UQ-Brisbane SISG

Module 9: Gene Expression and Epigenetic Profiling

Tuesday February 14, 2017

"Epigenomics"

Greg Gibson

Georgia Institute of Technology

greg.gibson@biology.gatech.com

Content of the Lecture

Epigenomics
ENCODE
Roadmap
ModENCODE
IHEC

Functional Annotation
Enrichment
Regulome
CADD
CATO

GWAS & EpiQTL
EpiWAS
meQTL
hQTL
ccQTL

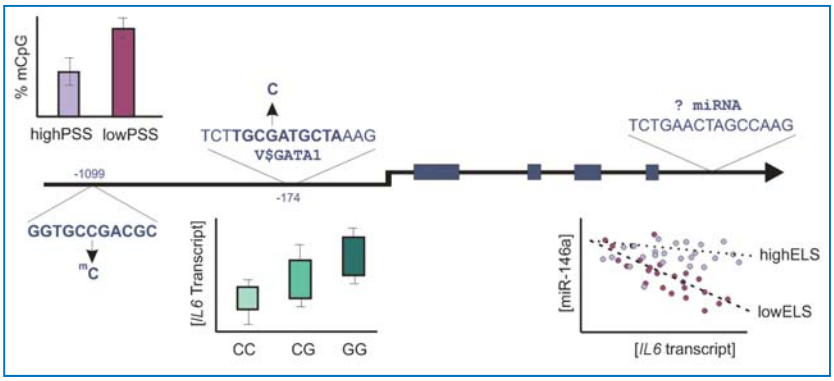
Association
CoLoc
FM QTL
Joint GWAS

1. Epigenome Projects from ENCODE to IHEC
2. Annotation of regulatory function
3. EpiWAS and the genetics of epigenome regulation

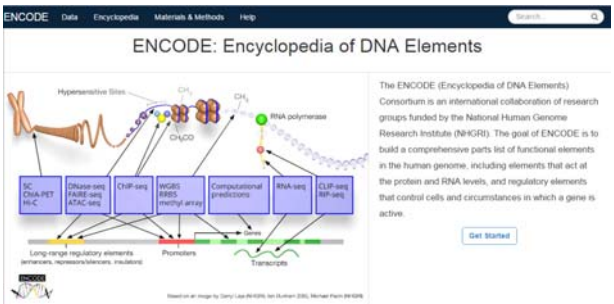
After the Break:

4. Colocalization of eSNPs and GWAS SNPs

The integrative nature of transcriptional regulation

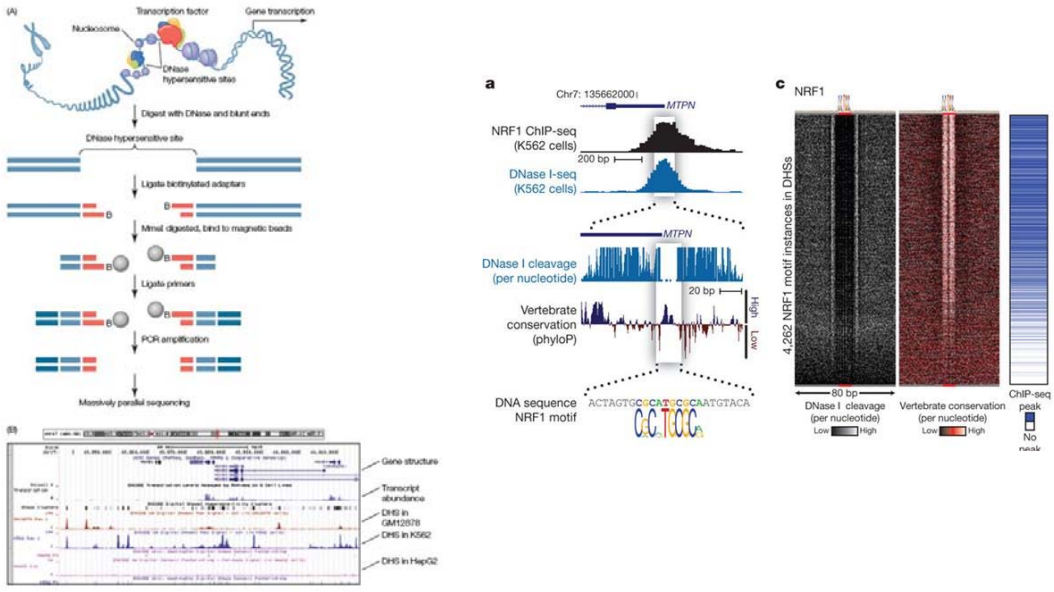


https://www.encodeproject.org/

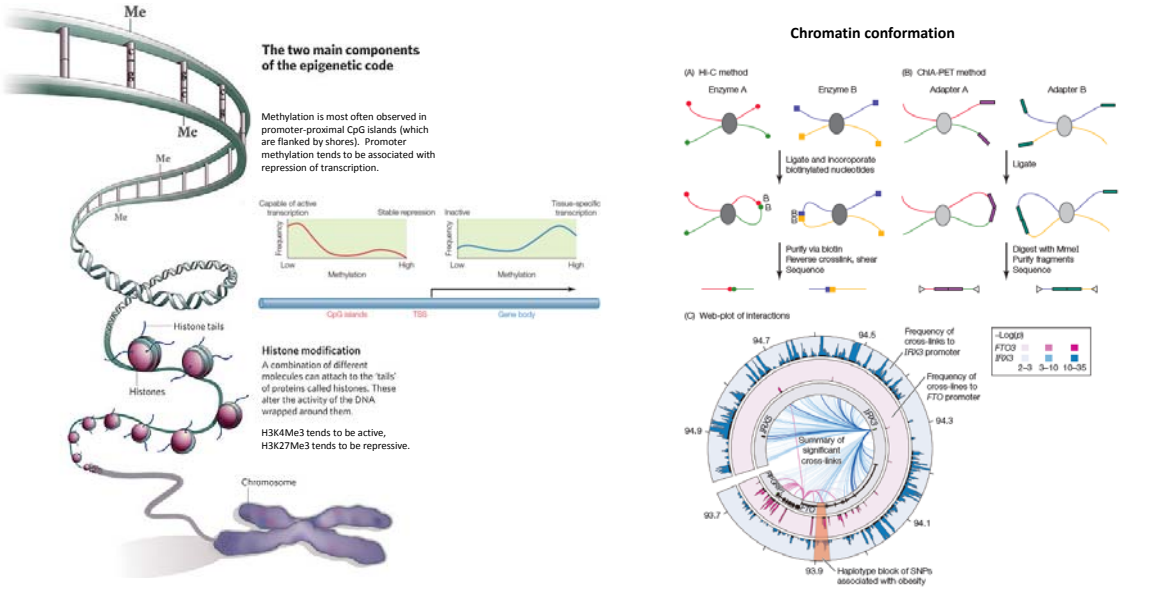


The ENCODE Project Consortium (2011) *PLOS Biology* **9**: 1001046

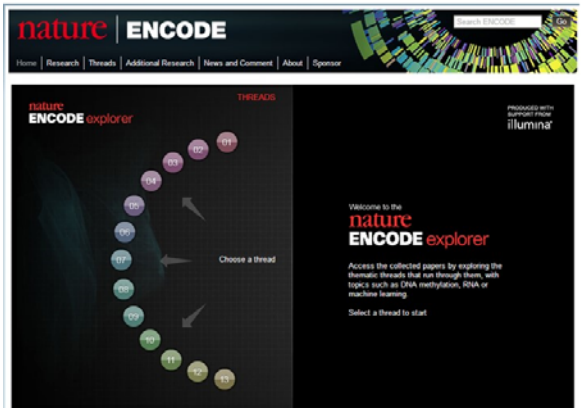
DHS and TFBS: DNAse hypersensitive sites and TF Binding



