Summer Institute
In Statistical Genetics
2017

Integrative Genomics
5a. Epigenetics and Single Cell RNAseq

ggibson.gt@gmail.com
http://www.gibsongroup.biology.gatech.edu

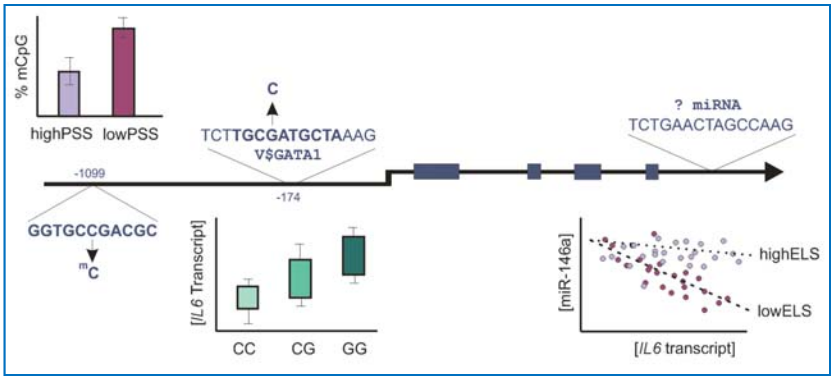
Georgia Tech

Center for Integrative Genomics

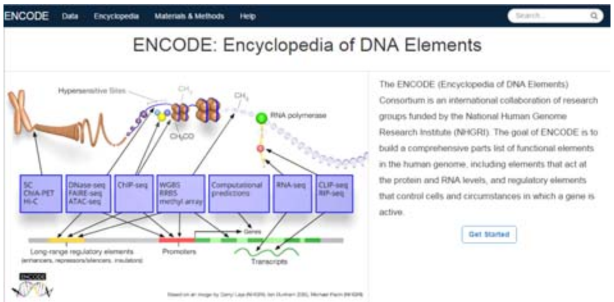
Content of the Lecture

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

The integrative nature of transcriptional regulation



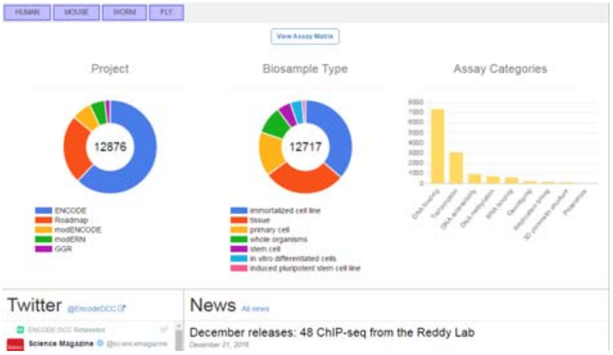
https://www.encodeproject.org/



ENCODE: Encyclopedia of DNA Elements

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

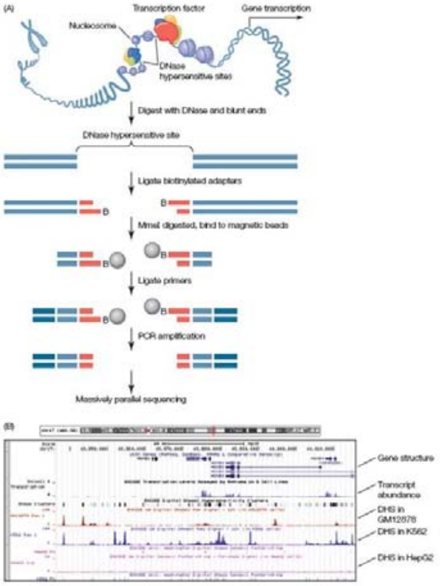
Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

