PENNSTATE
1855

Genomics of Gene Regulation

Genomic and Proteomic Approaches to Heart, Lung, Blood and Sleep Disorders

Jackson Laboratories
Ross Hardison

August 12, 2009

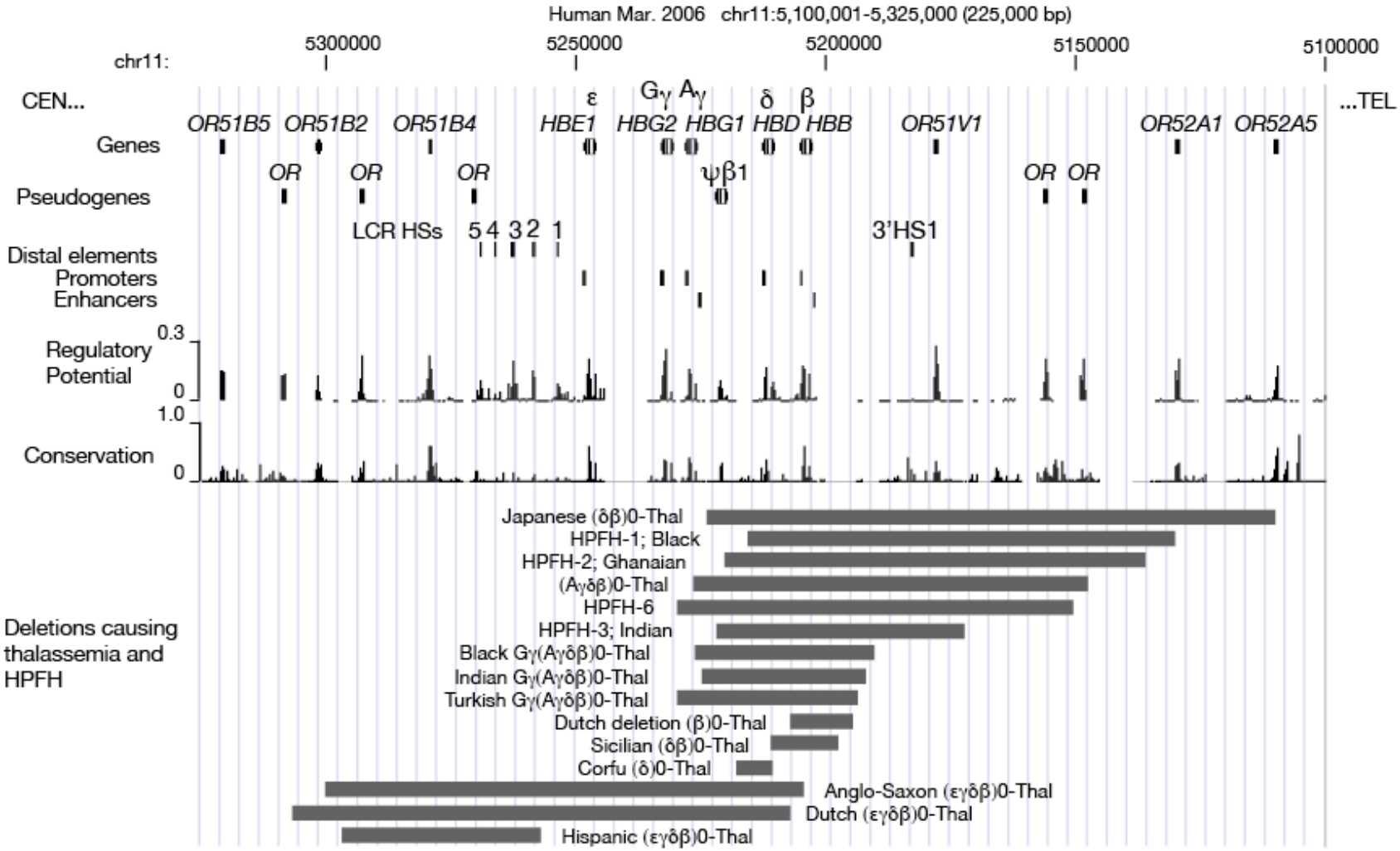
Heritable variation in gene regulation

"Simple" Mendelian traits, e.g. thalassemias

Variation in expression is common in normal individuals

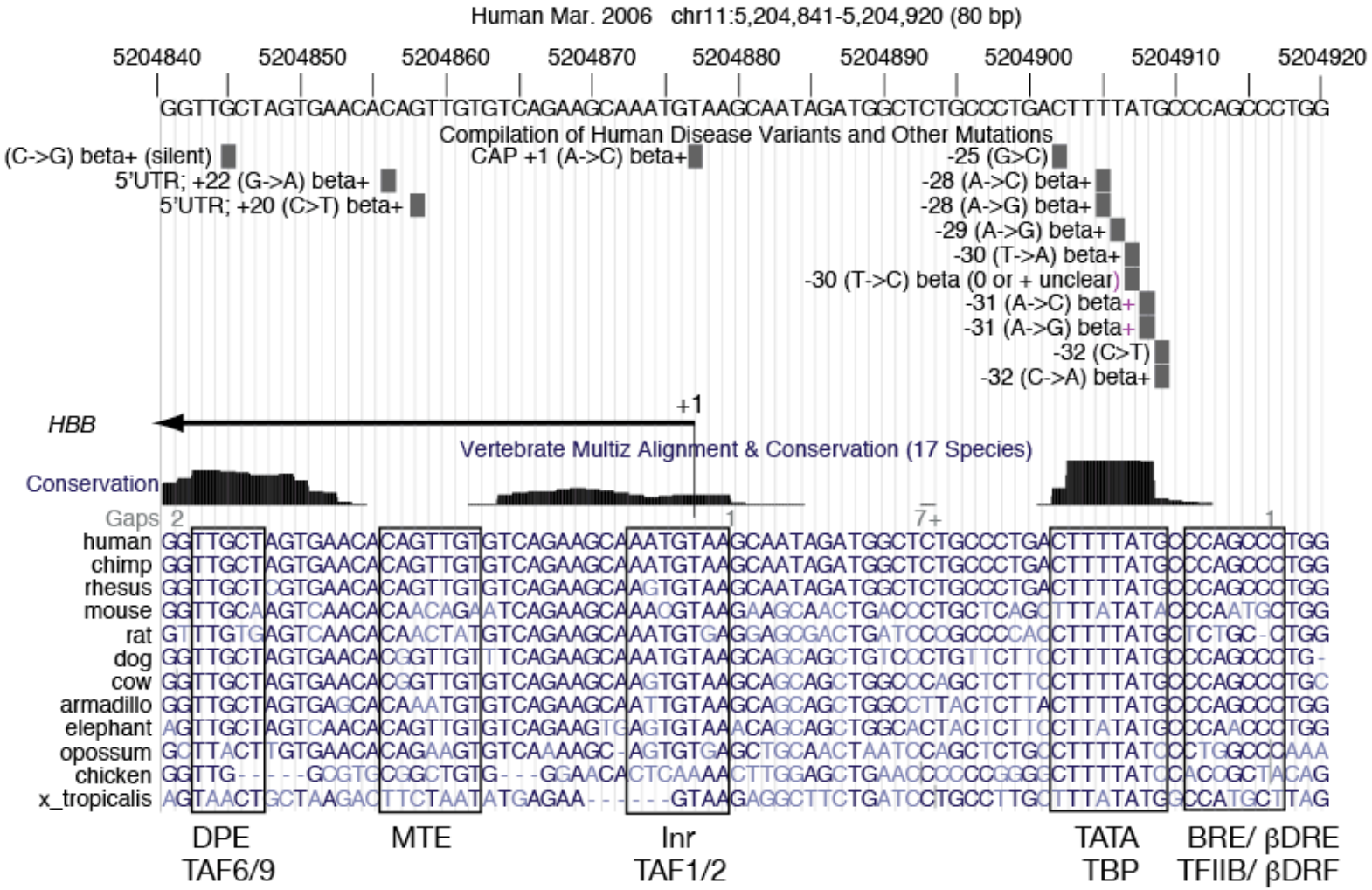
Variation in expression may be a major contributor to complex traits (including heart, lung, blood and sleep disorders)

Deletions of noncoding DNA can affect gene expression



Forget and Hardison, Chapter in Disorders of Hemoglobin, 2nd edition

Substitutions in promoters can affect expression



Forget and Hardison, Chapter in Disorders of Hemoglobin, 2nd edition

Variation of gene expression among individuals

- Levels of expression of many genes vary in humans (and other species)
- Variation in expression is heritable
- Determinants of variability map to discrete genomic intervals
- Often multiple determinants
- This variation indicates an abundance of *cis*-regulatory variation in the human genome
- "We predict that variants in regulatory regions make a greater contribution to complex disease than do variants that affect protein sequence" Manolis Dermitzakis, Science*Daily*
  - Microarray expression analyses of 3554 genes in 14 families
    - Morley M … Cheung VG (2004) Nature 430:743-747
  - Expression analysis of EBV-transformed lymphoblastoid cells from all 270 individuals genotypes in HapMap
    - Stranger BE … Dermitzakis E (2007) Nature Genetics 39:1217-1224

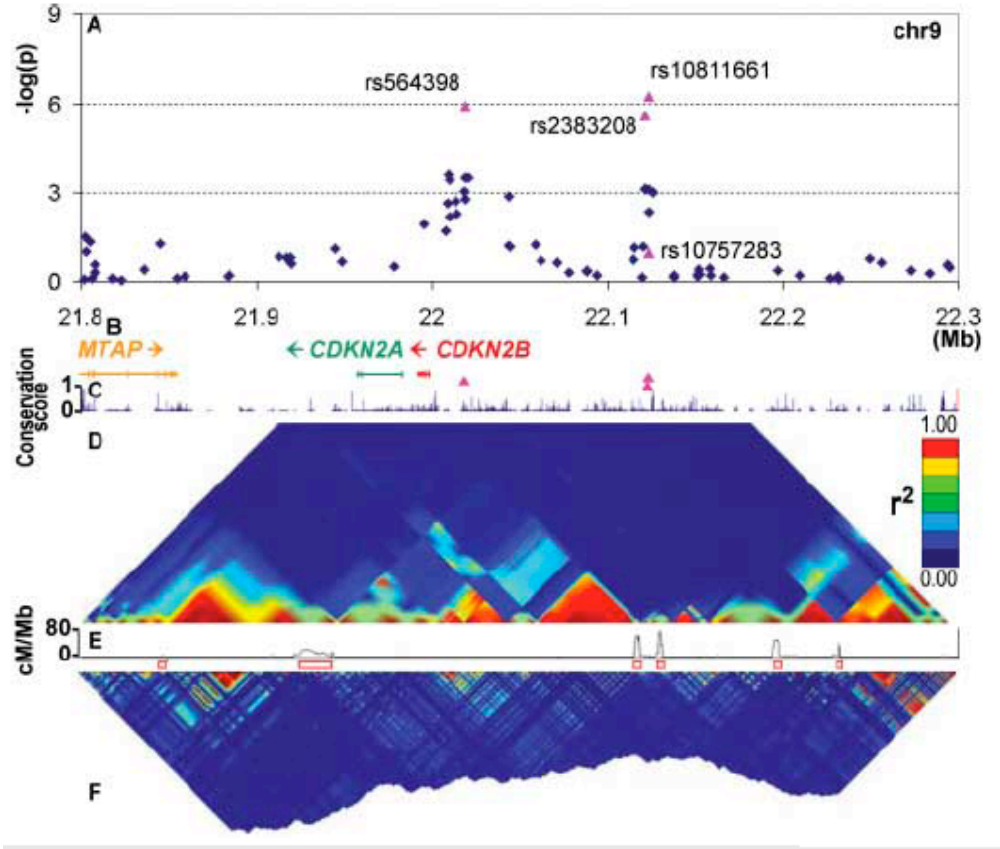
Risk loci in noncoding regions

Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes

Eleftheria Zeggini,[1,2]* Michael N. Weedon,[3,4]* Cecilia M. Lindgren,[1,2]* Timothy M. Frayling,[3,4]* Katherine S. Elliott,[2] Hana Lango,[3,4] Nicholas J. Timpson,[2,5] John R. B. Perry,[3,4] Nigel W. Rayner,[1,2] Rachel M. Freathy,[3,4] Jeffrey C. Barrett,[2] Beverley Shields,[4] Andrew P. Morris,[2] Sian Ellard,[4,6] Christopher J. Groves,[1] Lorna W. Harries,[4] Jonathan L. Marchini,[7] Katharine R. Owen,[1] Beatrice Knight,[4] Lon R. Cardon,[2] Mark Walker,[8] Graham A. Hitman,[9] Andrew D. Morris,[10] Alex S. F. Doney,[10] The Wellcome Trust Case Control Consortium (WTCCC),† Mark I. McCarthy,[1,2]‡§ Andrew T. Hattersley[3,4]‡

The molecular mechanisms involved in the development of type 2 diabetes are poorly understood. Starting from genome-wide genotype data for 1924 diabetic cases and 2938 population controls generated by the Wellcome Trust Case Control Consortium, we set out to detect replicated diabetes association signals through analysis of 3757 additional cases and 5346 controls and by integration of our findings with equivalent data from other international consortia. We detected diabetes susceptibility loci in and around the genes *CDKAL1*, *CDKN2A/CDKN2B*, and *IGF2BP2* and confirmed the recently described associations at *HHEX/IDE* and *SLC30A8*. Our findings provide insight into the genetic architecture of type 2 diabetes, emphasizing the contribution of multiple variants of modest effect. The regions identified underscore the importance of pathways influencing pancreatic beta cell development and function in the etiology of type 2 diabetes.

(2007) Science 316: 1336-1341



DNA sequences involved in regulation of gene transcription

Protein-DNA interactions

Chromatin effects

Distinct classes of regulatory regions

Act in *cis*, affecting expression of a gene on the same chromosome.

Cis-regulatory modules (**CRMs**)

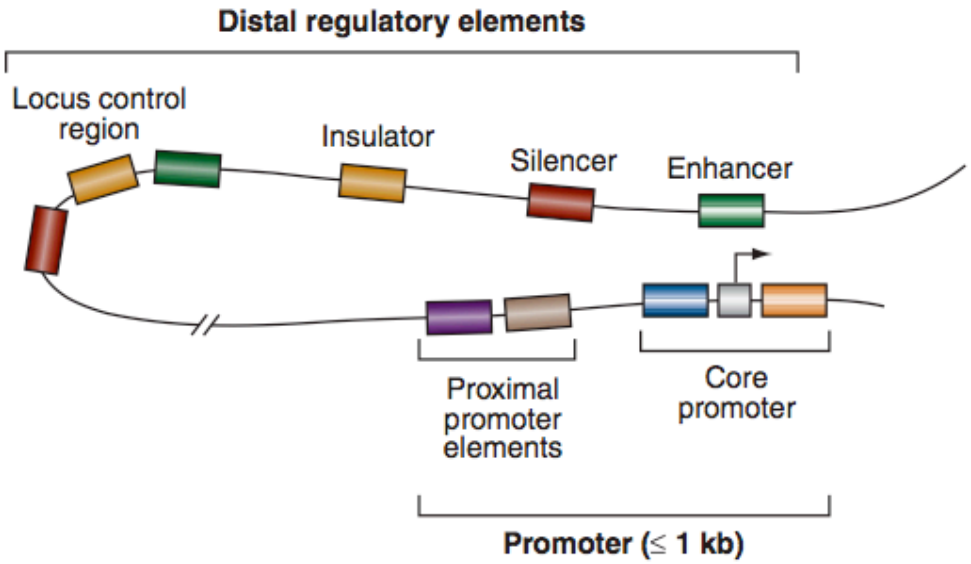


Figure 1

Schematic of a typical gene regulatory region. The promoter, which is composed of a core promoter and proximal promoter elements, typically spans less than 1 kb pairs. Distal (upstream) regulatory elements, which can include enhancers, silencers, insulators, and locus control regions, can be located up to 1 Mb pairs from the promoter. These distal elements may contact the core promoter or proximal promoter through a mechanism that involves looping out the intervening DNA.

Maston G, Evans S and Green M (2006) Annu Rev Genomics Hum Genetics 7:29-59

General features of promoters

- A promoter is the DNA sequence required for correct initiation of transcription
- It affects the **amount** of product from a gene, but does not affect the **structure** of the product.
- Most promoters are at the 5' end of the gene.

