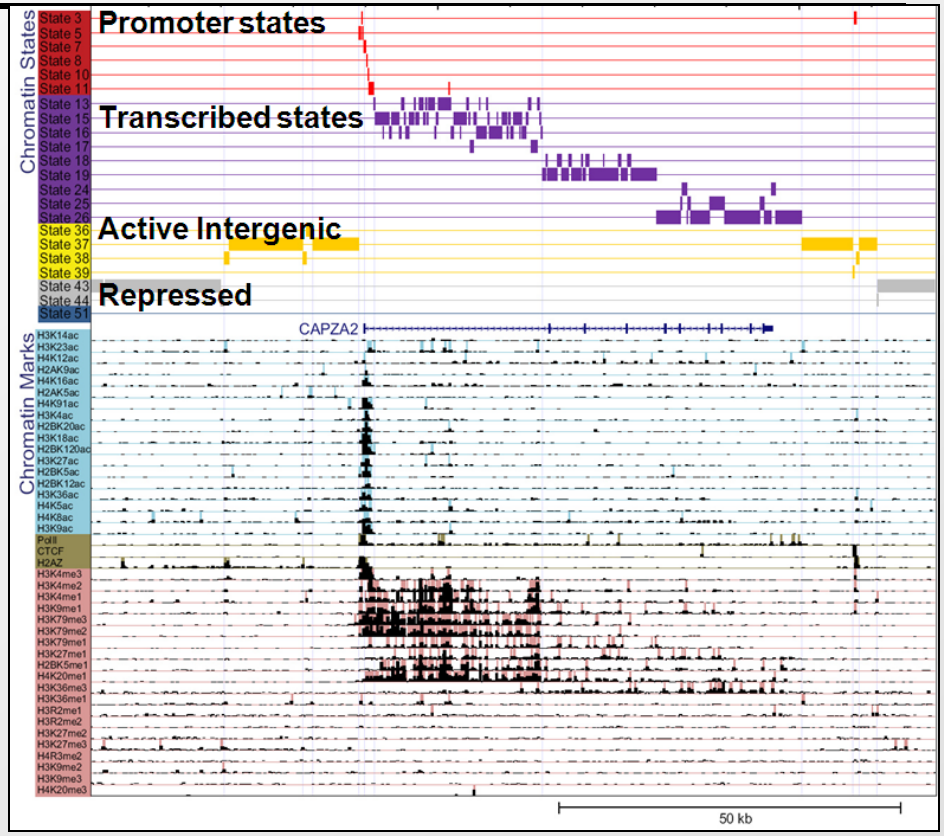
Functional genomics and GTEx for disease

1. **Reference Epigenomes → chromatin states, linking**
   - Annotate dynamic regulatory elements in multiple cell types
   - Activity-based linking of regulators → enhancers → targets
2. **Interpreting disease-associated sequence variants**
   - Mechanistic predictions for individual top-scoring SNPs
   - Functional roles of 1000s of disease-associated SNPs
3. **Multi-cell systems-level expression changes in GTEx**
   - Learn expression programs in 40 tissues for each person
   - Genetic basis of gene module membership changes
4. **Genetic / epigenomic variation in health and disease**
   - Genetic variation⇔Brain methylation⇔Alzheimer's disease
   - Global repression of distal enhancers. NRSF, ELK1, CTCF

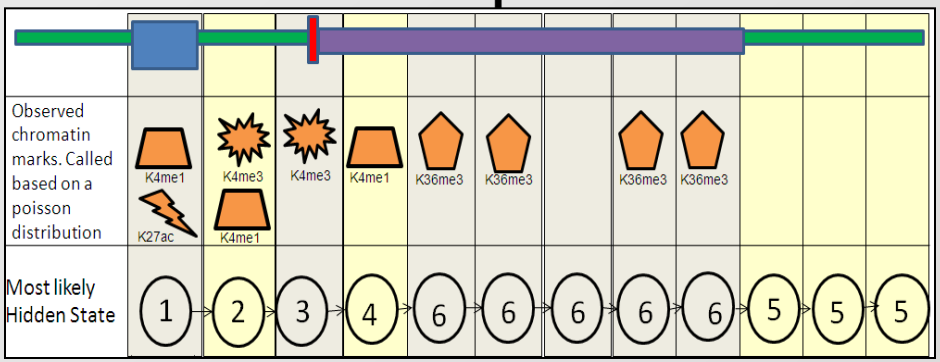

Comparative genomics maps
• Measure constraint across species and identify conserved regions
• Define evolutionary signatures for genes, ncRNAs, , miRNAs, motifs
• Measure lineage-specific constraint within the human population
1
29 mammals
Functional Genomics and Epigenomics in reference cell types
• Fine-map top-scoring loci
• Identify relevant cell types
• Identify relevant pathways
• Detect additional loci
ENCODE
Roadmap
Epigenomics
2
REFERENCE MAPS
CATGACTG
CATGCCTG
Genetic Variation
GWAS
• Top-scoring loci
• P-values, effect sizes
• Agnostic to mechanism
Disease