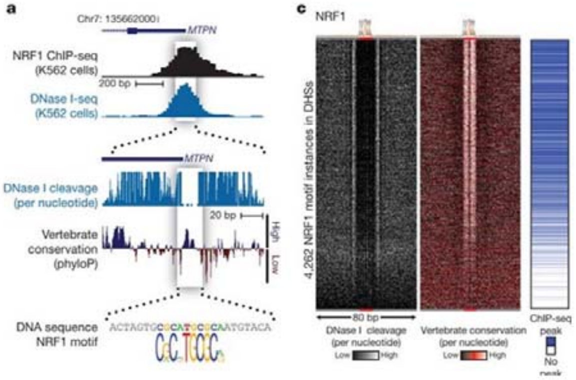


December releases: 48 ChIP-seq from the Reddy Lab
December 21, 2016

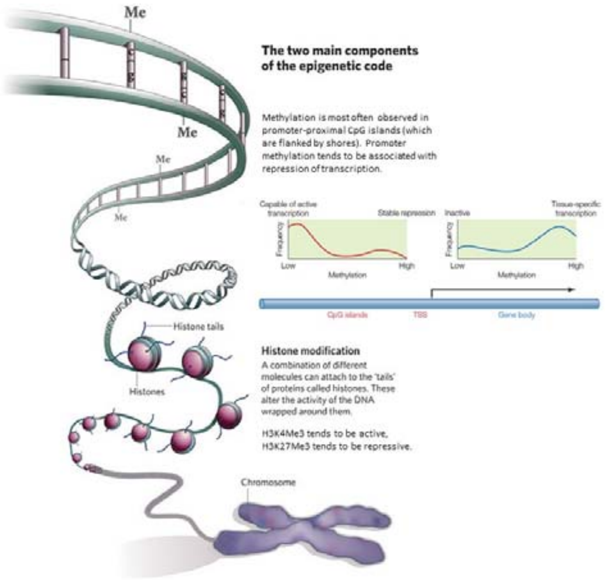
The ENCODE Project Consortium (2011) *PLOS Biology* **9**: 1001046

DHS and TFBS: DNAse hypersensitive sites and TF Binding


(A)
Transcription factor
Gene transcription
Nucleosome
DNase hypersensitive sites
Digest with DNase and blunt ends
DNase hypersensitive site
Ligate biotinylated adapters
B
MmeI digested, bind to magnetic beads
Ligate primers
PCR amplification
Massively parallel sequencing
(B)
Gene structure
Transcript abundance
DHS in GM12878
DHS in K562
DHS in HepG2



Three modes of epigenetic regulation



The two main components of the epigenetic code

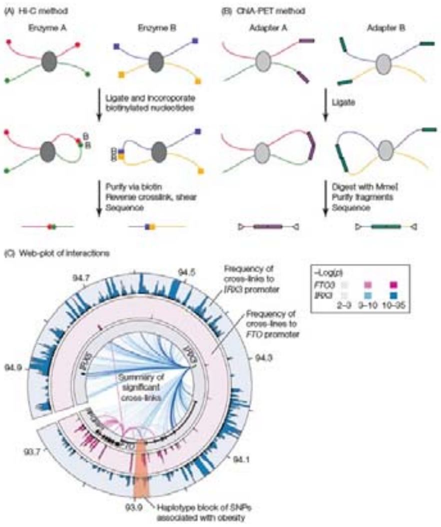
Methylation is most often observed in promoter-proximal CpG islands (which are flanked by shores). Promoter methylation tends to be associated with repression of transcription.

Histone modification

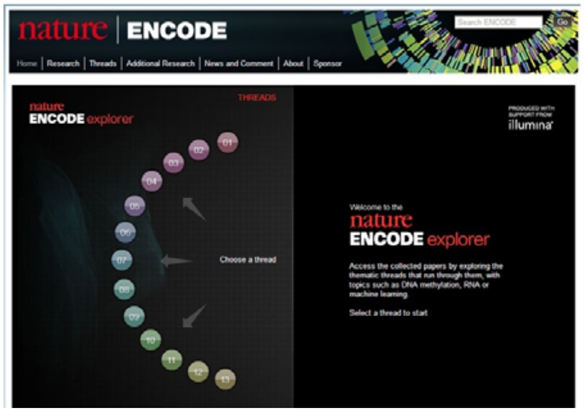
A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.

H3K4Me3 tends to be active, H3K27Me3 tends to be repressive.

Chromatin conformation


(A) Hi-C method
Enzyme A
Enzyme B
Ligate and incorporate biotinylated nucleotides
B
Purify via biotin
Reverse crosslink, shear
Sequence
(B) ChIA-PET method
Adapter A
Adapter B
Ligate
Digest with MmeI
Purify fragments
Sequence
(C) Web-plot of interactions
94.7
94.5
94.3
94.1
93.9
93.7
94.9
Frequency of cross-links to IRX3 promoter
Frequency of cross-lines to FTO promoter
Summary of significant cross-links
IRX3
FTO
Haplotype block of SNPs associated with obesity
-Log(p)
FTO3
IRX3
2–3 3–10 10–35

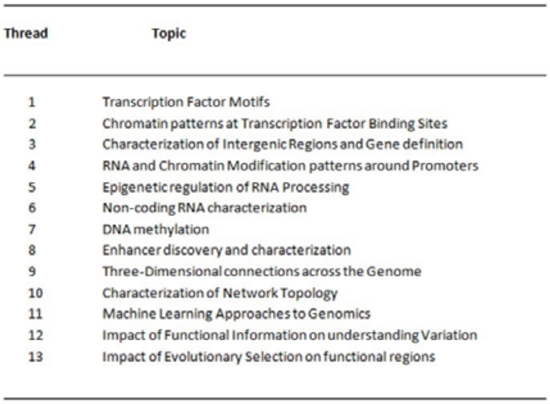
ENCODE *Nature* threads 2012


nature | ENCODE
Search ENCODE
Go
Home | Research | Threads | Additional Research | News and Comment | About | Sponsor
nature ENCODE explorer
THREADS
01
02
03
04
05
06
07
08
09
10
11
12
13
Choose a thread
PRODUCED WITH SUPPORT FROM illumina

Welcome to the nature ENCODE explorer

Access the collected papers by exploring the thematic threads that run through them, with topics such as DNA methylation, RNA or machine learning.

