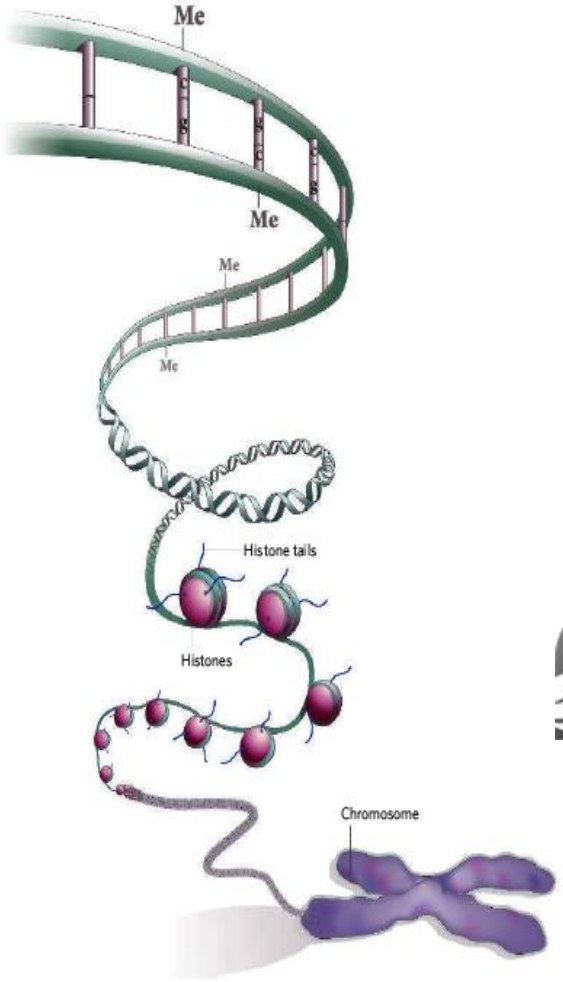
Charting human epigenomes:

Shedding light on the genome's 'dark matter'



Brad Bernstein

MASSACHUSETTS GENERAL HOSPITAL
HHMI HOWARD HUGHES MEDICAL INSTITUTE
HARVARD MEDICAL SCHOOL
BROAD INSTITUTE

ROADMAP epigenomics PROJECT

ENCODE

Charting human epigenomes

1. Epigenomic features and mapping technologies

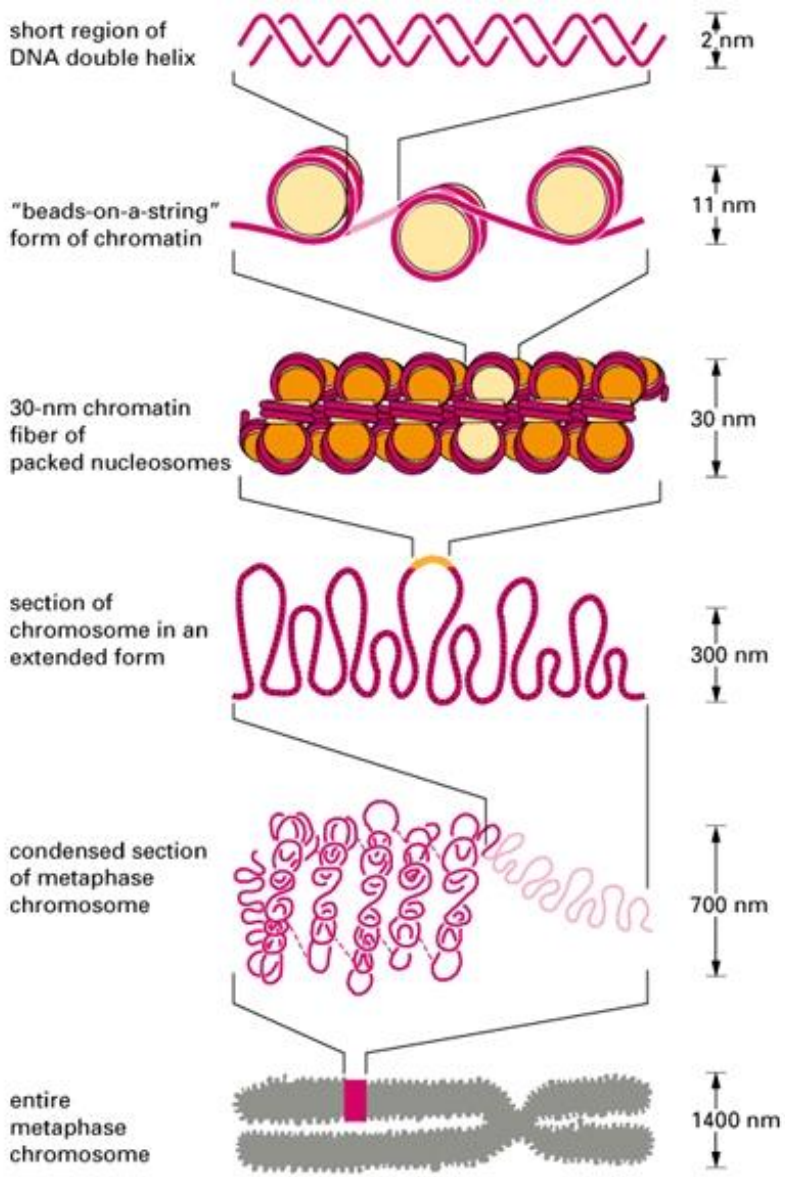
2. NIH Epigenome Mapping Centers

‘Reference epigenomes distinguish developmental stage & lineage’

3. ENCODE Project

‘Beyond the genes: shedding light on the genome’s dark matter’

3 billion base pairs of DNA

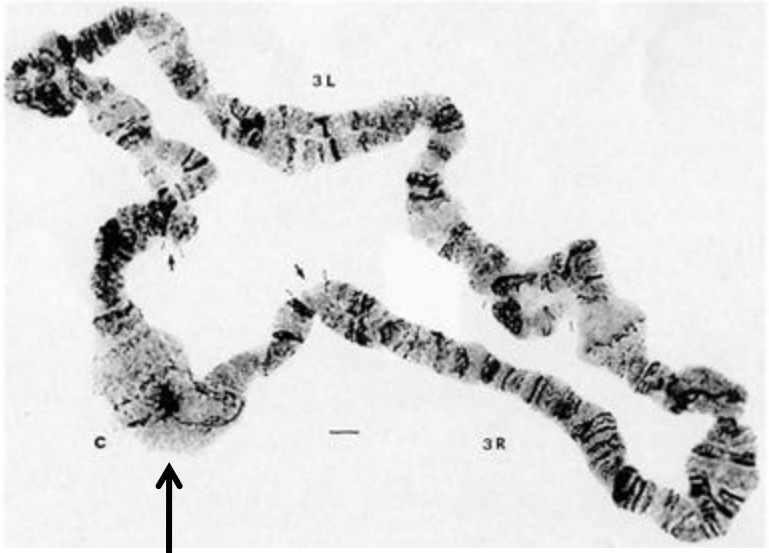


Alberts, Molecular Biology of the Cell

Two meters of DNA in a nucleus smaller than of the head of a pin

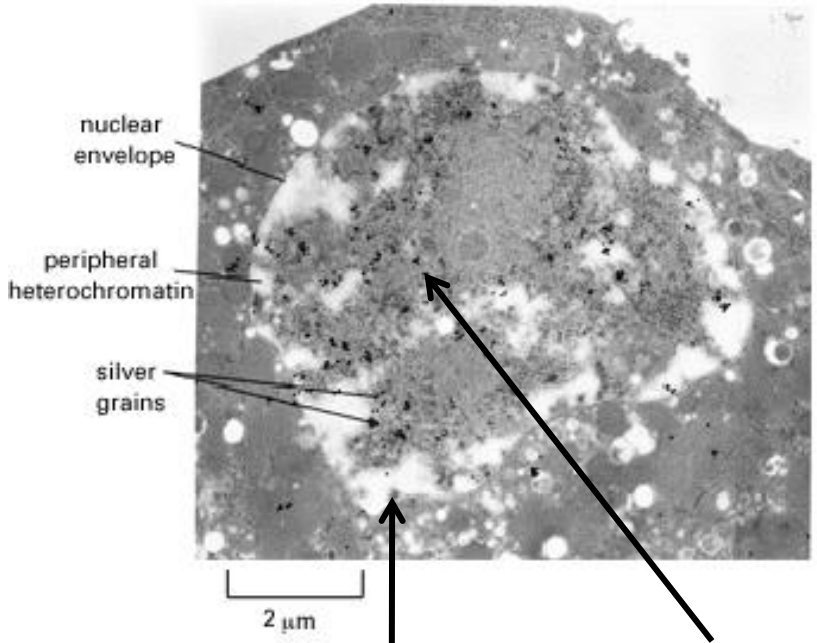
Genes and DNA elements in ‘open’ chromatin

Chromosome (polytene)



Open chromatin, transcribed gene

Nucleus

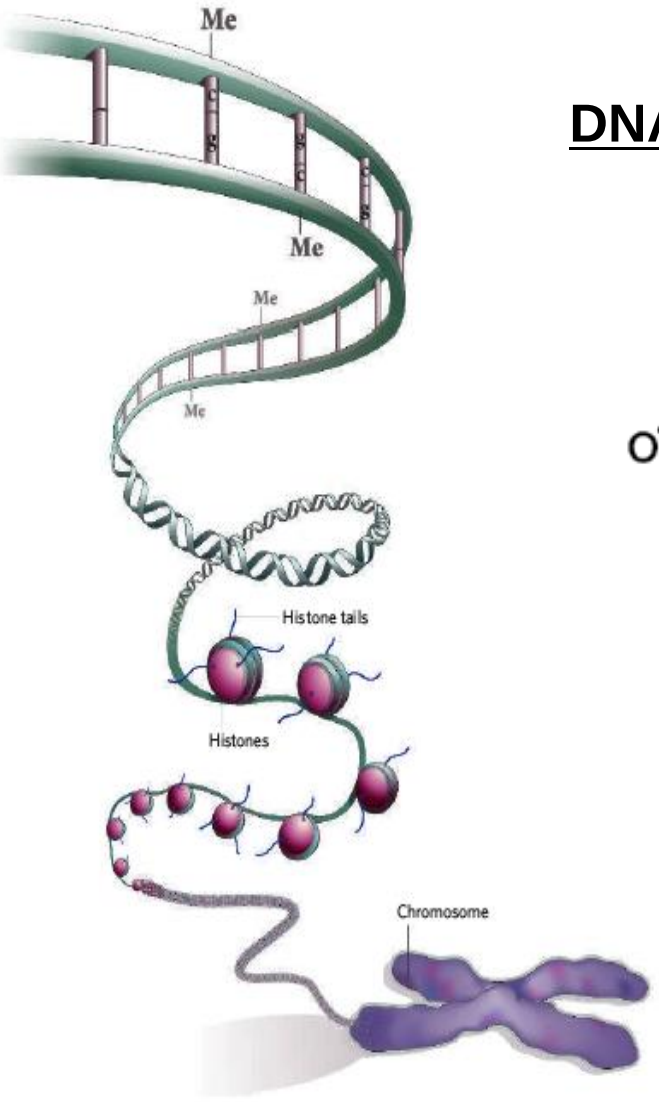


Compact

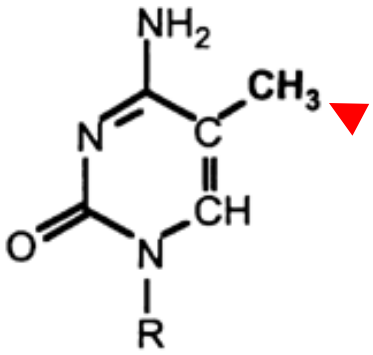
Active/transcribed

Genes and control elements are accessible to RNA polymerase and other regulatory proteins