Chromatin signatures for genome annotation



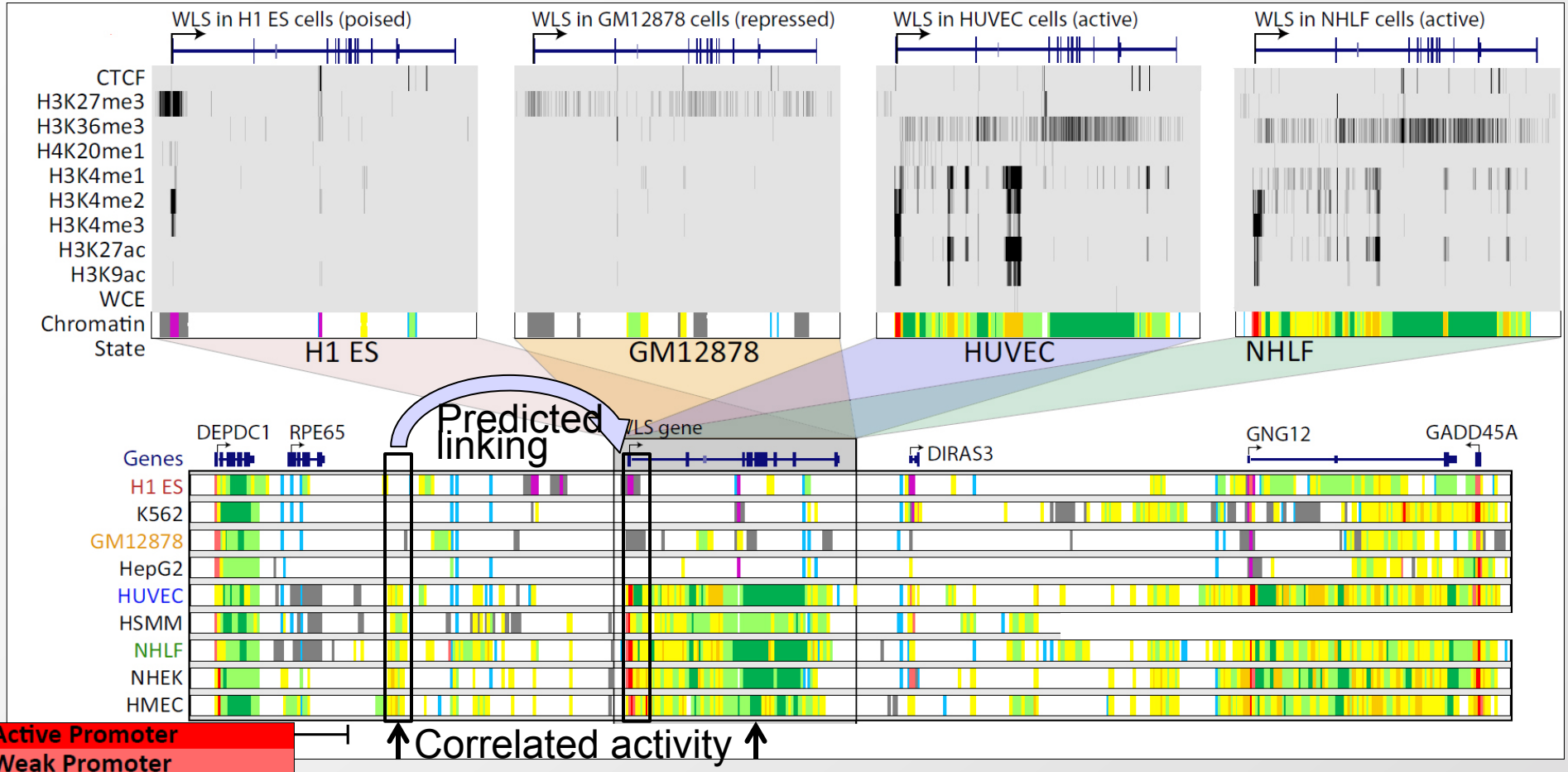


Multi-variate HMM captures combinations



Ernst & Kellis Nature Biotech 2010. Ernst & Kellis Nature Methods, 2012.ChromHMM

Chromatin state dynamics reveal linking/regulators

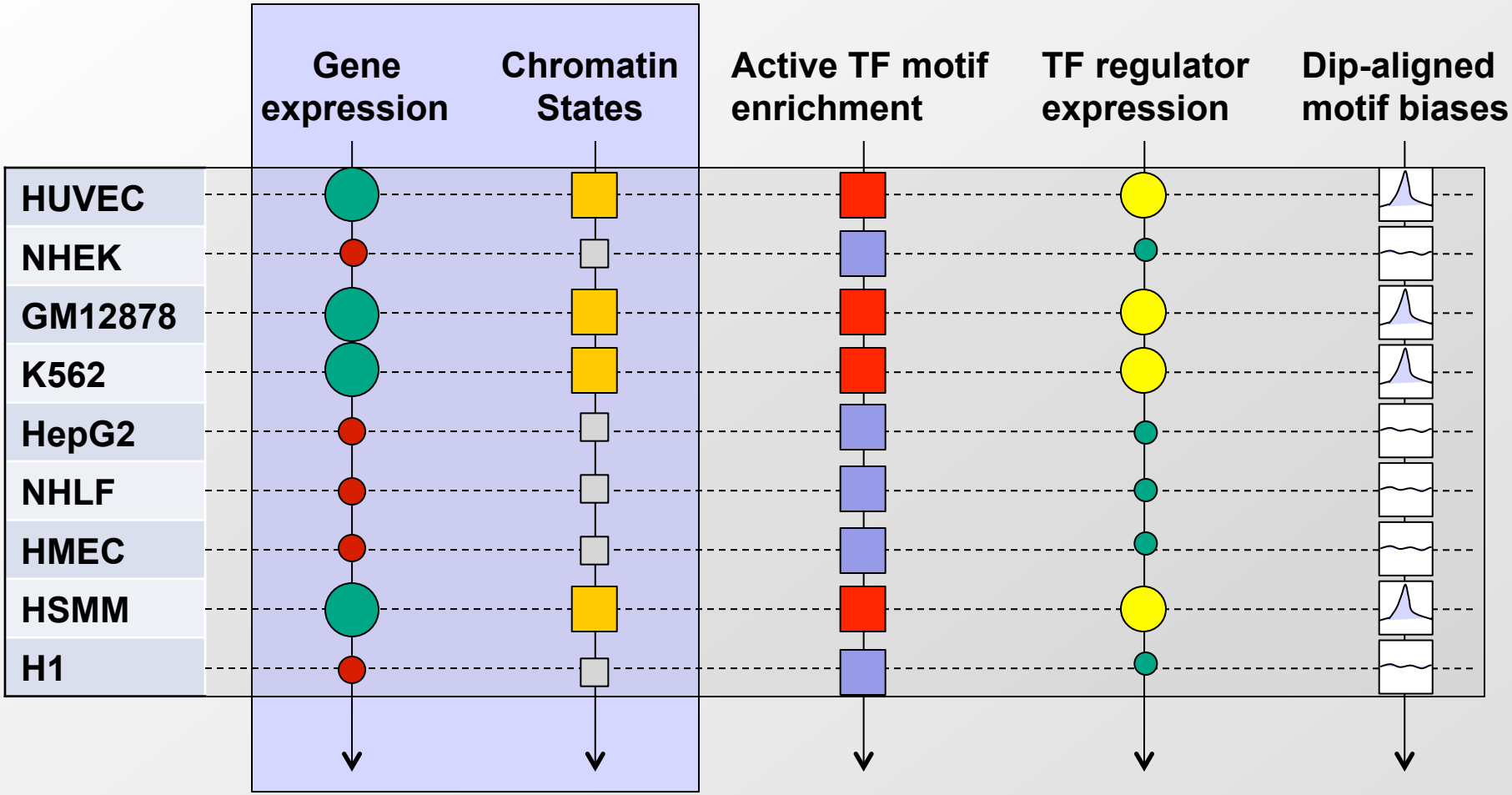


Active Promoter
Weak Promoter
Inactive/poised Promoter
Strong enhancer
Strong enhancer
Weak/poised enhancer
Weak/poised enhancer
Insulator
Transcriptional transition
Transcriptional elongation
Weak transcribed
Polycomb-repressed
Heterochrom; low signal

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

Ernst , Kheradpour, *et al*, Nature 2011

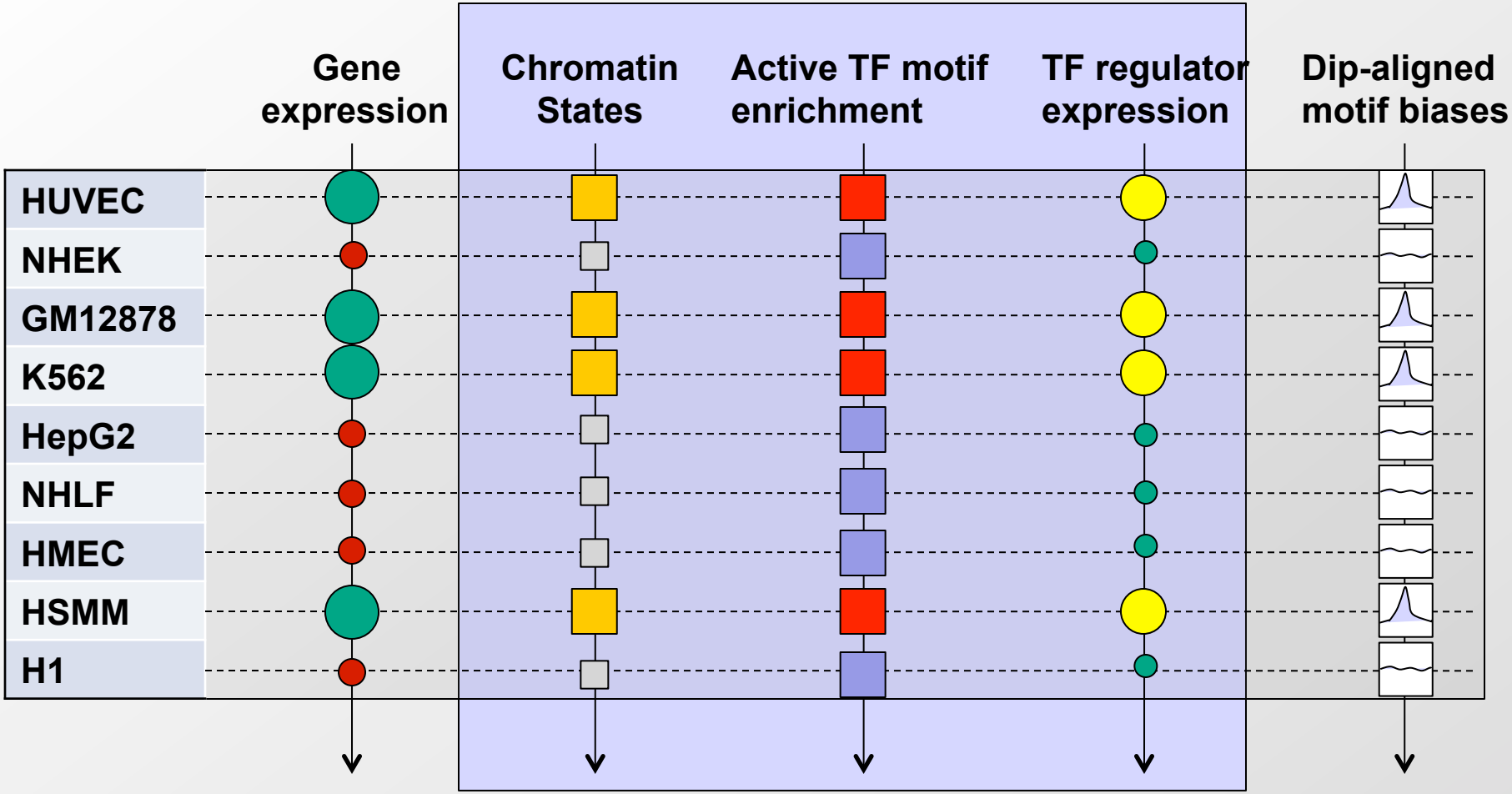
Multi-cell activity profiles connect enhancers