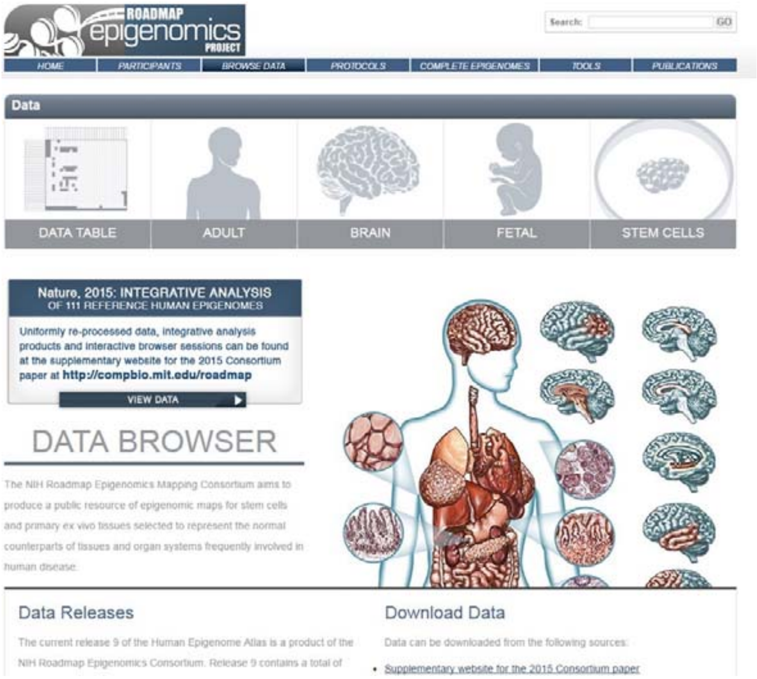
Select a thread to start



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

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

Roadmap Epigenomics Consortium



Nature, 2015: INTEGRATIVE ANALYSIS OF 111 REFERENCE HUMAN EPIGENOMES

Uniformly re-processed data, integrative analysis products and interactive browser sessions can be found at the supplementary website for the 2015 Consortium paper at **http://compbio.mit.edu/roadmap**

DATA BROWSER

The NIH Roadmap Epigenomics Mapping Consortium aims to produce a public resource of epigenomic maps for stem cells and primary ex vivo tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease.

Data Releases

The current release 9 of the Human Epigenome Atlas is a product of the NIH Roadmap Epigenomics Consortium. Release 9 contains a total of

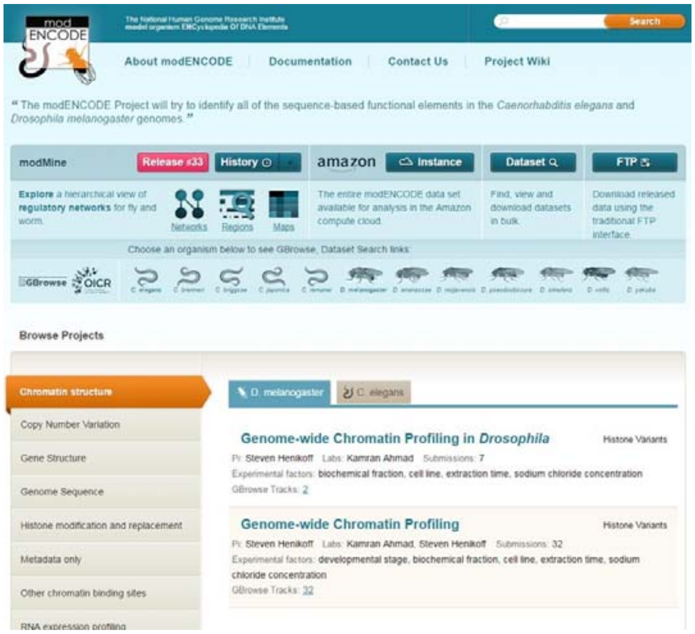
Download Data

Data can be downloaded from the following sources:

- Supplementary website for the 2015 Consortium paper

http://www.roadmapepigenomics.org/

Model Organism ENCODE



mod ENCODE
The National Human Genome Research Institute model organism ENCyclopedia Of DNA Elements
Search
About modENCODE
Documentation
Contact Us
Project Wiki

“ The modENCODE Project will try to identify all of the sequence-based functional elements in the *Caenorhabditis elegans* and *Drosophila melanogaster* genomes. ”

Explore a hierarchical view of **regulatory networks** for fly and worm.