RNA polymerase II

Upstream regulatory elements:
Regulate efficiency of utilization of minimal promoter

TATA box + Initiator:
Core or minimal promoter. Site of assembly of preinitiation complex

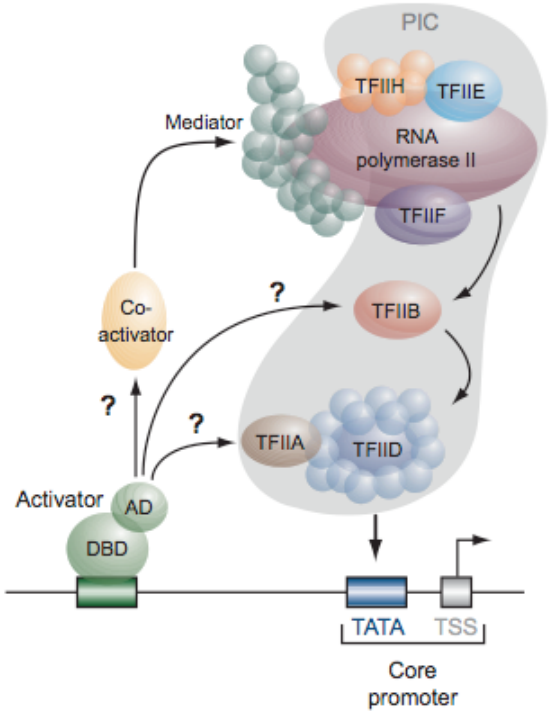
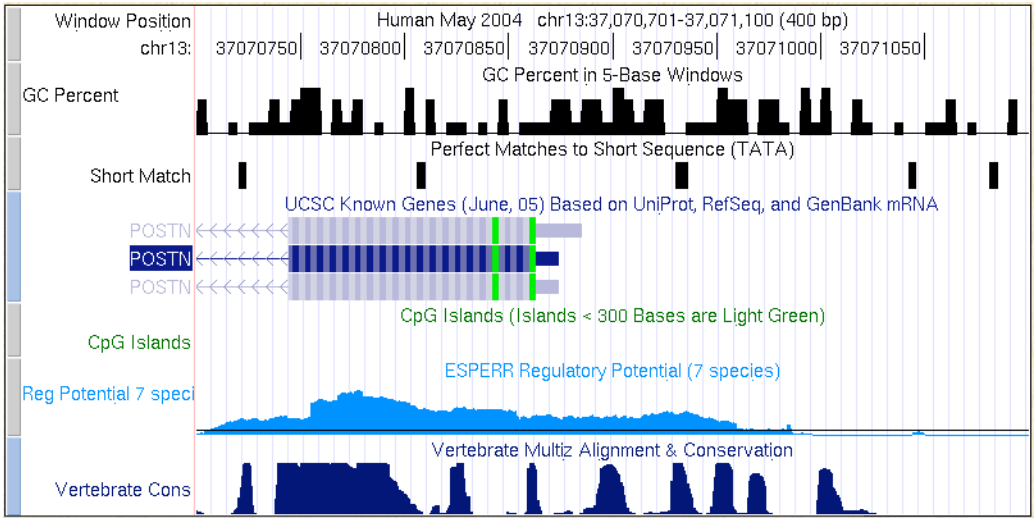


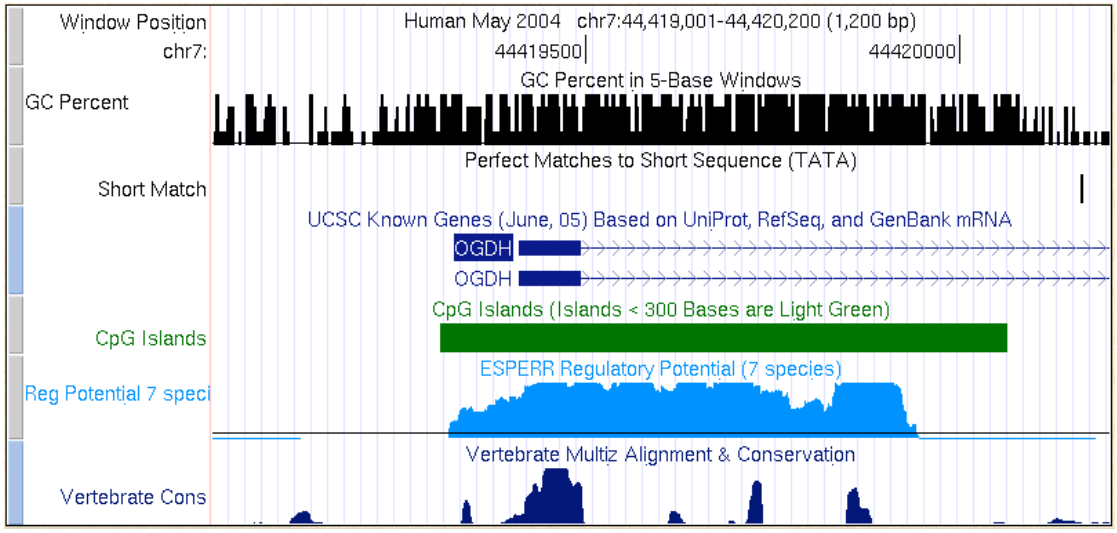
Figure 2
The eukaryotic transcriptional machinery. Factors involved in eukaryotic transcription by RNA polymerase II can be classified into three groups: general transcription factors (GTFs), activators, and coactivators. GTFs,

Maston, Evans & Green (2006) Ann Rev Genomics & Human Genetics, 7:29-59

Most promoters in mammals are CpG islands



TATA, no CpG island
About 10% of promoters

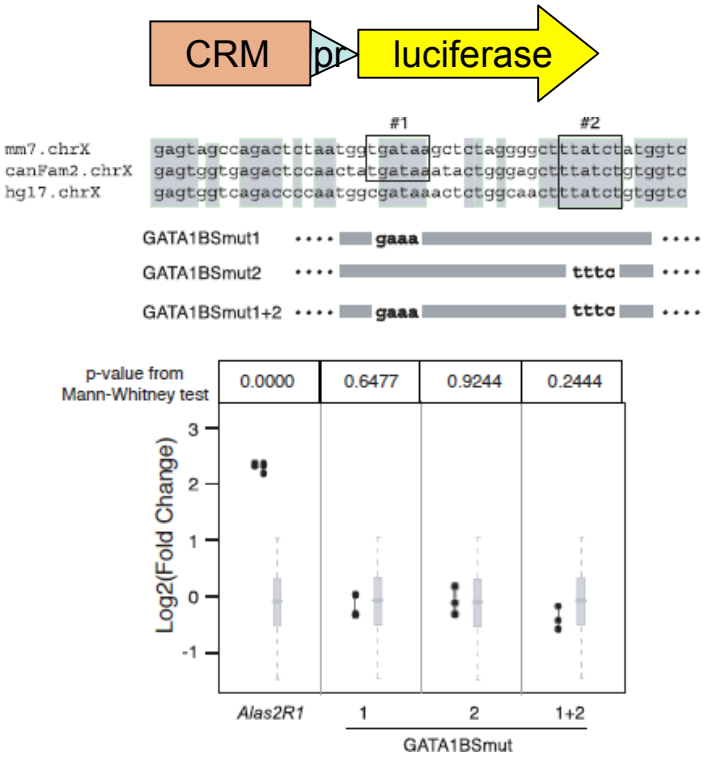


CpG island, no TATA
About 90% of promoters

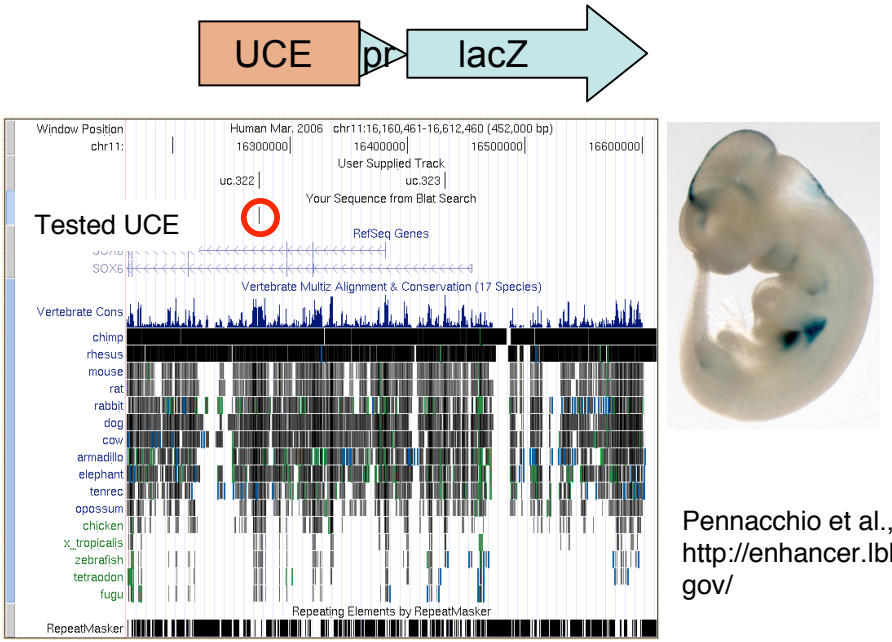
Carninci … Hayashizaki (2006)
Nature Genetics 38:626

Enhancers

- *Cis*-acting sequences that cause an **increase** in expression of a gene
- Act **independently** of **position** and **orientation** with respect to the gene.



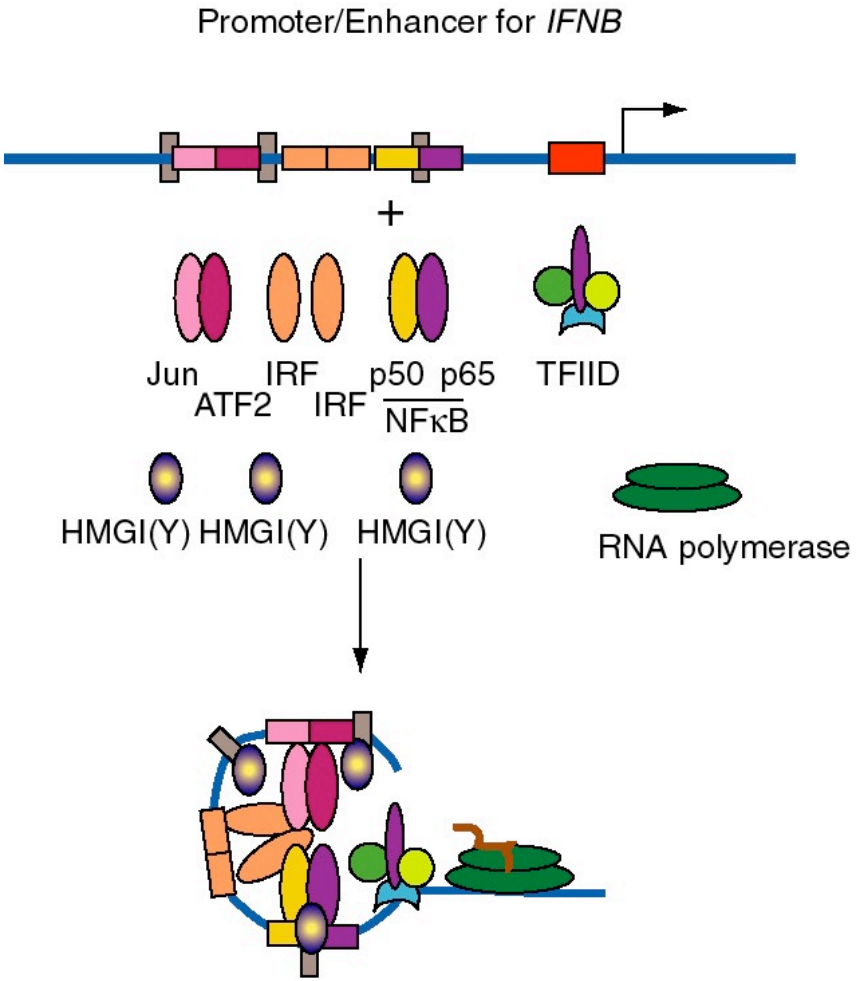
About half of the enhancers predicted by interspecies alignments are validated in erythroid cells
Wang et al. (2006) Genome Research 16:1480- 1492



Pennacchio et al., http://enhancer.lbl.gov/

Over half of ultraconserved noncoding sequences are developmental enhancers
Pennacchio et al. (2006) Nature 444:499-502

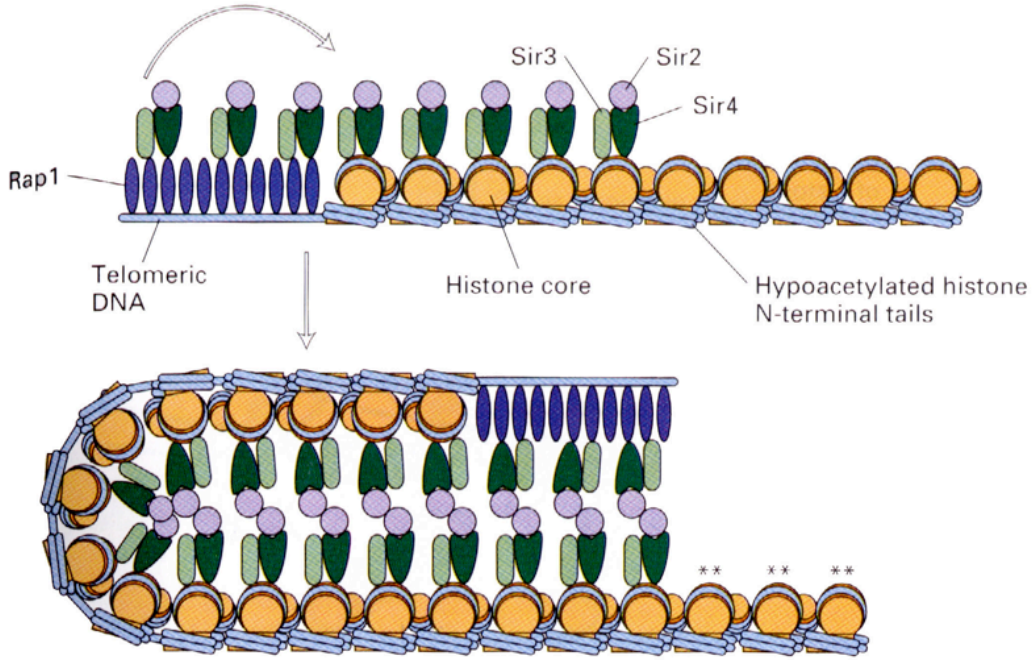
CRMs are clusters of specific binding sites for transcription factors



Hardison (2002) on-line textbook *Working with Molecular Genetics* http://www.bx.psu.edu/~ross/

Silencer

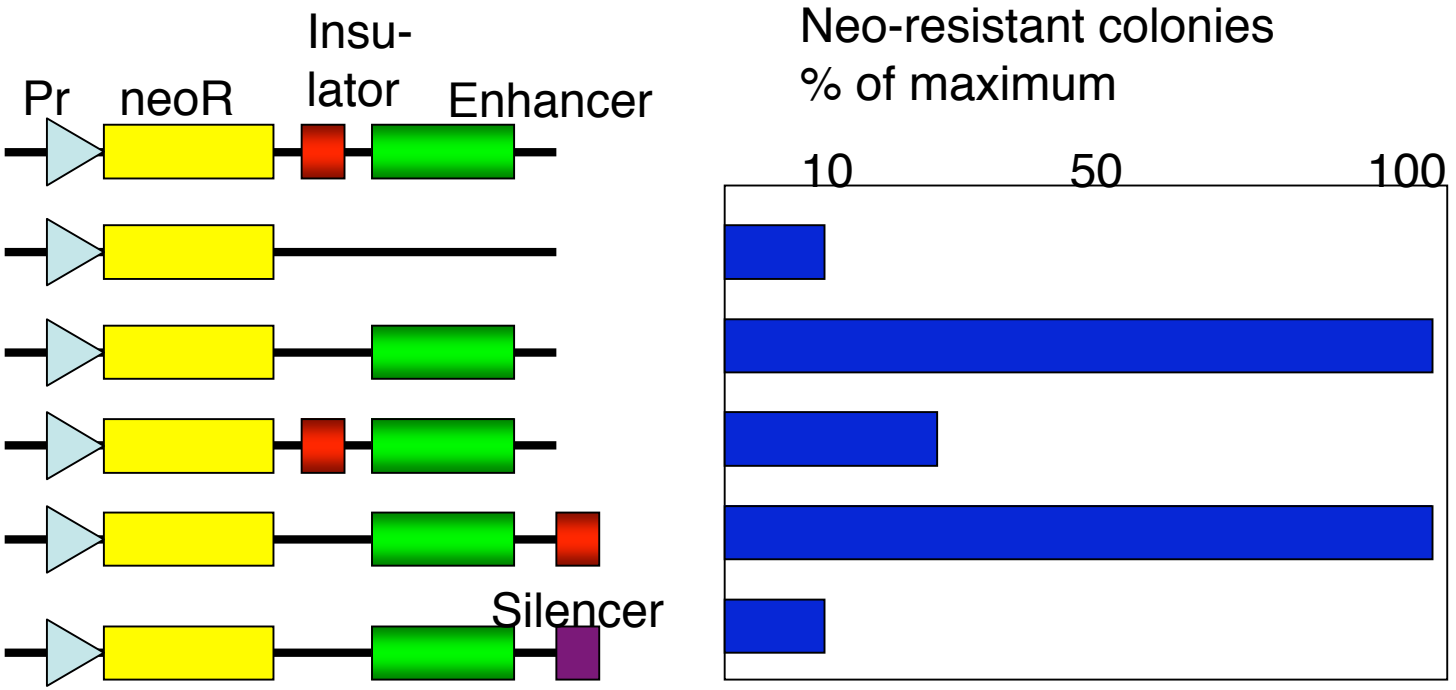
- *Cis*-acting sequences that cause a **decrease** in gene expression
- Similar to enhancer but has an opposite effect on gene expression
- Gene repression - inactive chromatin structure (heterochromatin)



- SIR proteins (Silent Information Regulators)
- Nucleates assembly of multi-protein complex
  - hypoacetylated N-terminal tails of histones H3 and H4
  - methylated N-terminal tail of H3 (Lys 9)

Insulators and boundaries

- A **boundary** in chromatin marks a transition from open to closed chromatin
- An **insulator** blocks activation of promoter by an enhancer
  - Requires CTCF
- Example: HS4 from chick *HBB* complex has both functions



Repression by PcG proteins via chromatin modification

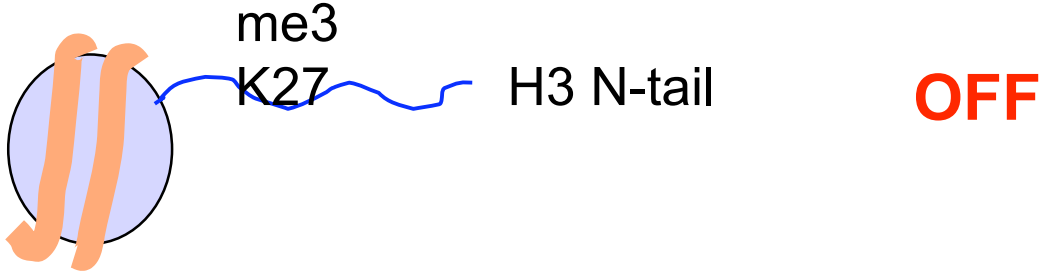
Cell, Vol. 111, 197–208, October 18, 2002, Copyright ©2002 by Cell Press

Histone Methyltransferase Activity of a *Drosophila* Polycomb Group Repressor Complex

Jürg Müller,[1,2,7,8] Craig M. Hart,[4,8,9] Nicole J. Francis,[5,6,8] Marcus L. Vargas,[4] Aditya Sengupta,[1,2] Brigitte Wild,[1] Ellen L. Miller,[4] Michael B. O'Connor,[3,4] Robert E. Kingston,[5,6,7] and Jeffrey A. Simon[4,7]