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

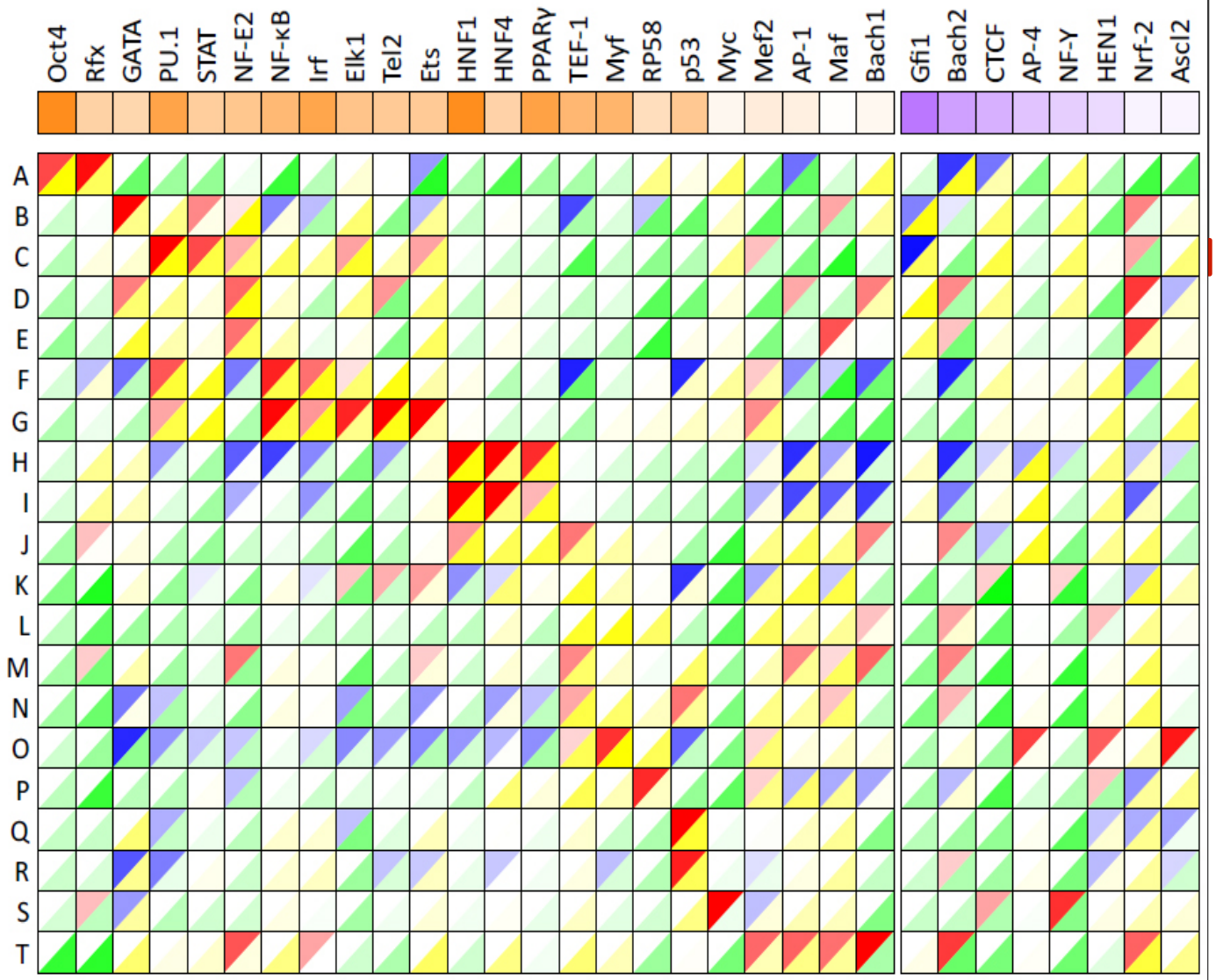
ON
OFF
Active enhancer
Repressed
Motif enrichment
Motif depletion
TF On
TF Off
Motif aligned
Flat profile

Coordinated activity reveals activators/repressors

Enhancer activity

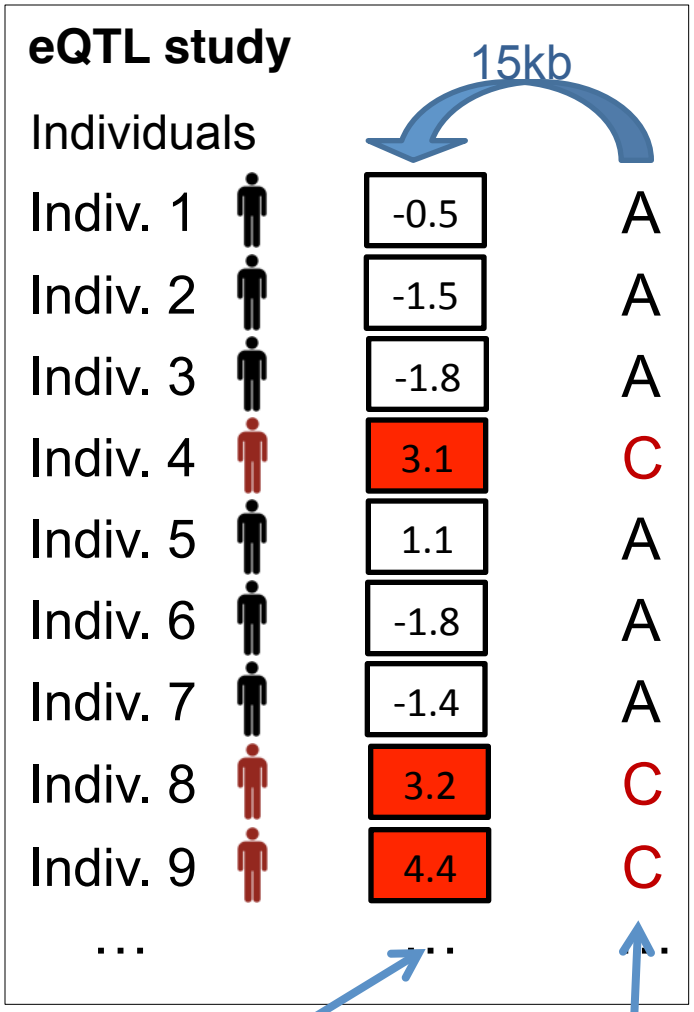
	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC
A	44	3	2	3	2	2	2	2	2
B	1	48	2	4	2	2	2	2	2
C	2	48	47	8	5	5	4	5	3
D	4	46	9	10	45	10	11	14	10
E	4	48	12	14	10	10	10	44	26
F	0	1	48	1	1	1	1	1	1
G	2	5	48	6	46	16	13	12	7
H	1	2	2	51	1	1	1	2	1
I	1	1	1	36	1	1	1	1	1
J	3	7	6	49	31	30	18	12	10
K	1	2	2	2	46	3	4	4	3
L	2	3	2	2	47	45	20	7	9
M	3	3	5	6	49	26	19	47	34
N	2	2	6	5	5	47	18	46	31
O	1	1	2	1	2	45	5	2	3
P	1	3	4	2	4	21	45	4	4
Q	2	1	5	5	3	3	4	49	45
R	1	1	3	2	3	3	3	45	8
S	2	4	4	4	8	6	5	8	43
T	11	40	27	25	45	41	39	45	41

Activity signatures for each TF



- Enhancer networks: Regulator → enhancer → target gene

Enhancer-gene links supported by eQTL-gene links



Expression level of gene ⇔ **Sequence variant at distal position**

Validation rationale:

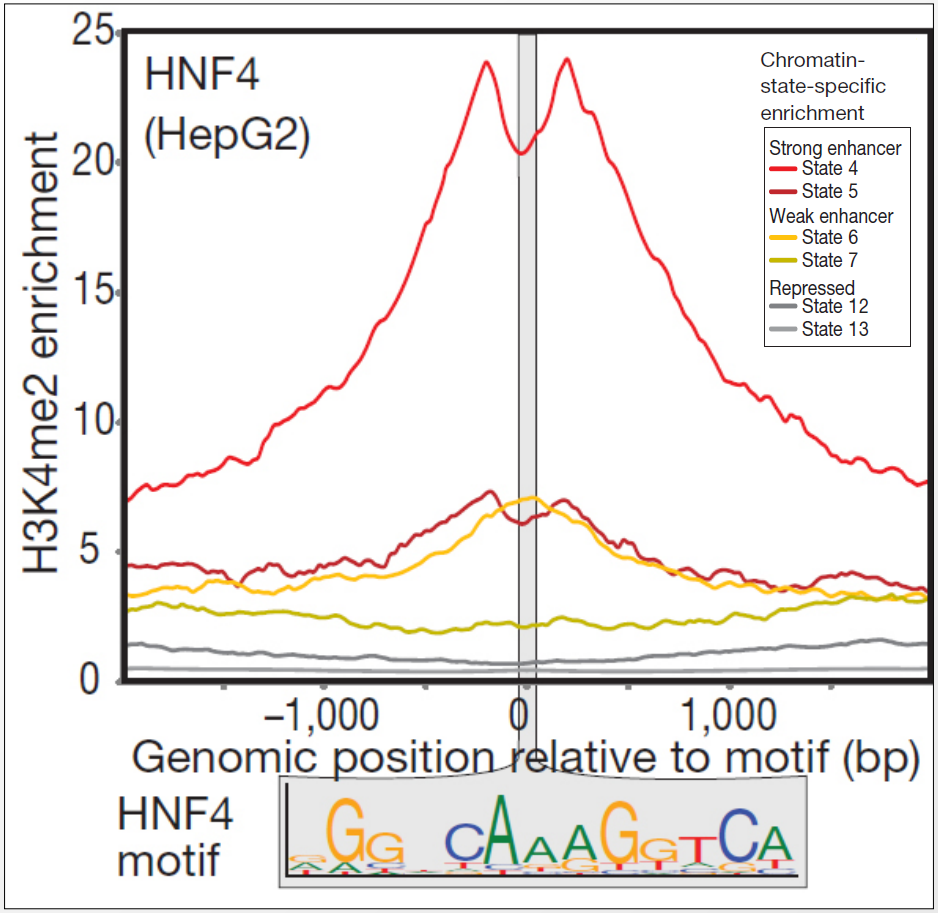
- Expression Quantitative Trait Loci (eQTLs) provide **independent SNP-to-gene** links
- Do they agree with activity-based links?

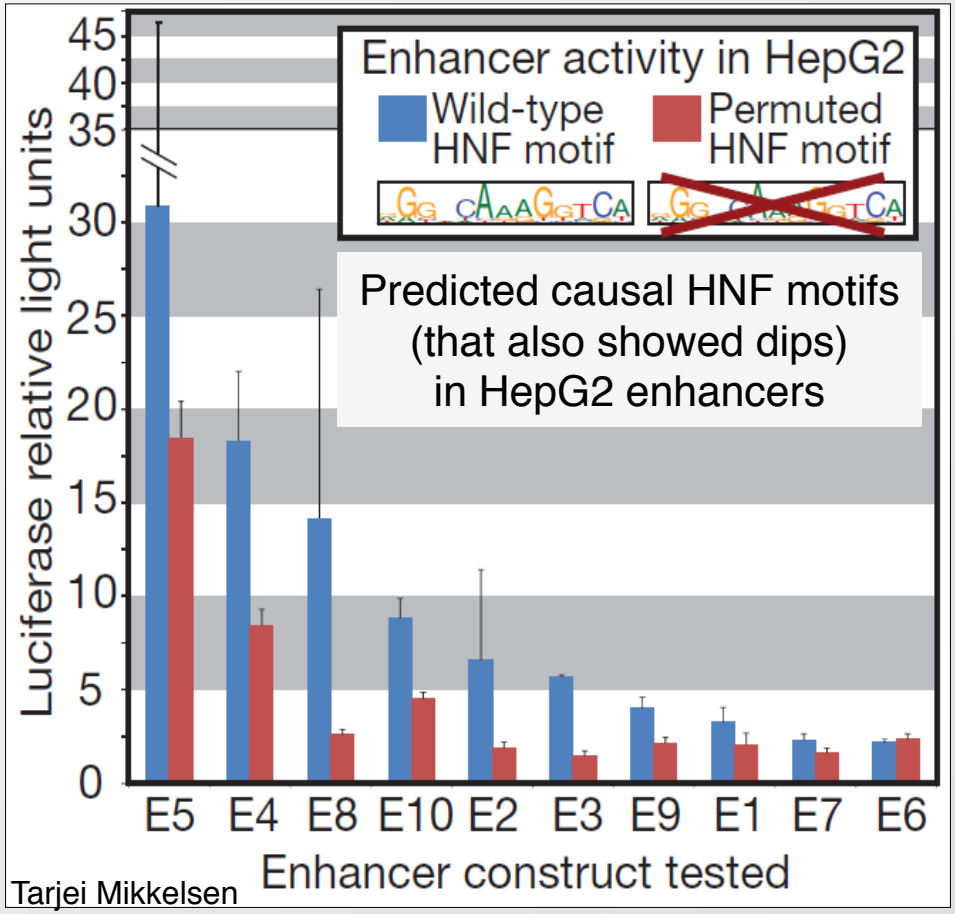
Example: Lymphoblastoid (GM) cells study

- **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.