The entire modENCODE data set available for analysis in the Amazon compute cloud.

Find, view and download datasets in bulk.

Download released data using the traditional FTP interface.

Choose an organism below to see GBrowse, Dataset Search links:

Browse Projects

Genome-wide Chromatin Profiling in *Drosophila* Histone Variants

PI: Steven Henikoff Labs: Kamran Ahmad Submissions: 7

Experimental factors: biochemical fraction, cell line, extraction time, sodium chloride concentration

GBrowse Tracks: 2

Genome-wide Chromatin Profiling Histone Variants

PI: Steven Henikoff Labs: Kamran Ahmad, Steven Henikoff Submissions: 32

Experimental factors: developmental stage, biochemical fraction, cell line, extraction time, sodium chloride concentration

GBrowse Tracks: 32

http://www.modencode.org/

International Human Epigenome Consortium

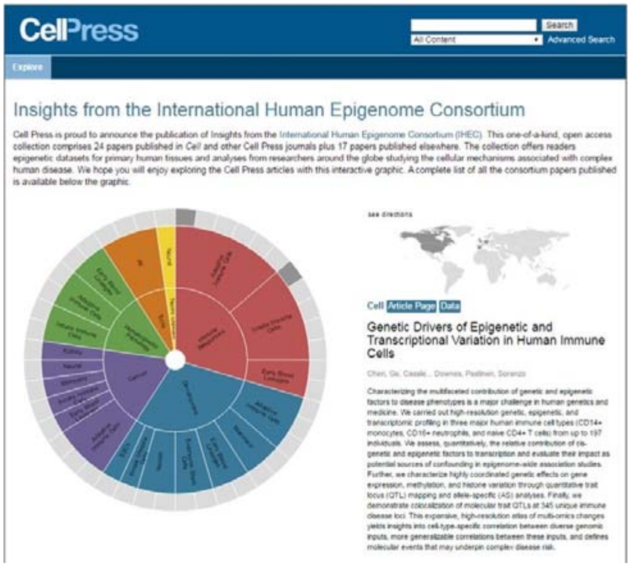
By setting quality standards and providing efficient communication structures IHEC fosters continuous exchange among scientists. It promotes rapid data sharing and minimizes redundancy between different individual research projects. Learn more about the research activities of IHEC.

Epigenomics research and human health are closely linked to each other. Progress in this field of research will thus add to an improved understanding of diseases, and how to better treat and prevent them. Find out what makes epigenomics and the endeavor of IHEC so fascinating.

IHEC makes available comprehensive sets of reference epigenomes relevant to health and disease. You may view, search and download the data already released by the different IHEC-associated projects via the IHEC Data Portal.

http://ihec-epigenomes.org/

IHEC *Cell* threads 2016



Insights from the International Human Epigenome Consortium

Cell Press is proud to announce the publication of Insights from the International Human Epigenome Consortium (IHEC). This one-of-a-kind, open access collection comprises 24 papers published in *Cell* and other Cell Press journals plus 17 papers published elsewhere. The collection offers readers epigenetic datasets for primary human tissues and analyses from researchers around the globe studying the cellular mechanisms associated with complex human disease. We hope you will enjoy exploring the Cell Press articles with this interactive graphic. A complete list of all the consortium papers published is available below the graphic.

Cell Article Page Data

Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells

Chen, Ge, Casale... Downes, Pastinen, Soranzo

Characterizing the multifaceted contribution of genetic and epigenetic factors to disease phenotypes is a major challenge in human genetics and medicine. We carried out high-resolution genetic, epigenetic, and transcriptomic profiling in three major human immune cell types (CD14+ monocytes, CD16+ neutrophils, and naive CD4+ T cells) from up to 197 individuals. We assess, quantitatively, the relative contribution of cis-genetic and epigenetic factors to transcription and evaluate their impact as potential sources of confounding in epigenome-wide association studies. Further, we characterize highly coordinated genetic effects on gene expression, methylation, and histone variation through quantitative trait locus (QTL) mapping and allele-specific (AS) analyses. Finally, we demonstrate colocalization of molecular trait QTLs at 345 unique immune disease loci. This expansive, high-resolution atlas of multi-omics changes yields insights into cell-type-specific correlation between diverse genomic inputs, more generalizable correlations between these inputs, and defines molecular events that may underpin complex disease risk.