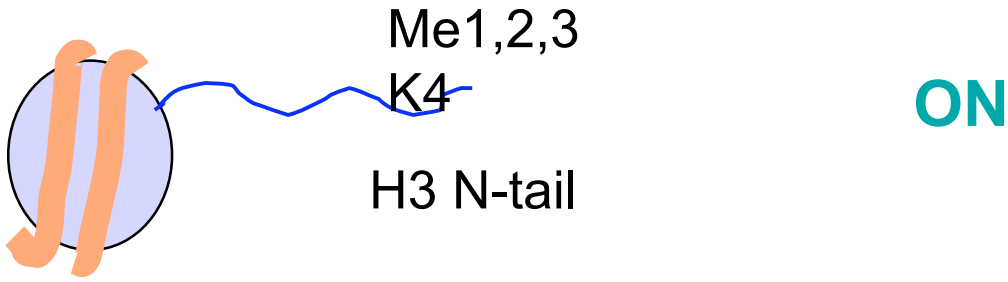
Polycomb Group (PcG) Repressor Complex 2:
ESC, E(Z), NURF-55, and PcG repressor SU(Z)12

Methylates K27 of Histone H3 via the SET domain of E(Z)

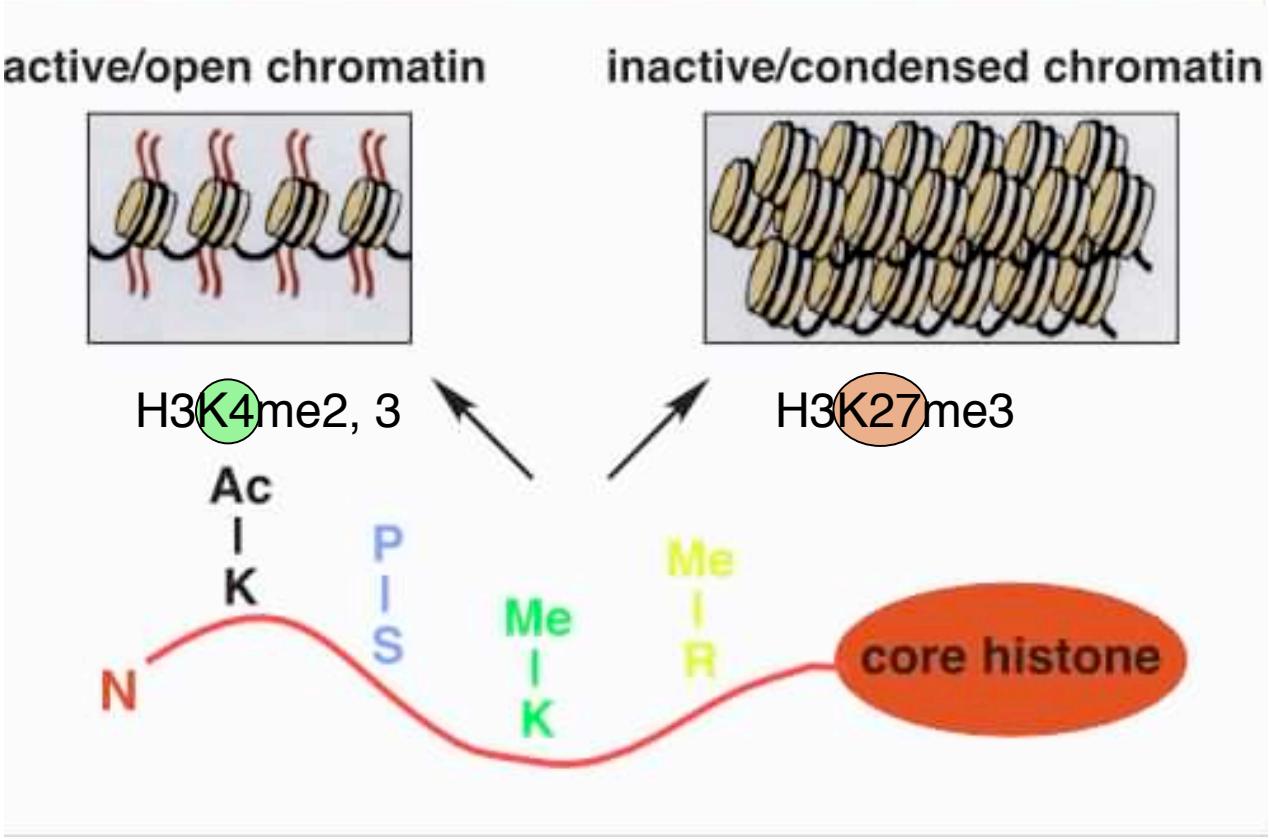


trx group (trxG) proteins activate via chromatin changes

- SWI/SNF nucleosome remodeling
- Histone H3 and H4 acetylation
- Methylation of K4 in histone H3
  - Trx in Drosophila, MLL in humans
- http://www.igh.cnrs.fr/equip/cavalli/link.PolycombTeaching.html#Part_3



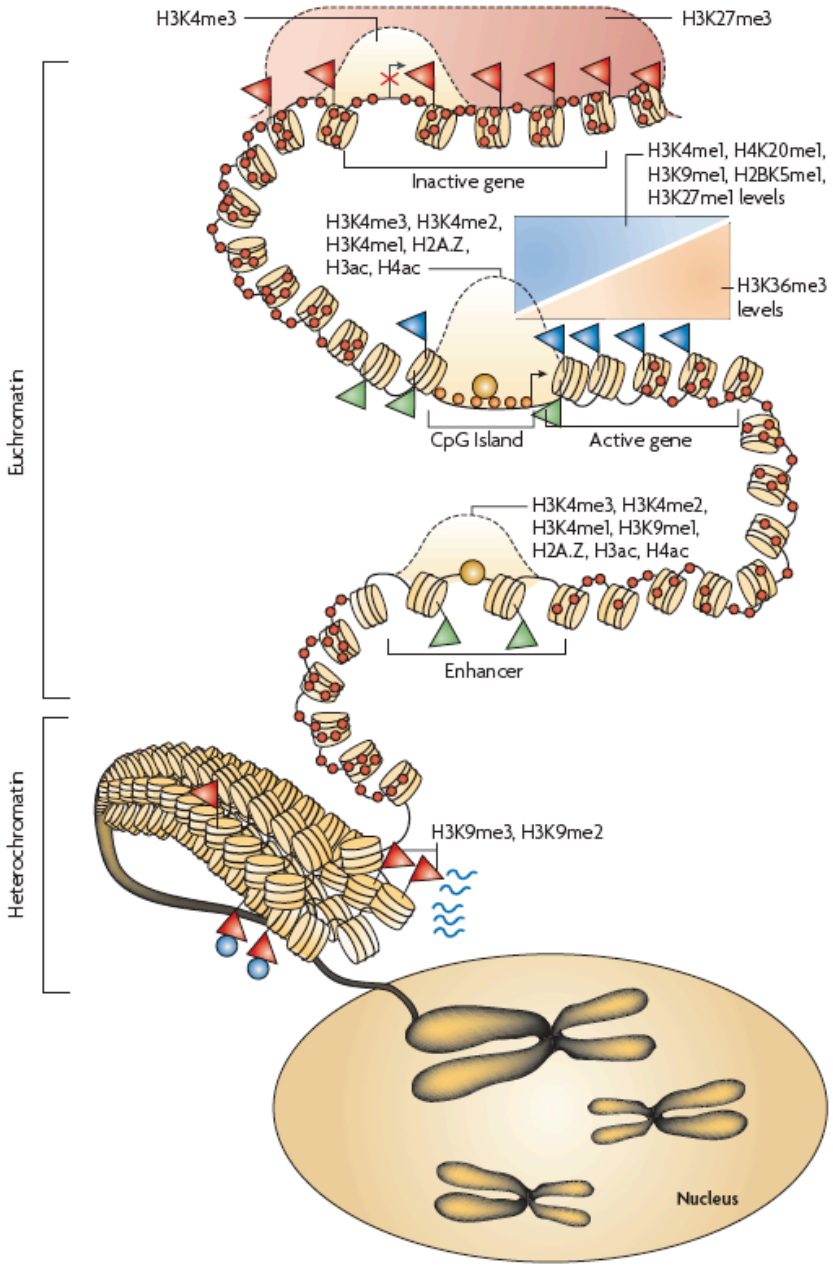
Histone modifications modulate chromatin structure



http://www.imt.uni-marburg.de/bauer/images/fig2.jpg

Uta-Maria Bauer

Repressed and active chromatin



Dustin Schones and Keiji Zhao (2008) Nature Reviews Genetics 9: 179

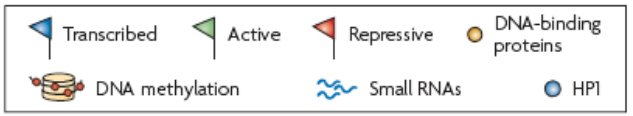


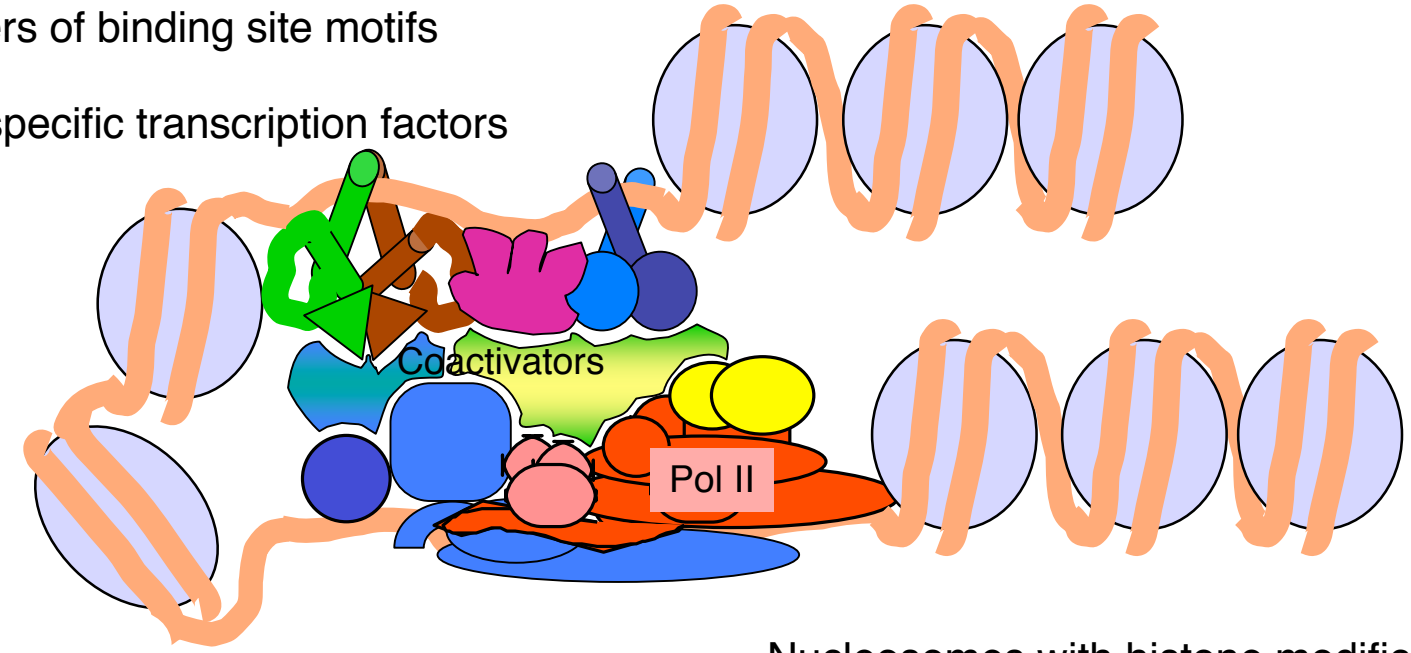
Figure 5 | **Characteristics of epigenomes.** The interaction of DNA methylation, histone modification, nucleosome positioning and other factors such as small RNAs contribute to an overall epigenome that regulates gene expression and allows cells to remember their identity. Chromosomes are divided into accessible regions of euchromatin and poorly accessible regions of heterochromatin. Heterochromatic regions are marked with histone H3 lysine 9 di- and trimethylation (H3K9me2 and H3K9me3), which serve as a platform for HP1 (heterochromatic protein 1) binding. Small RNAs have been implicated in the maintenance of heterochromatin. DNA methylation is persistent throughout genomes, and is missing only in regions such as CpG islands, promoters and possibly enhancers. The H3K27me3 modification is present in broad domains that encompass inactive genes. Histone modifications including H3K4me3, H3K4me2, H3K4me1 as well as histone acetylation and histone variant H2A.Z mark the transcription start site regions of active genes. The monomethylations of H3K4, H3K9, H3K27, H4K20 and H2BK5 mark actively transcribed regions, peaking near the 5′ end of genes. The trimethylation of H3K36 also marks actively transcribed regions, but peaks near the 3′ end of genes.

Biochemical features of DNA in CRMs

Accessible to cleavage:
DNase hypersensitive site

Clusters of binding site motifs

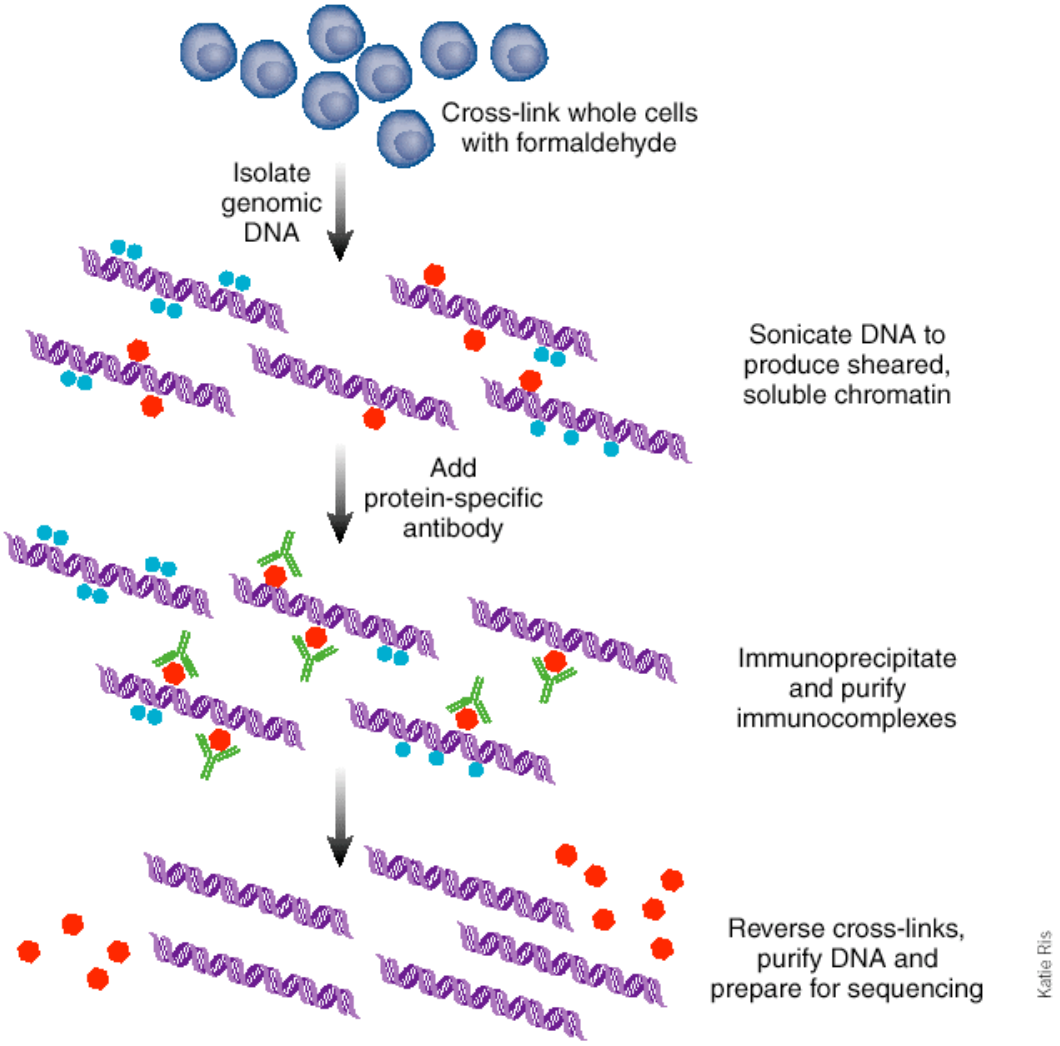
Bound by specific transcription factors



Associated with RNA polymerase
and general transcription factors

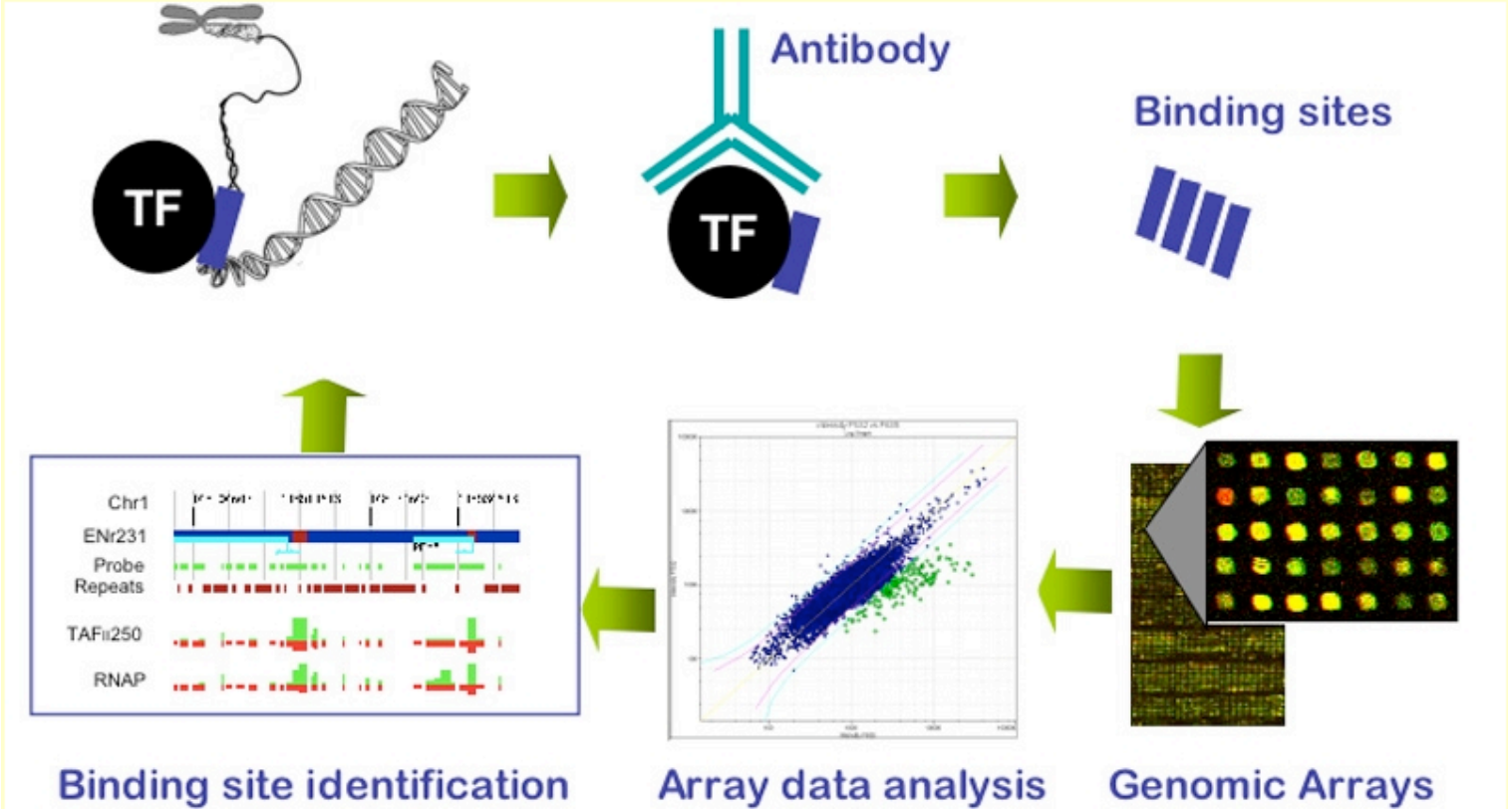
Nucleosomes with histone modifications:
Acetylation of H3 and H4
Methylation of H3K4