Causal motifs supported by dips & enhancer assays



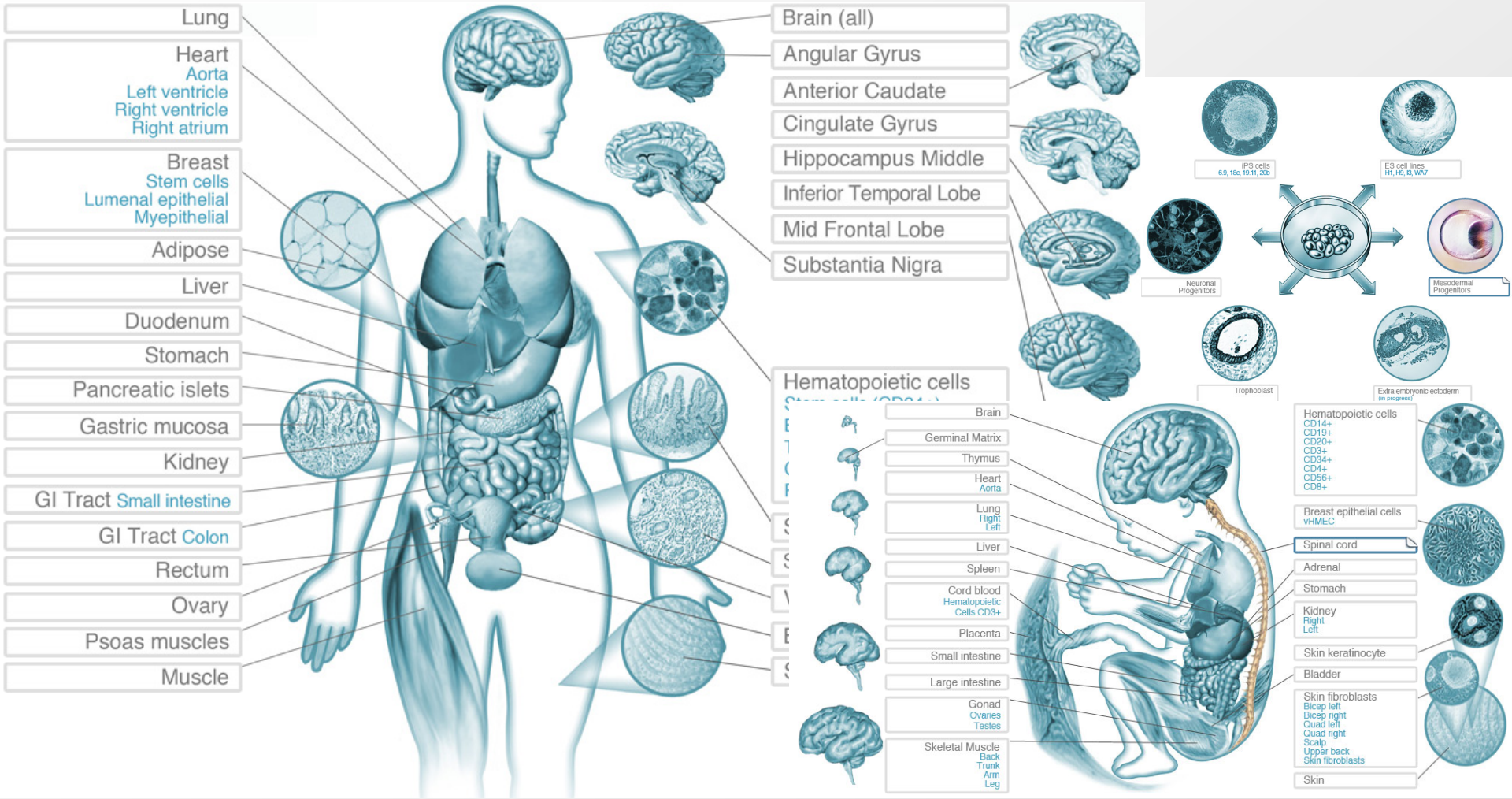


Dip evidence of TF binding (nucleosome displacement)

Enhancer activity halved by single-motif disruption

➔ Motifs bound by TF, contribute to enhancers

Epigenomics Roadmap across 110 tissues/cell types



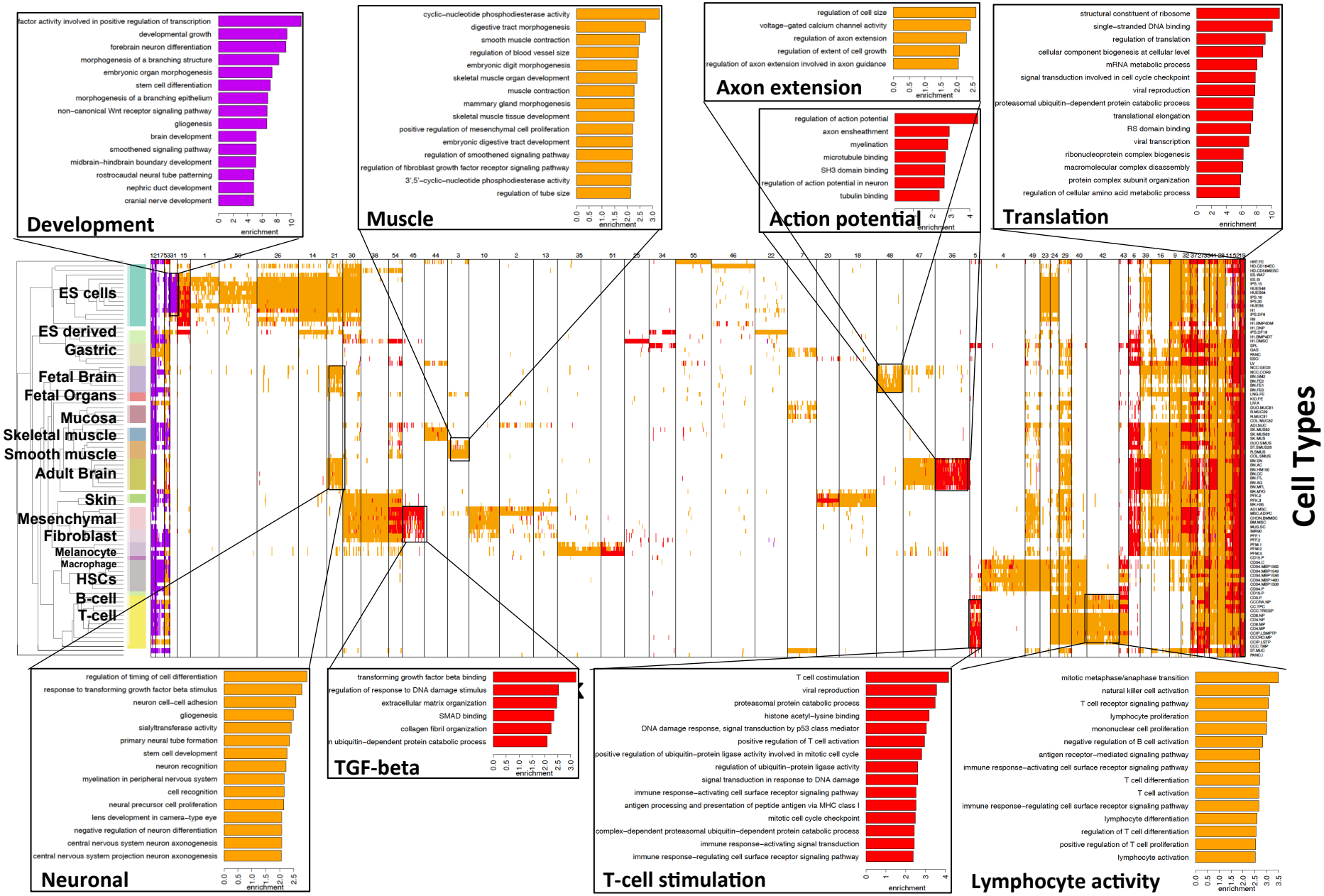
- **ChIP-Seq: 8-20 histone marks**
- **Methyl: WGBS, RRBS, MeDip, MRE**
- **Accessibility: DNase, Footprints**
- **RNA: mRNA, smRNA, Exons**

Integration: chromatin states, regulatory regions, hi-res, high-coverage

Epigenomics Roadmap: 90 reference epigenomes

Interpret GWAS, global effects, reveal relevant cell types

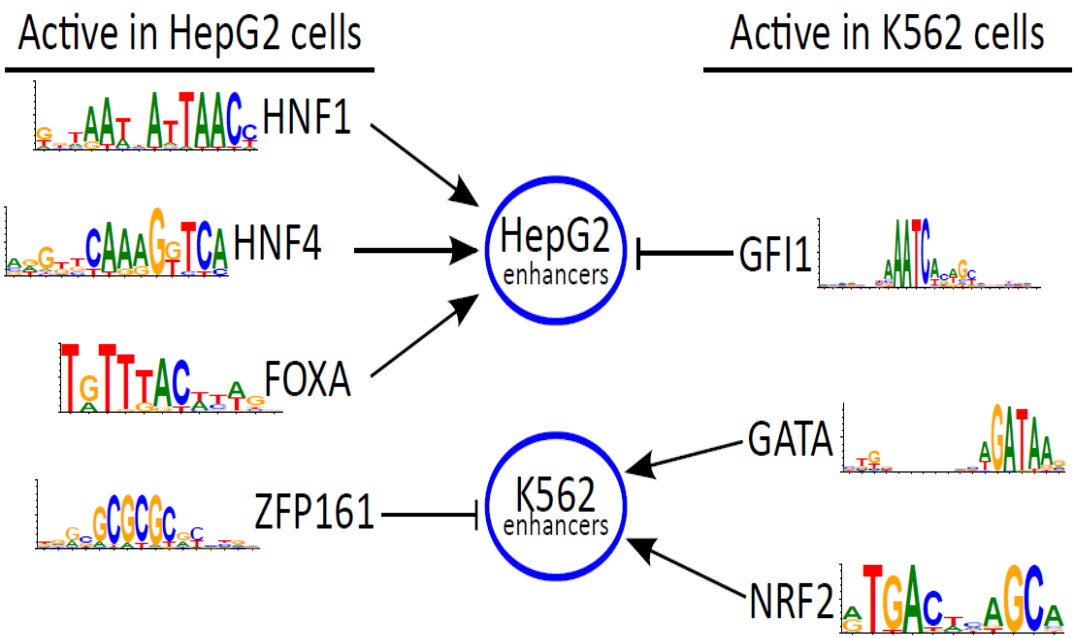
Cell-type specific functional enrichments for cell-type specific enhancers