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

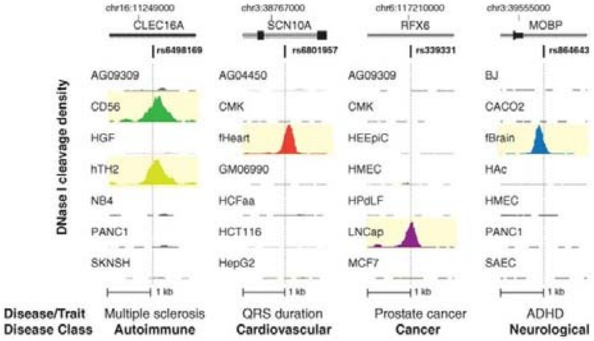
93% of GWAS peak SNPs are located in regulatory regions rather than affecting the protein sequence

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



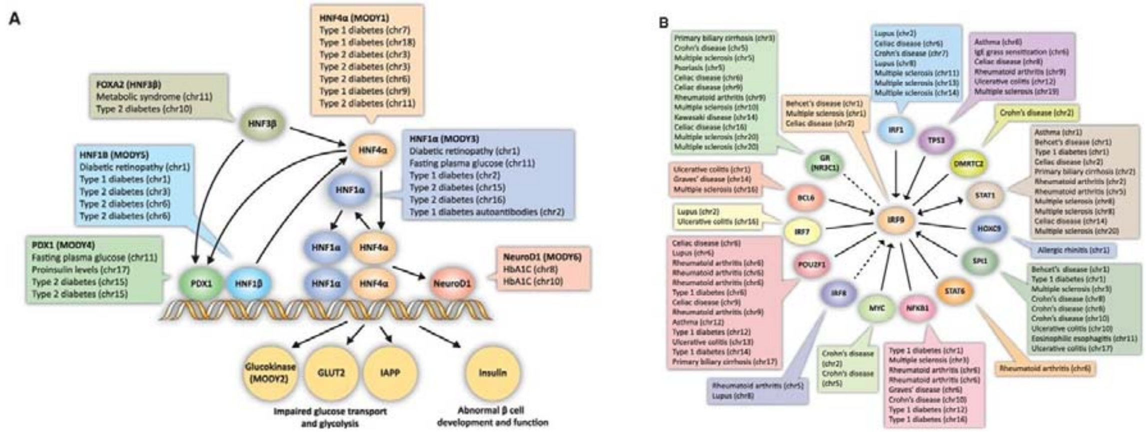
Maurano et al (2012) *Science* **337**: 1190-1195

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

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

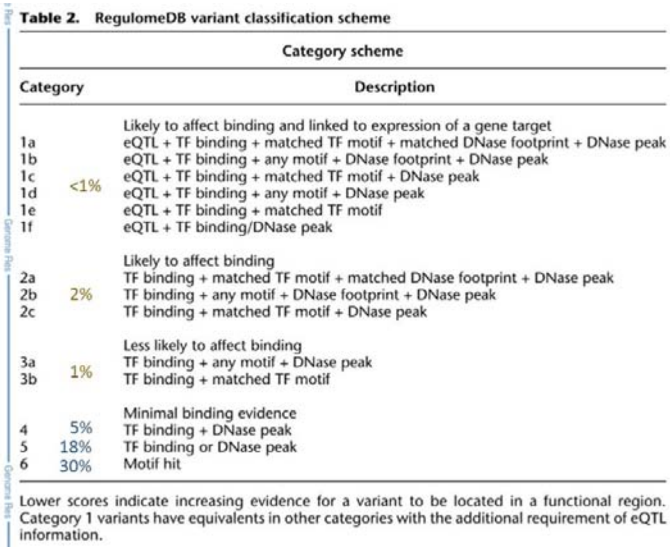


Table 2. RegulomeDB variant classification scheme

Category scheme		
Category		Description
		Likely to affect binding and linked to expression of a gene target
1a	<1%	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		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	2%	TF binding + matched TF motif + matched DNase footprint + DNase peak
2b		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