Three modes of epigenetic regulation



ENCODE *Nature* threads 2012

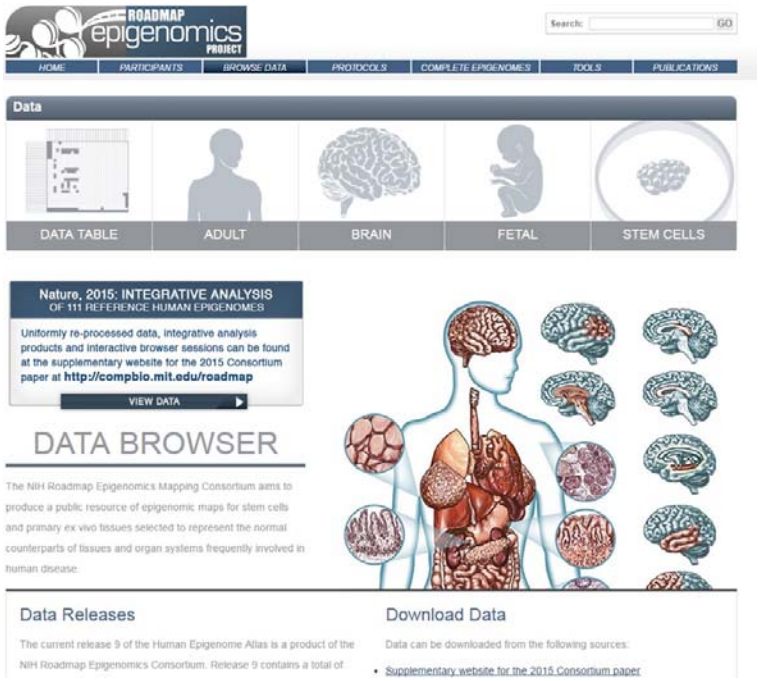


Thread	Topic
1	Transcription Factor Motifs
2	Chromatin patterns at Transcription Factor Binding Sites
3	Characterization of Intergenic Regions and Gene definition
4	RNA and Chromatin Modification patterns around Promoters
5	Epigenetic regulation of RNA Processing
6	Non-coding RNA characterization
7	DNA methylation
8	Enhancer discovery and characterization
9	Three-Dimensional connections across the Genome
10	Characterization of Network Topology
11	Machine Learning Approaches to Genomics
12	Impact of Functional Information on understanding Variation
13	Impact of Evolutionary Selection on functional regions