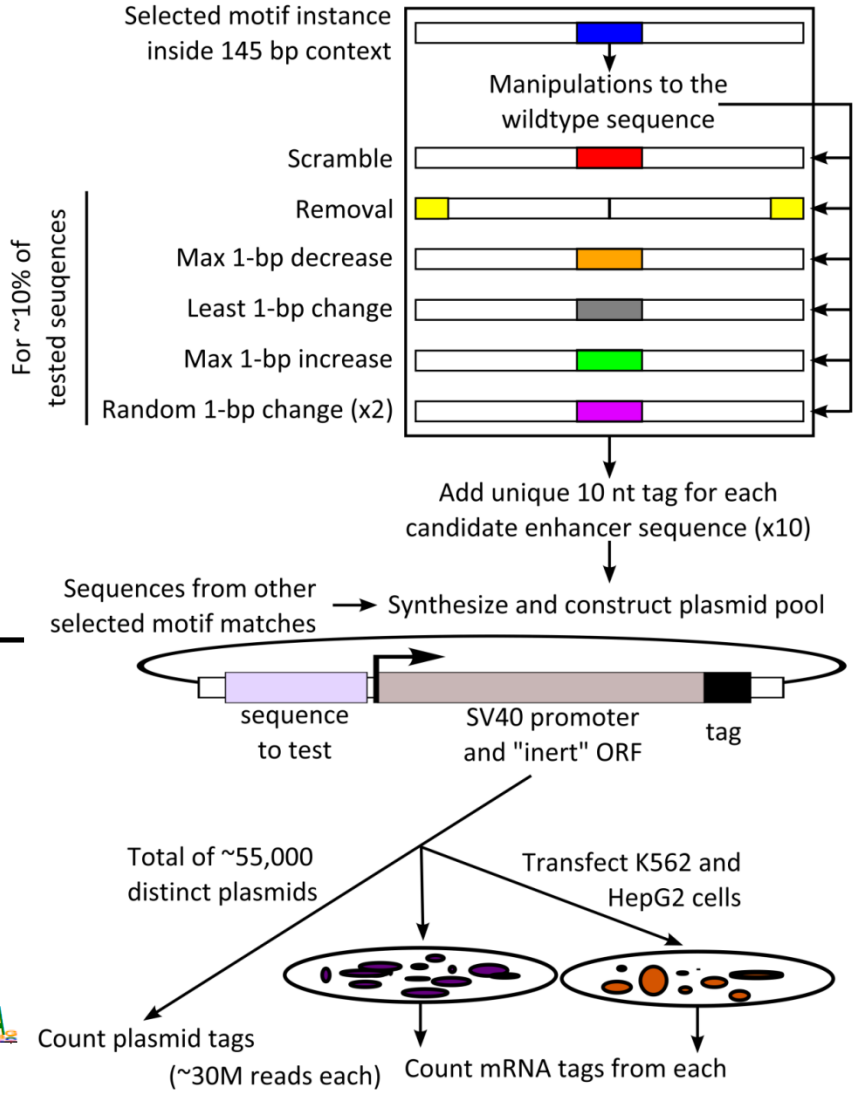


Wouter Meuleman, Epigenomics Roadmap project, in preparation

Systematic motif disruption for 5 activators and 2 repressors in 2 human cell lines

	Motif-motif similarity							Motif enrichment in enhancers				Factor expression			
								HepG2		K562		HepG2		K562	
	HNF1	HNF4	FOXA	GATA4	NRF2	ZFP161	GFI1	0.0	0.4	0.0	0.4	rep1	rep2	rep1	rep2
HNF1	1.0	0.4	0.4	0.4	0.4	0.1	0.4	1.5	2.3	1.0	1.0	0.8	0.5	-0.1	-0.2
HNF4	0.4	1.0	0.4	0.3	0.3	0.2	0.3	1.7	2.1	1.0	1.0	1.0	0.5	-0.0	-0.1
FOXA	0.4	0.4	1.0	0.3	0.5	0.1	0.4	1.4	1.7	1.0	1.0	2.2	2.1	-0.4	-0.4
GATA	0.4	0.3	0.3	1.0	0.3	0.1	0.5	1.0	1.0	2.1	2.8	0.1	0.3	0.4	0.4
NRF2	0.4	0.3	0.5	0.3	1.0	0.2	0.4	1.0	1.1	1.5	1.8	0.3	0.7	-0.1	-0.3
ZFP161	0.1	0.2	0.1	0.1	0.2	1.0	0.1	0.8	0.5	1.2	1.0	0.0	0.0	0.1	0.1
GFI1	0.4	0.3	0.4	0.5	0.4	0.1	1.0	1.0	1.0	0.6	0.5	0.4	0.3	1.3	1.1





54000+ measurements (x2 cells, 2x repl)

Kheradpour *et al* Genome Research 2013

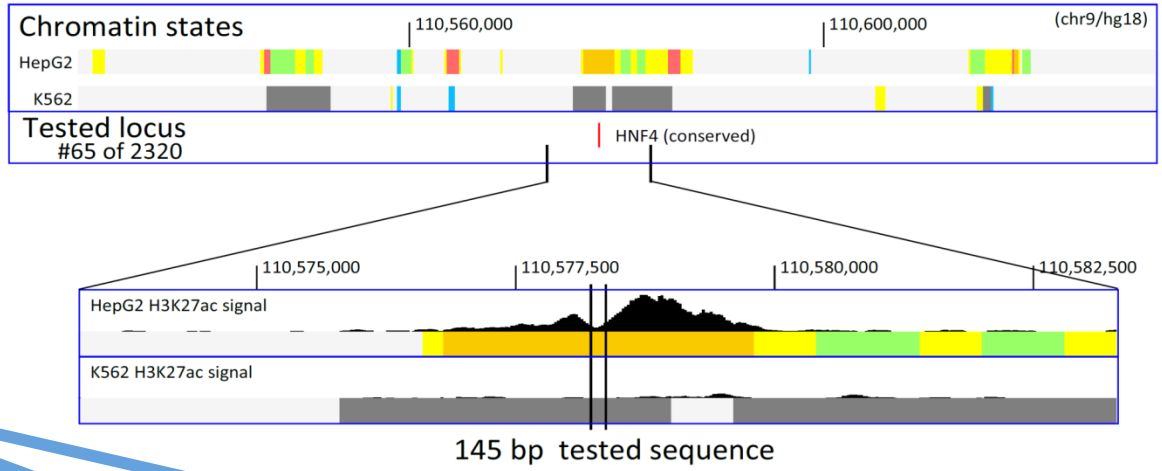
Example activator: conserved HNF4 motif match