Chemical ‘tags’ underlie chromatin organization

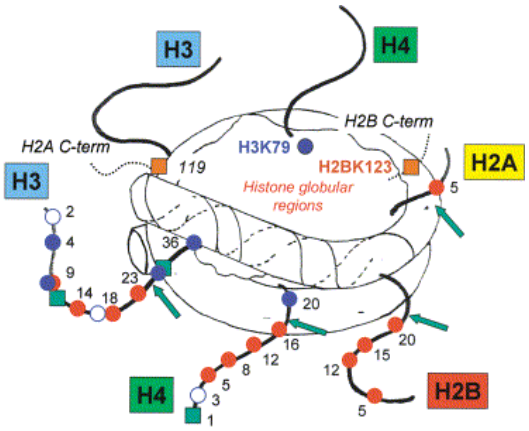


DNA methylation



5-methyl-cytidine

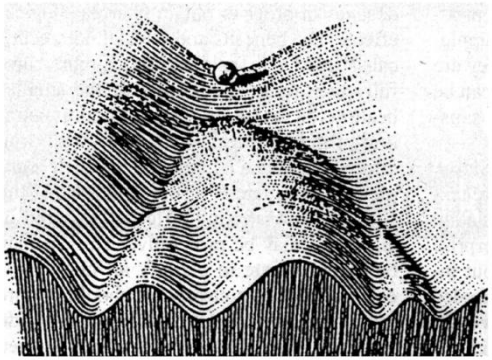
Histone modifications

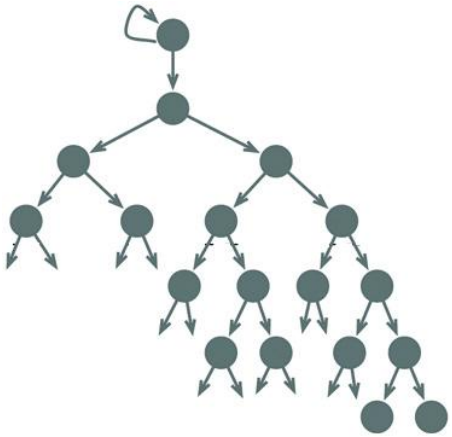


- Lysine acetylation
- Lysine methylation
- Arginine methylation
- Serine phosphorylation
- Lysine ubiquitination

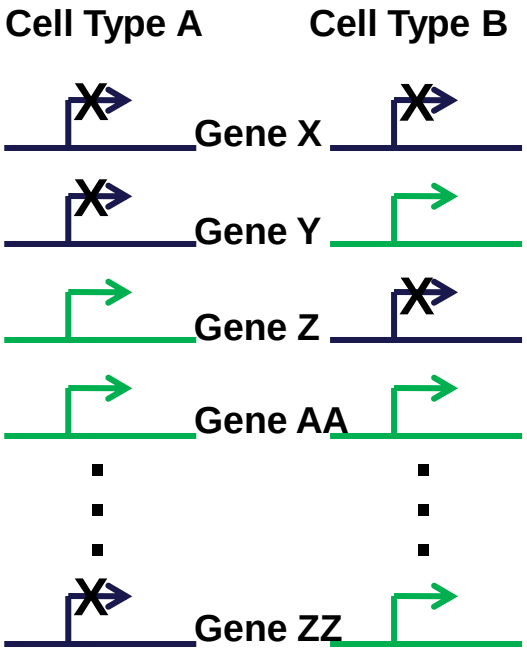
Epigenetic regulation of development

Development & Lineage-Specification

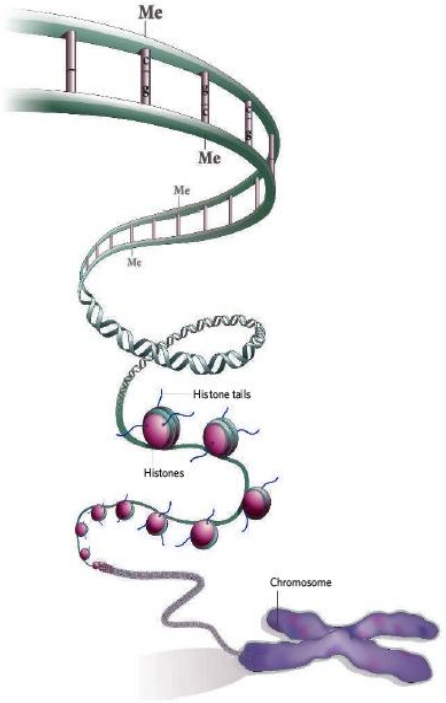




Cell Type-Specific Gene Expression Programs



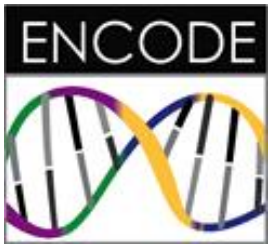
Chromatin Structure & the Epigenome



Epigenomics of human disease

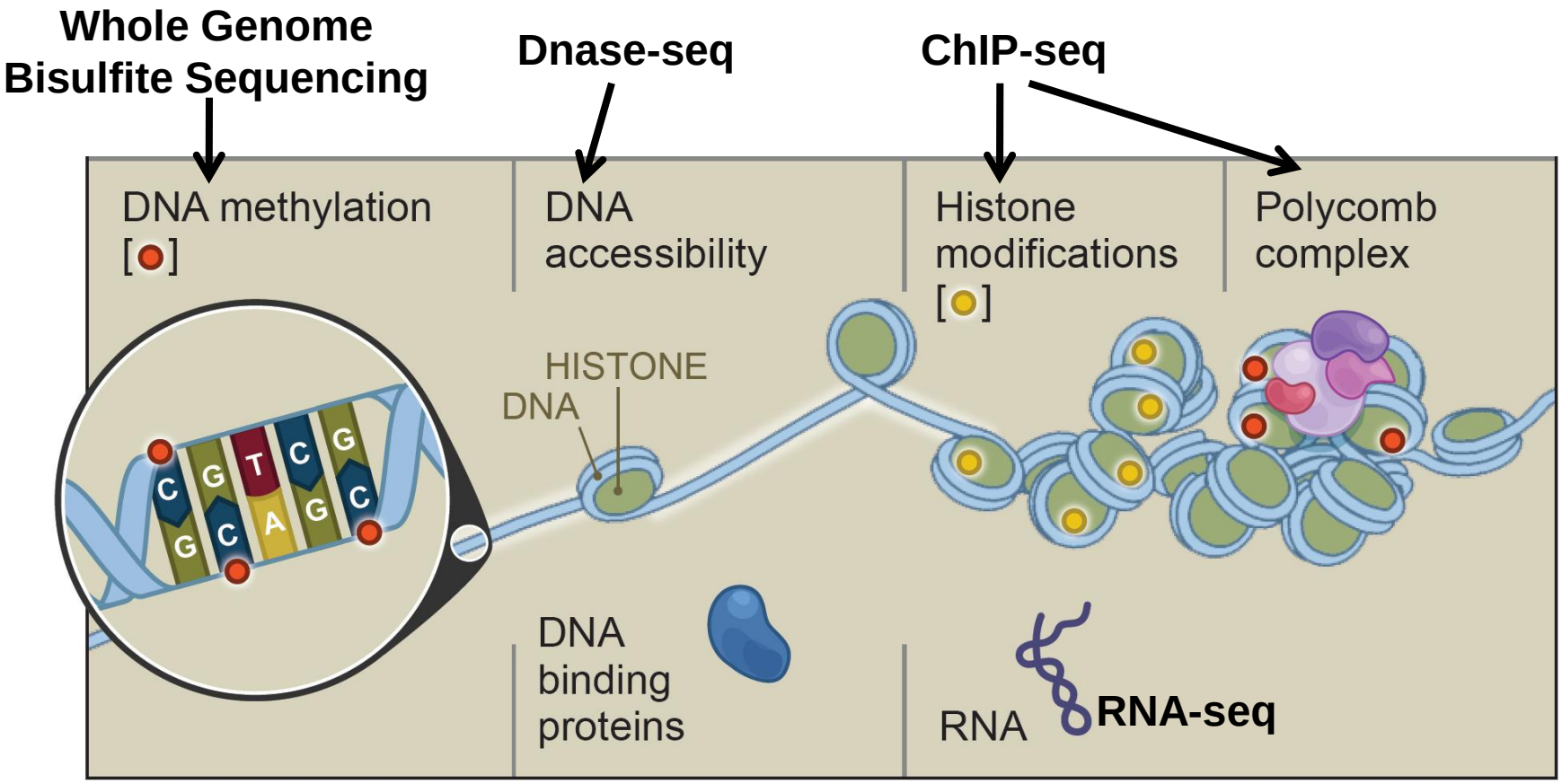
- Cancer is a genetic and epigenetic disease
  - Aberrant DNA methylation is a hallmark
  - Prevalent mutations in wide range of chromatin enzymes
- Neuropsychiatric, metabolic, developmental disorders
  - Mutations in chromatin regulators (MeCP2), aberrant epigenetics
  - Long-term consequences of early environmental exposures
- Functional annotation encoded in epigenome vital to understand how genotype gives rise to phenotype in any tissue or disease