30 Papers published in June 2012 (Nature, Genome Biology, Genome Research)

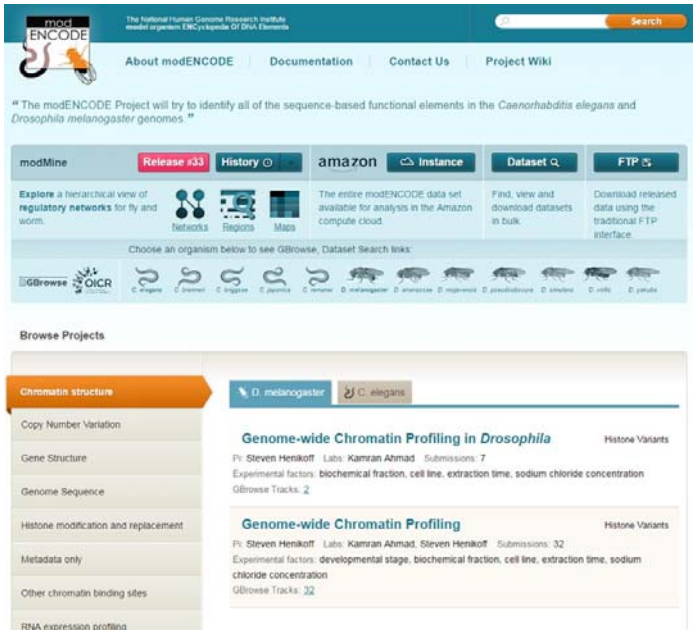
http://www.nature.com/encode/#/threads

Roadmap Epigenomics Consortium



http://www.roadmapepigenomics.org/

Model Organism ENCODE



http://www.modencode.org/

International Human Epigenome Consortium

http://ihec-epigenomes.org/

IHEC *Cell* threads 2016

24 Papers published in Nov 2016 (Cell, Cell Reports, Cell Stem Cell, Cancer Cell)

http://www.cell.com/consortium/IHEC

Enrichment of regulatory elements at GWAS loci

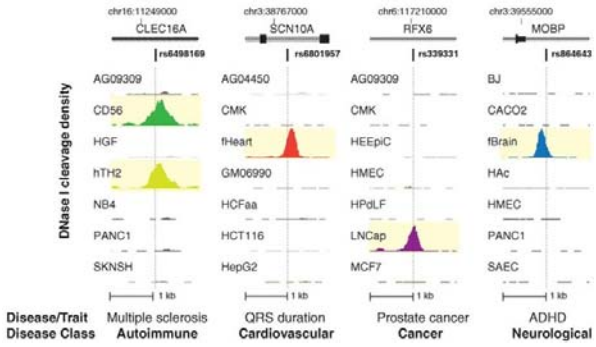
93% of GWAS peak SNPs are located in regulatory regions rather than affecting the protein sequence

Maurano et al performed DNAse-Seq on 349 cell and tissue samples, identifying ~ 200,000 DHS per sample (2% of DNA)

75% of 5,130 GWAS peak SNPs are in a DHS, many specifically in a tissue expected to relate to pathology

419 of these pair with active promoters by Chia-PET, 40% acting over 250kb and 80% not with the closest gene

20% - 40% show allelic imbalance for chromatin accessibility



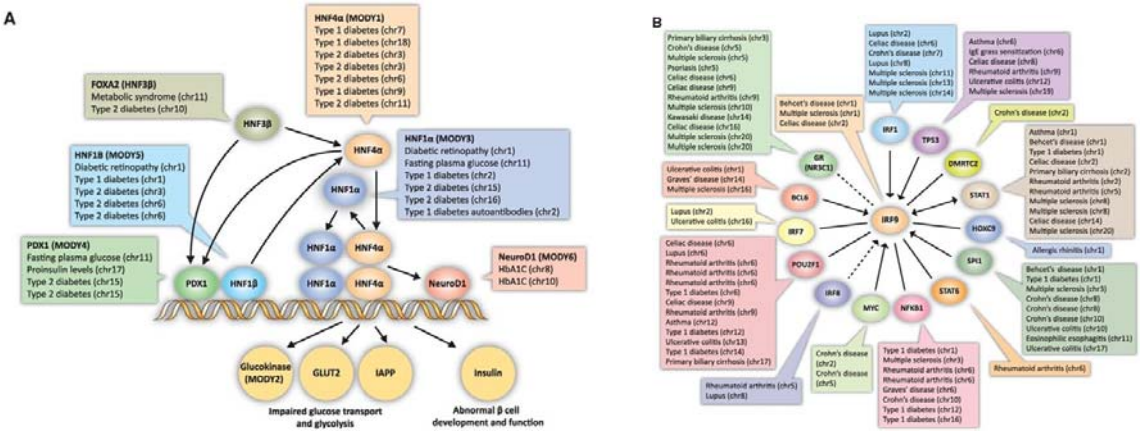
Maurano et al (2012) *Science* **337**: 1190-1195 "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA"

Disease associations cluster in regulatory pathways

(A) Monogenic diabetes locus TFBS are enriched at GWAS / DHS sites for Types 1 and 2 diabetes

(B) Transcription factors associated with multiple autoimmune diseases are enriched at GWAS / DHS sites

Similar results observed for several types of cancer and neurological disorders



Maurano et al (2012) *Science* **337**: 1190-1195 "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA"

RegulomeDB annotation of likely regulatory function

http://regulome.stanford.edu/index

RegulomeDB is an index from the Snyder lab at Stanford that summarizes evidence from:

- eQTL
- TF binding (ChIP data)
- TF motif informatics
- DHS footprints or peaks

The average human genome has ~25,000 homozygous Category 1 or 2 variants that potentially affect gene expression

The score can be used to refine credible intervals by focusing on a few percent of the candidate SNPs in a locus