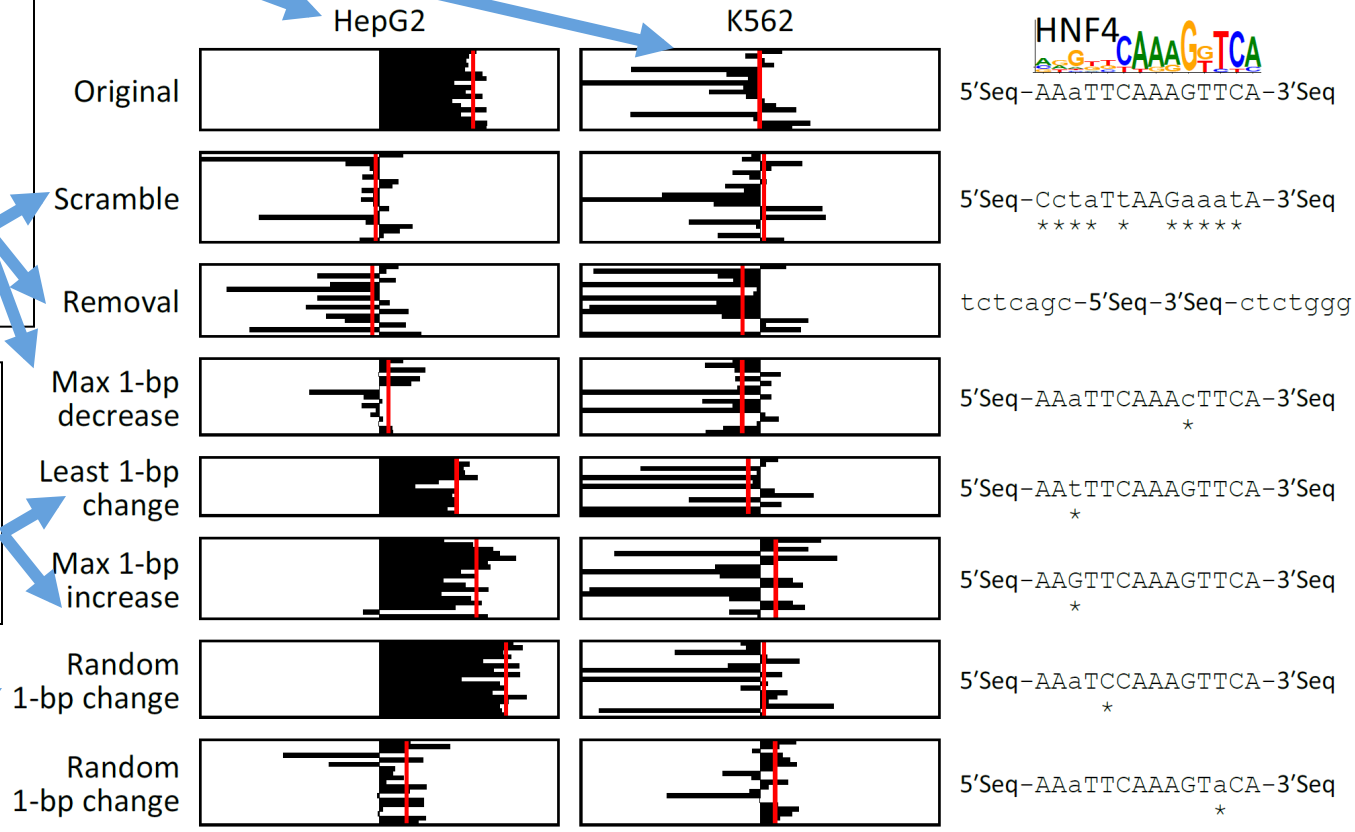
WT expression specific to HepG2

Motif match disruptions reduce expression to background

Non-disruptive changes maintain expression

Random changes depend on effect to motif match





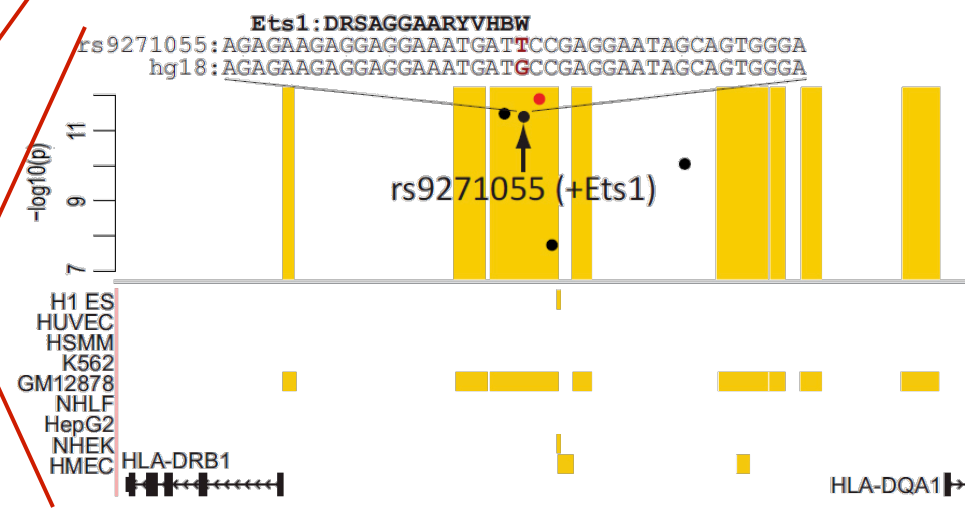
Functional genomics and GTEx for disease

1. **Reference Epigenomes → chromatin states, linking**
   - Annotate dynamic regulatory elements in multiple cell types
   - Activity-based linking of regulators → enhancers → targets
2. **Interpreting disease-associated sequence variants**
   - Mechanistic predictions for individual top-scoring SNPs
   - Functional roles of 1000s of disease-associated SNPs
3. **Multi-cell systems-level expression changes in GTEx**
   - Learn expression programs in 40 tissues for each person
   - Genetic basis of gene module membership changes
4. **Genetic / epigenomic variation in health and disease**
   - Genetic variation⇔Brain methylation⇔Alzheimer's disease
   - Global repression of distal enhancers. NRSF, ELK1, CTCF

Revisiting disease-associated variants

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

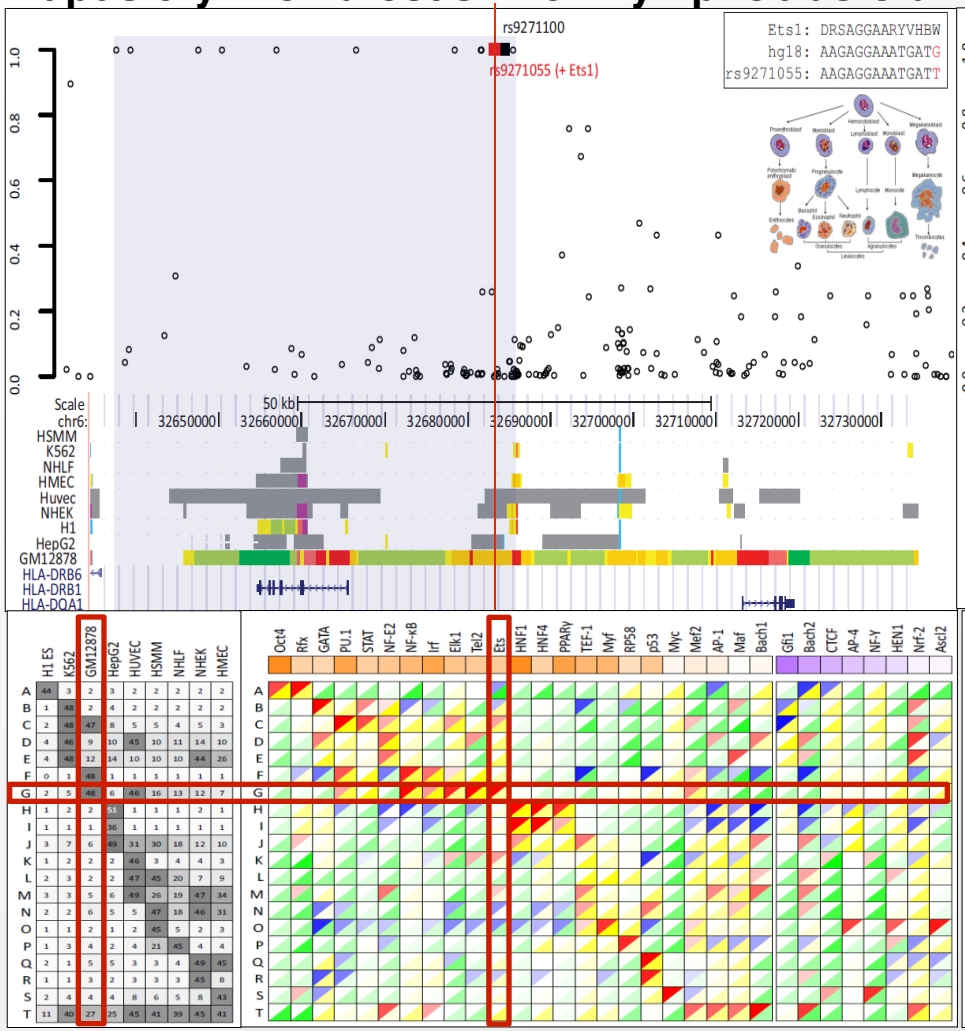
SNP	H1 ES	K562	GM	HepG2	Huvec	HSMM	NHLF	NHEK	HMEC	Chrom. Band	Gene	Link Sc	Distanc
rs13385731										2p22			
rs10036748										5q33			
rs1385374										12q24	MGC16384	-	1
rs2230926										6q23	TNFAIP3	3.7	7
rs4728142										7q32	IRF5	-	4
rs9271100										6p21	HLA-DRB1	4.5	19
rs4917014			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