Urgent need for reference human epigenomes and toolkits to enable disease researchers to characterize and understand these defects



Genome-wide maps of epigenomic features

Next-generation sequencing has transformed epigenomics research



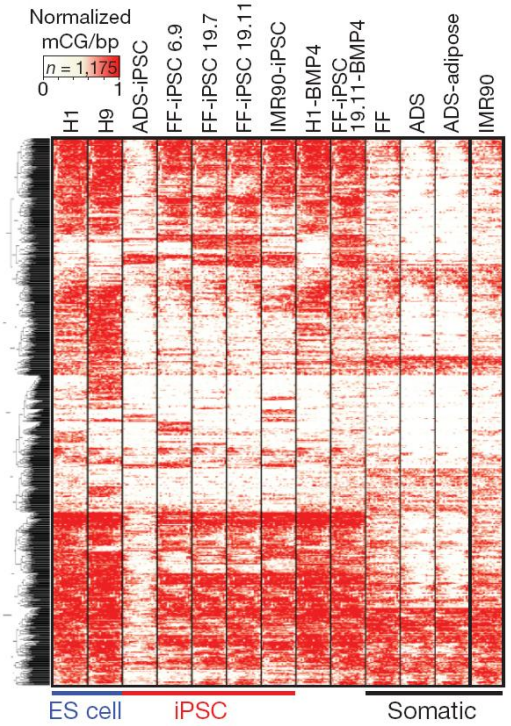
But many challenges: direct mC detection, single cell analysis, tissues, etc

Methylomes of pluripotent stem cells

1. Whole methylomes for ES cells by bisulfite sequencing (Lister *et al*, *Nature* **2009**)

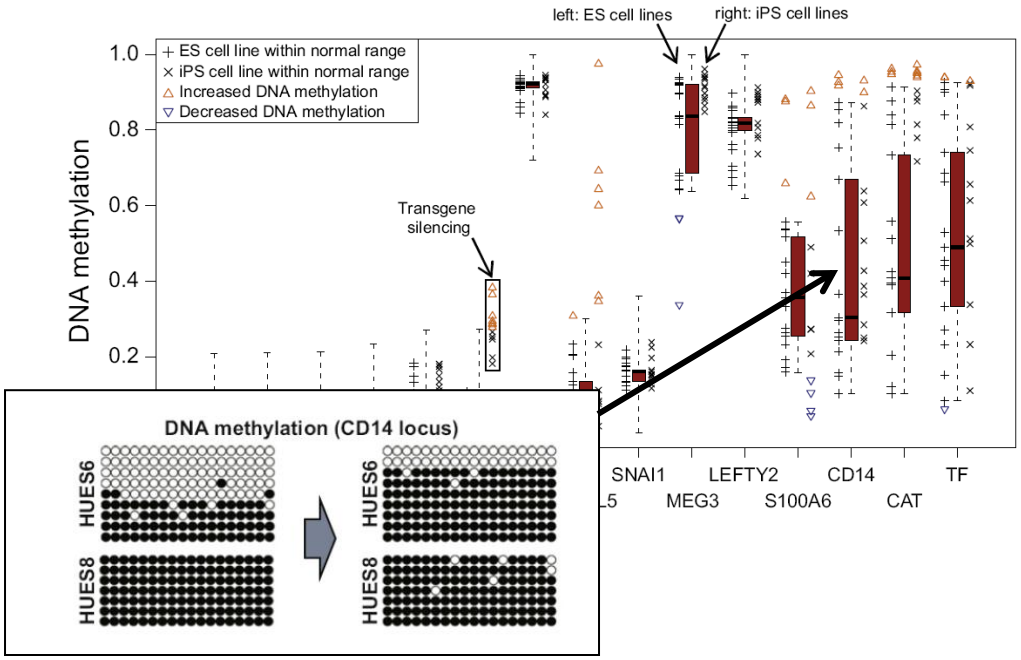
ACT^{m}CGT ⇨ ATTCGT ⇨ Deep sequence (20x genome coverage)

2. Epigenomic defects in iPSCs



Ecker, Thomson & Colleagues, *Nature* **2011**

3. Epigenome -> developmental potential



HUES8 can't make macrophages

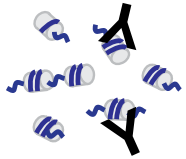
Meissner, Eggan & Colleagues *Cell* **2011**

Mapping histone modifications

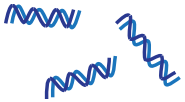
ROADMAP epigenomics PROJECT

ChIP-seq:

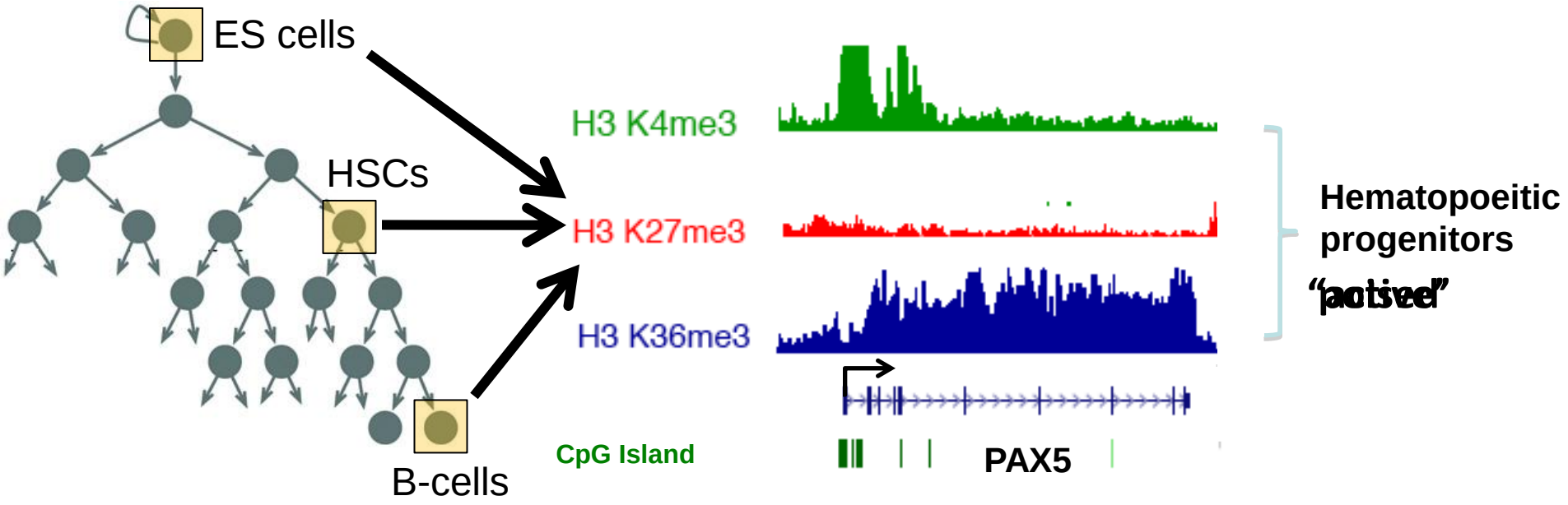
1. Chromatin IP for modified histones



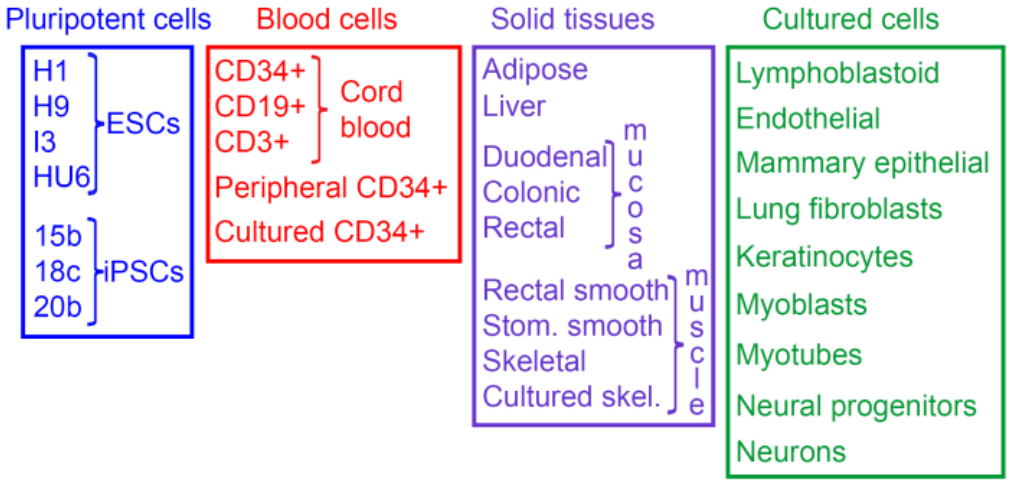
2. Deep sequence enriched DNA



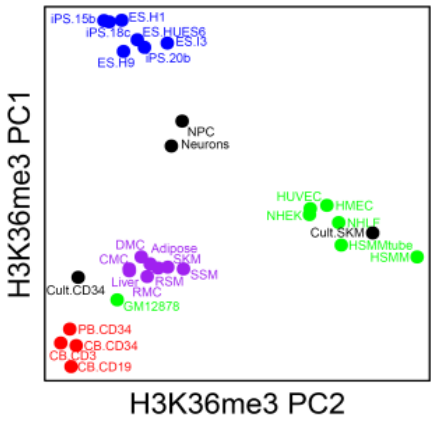
3. Integrate reads into density profiles

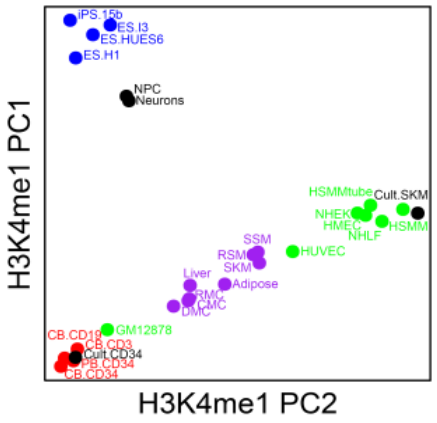


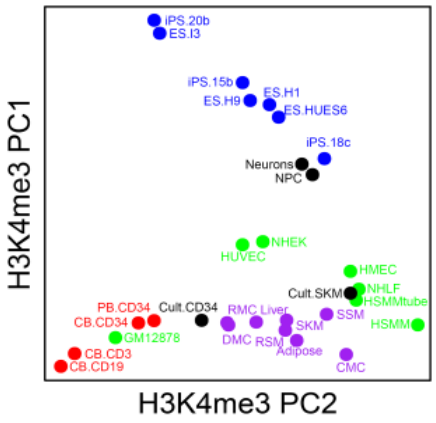
Epigenomes and cell phenotypes

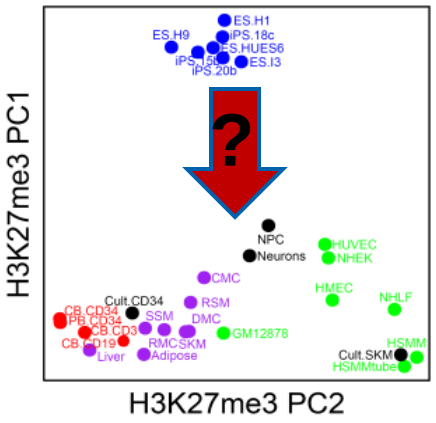


Principle component analysis (PCA)



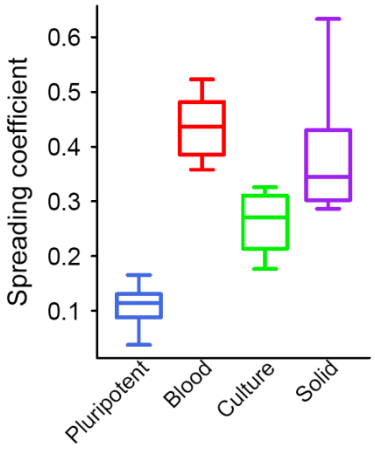


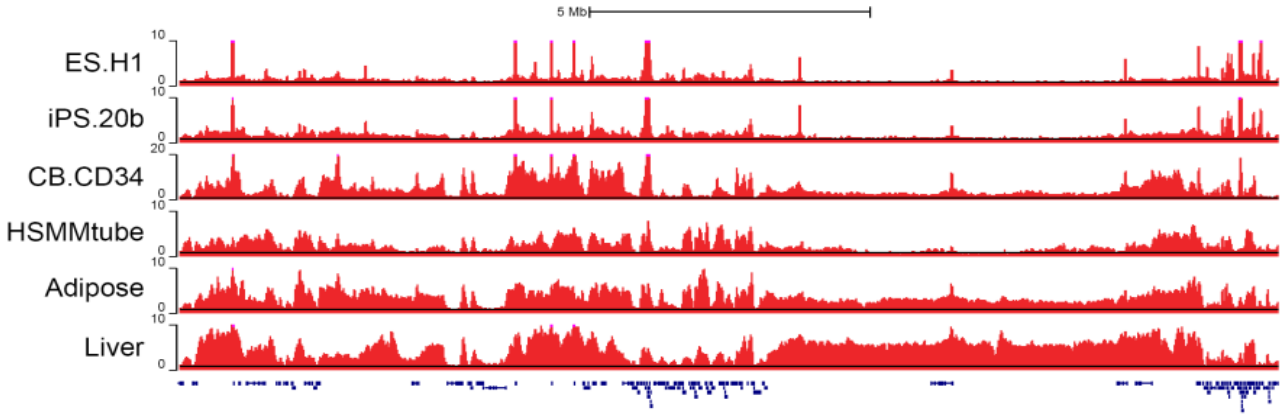


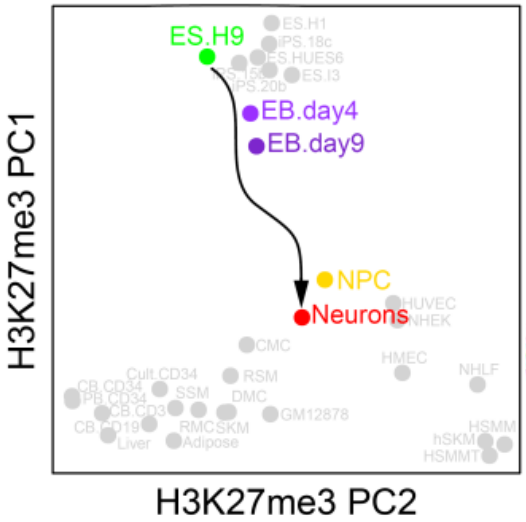


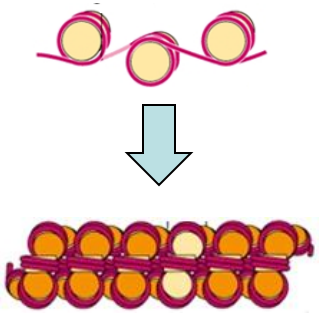
Jiang Zhu, Mazhar Adli, Noam Shoresh, Chuck Epstein, Xiaolan Zhang (Broad/MGH)

Large domains of repressive chromatin formed during lineage-commitment





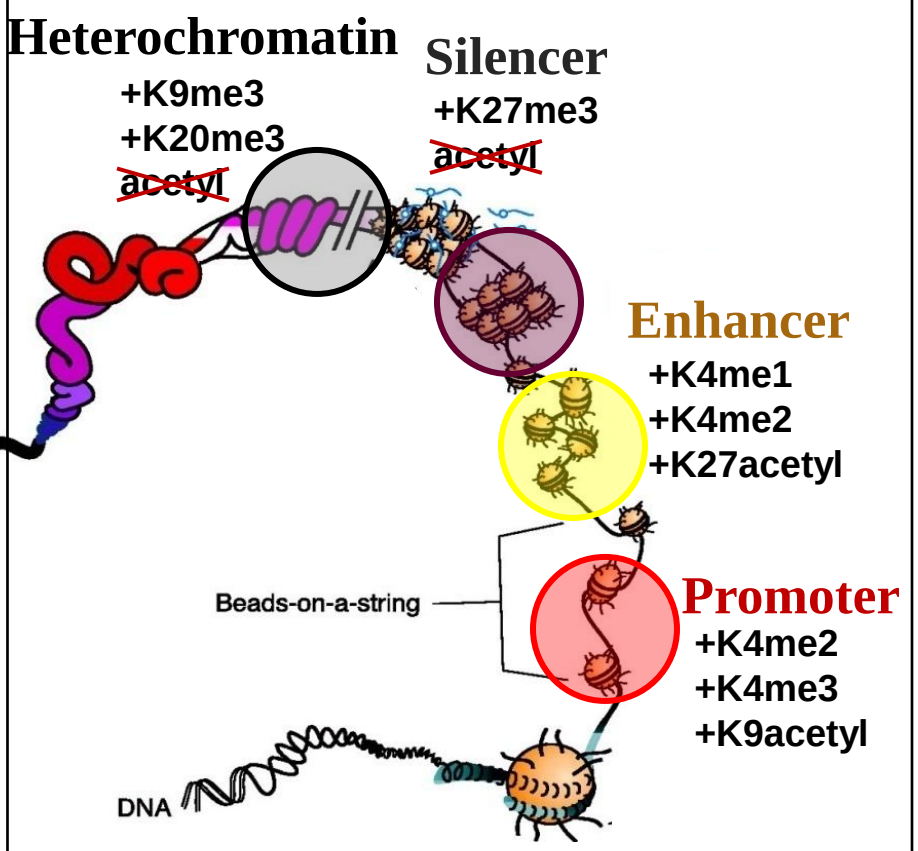




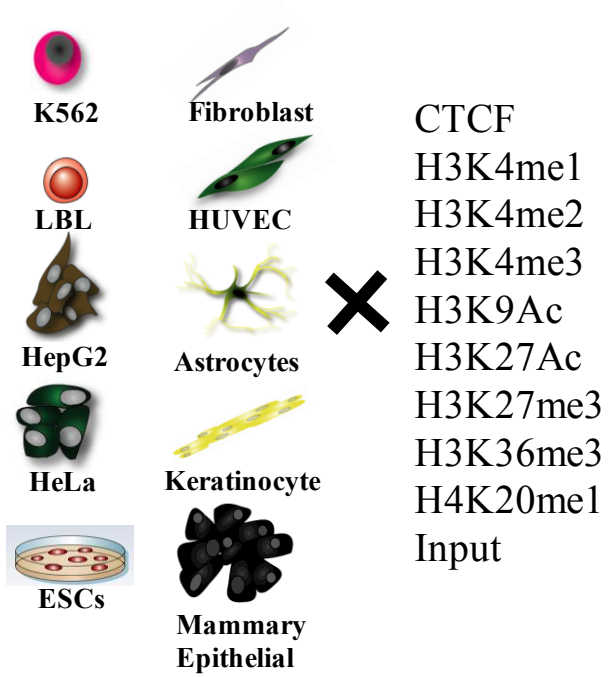
- Epigenome mapping centers highlight prominent role of epigenetic repression in lineage-restriction
- Rich resource for development, regenerative medicine and disease research

Signature histone modification patterns identify non-coding DNA elements

Combinatorial patterns



Compendium of chromatin maps



Jason Ernst, **Manolis Kellis**, Pouya Kheradpour, Tarjei Mikkelsen, Noam Shoresh, Tim Durham
Chuck Epstein, Xiaolan Zhang, Mike Coyne, Li Wang, Robbyn Issner, Luke Ward, Manching Ku

Genome annotation by chromatin state

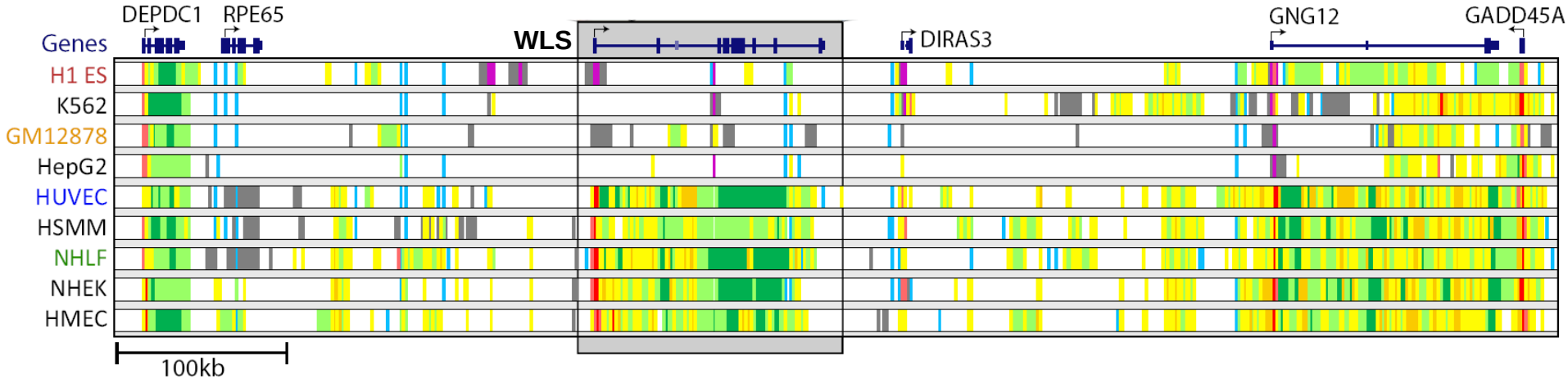
- **Learn recurrent modification patterns or ‘states’ by HMM**
- **Annotate genome by chromatin states -> enhancers, other elements**

Chromatin States: State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Coverage: Median	Coverage: H1 ES	Coverage: GM	Median Length	+/-2kb TSS	Conserved non-exon	DNase (K562)	C-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear Lamina (NHLF)	Candidate state annotation
1	16	2	2	6	17	93	99	96	98	2	0.6	0.5	1.2	1.0	83	3.8	23.3	82.0	40.7	0.2	0.15	Active Promoter
2	12	2	6	9	53	94	95	14	44	1	0.5	1.2	1.3	0.4	58	2.8	15.3	12.6	5.8	0.6	0.30	Weak Promoter
3	13	72	0	9	48	78	49	1	10	1	0.2	4.0	1.0	0.6	49	4.3	10.8	3.1	1.0	0.4	0.68	Inactive/poised Promoter
4	11	1	15	11	96	99	75	97	86	4	0.7	0.1	1.1	0.6	23	2.7	23.1	31.8	49.0	1.3	0.05	Strong enhancer
5	5	0	10	3	88	57	5	84	25	1	1.2	0.2	0.7	0.6	3	1.8	13.6	6.3	15.8	1.4	0.10	Strong enhancer
6	7	1	1	3	58	75	8	6	5	1	0.9	1.3	1.0	0.2	17	2.4	11.9	5.7	7.0	1.1	0.31	Weak/poised enhancer
7	2	1	2	1	56	3	0	6	2	1	1.9	1.2	1.1	0.4	4	1.5	5.1	0.6	2.4	1.3	0.20	Weak/poised enhancer
8	92	2	1	3	6	3	0	0	1	1	0.5	1.4	1.0	0.4	3	1.5	12.8	2.5	1.2	1.1	0.61	Insulator
9	5	0	43	43	37	11	2	9	4	1	0.7	1.3	1.0	0.8	4	1.1	4.5	0.7	0.8	2.4	0.02	Transcriptional transition
10	1	0	47	3	0	0	0	0	0	1	4.3	0.6	1.2	3.0	1	0.9	0.3	0.0	0.0	2.5	0.11	Transcriptional elongation
11	0	0	3	2	0	0	0	0	0	0	12.5	1.3	0.8	2.6	2	0.9	0.3	0.0	0.1	1.9	0.24	Weak transcribed
12	1	27	0	2	0	0	0	0	0	0	4.1	0.3	0.7	2.8	5	1.4	0.3	0.0	0.1	0.8	0.63	Polycomb-repressed
13	0	0	0	0	0	0	0	0	0	0	71.4	1.0	1.0	10.0	1	0.9	0.1	0.0	0.0	0.7	1.30	Heterochrom; low signal
14	22	28	19	41	6	5	26	5	13	37	0.1	0.9	1.2	0.6	3	0.4	1.9	0.3	0.2	0.4	1.44	Repetitive/CNV
15	85	85	91	88	76	77	91	73	85	78	0.1	0.9	1.0	0.2	1	0.2	5.9	9.5	7.4	0.4	1.30	Repetitive/CNV

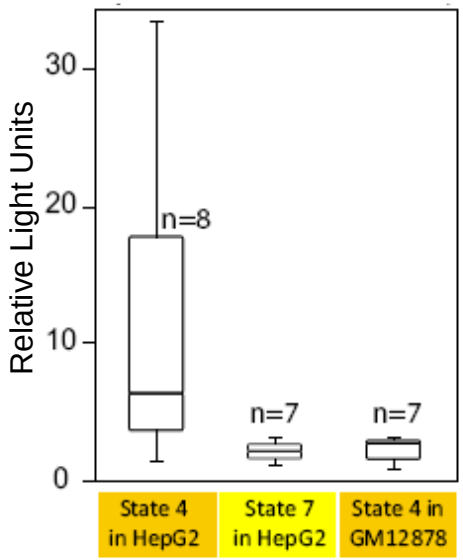
~**180** ChIP-seq exps
~**2.4 billion** reads
~**100 billion** bases

Single genome-wide annotation for each cell type

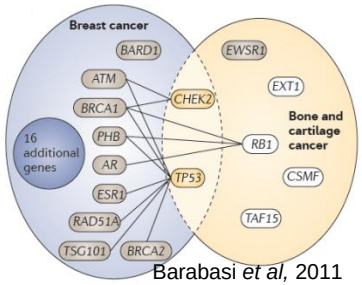
Regulatory elements are cell type-specific



Cell type-specific enhancers validated



Genome regulatory network

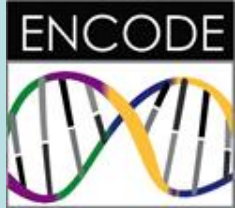


Barabasi *et al*, 2011

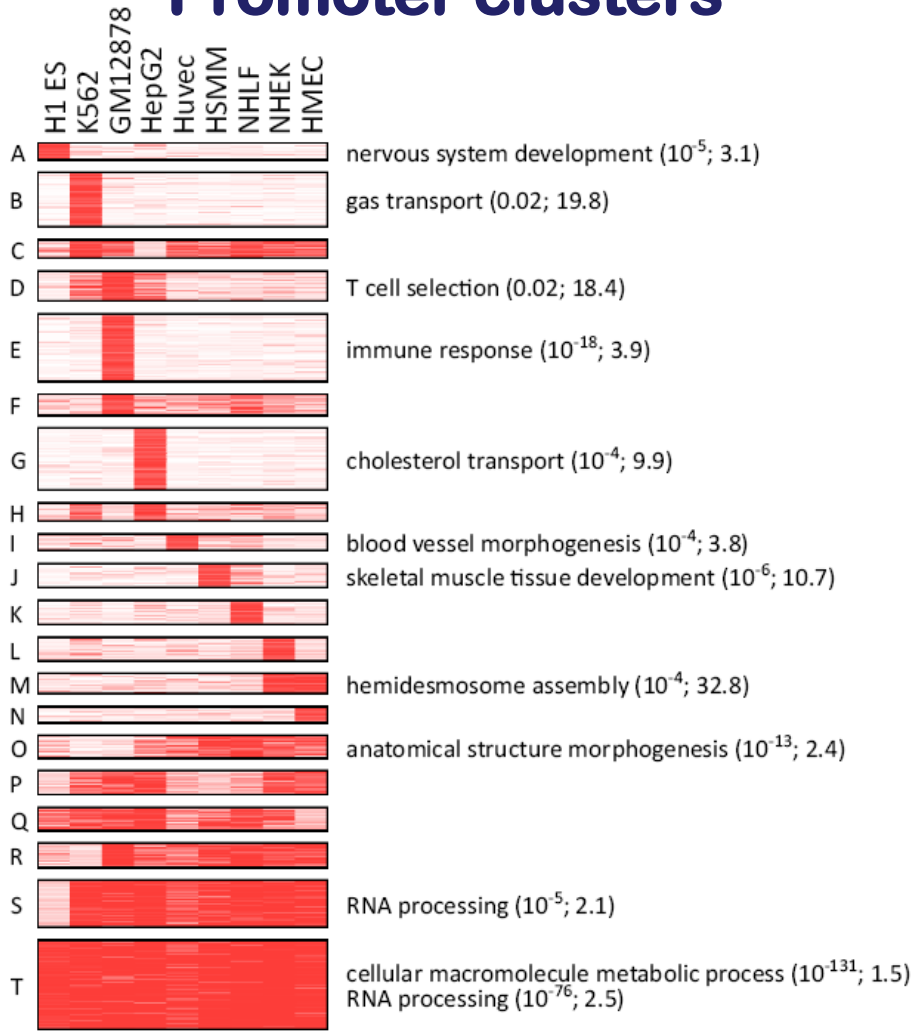
- Gene interactions inferred from co-variation
- Ignored non-coding regulatory elements

Can we incorporate enhancers into a regulatory network by using variations in chromatin state?

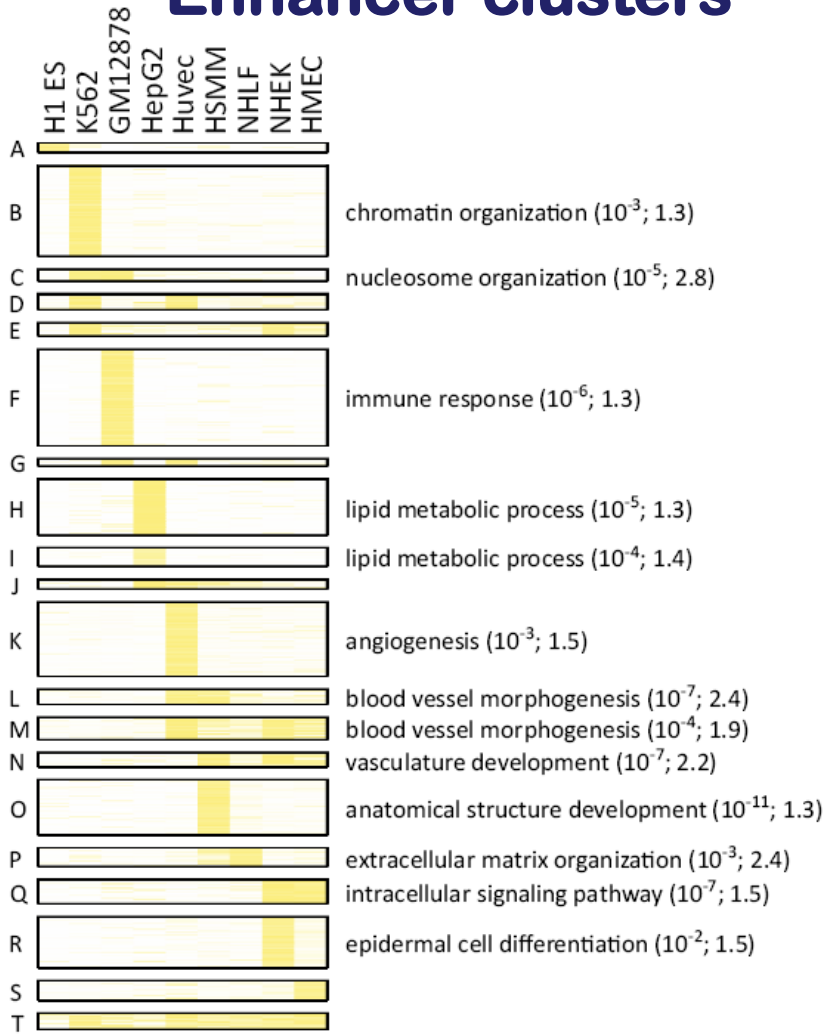
Regulatory elements can be grouped based on their cell type-specific activity



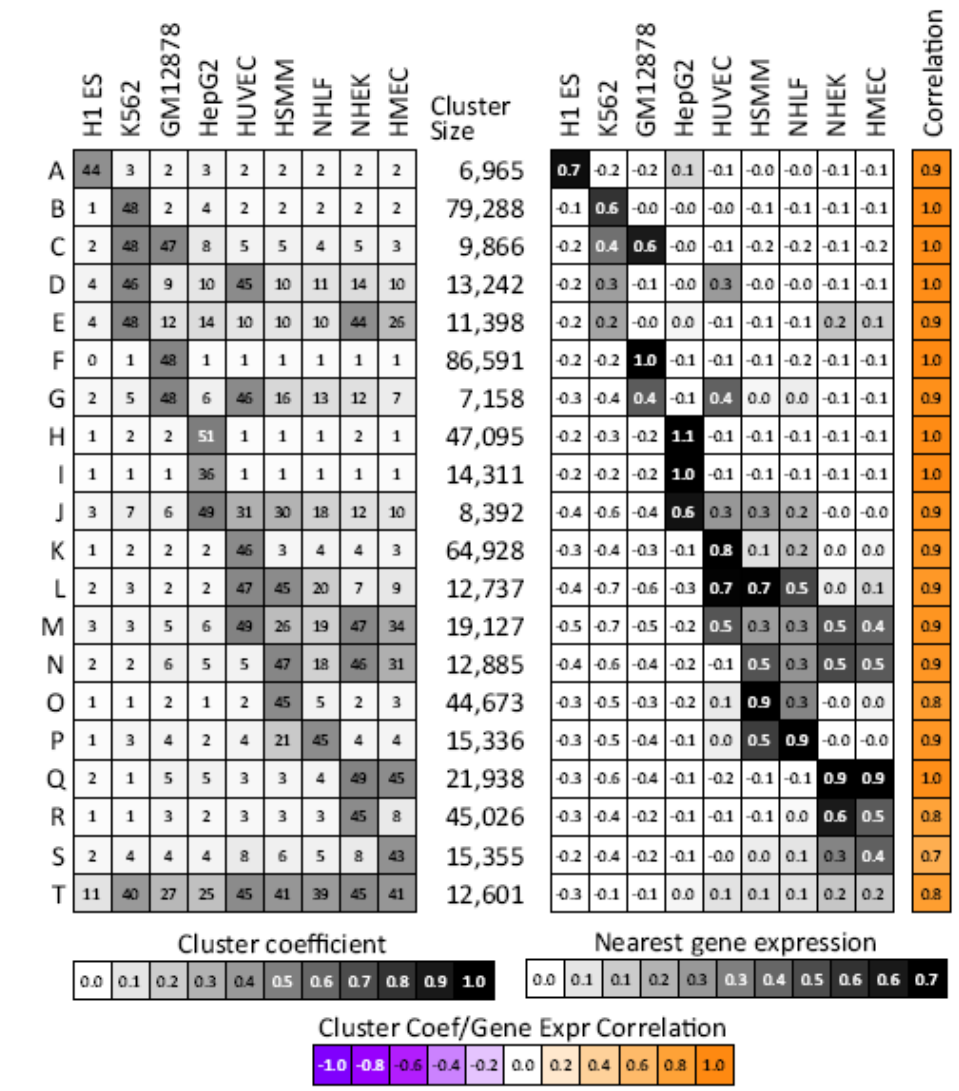
Promoter clusters

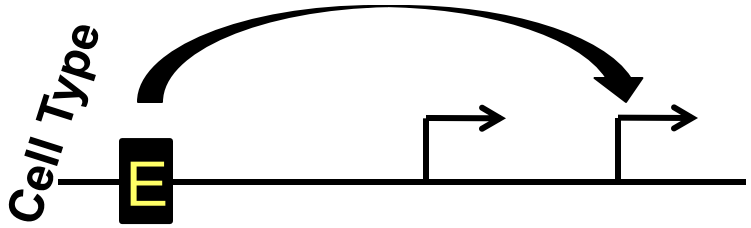


Enhancer clusters



Enhancers can be linked to target genes based on correlated activity profiles





A	ON	OFF	ON
B	ON	ON	ON
C	OFF	ON	OFF
.			
.			
.			
X	OFF	OFF	OFF

Enhancer-gene linkages supported by QTLs

Enhancer clusters **Proximal genes**

	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC	Cluster Size	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC	Correlation
A	44	3	2	3	2	2	2	2	2	6,965	0.7	-0.2	-0.2	0.1	-0.1	-0.0	-0.0	-0.1	-0.1	0.9
B	1	48	2	4	2	2	2	2	2	79,288	-0.1	0.6	-0.0	-0.0	-0.0	-0.1	-0.1	-0.1	-0.1	1.0
C	2	48	47	8	5	5	4	5	3	9,866	-0.2	0.4	0.6	-0.0	-0.1	-0.2	-0.2	-0.1	-0.2	1.0
D	4	46	9	10	45	10	11	14	10	13,242	-0.2	0.3	-0.1	-0.0	0.3	-0.0	-0.0	-0.1	-0.1	1.0
E	4	48	12	14	10	10	10	44	26	11,398	-0.2	0.2	-0.0	0.0	-0.1	-0.1	-0.1	0.2	0.1	0.9
F	0	1	48	1	1	1	1	1	1	86,591	-0.2	-0.2	1.0	-0.1	-0.1	-0.1	-0.2	-0.1	-0.1	1.0
G	2	5	48	6	46	16	13	12	7	7,158	-0.3	-0.4	0.4	-0.1	0.4	0.0	0.0	-0.1	-0.1	0.9
H	1	2	2	51	1	1	1	2	1	47,095	-0.2	-0.3	-0.2	1.1	-0.1	-0.1	-0.1	-0.1	-0.1	1.0
I	1	1	1	36	1	1	1	1	1	14,311	-0.2	-0.2	-0.2	1.0	-0.1	-0.1	-0.1	-0.1	-0.1	1.0
J	3	7	6	49	31	30	18	12	10	8,392	-0.4	-0.6	-0.4	0.6	0.3	0.3	0.2	-0.0	-0.0	0.9
K	1	2	2	2	46	3	4	4	3	64,928	-0.3	-0.4	-0.3	-0.1	0.8	0.1	0.2	0.0	0.0	0.9
L	2	3	2	2	47	45	20	7	9	12,737	-0.4	-0.7	-0.6	-0.3	0.7	0.7	0.5	0.0	0.1	0.9
M	3	3	5	6	49	26	19	47	34	19,127	-0.5	-0.7	-0.5	-0.2	0.5	0.3	0.3	0.5	0.4	0.9
N	2	2	6	5	5	47	18	46	31	12,885	-0.4	-0.6	-0.4	-0.2	-0.1	0.5	0.3	0.5	0.5	0.9
O	1	1	2	1	2	45	5	2	3	44,673	-0.3	-0.5	-0.3	-0.2	0.1	0.9	0.3	-0.0	0.0	0.8
P	1	3	4	2	4	21	45	4	4	15,336	-0.3	-0.5	-0.4	-0.1	0.0	0.5	0.9	-0.0	-0.0	0.9
Q	2	1	5	5	3	3	4	49	45	21,938	-0.3	-0.6	-0.4	-0.1	-0.2	-0.1	-0.1	0.9	0.9	1.0
R	1	1	3	2	3	3	3	45	8	45,026	-0.3	-0.4	-0.2	-0.1	-0.1	-0.1	0.0	0.6	0.5	0.8
S	2	4	4	4	8	6	5	8	43	15,355	-0.2	-0.4	-0.2	-0.1	-0.0	0.0	0.1	0.3	0.4	0.7
T	11	40	27	25	45	41	39	45	41	12,601	-0.3	-0.1	-0.1	0.0	0.1	0.1	0.1	0.2	0.2	0.8

Cluster coefficient

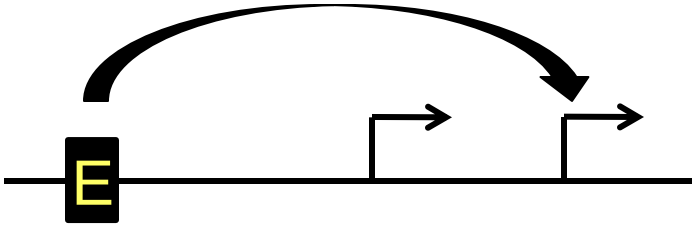
0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Nearest gene expression

0.0	0.1	0.1	0.2	0.3	0.3	0.4	0.5	0.6	0.6	0.7

Cluster Coef/Gene Expr Correlation

-1.0	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	1.0



Chromatin dynamics ~ Genetic variation

Schadt Liver - HepG2

Window Expansion	Fold (distance controlled)	p-val	#eQTL Pairs Linked	#Link Eligible	Expected (random)	Expected by Distance	**Observed Fraction Linked**
0	1.9	0.03	11	49	0.06	0.12	**0.22**
600	1.7	0.02	16	73	0.06	0.13	**0.22**
1000	1.7	0.01	21	98	0.06	0.13	**0.21**

Pickrell LCL - GM12878

Window Expansion	Fold (distance controlled)	p-val	#eQTL Pairs Linked	#Link Eligible	Expected (random)	Expected by Distance	**Observed Fraction Linked**
0	2.3	1.4E-02	9	32	0.07	0.12	**0.28**
600	2.3	4.0E-03	12	43	0.07	0.12	**0.28**
1000	1.9	2.0E-02	12	51	0.07	0.12	**0.24**

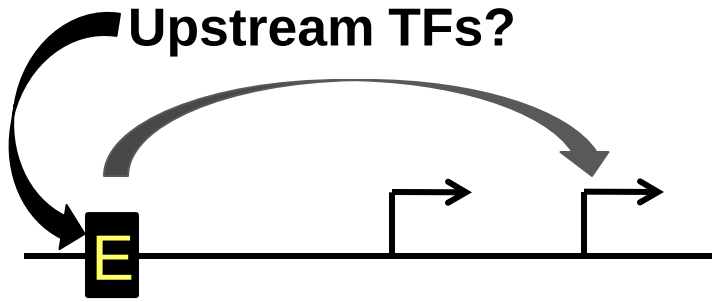
Predicting upstream regulators of enhancers

Enhancer clusters

	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC	Cluster Size
A	44	3	2	3	2	2	2	2	2	6,965
B	1	48	2	4	2	2	2	2	2	79,288
C	2	48	47	8	5	5	4	5	3	9,866
D	4	46	9	10	45	10	11	14	10	13,242
E	4	48	12	14	10	10	10	44	26	11,398
F	0	1	48	1	1	1	1	1	1	86,591
G	2	5	48	6	46	16	13	12	7	7,158
H	1	2	2	51	1	1	1	2	1	47,095
I	1	1	1	36	1	1	1	1	1	14,311
J	3	7	6	49	31	30	18	12	10	8,392
K	1	2	2	2	46	3	4	4	3	64,928
L	2	3	2	2	47	45	20	7	9	12,737
M	3	3	5	6	49	26	19	47	34	19,127
N	2	2	6	5	5	47	18	46	31	12,885
O	1	1	2	1	2	45	5	2	3	44,673
P	1	3	4	2	4	21	45	4	4	15,336
Q	2	1	5	5	3	3	4	49	45	21,938
R	1	1	3	2	3	3	3	45	8	45,026
S	2	4	4	4	8	6	5	8	43	15,355
T	11	40	27	25	45	41	39	45	41	12,601

Cluster coefficient

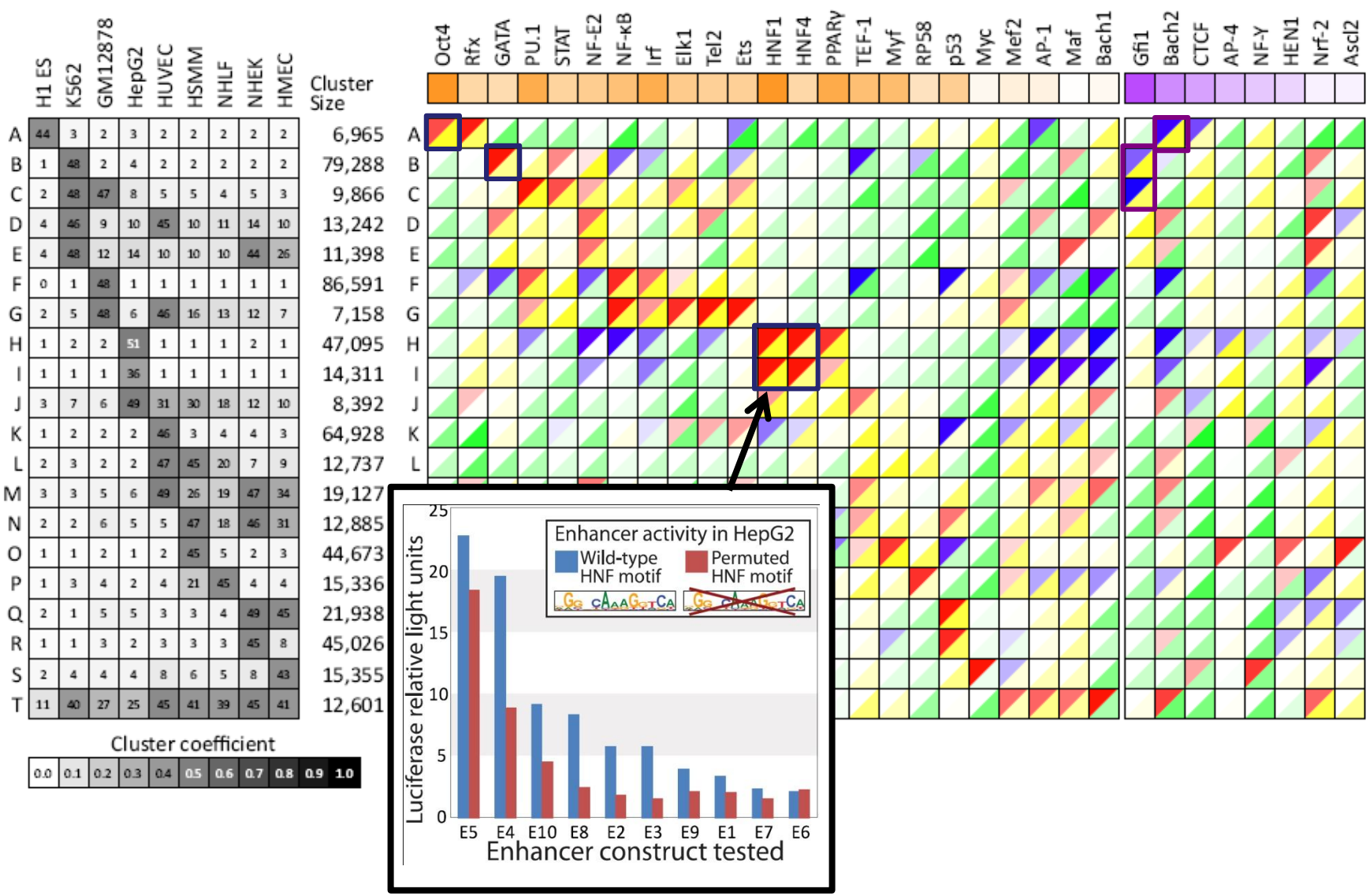
0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0



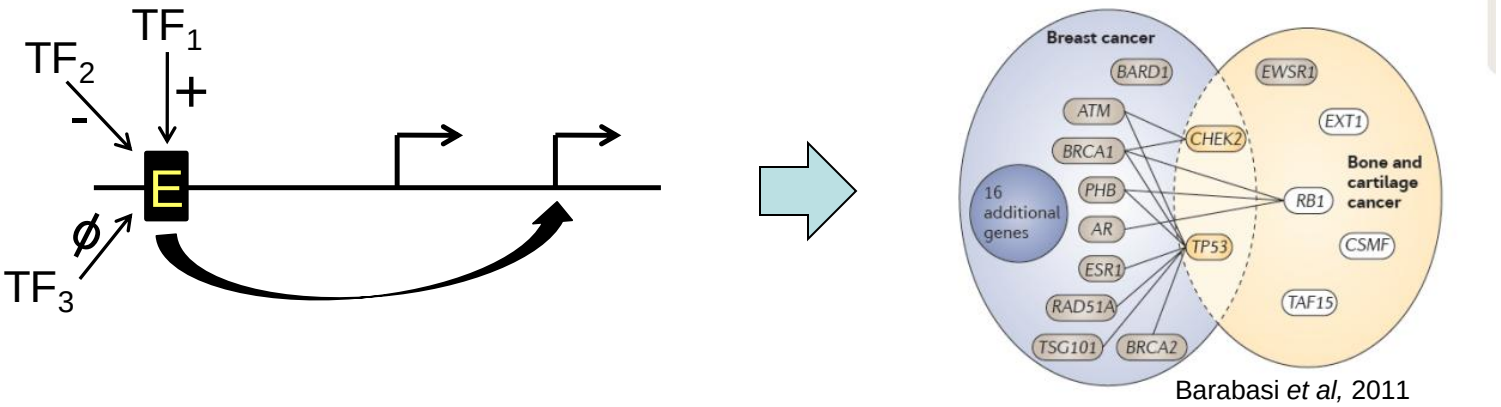
Approach:

1. Identify enriched motifs
2. Evaluate expression of cognate TFs
3. Identify signatures for ‘activators’ and ‘repressors’ of enhancer clusters

Predicting upstream regulators of enhancers



Incorporating distal enhancers into a genome regulatory network

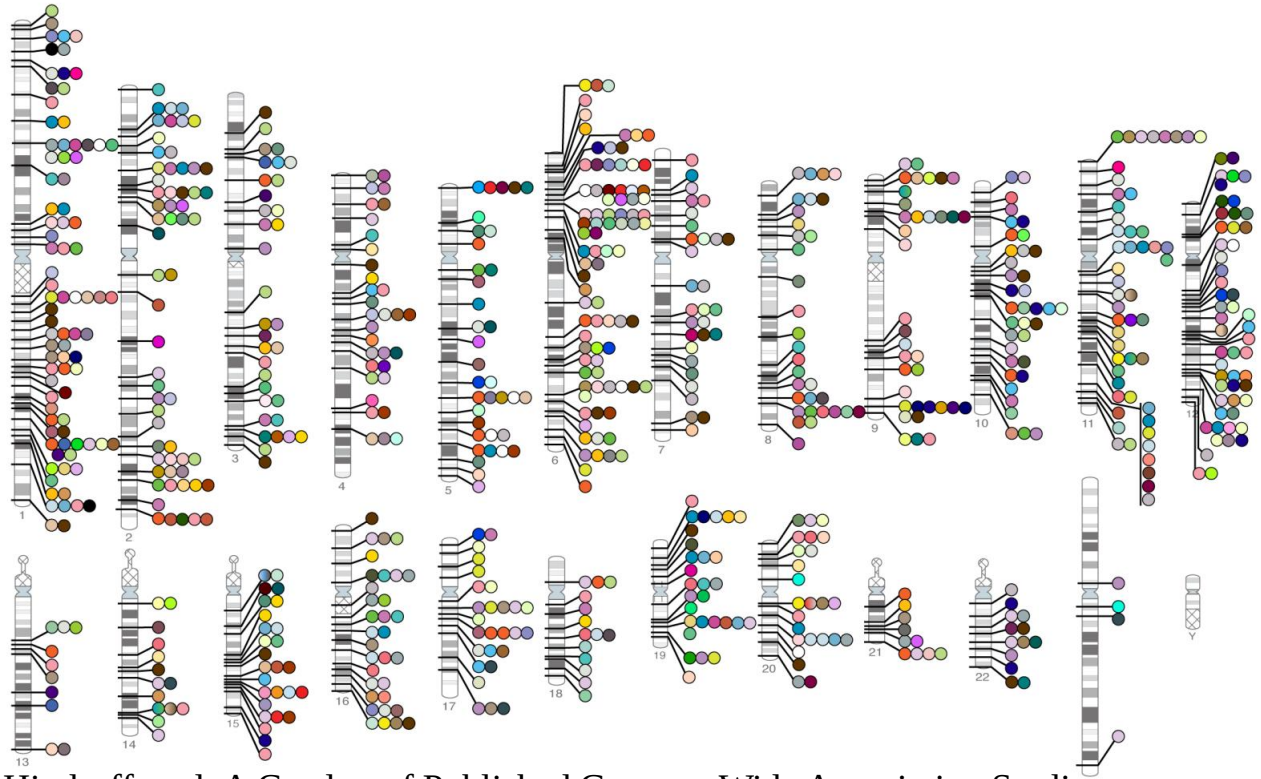


Barabasi *et al,* 2011

- >100,000 predicted enhancers
- ~10% linked to candidate target genes
- ~25 cell type-specific TFs as upstream regulators

Disease SNPs enriched in enhancer states

GWAS: Most disease-associated variants are non-coding



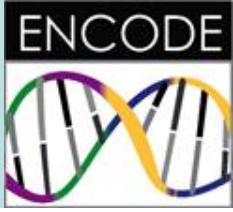
- Acute lymphoblastic leukemia
- Adhesion molecules
- Adiponectin levels
- Age-related macular degeneration
- AIDS progression
- Alcohol dependence
- Alzheimer disease
- Amyotrophic lateral sclerosis
- Angiotensin-converting enzyme activity
- Ankylosing spondylitis
- Arterial stiffness
- Asthma
- Atherosclerosis in HIV
- Atrial fibrillation
- Attention deficit hyperactivity disorder
- Autism
- Basal cell cancer
- Bipolar disorder
- Bilirubin
- Bladder cancer
- Blond or brown hair
- Blood pressure
- Blue or green eyes
- BMI, waist circumference
- Bone density
- Breast cancer
- C-reactive protein
- Cardiac structure/function
- Carnitine levels
- Carotenoid/tocopherol levels
- Celiac disease
- Chronic lymphocytic leukemia
- Cleft lip/palate
- Cognitive function
- Colorectal cancer
- Coronary disease
- Creutzfeldt-Jakob disease
- Crohn's disease
- Cutaneous nevi
- Dermatitis
- Drug-induced liver injury
- Eosinophil count
- Eosinophilic esophagitis
- Erythrocyte parameters
- Esophageal cancer
- Essential tremor
- Exfoliation glaucoma
- F cell distribution
- Fibrinogen levels
- Folate pathway vitamins
- Freckles and burning
- Gallstones
- Glioma
- Glycemic traits
- Hair color
- Hair morphology
- HDL cholesterol
- Heart rate
- Height
- Hemostasis parameters
- Hepatitis
- Hirschsprung's disease
- HIV-1 control
- Homocysteine levels
- Idiopathic pulmonary fibrosis
- IgE levels
- Inflammatory bowel disease
- Intracranial aneurysm
- Iris color
- Iron status markers
- Ischemic stroke
- Juvenile idiopathic arthritis
- Kidney stones
- LDL cholesterol
- Leprosy
- Leptin receptor levels

Hindorff et al: A Catalog of Published Genome-Wide Association Studies.

Disease SNPs enriched >2-fold in enhancer chromatin states

(control SNPs do not show any preference)

Disease SNPs correlate with enhancers that are active in related cell types



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

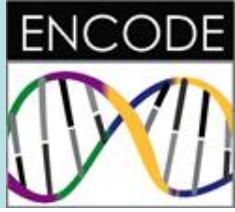
Disease/phenotype **Enhancer specificity**

Erythrocyte phenotypes ⟷ Erythrocytic cells

Lipids, cholesterol ⟷ Hepatic cells

Lupus, arthritis ⟷ Lymphoid cells

GWAS datasets intersected with chromatin states



Blood lipids

SNP	H1 ES	K562	GM12878	HepG2	Huvec	HSMM	NHLF	NHEK	HMEC	Chrom. Band	Gene	Link Score	Distance (kb)
rs2479409										1p32			
rs10761731										10q21			
rs1169288										12q24			
rs2072183										7p13			
rs2923084										11p15			
rs629301										1p13	CELSR2	2.1	26
rs4846914										1q42			
rs514230										1q42			
rs1367117										2p24	APOB	-	3
rs581080										9p22	TTC39B	-	2
rs3136441										11p11	F2	-	2
rs16942887										16q22	PSKH1	-	1
rs6511720										19p13	LDLR	-	2
rs737337										19p13	LOC55908	-	1
rs439401										19q13	APOC1	-	4
rs605066										6q24			
rs2000999										16q22	HPR	5.5	11
rs1800961										20q13	HNF4A	2.3	12
rs9686661										5q11			
rs2293889										8q23			
rs1883025										9q31	ABCA1	1.9	26
rs2068888										10q23	CYP26A1	1.6	6
rs4765127										12q24	ZNF664	-	2
rs838880										12q24			
rs6065906										20q13	CTSA	1.6	35
rs1689800										1q25			
rs2412710										15q15			
rs11649653										16p11			

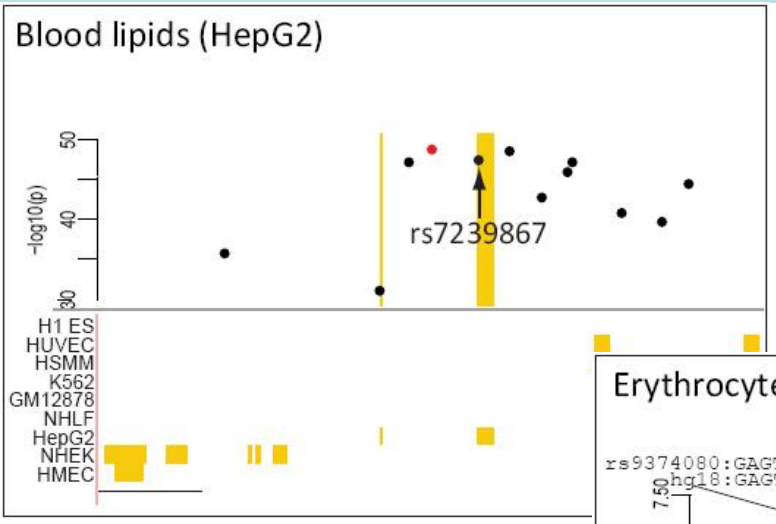
Erythrocyte phenotypes

SNP	H1 ES	K562	GM12878	HepG2	Huvec	HSMM	NHLF	NHEK	HMEC	Chrom. Band	Gene	Link Score	Distance (kb)
rs11915082										3q29			
rs1122794										16p13	HBQ1	2.2	79
rs11085824										19p13	GCDH	-	1
rs643381										6q24			
rs9349205										6p21	BYSL	2.2	36
rs12718597										7p12	IKZF1	1.9	81
rs11065987										12q24	BRAP	1.8	51
rs4895441										6q23	HBS1L	1.7	51
rs628751										6q24			
rs10758658										9p24	RCL1	1.6	64
rs172629										4q12			
rs7776054										6q23	HBS1L	1.6	43
rs131794										22q13	ODF3B	-	1
rs9374080										6q21			
rs7786877										7q22	MOSPD3	-	4
rs7385804										7q22	TFR2	-	3

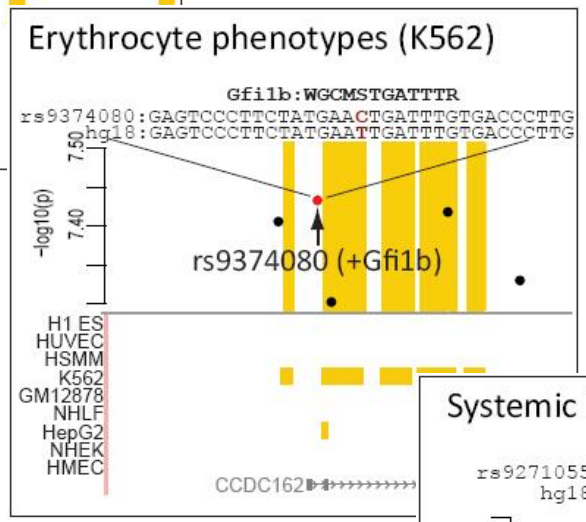
Systemic lupus erythematosus

SNP	H1 ES	K562	GM12878	HepG2	Huvec	HSMM	NHLF	NHEK	HMEC	Chrom. Band	Gene	Link Score	Distance (kb)
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			

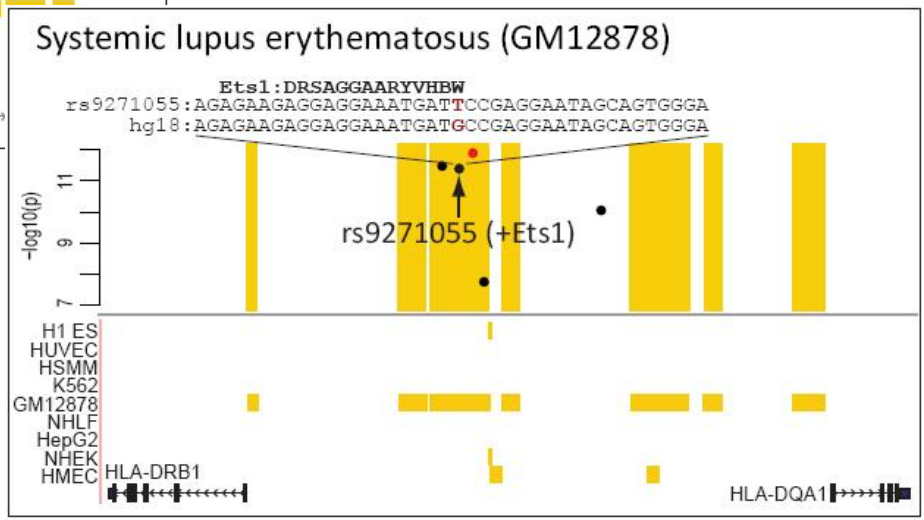
Annotations & regulatory predictions for GWAS



Associated SNP coincides with cell type-specific enhancer



SNP affects predicted regulatory interactions



Genome annotations & regulatory predictions for biomedical research

- Epigenomic maps reveal non-coding elements & cell type-specificities
- Chromatin dynamics link enhancers, regulators & target genes
- Annotations & regulatory predictions for GWAS
- Integration of ENCODE (TFs, RNAs) & Epigenomics (*in vivo* tissues) data will provide a rich resource for interpreting human disease