Table 2. RegulomeDB variant classification scheme

Category scheme		
Category		**Description**
		Likely to affect binding and linked to expression of a gene target
1a		eQTL + TF binding + matched TF motif + matched DNase footprint + DNase peak
1b		eQTL + TF binding + any motif + DNase footprint + DNase peak
1c		eQTL + TF binding + matched TF motif + DNase peak
1d	<1%	eQTL + TF binding + any motif + DNase peak
1e		eQTL + TF binding + matched TF motif
1f		eQTL + TF binding/DNase peak
		Likely to affect binding
2a		TF binding + matched TF motif + matched DNase footprint + DNase peak
2b	2%	TF binding + any motif + DNase footprint + DNase peak
2c		TF binding + matched TF motif + DNase peak
		Less likely to affect binding
3a	1%	TF binding + any motif + DNase peak
3b		TF binding + matched TF motif
		Minimal binding evidence
4	5%	TF binding + DNase peak
5	18%	TF binding or DNase peak
6	30%	Motif hit

Lower scores indicate increasing evidence for a variant to be located in a functional region. Category 1 variants have equivalents in other categories with the additional requirement of eQTL information.

Boyle et al (2012) *Genome Research* **22**: 1790-1797 "Annotation of functional variation in personal genomes using RegulomeDB"

CADD score annotation of likely deleteriousness

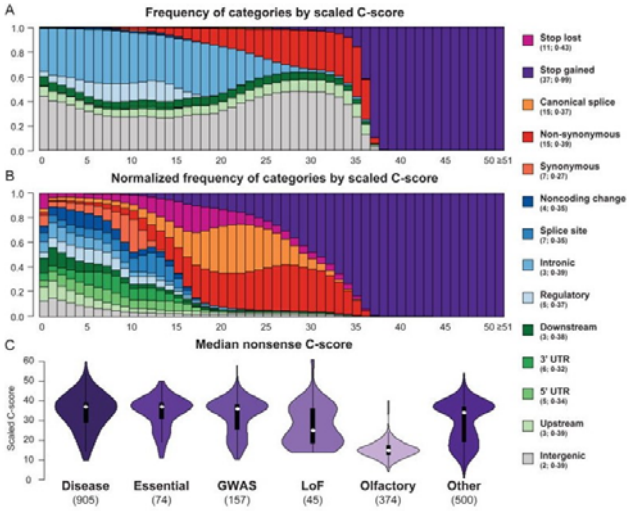
http://cadd.gs.washington.edu/

CADD (combined annotation dependent depletion) is an index from the Shendure lab at UW that summarizes evidence from 63 annotations encompassing:

- Functional or regulatory annotation
- Allele frequency and diversity
- Evolutionary conservation

The raw C-score is scaled to a relative CADD score as the –10*log10(rank/total), namely:
30 is the top 0.1% of likely deleterious
20 is in the top 1%
10 is in the top 10%

The score attempts unbiased prediction of "deleteriousness", based on machine learning comparison of 15M observed and simulated human variants



Kircher et al (2014) *Nature Genetics* **46**: 310-315 "A general framework for estimating the relative pathogenicity of human genetic variants"

CATO annotation of likely regulatory function - I

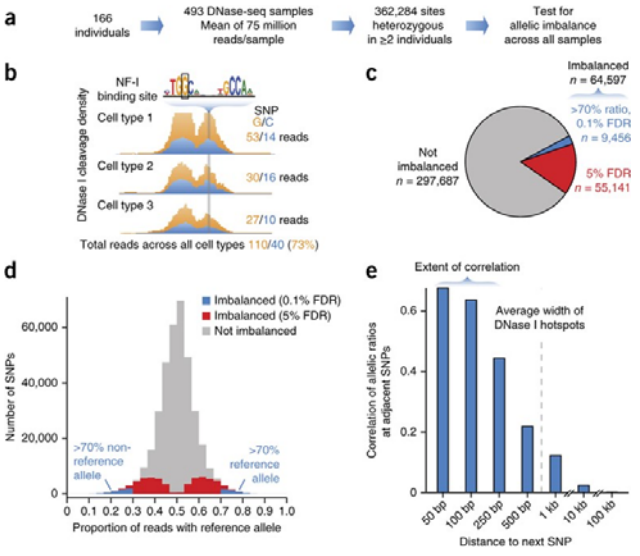
http://www.uwencode.org/proj/CATO/

CATO (contextual analysis of transcription factor occupancy) is an index from the Stomatoyannopoulos lab at UW that summarizes evidence from allelic imbalance of DHS assays for 493 samples