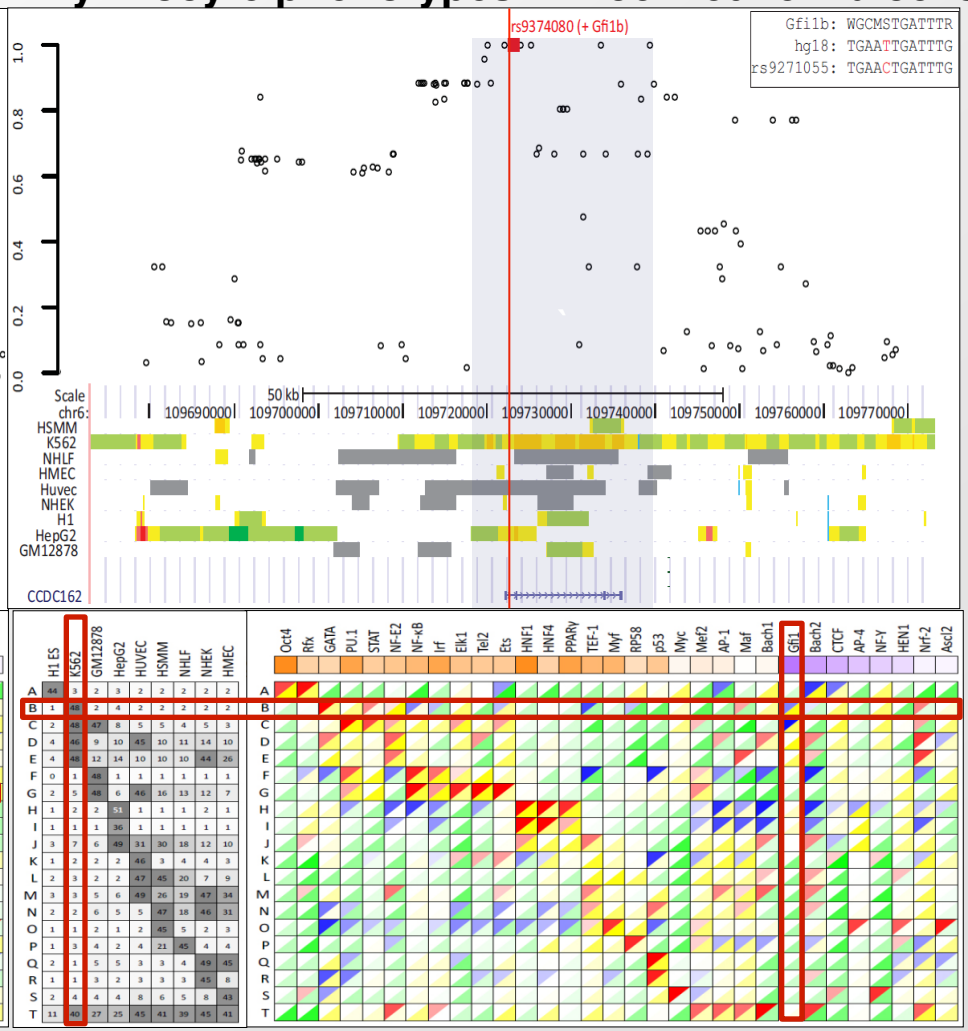
- **Disease-associated SNPs enriched for enhancers in relevant cell types**
- **E.g. lupus SNP in GM enhancer disrupts Ets1 predicted activator**

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

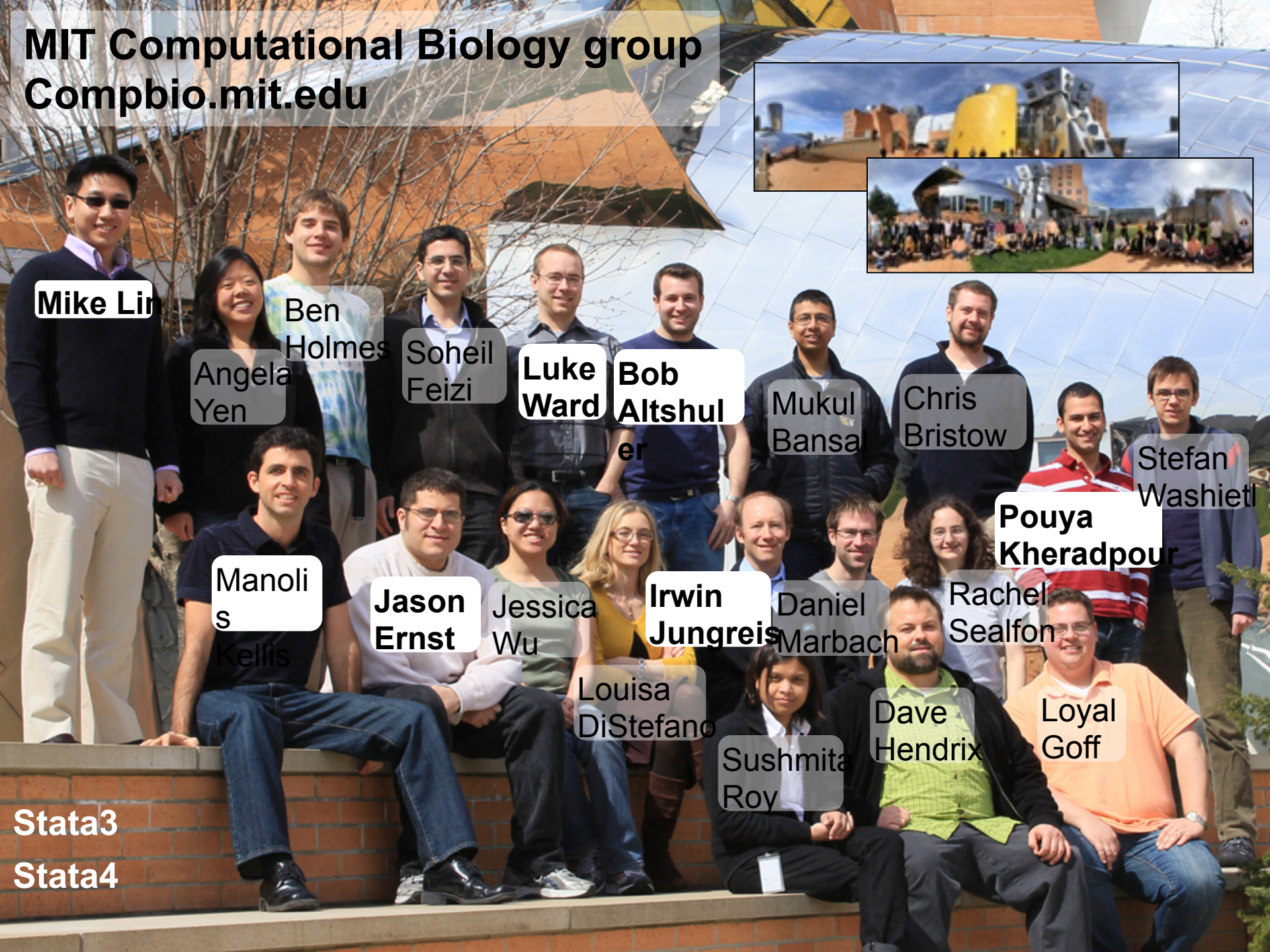
Acknowledgements and Collaborators

1. **Epigenomics Roadmap Project and ENCODE**
   - Brad Bernstein, Bing Ren, John Stam, Joe Costello
   - Anshul Kundaje, Wouter Meuleman, Matt Eaton, Pouya Kheradpour
2. **GWAS interpretation: HaploReg and whole-genome**
   - Luke Ward, Abhishek Sarkar
   - Collaborat.: Vineeta Agarwala, David Altshuler, Martin Aryee
3. **Natural molecular variation: network QTLs**
   - GTEx: Kristin Ardlie, Gaddy Getz, Manolis Dermitzakis
   - Benjamin Iriarte
4. **Molecular variation in disease: AD brain methylation**
   - David Bennett, Phil De Jager, Memory and Aging Project (MAP), and Religious Order Study (ROS) cohorts
   - Matthew Eaton

Collaborators and Acknowledgements

- **ENCODE**
  - Brad Bernstein, Tarjei Mikkelsen, Noam Shoresh, David Epstein
- **Massively parallel enhancer reporter assays**
  - Tarjei Mikkelsen, Broad Institute
- **Epigenome Roadmap**
  - Bing Ren, Brad Bernstein, John Stam, Joe Costello
- **2X mammals**
  - Kerstin Lindblad-Toh, Eric Lander, Manuel Garber, Or Zuk
- **Funding**
  - NHGRI, NIH, NSF
    Sloan Foundation

MIT Computational Biology group
Compbio.mit.edu



Stata3

Stata4