Chromatin immunoprecipitation: Greatly enrich for DNA occupied by a protein



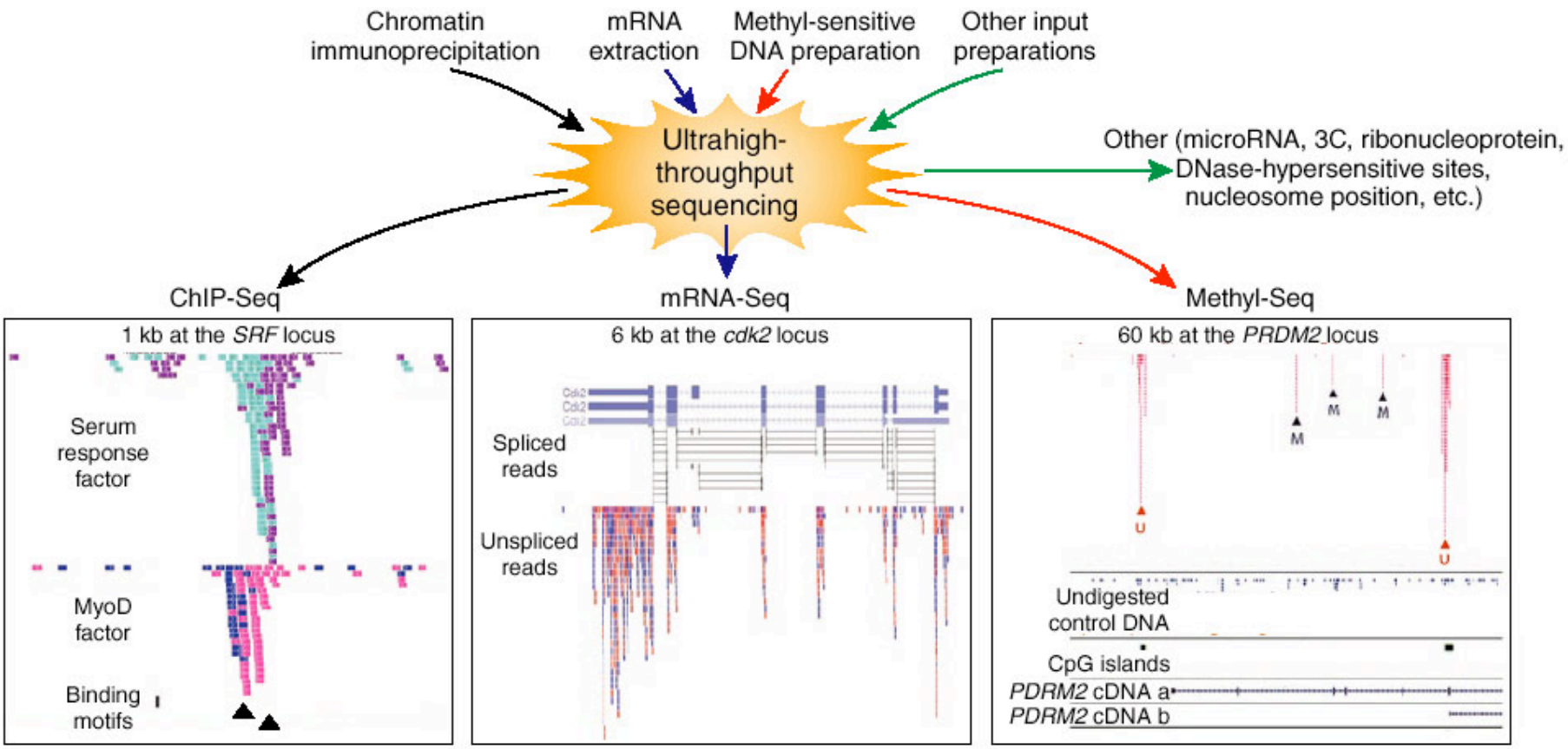
Elaine Mardis (2007) Nature Methods 4: 613-614

ChIP-chip: High throughput mapping of DNA sequences occupied by protein



http://www.chiponchip.org Bing Ren's lab

Enrichment of sequence tags reveals function



Barbara Wold & Richard M Myers (2008) “Sequence Census Methods” Nature Methods 5:19-21

ChIP-seq for chromatin modifications

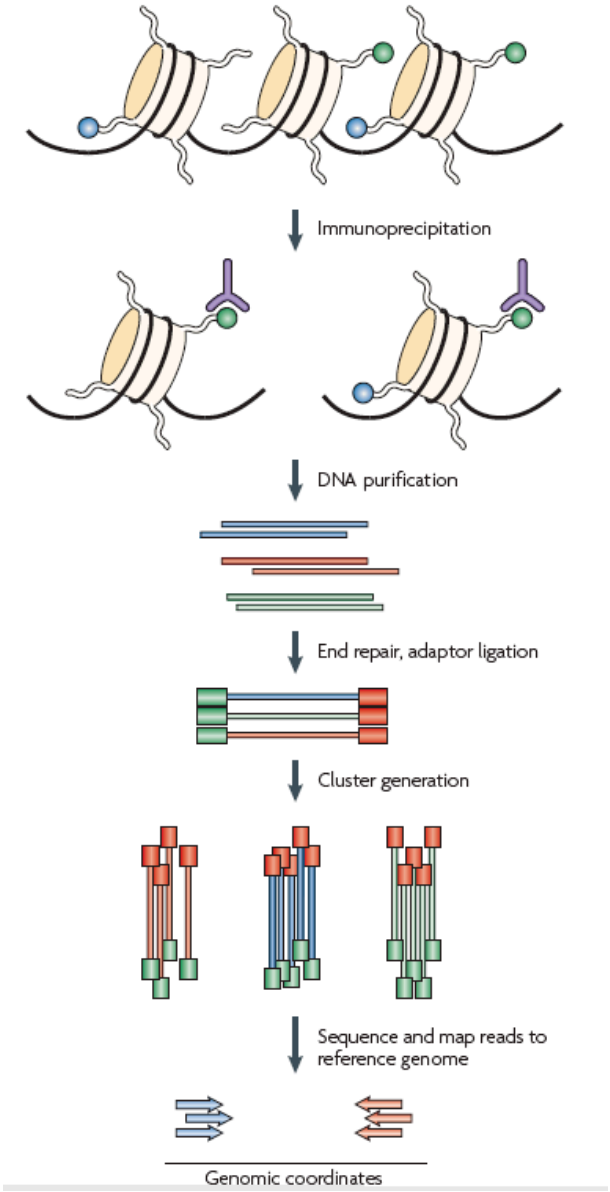
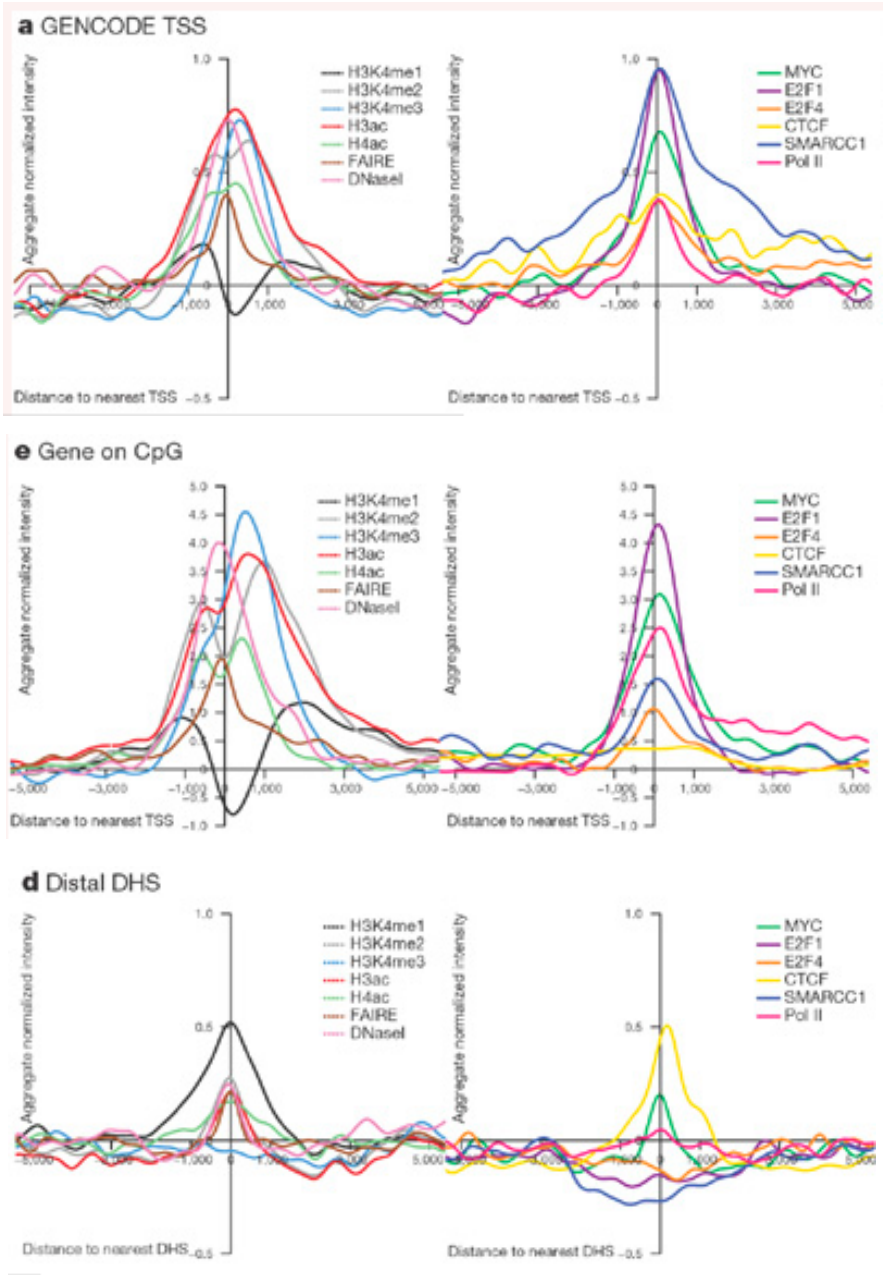


Figure 4 | **Chromatin immunoprecipitation combined with high-throughput sequencing techniques (ChIP–Seq).** One of the most exciting recent advances in technologies for studying epigenetic phenomena at a genomic scale relies on the combination of ChIP experiments with high-throughput sequencing. The procedure that is outlined here is specific to the Illumina Genome Analyzer using Solexa technology, although other high-throughput sequencing techniques would also work in principle. The first step is the purification of modified chromatin by immunoprecipitation using an antibody that is specific to a particular histone modification (shown in green). The ChIP DNA ends are repaired and ligated to a pair of adaptors, followed by limited PCR amplification. The DNA molecules are bound to the surface of a flow cell that contains covalently bound oligonucleotides that recognize the adaptor sequences. Clusters of individual DNA molecules are generated by solid-phase PCR and sequencing by synthesis is performed. The resulting sequence reads are mapped to a reference genome to obtain genomic coordinates that correspond to the immunoprecipitated fragments.

Dustin Schones and Keiji Zhao (2008)
Nature Reviews Genetics 9: 179

Distribution of histone modifications and factor binding around regulatory regions

- Symmetrical
- Promoters:
  - **H3K4me3**, H3K4me2
  - E2F1, E2F4, Myc, Pol II
- Distal HSs
  - **H3K4me1**: enhancers
  - CTCF: insulators

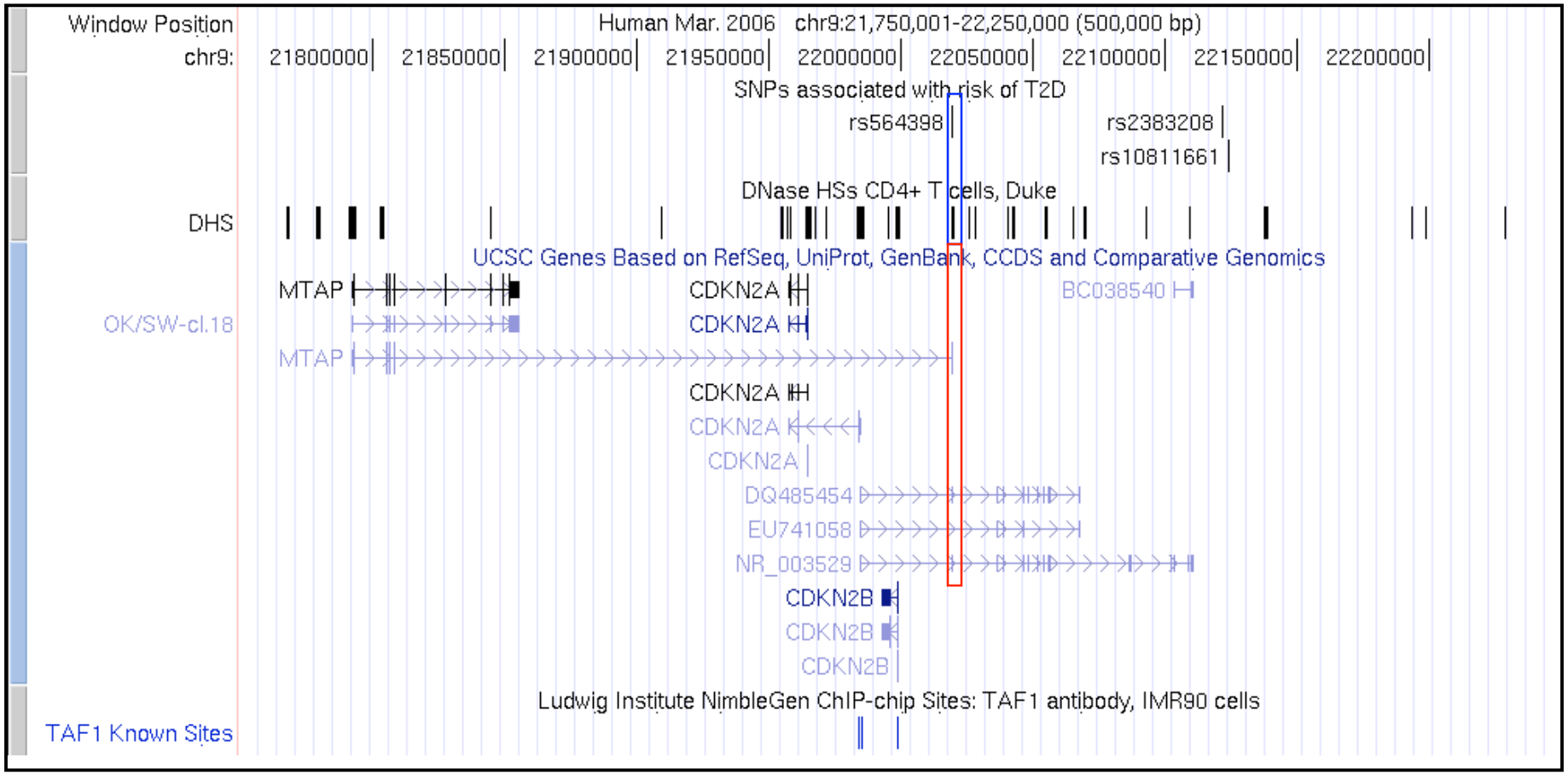


Birney et al. (2007) Nature, 447:799-816

Examples of genome-wide data on CRM features

- RNA polymerase II, preinitiation complex
  - IMR90 cells: Kim TH …Ren B (2005) Nature 436: 876-880
- Start sites for transcription
  - Carninci et al. (2006) Nature Genetics 38:626-635
- Histone modifications
  - T cells: Roh ... Zhao K (2006) PNAS 103:15782-15878
- Insulator protein CTCF
  - Primary fibroblasts: Kim TH … Ren B (2007) Cell 128:1231-1245
- DNase hypersensitive sites
  - CD4+ T cells: Boyle… Crawford G (2008) Cell 132:311-322
- Data for many on hg18 human genome assembly
  - http://www.bx.psu.edu
  - Go to Hardison lab
- Many datastreams: ENCODE project
  - Birney et al. (2007) Nature 477:799-816
  - http://genome.ucsc.edu
  - http://genome-test.cse.ucsc.edu

Genomic features at T2D risk variants



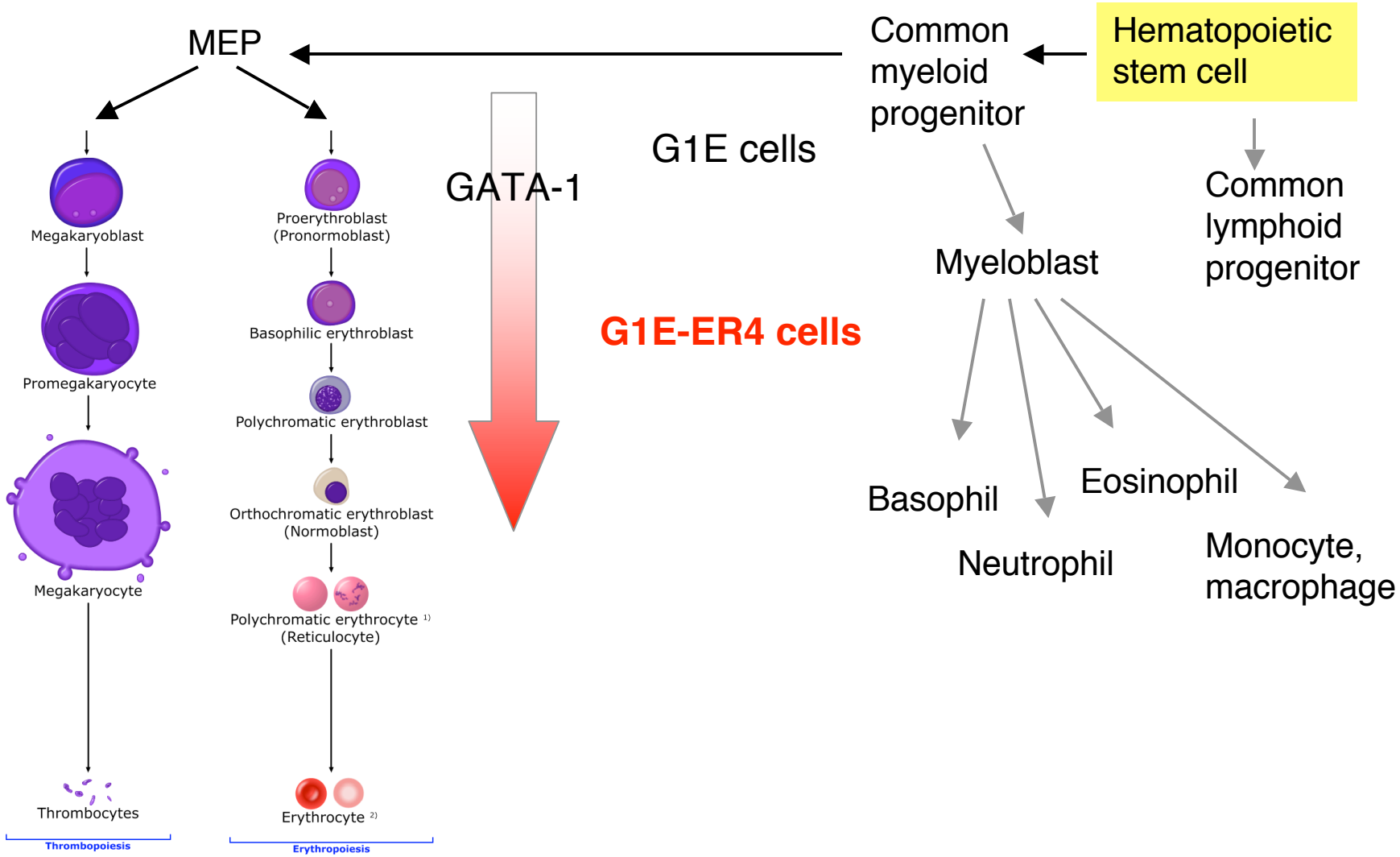
Overlap of SNP rh564398 with DHS suggests a role in transcriptional regulation, but overlap with an exon of a noncoding RNA suggests a role in post-transcriptional regulation. Different hypotheses to test in future work.

Occupancy by GATA1 and other TFs plus histone modifications lead to global insights for erythroid gene regulation

Weisheng Wu, Yong Cheng, Demesew Abebe, Cheryl Keller Capone,Ying Zhang, Ross, Swathi Ashok Kumar, Christine Dorman, David King

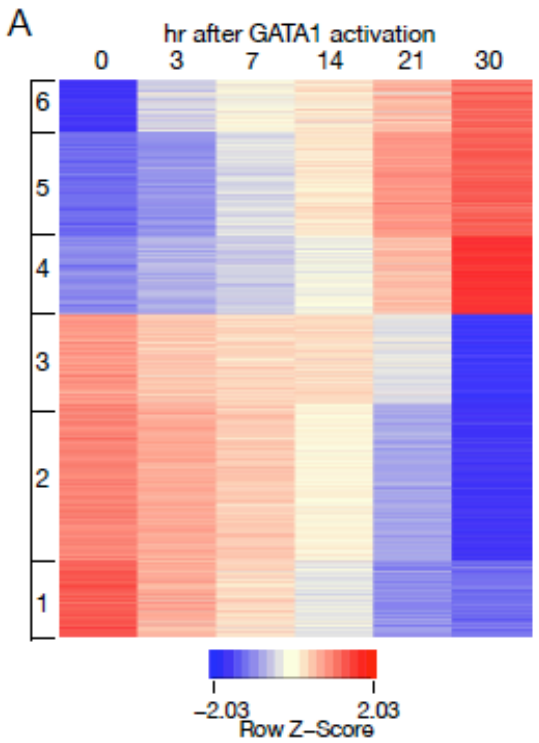
Collaborating labs: Mitch Weiss and Gerd Blobel (Childrens' Hospital of Philadelphia), James Taylor (Emory) Webb Miller, Francesca Chiaromonte, Yu Zhang, Stephan Schuster, Frank Pugh (PSU), Greg Crawford (Duke)

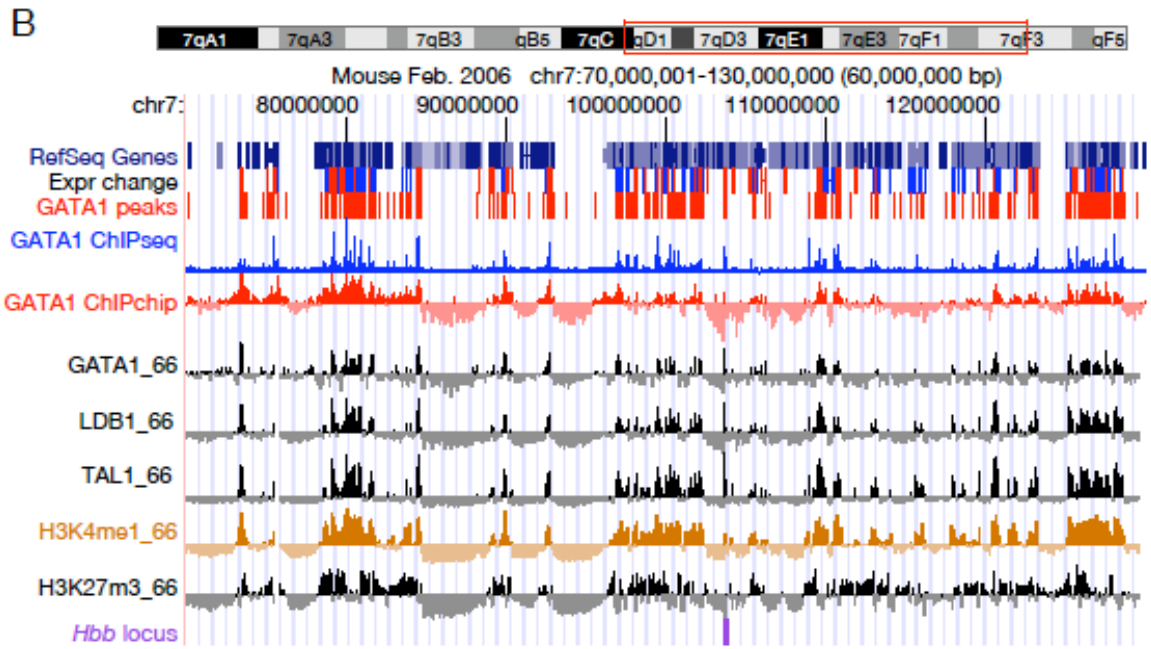
GATA-1 is required for erythroid maturation



Aria Rad, 2007 http://commons.wikimedia.org/wiki/Image:Hematopoiesis_(human)_diagram.png

GATA1-induced changes in gene expression and occupancy genome-wide

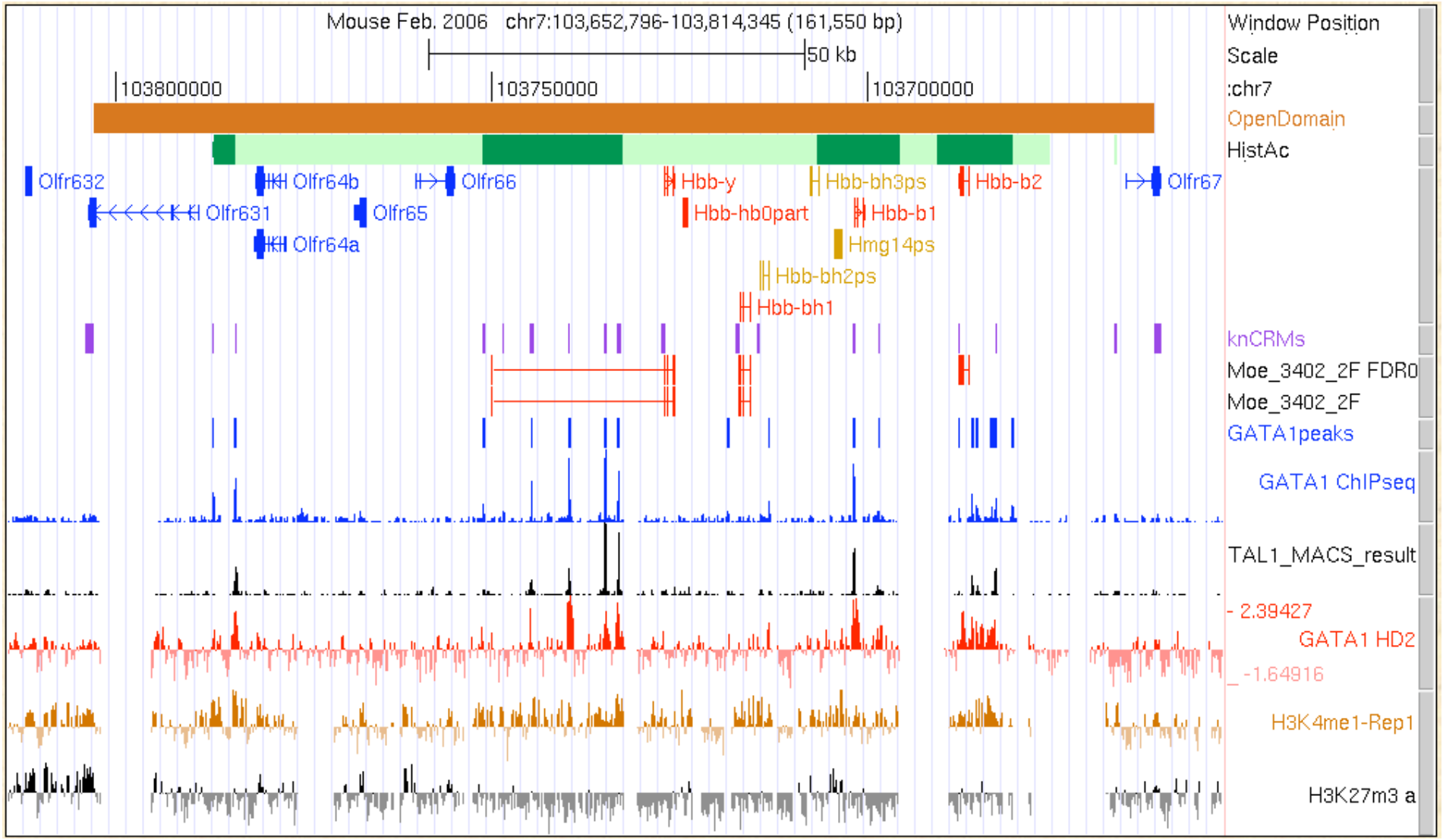




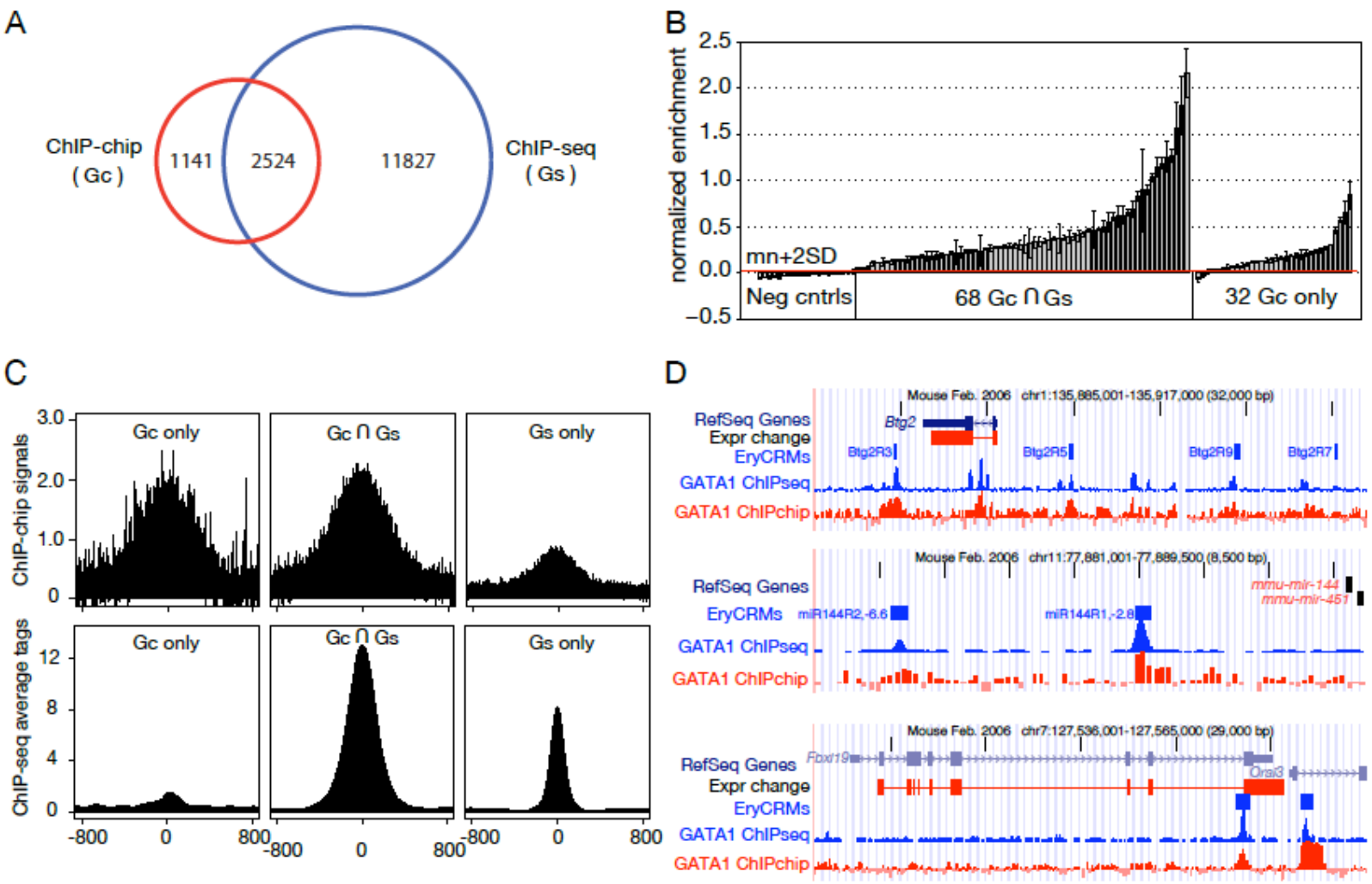
Genes induced or repressed after restoration of GATA1

Occupancy by TFs and histone modifications along a 60 Mb region

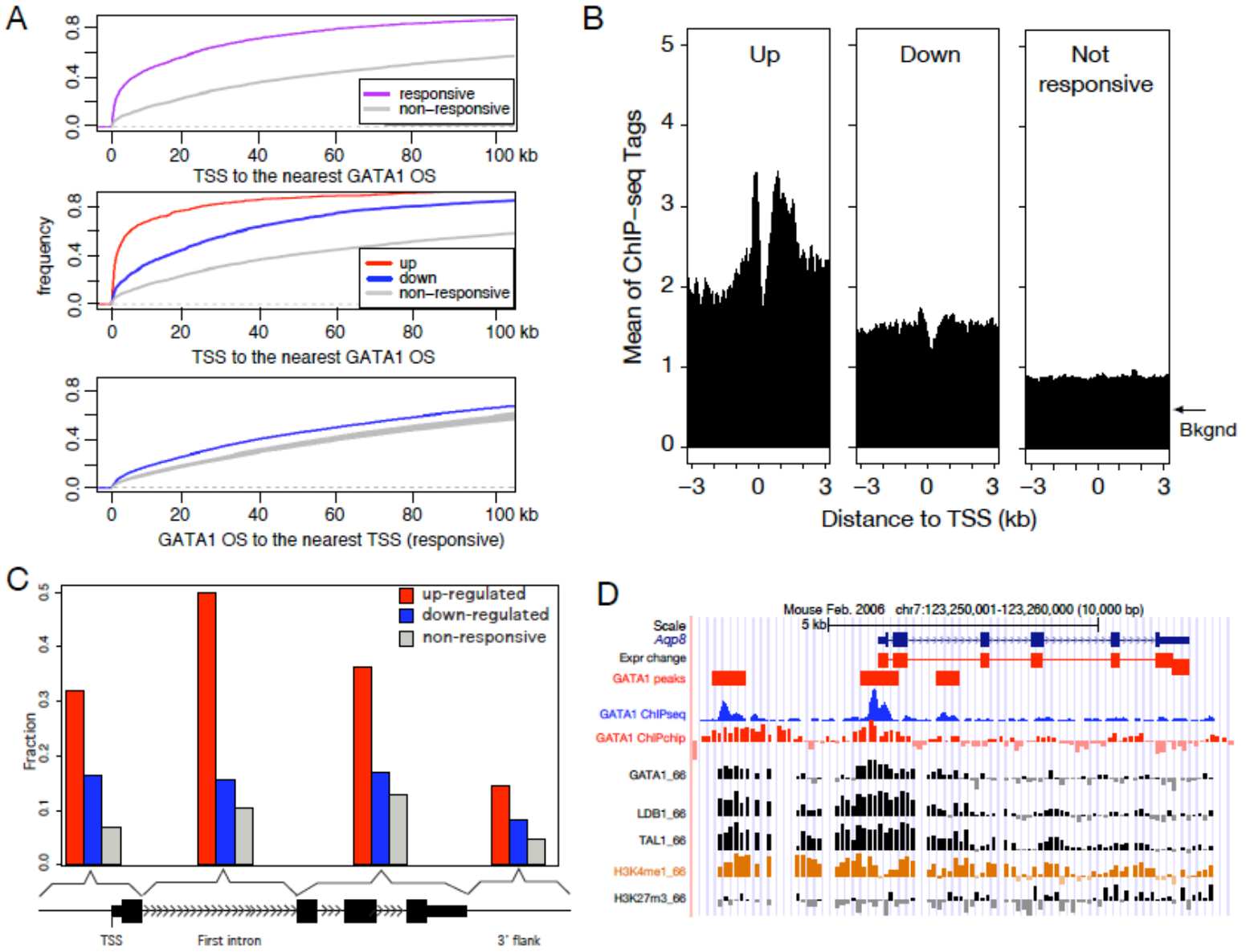
High throughput occupancy matches known CRMs at *Hbb* locus



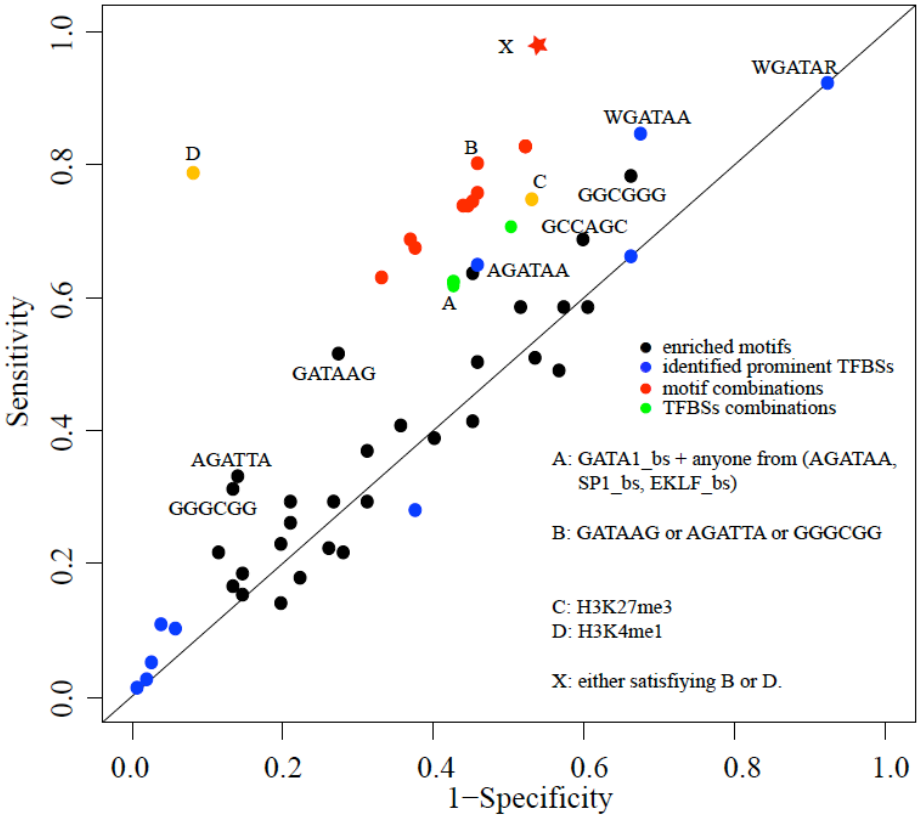
High sensitivity and specificity of high throughput occupancy data

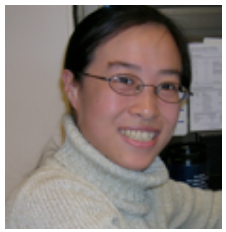


Induced genes have GATA1 occupied segments close to their TSS

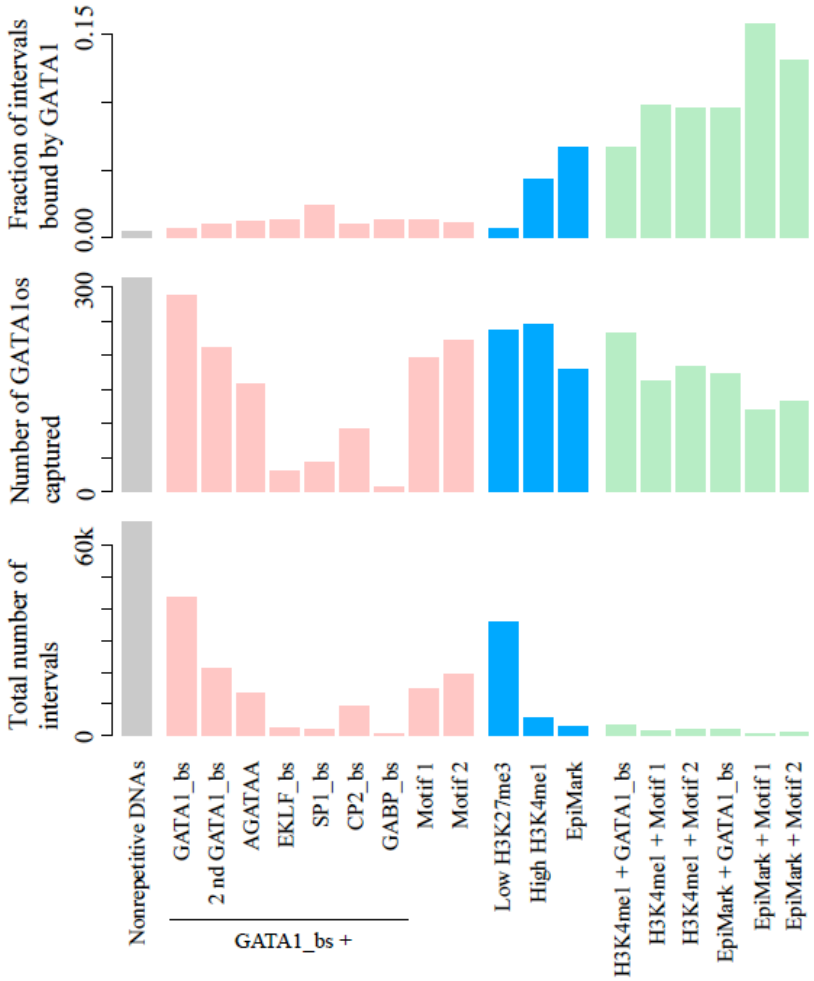


Determinants of occupancy by GATA1: Binding site motif WGATAR and H3K4me1





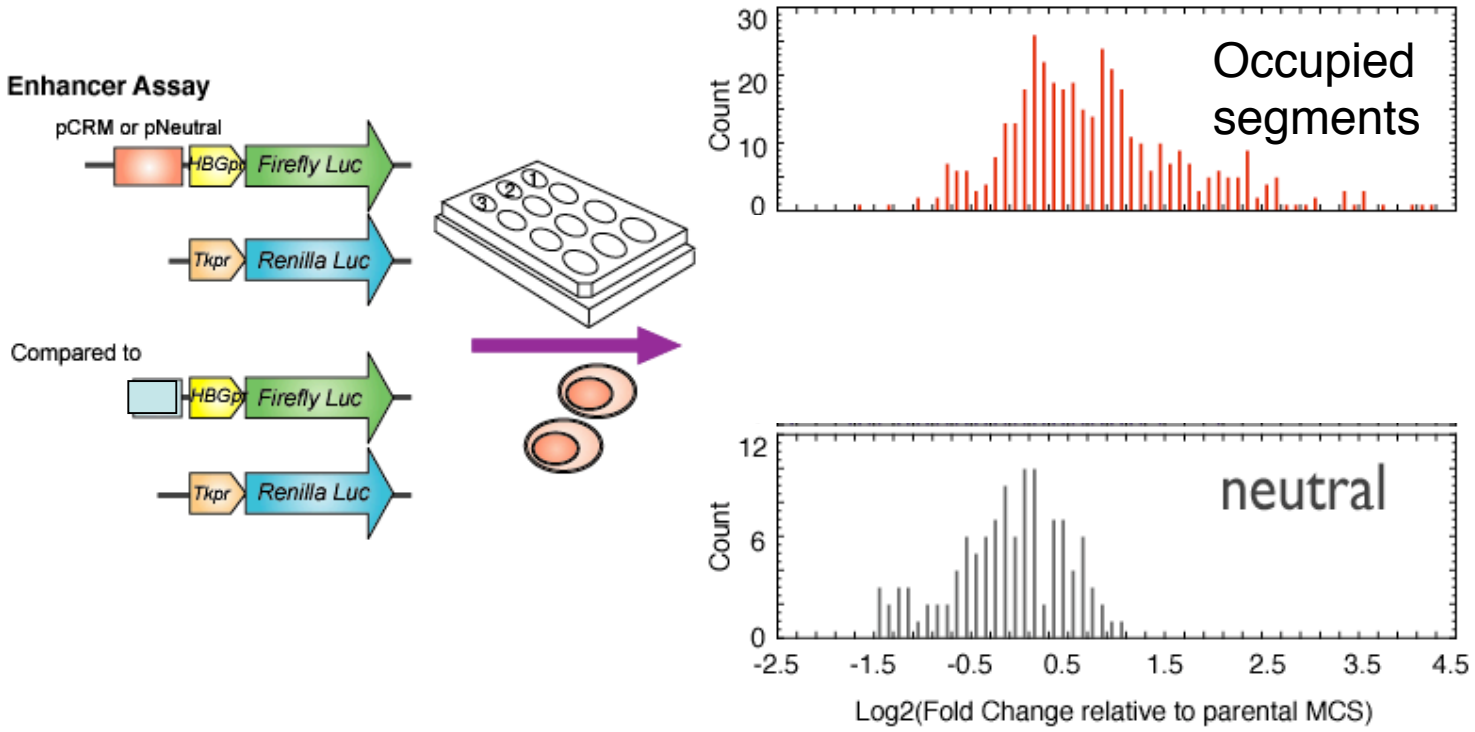
Ying Zhang



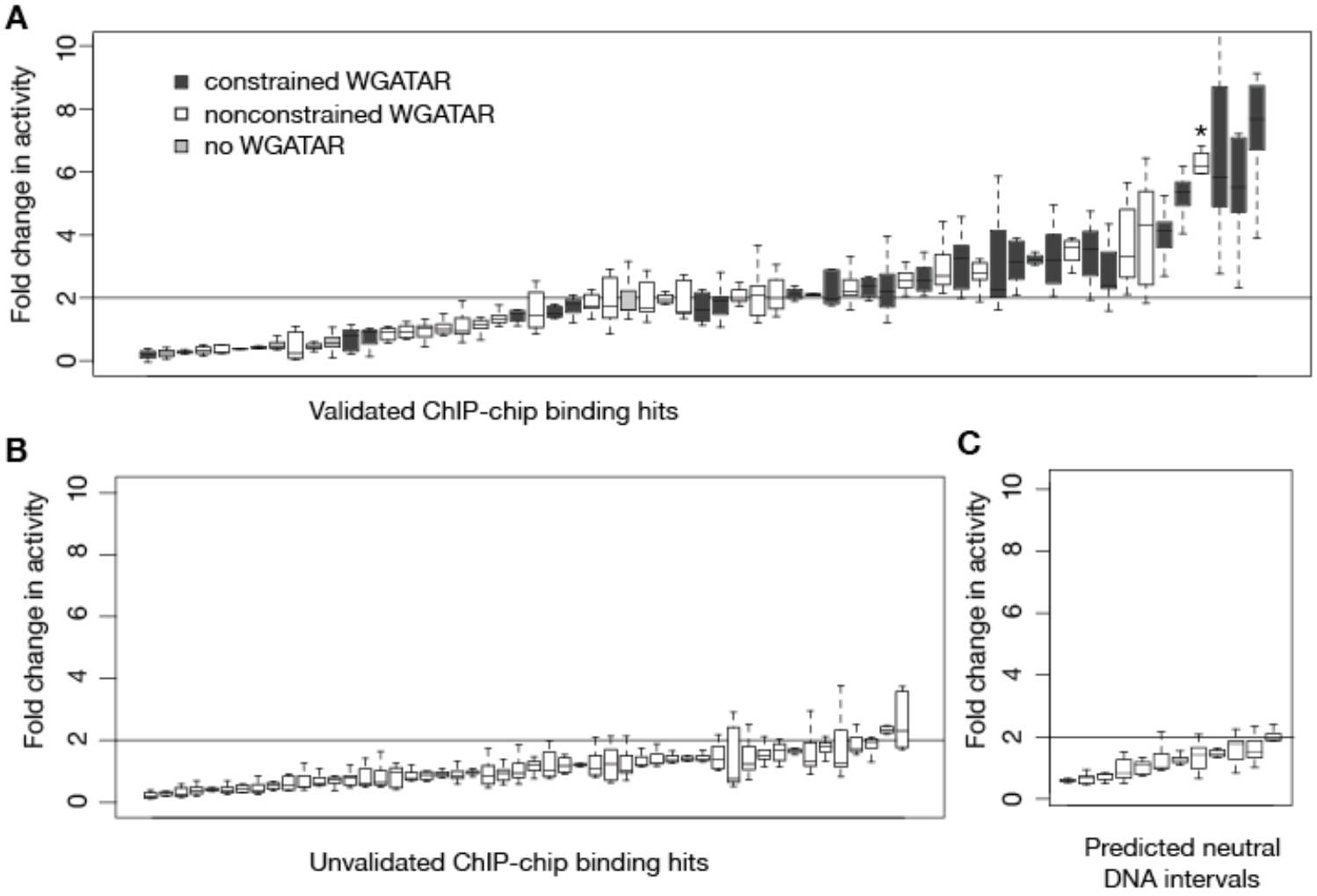
Motif 1: GATA1_bs paired with a 2nd bs_motif of GATA1, EKLF or SP1.
Motif 2: GATA1_bs paired with a 2nd bs_motif of GATA1, EKLF, SP1, CP2 or GABP.
EpiMark: Both Low H3K27me3 and High H3K4me1

DNA segments occupied by GATA-1 were tested for enhancer activity on transfected plasmids

Transiently transfected K562 cells



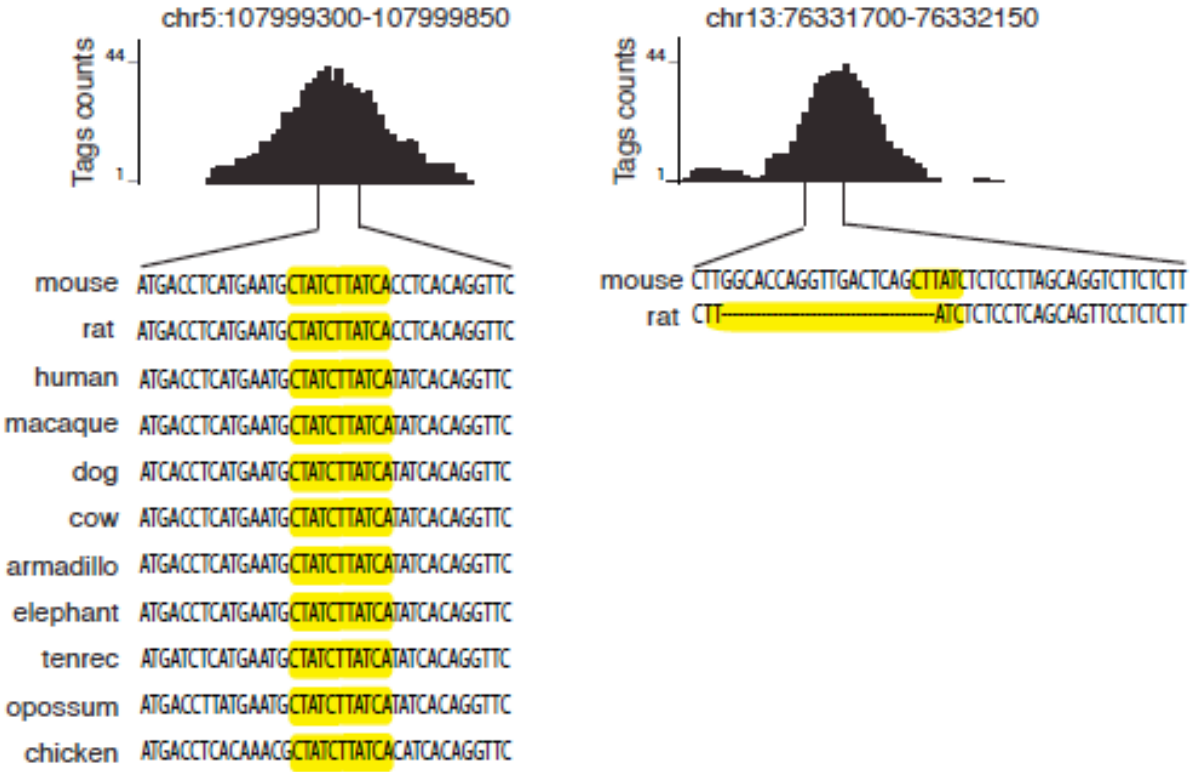
Some of the DNA segments occupied by GATA-1 are active as enhancers



Cheng et al. (2008) *Genome Research* 18:1896-1905

Binding site motifs in occupied DNA segments can be deeply preserved during evolution

Consensus binding site motif for GATA-1: WGATAR or YTATCW

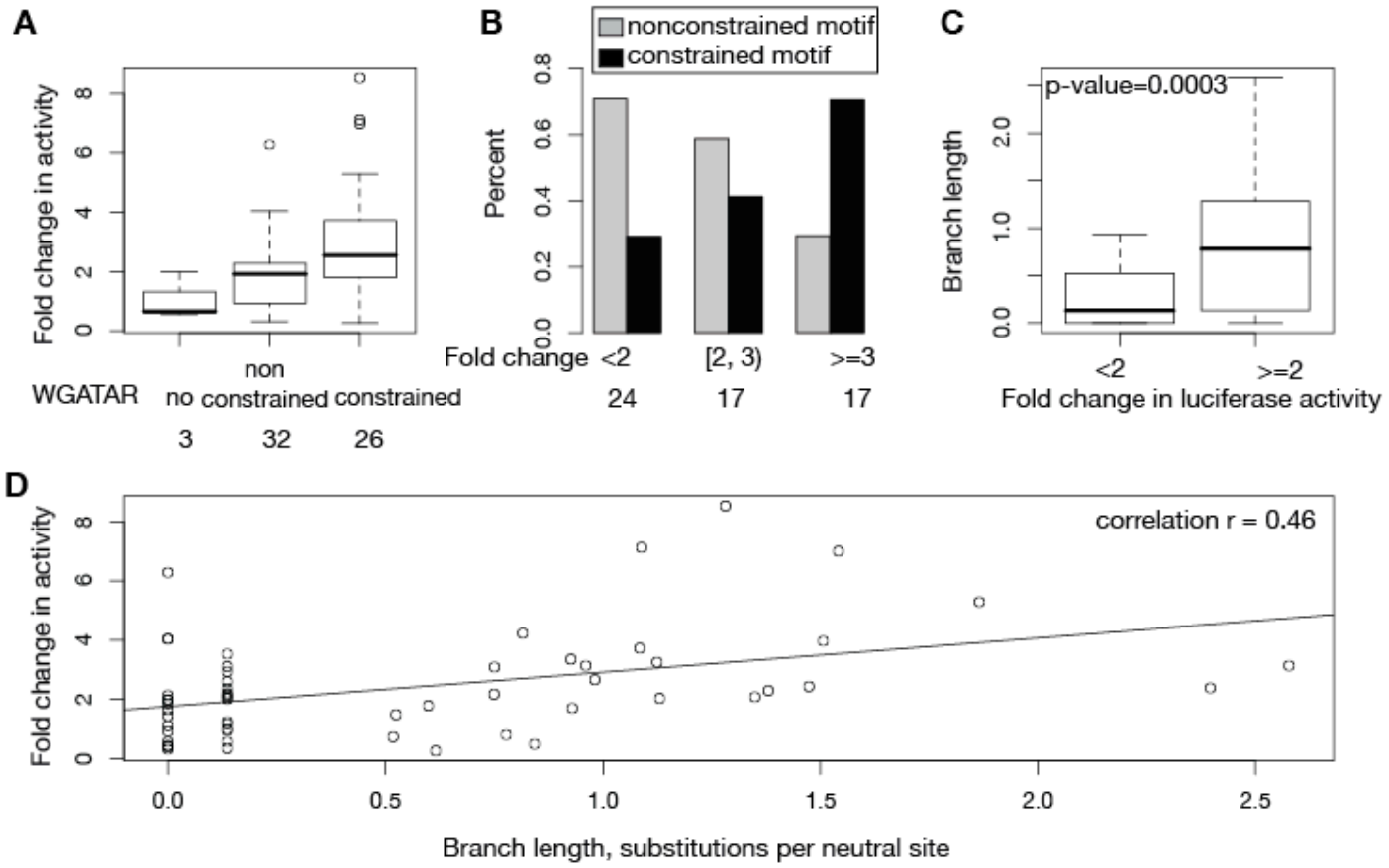


5997
constrained

7308
not constrained

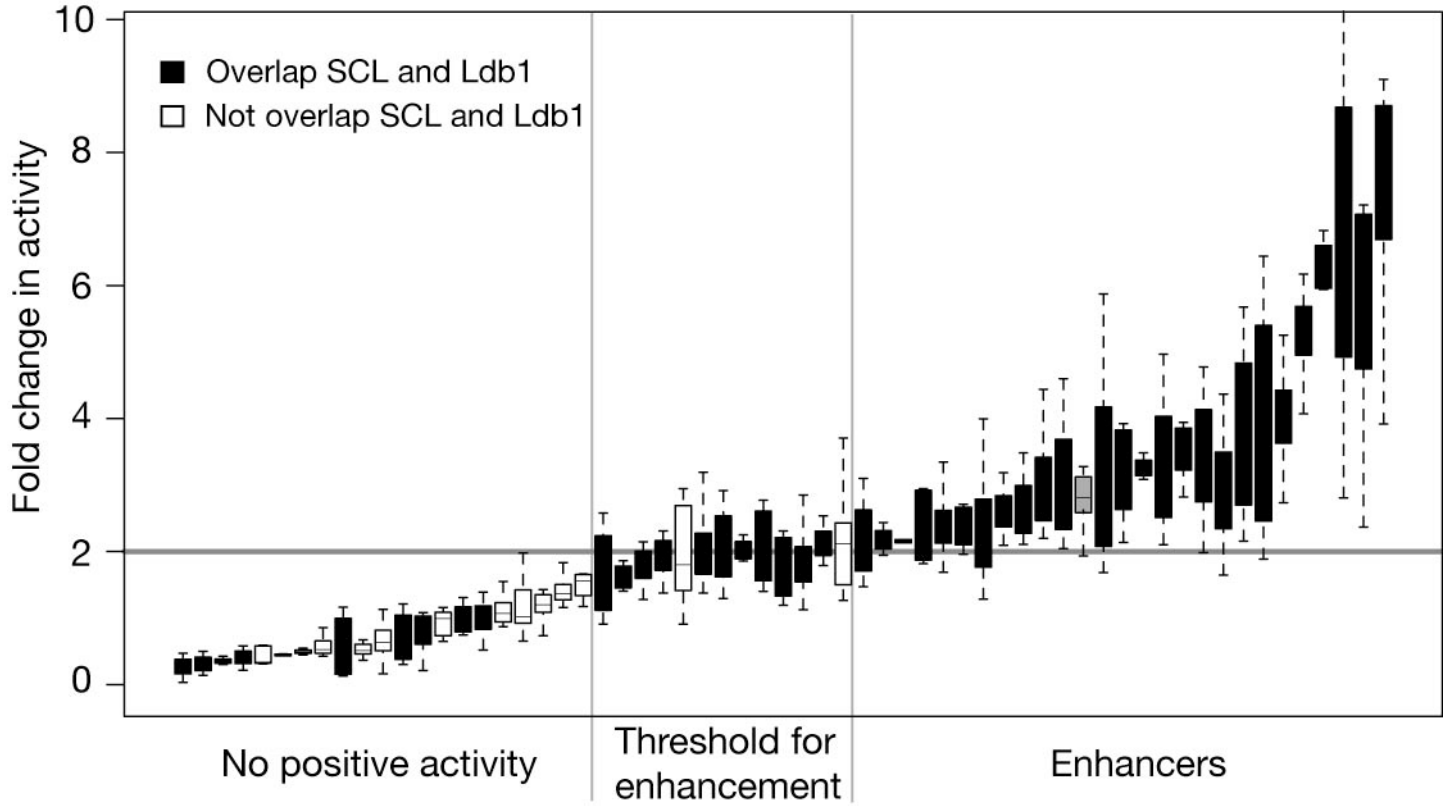
2055
no motif

Constraint on a binding site motif in an occupied DNA segment strongly correlates with enhancement

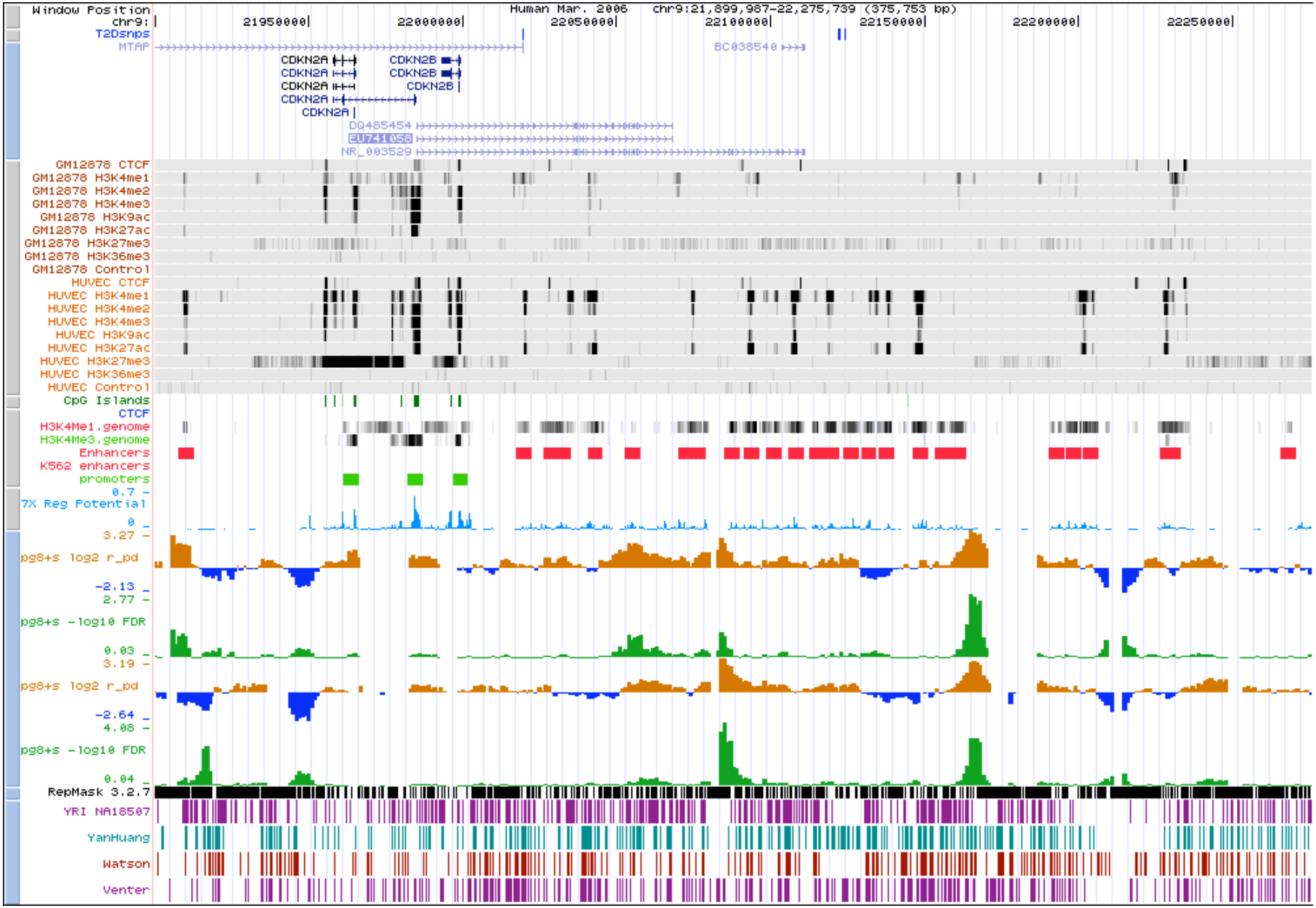


Cheng et al. (2008) *Genome Research* 18:1896-1905

All GATA1-occupied segments active as enhancers are also occupied by SCL and LDB1



Candidate functions in T2D SNP intervals



ENCODE data generates hypotheses for GWAS: *BCL11A* and fetal Hb

Therapeutic value of gene reactivation

- Most hemoglobinopathies results from mutations in the adult form of hemoglobin, HbA
- The fetal form, HbF, works in adults
- Level of HbF is a quantitative trait that is variable in humans
- Individuals that are both homozygous for the sickle cell allele (*HBB-S*) and have high levels of HbF have significantly milder disease
- Can we engineer reactivation of HbF?

QTLs for HbF map to *HBG*, *HBG1L_MYB*, and *BCL11A*

A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15

Stephan Menzel[1], Chad Garner[2], Ivo Gut[3], Fumihiko Matsuda[3], Masao Yamaguchi[3], Simon Heath[3], Mario Foglio[3], Diana Zelenika[3], Anne Boland[3], Helen Rooks[1], Steve Best[1], Tim D Spector[4], Martin Farrall[5], Mark Lathrop[3] & Swee Lay Thein[1,6]

F cells measure the presence of fetal hemoglobin, a heritable quantitative trait in adults that accounts for substantial phenotypic diversity of sickle cell disease and β thalassemia. We applied a genome-wide association mapping strategy to individuals with contrasting extreme trait values and mapped a new F cell quantitative trait locus to *BCL11A*, which encodes a zinc-finger protein, on chromosome 2p15. The 2p15 *BCL11A* quantitative trait locus accounts for 15.1% of the trait variance.

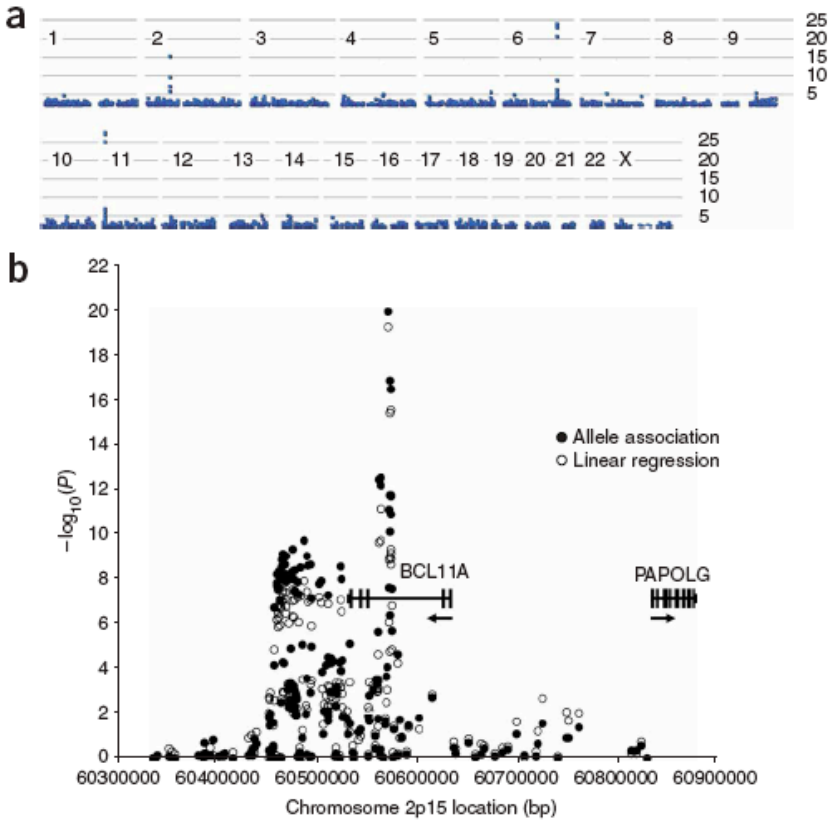


Figure 1 Association statistics ($-\log_{10}(P)$) for individuals included in the genome-wide screening panel. (**a**) Association statistics for 3,225 markers genome-wide with $P < 10^{-2}$. (**b**) Association statistics for 211 markers across the 2p15 region of association.

SNP in *BCL11A* associated with F-cells

BCL11A is a major HbF quantitative trait locus in three different populations with β-hemoglobinopathies ☆

Amanda E. Sedgewick[a, 1], Nadia Timofeev[a, 1], Paola Sebastiani[a], Jason C.C. So[b], Edmond S.K. Ma[b], Li Chong Chan[b], Goonnapa Fucharoen[c], Supan Fucharoen[c], Cynara G. Barbosa[d], Badri N. Vardarajan[e], Lindsay A. Farrer[a, e, f, g, h], Clinton T. Baldwin[e, i], Martin H. Steinberg[d] and David H.K. Chui[d],

(2008) Blood Cells, Molecules and Diseases 41:255-258

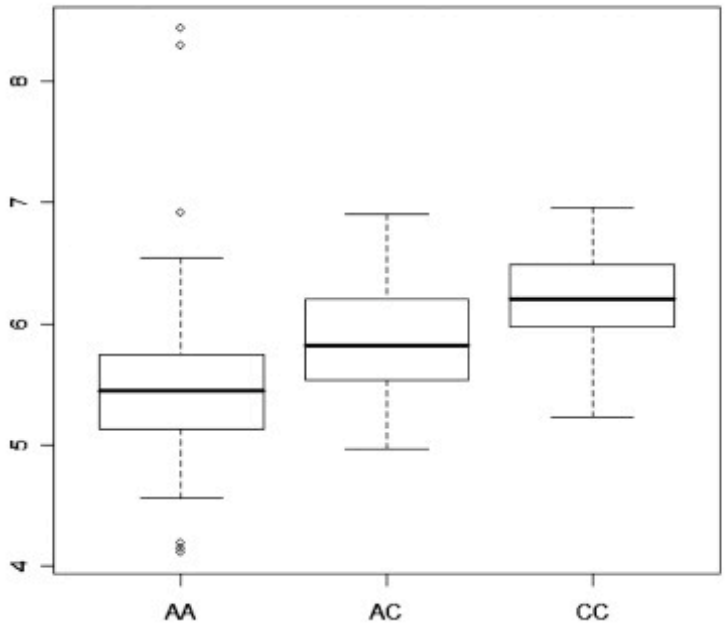


Fig. 1. Box plots showing the complete distribution of F-cells, expressed as 10^9 F-cells per liter of blood in log scale, on the *y*-axis among Chinese adult β-thalassemia heterozygotes, excluding those who were heterozygotes for the β-globin gene promoter nt − 28 A > G β^+-thalassemia mutation and those who were heterozygous for the C > T polymorphism (rs7482144) at the *HBG2* promoter nt − 158 bp. AA, AC, and CC represent the SNP genotypes at rs766432. Each rectangle shows the data between the 25th and 75th quartiles, and the bar in each rectangle is the median value for the F-cells in log scale.

Human Fetal Hemoglobin Expression Is Regulated by the Developmental Stage-Specific Repressor *BCL11A*

Vijay G. Sankaran,[1,2] Tobias F. Menne,[1] Jian Xu,[1] Thomas E. Akie,[1] Guillaume Lettre,[3,4] Ben Van Handel,[5] Hanna K. A. Mikkola,[5] Joel N. Hirschhorn,[3,4] Alan B. Cantor,[1] Stuart H. Orkin[1,2,6]*

Differences in the amount of fetal hemoglobin (HbF) that persists into adulthood affect the severity of sickle cell disease and the β-thalassemia syndromes. Genetic association studies have identified sequence variants in the gene *BCL11A* that influence HbF levels. Here, we examine *BCL11A* as a potential regulator of HbF expression. The high-HbF *BCL11A* genotype is associated with reduced *BCL11A* expression. Moreover, abundant expression of full-length forms of *BCL11A* is developmentally restricted to adult erythroid cells. Down-regulation of *BCL11A* expression in primary adult erythroid cells leads to robust HbF expression. Consistent with a direct role of *BCL11A* in globin gene regulation, we find that *BCL11A* occupies several discrete sites in the β-globin gene cluster. *BCL11A* emerges as a therapeutic target for reactivation of HbF in β-hemoglobin disorders.

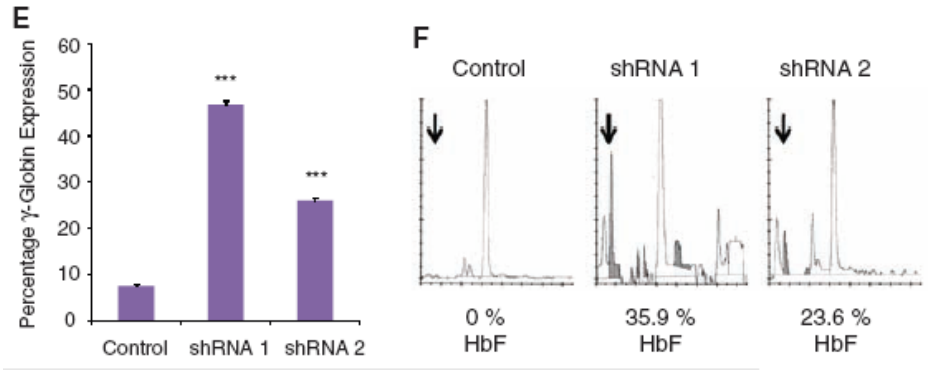
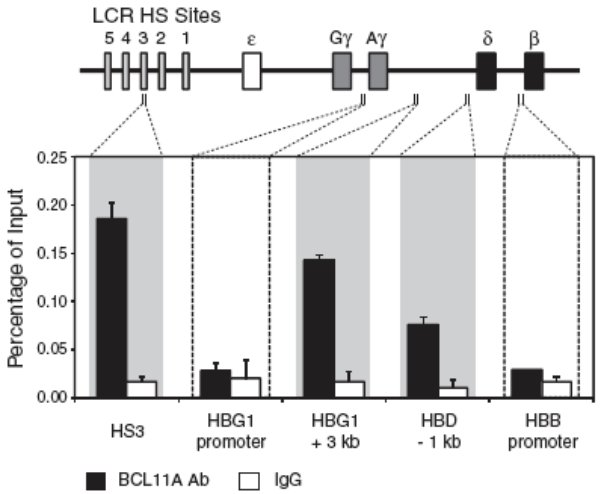


Fig. 4. *BCL11A* occupies discrete regions in the human β-globin locus in adult erythroid progenitors. The human β-globin locus is depicted at the top with regions showing significant binding shaded in gray in the histogram plot below. The results are means ± SD (n = 3 per group).



(2008) Science 322:1839

ENCODE data: Hints for regulation of *BCL11A*

SNPs associated with high HbF
Red=yes, blue=no

GenCode *BCL11A*

RNA seq

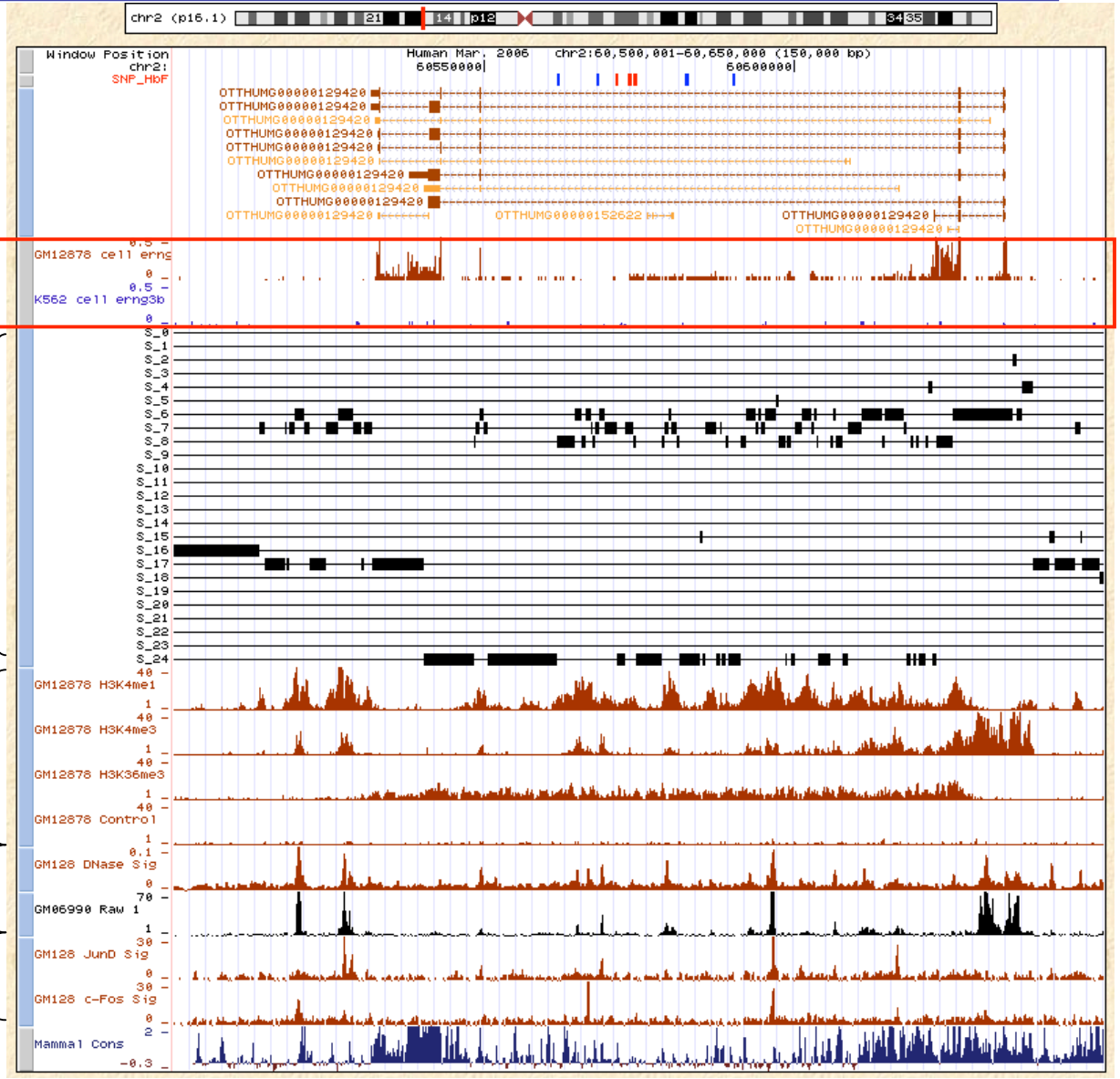
25-state HMM model integrating
33 ENCODE consortium datasets

Histone H3 modifications

DNase HSs

ChIPseq TF binding

Mammalian conservation



Transcription is high in lymphoid GM12878 cell and low in K562, consistent with gamma-globin expression in K562

Antisense transcript in *BCL11A*

SNPs associated with high HbF
Red=yes, blue=no

GenCode *BCL11A*

RNA seq

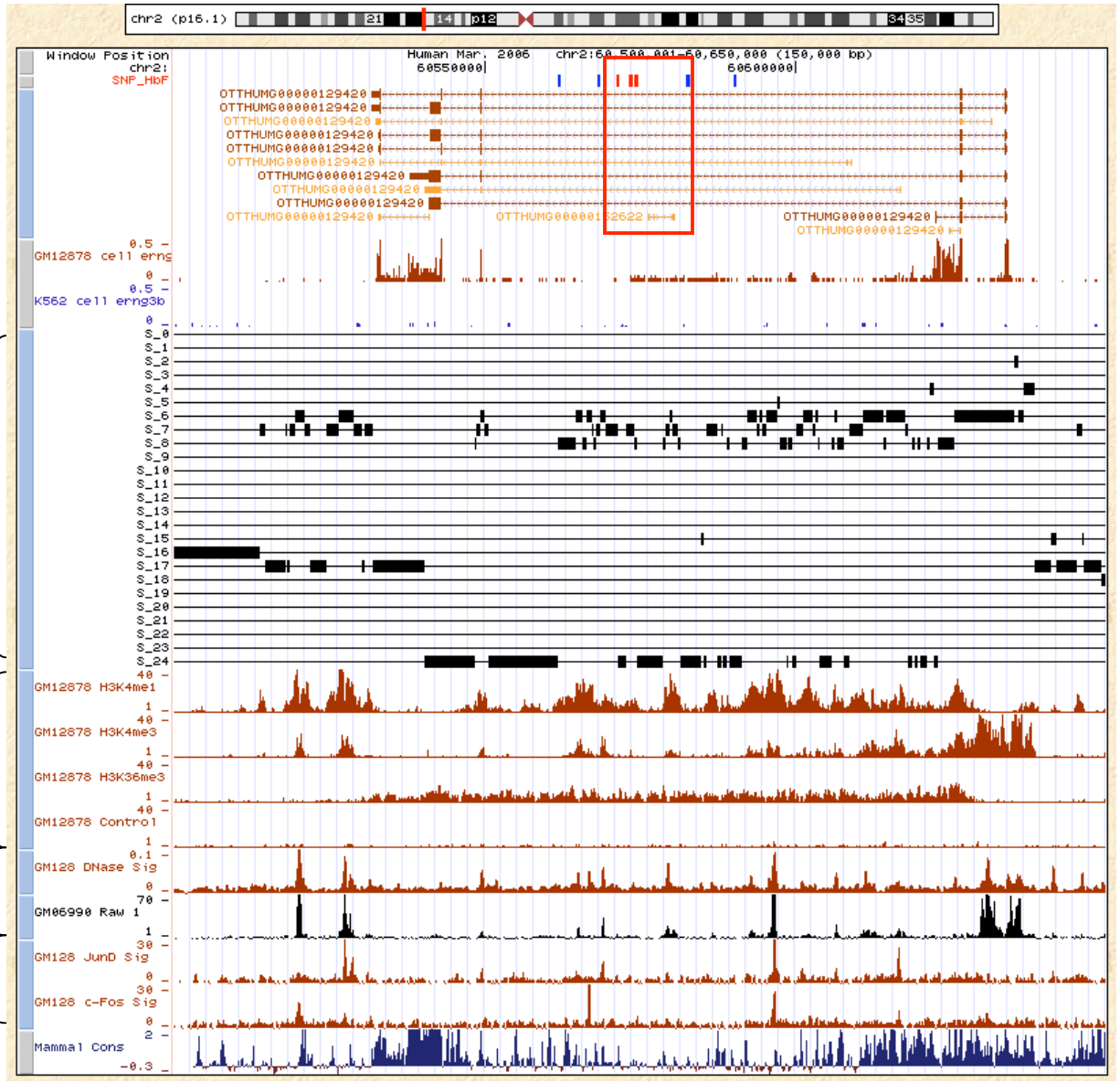
25-state HMM model integrating
33 ENCODE consortium datasets

Histone H3 modifications

DNase HSs

ChIPseq TF binding

Mammalian conservation



Intronic SNPs are close to predicted enhancers

SNPs associated with high HbF
Red=yes, blue=no

GenCode *BCL11A*

RNA seq

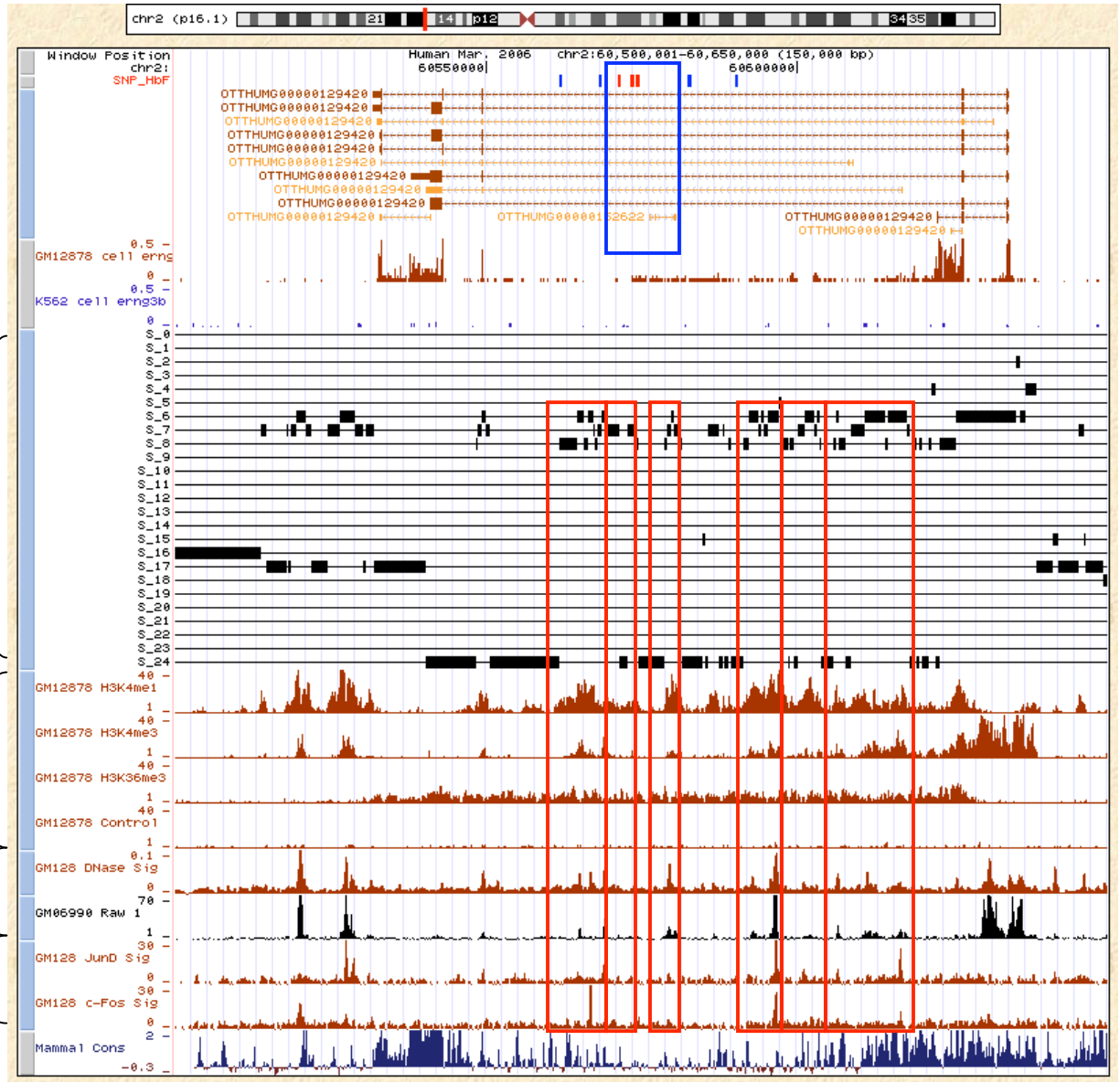
25-state HMM model integrating
33 ENCODE consortium datasets

Histone H3 modifications

DNase HSs

ChIPseq TF binding

Mammalian conservation



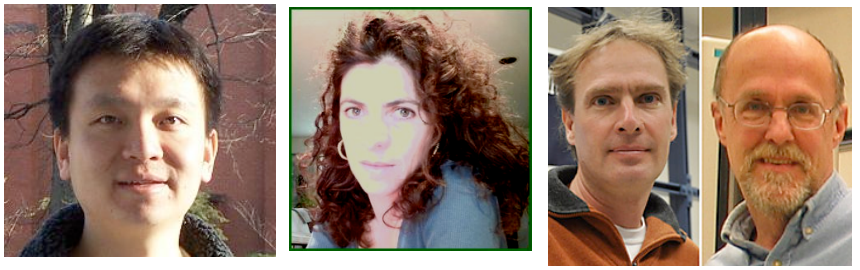
The intronic enhancers are predicted by H3K4me1, DNase HSs and occupancy by Jun and Fos.

Summary: Genomics of Gene Regulation

- Genetic determinants of variation in expression levels may contribute to complex traits - phenotype is not just determined by coding regions
- Biochemical features associated with cis-regulatory modules are being determined genome-wide for a range of cell types.
- These can be used to predict CRMs, but occupancy alone does not necessarily mean that the DNA is actively involved in regulation.
- Evolutionary preservation of binding site motifs within regions containing other indicators of CRMs (e.g. regulatory potential or protein occupancy) is a good predictor of function.
- Genome-wide data on biochemical signatures of functional sequences (DHS, chromatin modifications, transcription factor occupancy, transcripts, etc.) provide candidates for explaining how variants in noncoding regions contribute to phenotypes

Many thanks …

Ying Zhang, Swathi Kumar, Weisheng Wu, David King, Kuan-Bei Chen, Yong Cheng, Belinda Giardine, Cathy Riemer, Demesew Abebe, Christine Dorman, Cheryl Keller Capone



Yu Zhang, Francesca Chiaromonte, Stephan Schuster, Webb Miller
Collaborators in CCGB, PSU

James Taylor, Anton Nekrutenko
Galaxy

Mitch Weiss, Gerd Blobel
Childrens' Hospital of Philadelphia

ENCODE
consortium

Funding from NIDDK, NHGRI, Huck Institutes of Life Sciences at PSU