- 114 cell types in 166 people
- 60,000 variants imbalanced at 5% FDR
- > 2/5 of sites cell-type dependent

Matched imbalance to 2203 TFBS motifs for 825 genes, 44 of which are enriched

Interestingly, most SNPs do not disrupt DHS, implying that there is buffering that reduces the impact of polymorphism, particularly near the TSS



Maurano et al (2015) *Nature Genetics* **47**: 1393-1402 "Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo"

CATO annotation of likely regulatory function - II

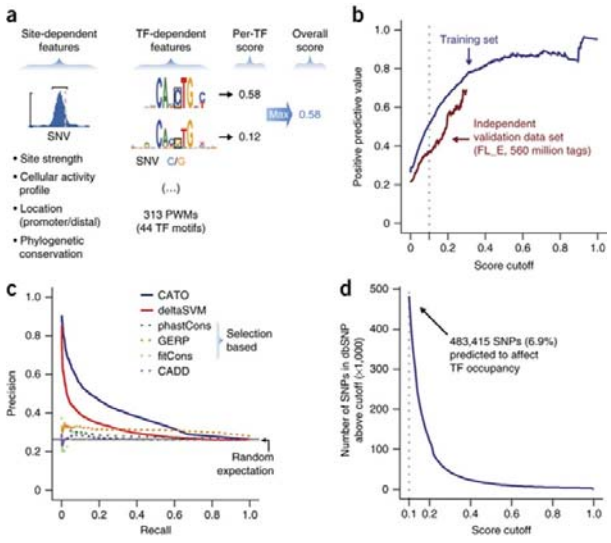
http://www.uwencode.org/proj/CATO/

Based on the training set of SNPs in TFBS that show allelic imbalance, Maurano et al used machine learning to predict the likelihood that regulatory SNPs affect enhancer occupancy.

- Cell-type specific imbalance
- Location of DHS
- Evolutionary conservation
- TF-specific profiles

Used this to predict almost 500,000 SNPs genome-wide that are likely to affect TF occupancy and hence influence transcription

The CATO (contextual analysis of transcription factor occupancy) score highlights about 1.5% of all non-coding SNPs, but has not yet been validated with respect to RNASeq data and GWAS



Maurano et al (2015) *Nature Genetics* **47**: 1393-1402 "Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo"

Some (concise) definitions

GWAS: Genome-wide association study – search for SNPs significantly associated with a trait (eSNPs)

TWAS: Transcriptome-wide association study – search for transcripts significantly associated with a trait (QTT)

EpiWAS: Epigenome-wide association study – search for epigenetic marks significantly associated with a trait (EWAS also used, but earlier used to refer to Environment-wide association study)

eQTL: a SNP which influences the abundance of a transcript. Cis-eQTL act locally (~ within ± 500kb)

eGene: a gene whose transcript abundance is regulated by a locally-acting SNP

meQTL: a genotype which is associated with the degree of methylation at a CpG site

Methyl ß: typical measure of the degree of methylation, ranging from 0 to 1 (none to complete)

hQTL: a genotype that is associated with the intensity of a histone mark (may be acetylation or methylation)

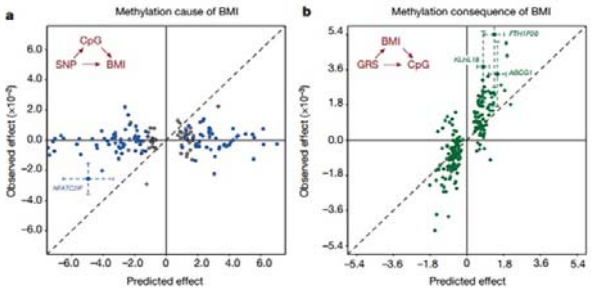
ccQTL: a genotype that influences the level of chromatin conformation / cross-linking