CADD score annotation of likely deleteriousness

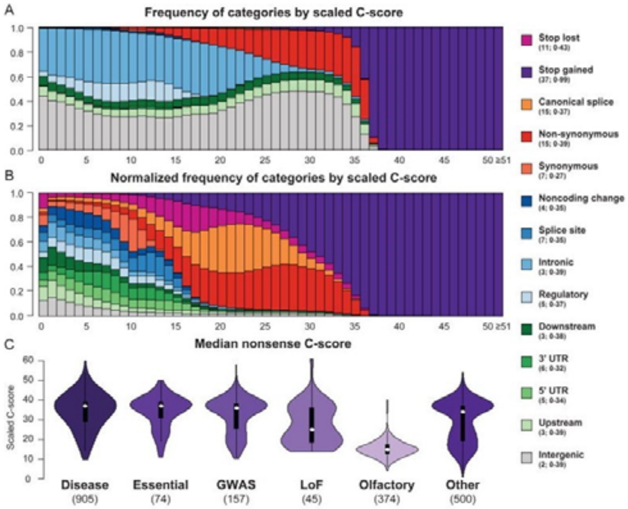
http://cadd.gs.washington.edu/

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

CATO annotation of likely regulatory function

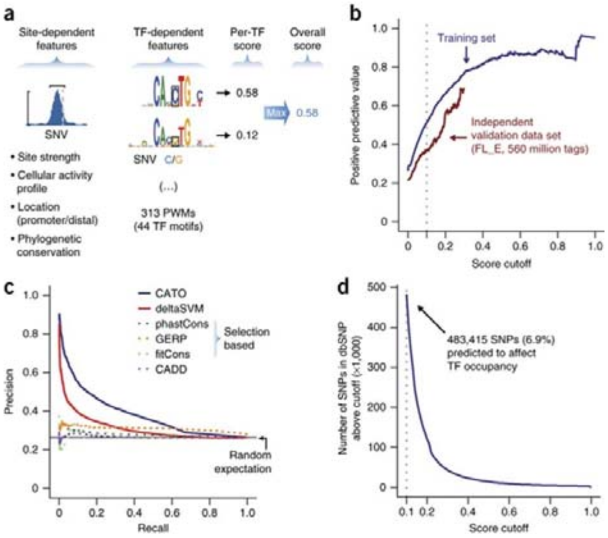
http://www.uwencode.org/proj/CATO/

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 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

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)

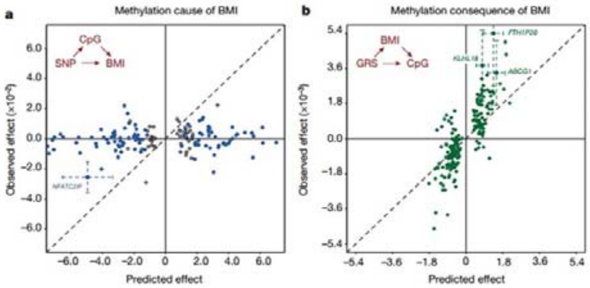
ccQTL: a genotype that influences the level of chromatin conformation / cross-linking

Epigenome-Wide Association Studies (EpiWAS) for Metabolic Disease

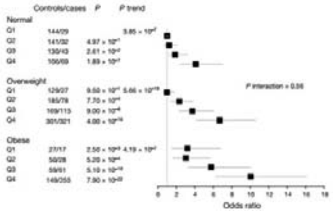
Methyl450 array study of whole blood DNA for 5,387 Europeans and Asians
Identified 278 CpG sites in 207 genes associated with BMI at $p<10^{-7}$: consistent across ethnicities, 90% replicated

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

meQTL for Inflammatory Bowel Disease - I

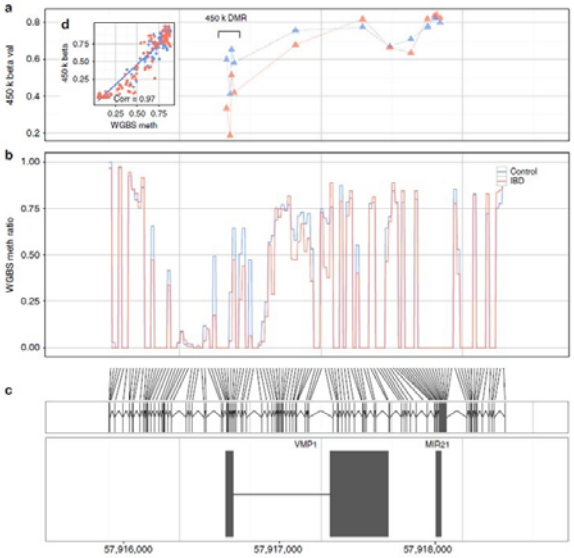
121 CD, 119 UC, 191 Healthy whole blood samples

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



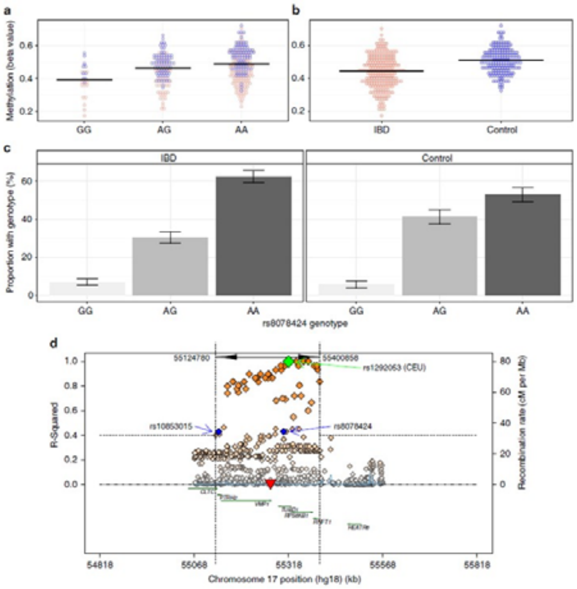
Ventham et al (2016) *Nature Communications* **7**: 13507

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



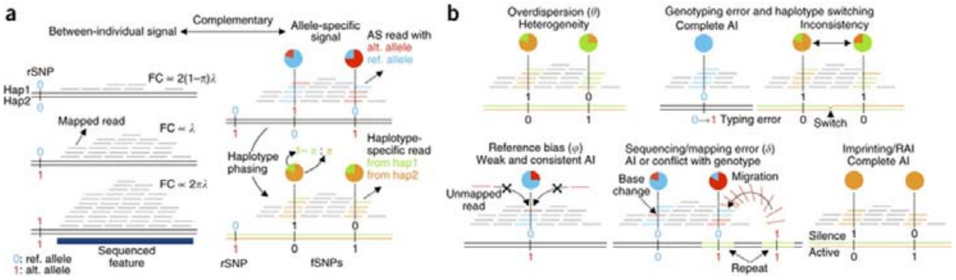
Ventham et al (2016) *Nature Communications* **7**: 13507

ATAC-Seq and enhancer detection

There are three basic approaches for detecting active chromatin, which is interpreted as enhancers:

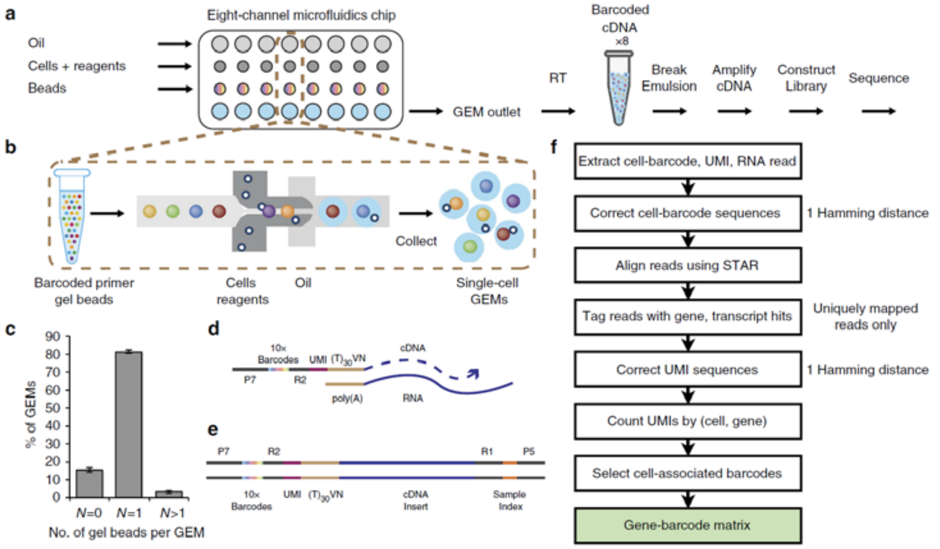
- DNAse Hypersensitivity Site Sequencing (DNaseSeq)
- Chromatin immunoprecipitation Sequencing with CTCF, other TFs (ChIP-Seq)
- Assay for Transcriptionally Active Chromatin (ATAC-Seq)

An emerging software for allele-specific ATAC-Seq (and RNASeq) analysis is RASQUAL
(Robust Allele-Specific Quantitation and Quality Control)


a
Between-individual signal
Complementary
Allele-specific signal
AS read with alt. allele ref. allele
rSNP
Hap1
Hap2
FC ≈ 2(1−π)λ
Mapped read
FC ≈ λ
FC ≈ 2πλ
Haplotype phasing
Haplotype-specific read from hap1 from hap2
Sequenced feature
0: ref. allele
1: alt. allele
rSNP
fSNPs
b
Overdispersion (θ) Heterogeneity
Genotyping error and haplotype switching
Complete AI
Inconsistency
0→1 Typing error
Switch
Reference bias (φ) Weak and consistent AI
Unmapped read
Sequencing/mapping error (δ) AI or conflict with genotype
Base change
Migration
Repeat
Imprinting/RAI Complete AI
Silence
Active

Kumasaka, Knights and Gaffney (2015) *Nature Genetics* **48**: 206-13

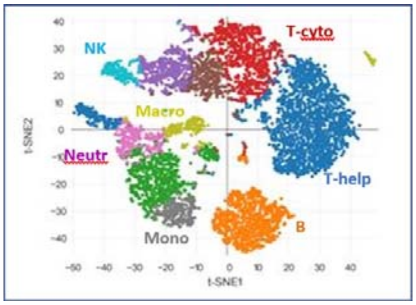
Drop Digital sc-RNASeq



Zheng et al (2017) *Nature Communications* **8**: 14049

Single Cell RNASeq

Peripheral Blood Monocytes

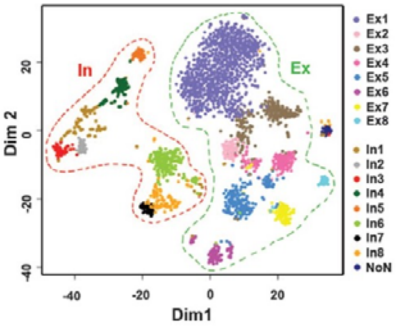


dd scRNASeq

10X Genomics
Illumina/BioRad
Dolomite Bio

Zheng *et al* (2017) *Nature Comm* **8**: 14049

Neuronal nuclei



Smart-Seq2

Fluidigm
Becton Dickinson

Lake *et al* (2016) *Science* **352**: 1586-1590
Ramsköld *et al* (2016) *Nat Biotrch* **30**: 777-782

Three types of Barcode

Each sequencing lane has:

4 or 8 Samples

1,000 – 10,000 single cells

Up to 50,000 reads per call, consisting of:

up to 10,000 different UMI (avg 5-1o reads per UMI)

up to 2-4 K different mRNA species (avg 1-5 UMI per transcript

The aim is to count the UMI, not the reads

Statistical Issues for Single Cell RNASeq

Comparison of SmartSeq2 and ddSeq results

Effect of three barcodes (sample, cell and UMI), in particular linearity of the UMI

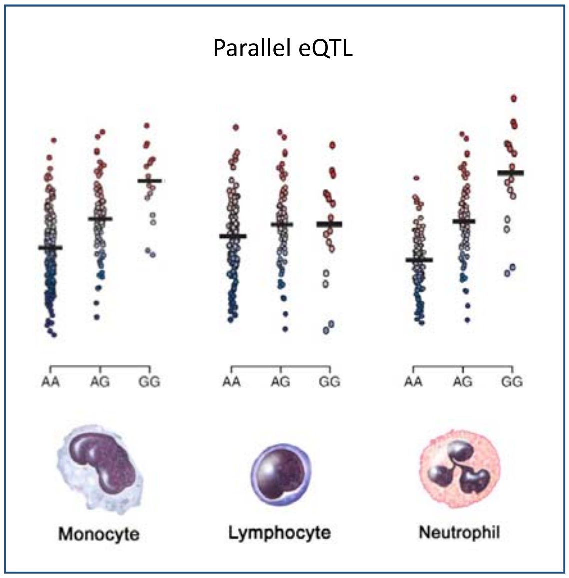
Correct identification of doublet cells

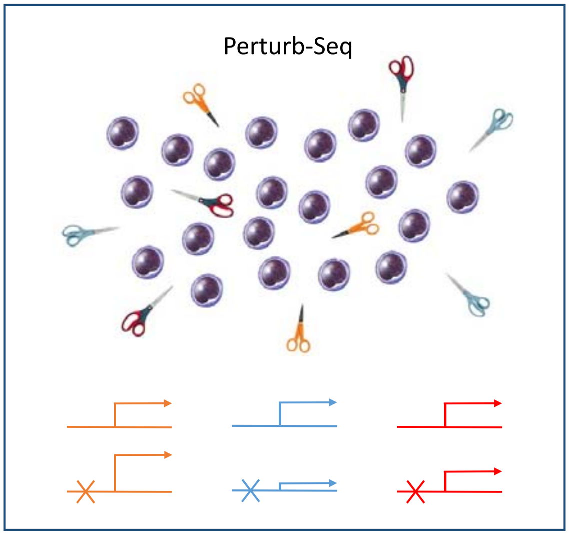
Expanding sample size by inferring individual identity from SNP data

Normalization adjusting for read depth and missing data

Robust and efficient clustering and definition of cellular identities

Single Cell Genetics


Parallel eQTL


Perturb-Seq

Adamson *et al* (2016) *Cell* **167**: 1867-1882
Datlinger *et al* (2017) *Nat Methods* **14**: 297-301