Epigenome-Wide Association Studies (EpiWAS) for Metabolic Disease

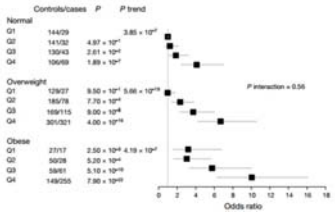
Methyl450 array study of whole blood DNA for 5,387 Europeans and Asians
Identified 278 CpG sites in 207 genes associated with BMI at $p<10^{-7}$: consistent across ethnicities, 90% replicated

Similar effects observed in T cells and neutrophils in independent sample of 60 adults,
about half of the sites also associated with BMI in fat, liver, muscle

However, Mendelian randomization of SNPs that associate with both BMI and methylation level (meQTL)
implies that only a single site is causal – the majority are responsive to obesity
and in turn are explained by variation in blood glucose and lipids which may mediate the methylation



Methylation Risk Score predicts T2D somewhat independent of classical risk factors



Wahl et al (2016) *Nature* **541**: 81-85 "Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity"

meQTL for Inflammatory Bowel Disease - I

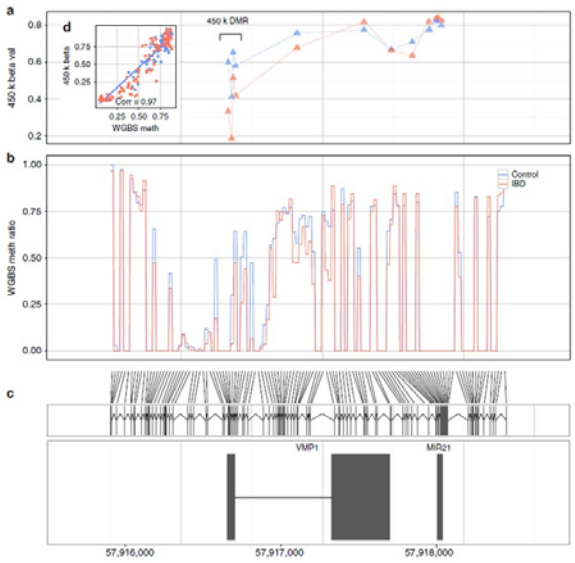
121 CD, 119 UC, 191 Healthy whole blood samples

Whole genome bisulfite sequencing (WGBS) of significantly improves resolution over arrays, and contrasts DMRs (regions with >2 CpG within 2kb) with DMPs (CpG positions) at the VMP1 locus

This association was not alleviated by immunotherapy treatment

There was a significant enrichment of DMPs in the vicinity of IBD GWAS loci, and 74 of the 439 DMPs have meQTL (next slide), some of which are cell-type specific

Multi-CpG composite Methylation Risk Scores strongly predicts CD



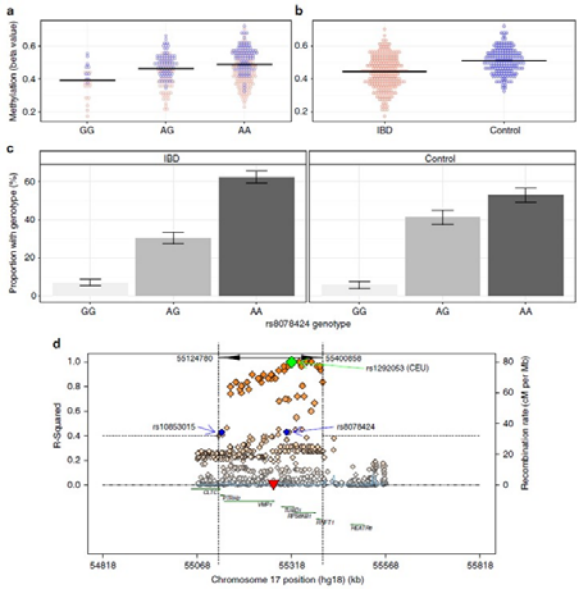
Ventham et al (2016) *Nature Communications* **7**: 13507 "Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in IBD"

meQTL for Inflammatory Bowel Disease - II

VMP1 methylation is influenced by an meQTL, and associates with IBD

An meQTL SNP associates with IBD

Two meQTL SNPs are in mild LD with the GWAS SNP, and flank the CpG site



Ventham et al (2016) *Nature Communications* **7**: 13507 "Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in IBD"

Epigenetics of Blood Gene Expression

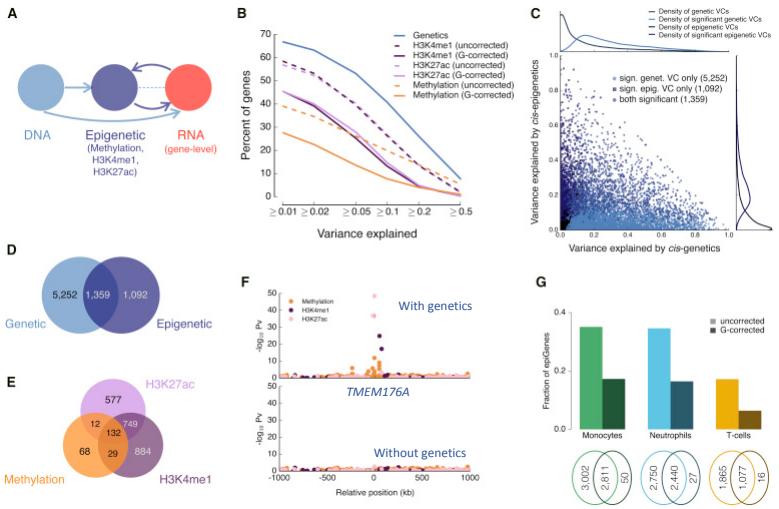
"Variance Decomposition Models": → Genetic effects explain more of the gene expression variance (14% on avg) than epigenetic effects (2% me, 3% ha, 3% hm) At around a third of all loci

EpiWAS asspociations are thus slightly more than half the time due to genotypes influencing both the epi mark and the phenotype

eQTL tend to map near TSS
meQTL tend to map near repressed TSS
haQTL tend to map near active enhancers

Chen et al (2016) *Cell* **167**: 1398-1414 "Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells"

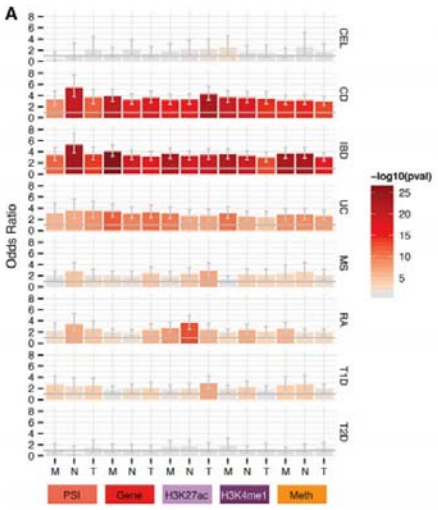
epiQTL for Gene Expression in monocytes, neutrophils, and T-cells



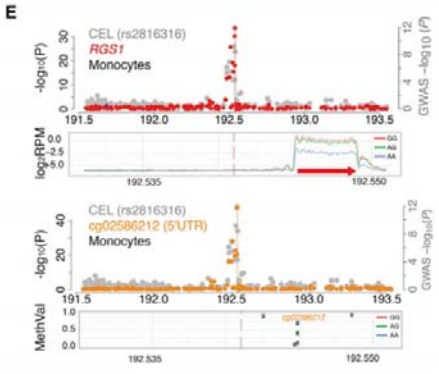
Chen et al (2016) *Cell* **167**: 1398-1414 "Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells"

epiQTL for autoimmune disease in monocytes, neutrophils, and T-cells

eQTL and epiQTL are enriched in cell types with autoimmune disease risk



Use "CoLoc" to map SNPs that jointly associate with the disease or with an eQTL or epiQTL



GWAS

eQTL

meQTL

At least two thirds of 345 disease-colocalized loci involved a DNA methylation or histone modification QTL without a corresponding Eqtl – implies independent regulatory mechanisms, but how?!

Chen et al (2016) *Cell* **167**: 1398-1414 "Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells"