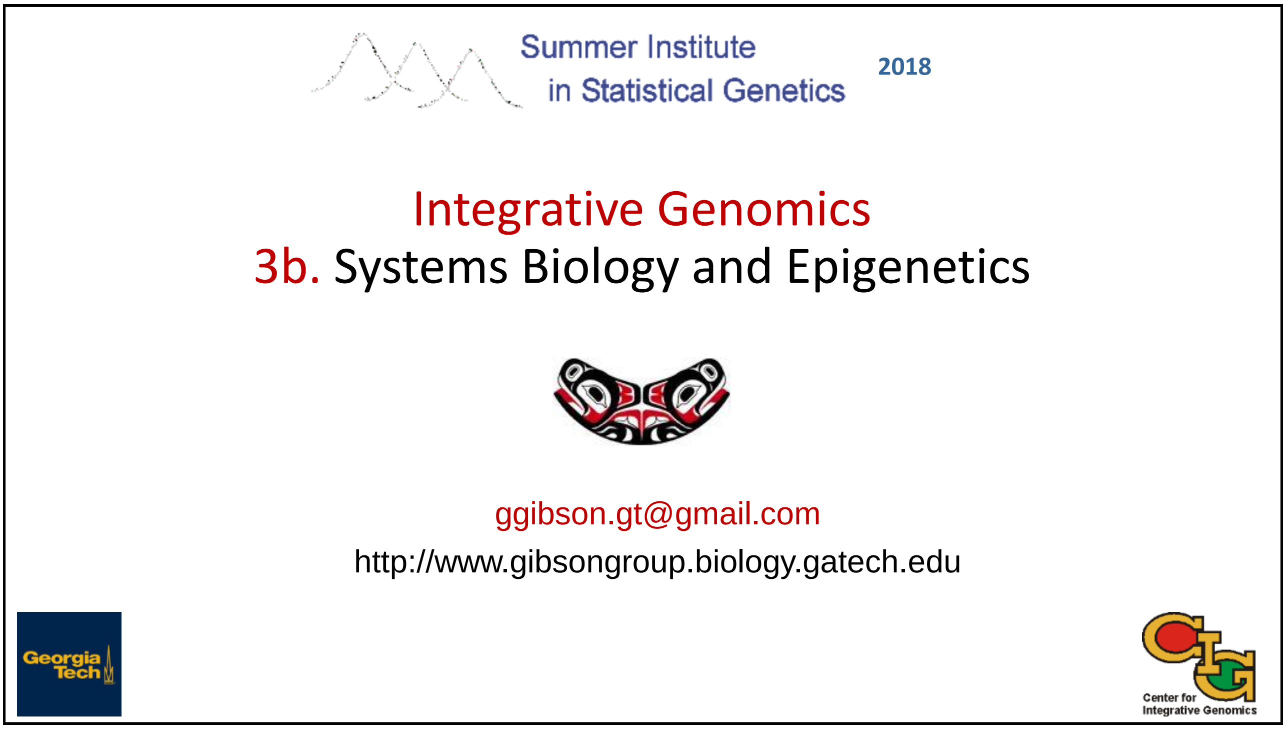


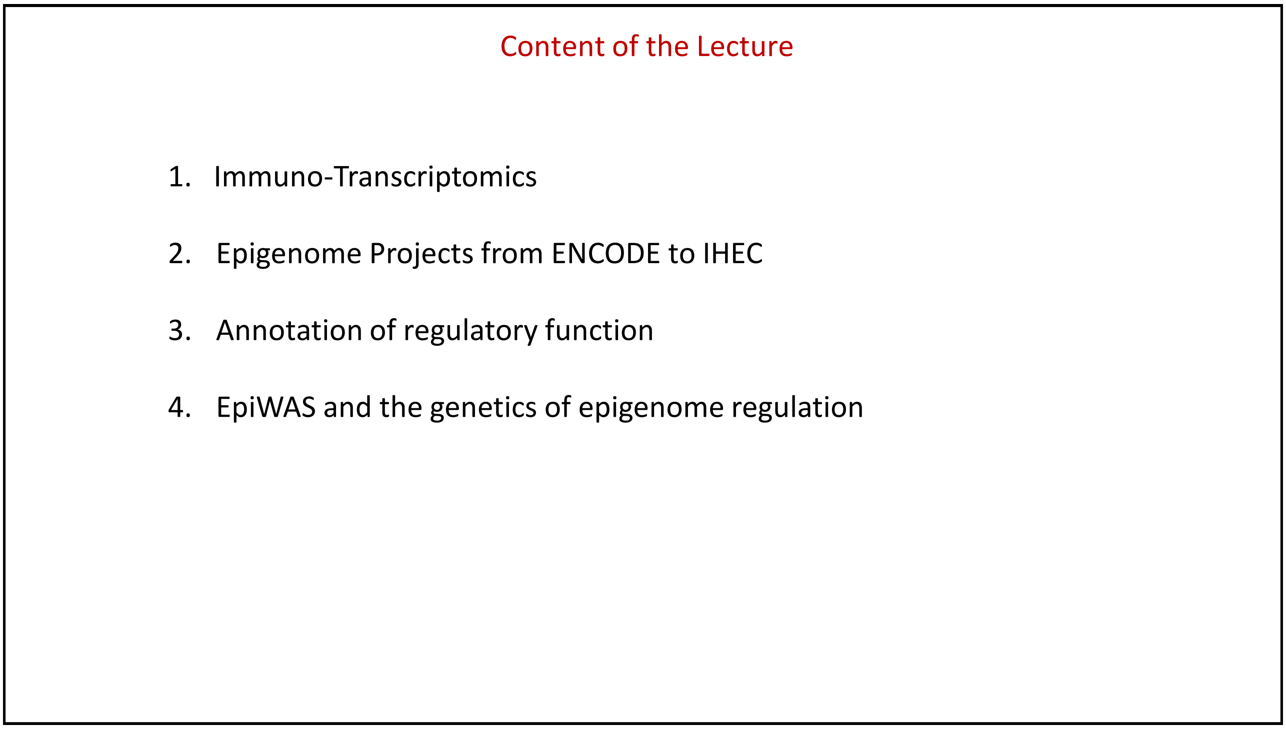
Summer Institute in Statistical Genetics 2018

Integrative Genomics
3b. Systems Biology and Epigenetics

ggibson.gt@gmail.com

http://www.gibsongroup.biology.gatech.edu

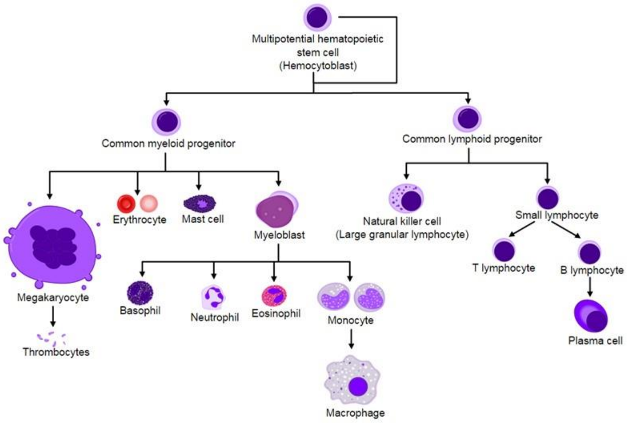
Georgia Tech
Center for Integrative Genomics



Content of the Lecture

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

Why Blood Gene Expression has such a high correlation structure


Multipotential hematopoietic stem cell (Hemocytoblast)
Common myeloid progenitor
Common lymphoid progenitor
Megakaryocyte
Thrombocytes
Erythrocyte
Mast cell
Myeloblast
Basophil
Neutrophil
Eosinophil
Monocyte
Macrophage
Natural killer cell (Large granular lymphocyte)
Small lymphocyte
T lymphocyte
B lymphocyte
Plasma cell

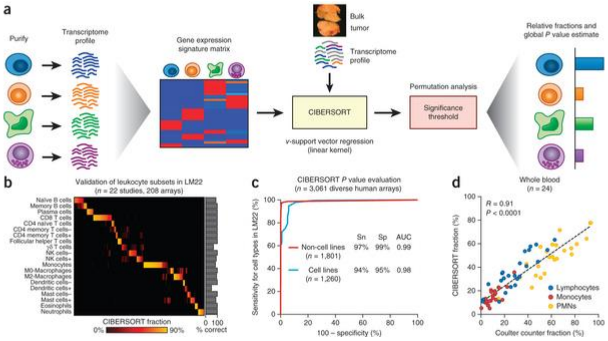
1. Because there are 3 common and dozens of rare blood cell types, and any cell-type biased gene expression correlates with abundance of the cell-type.
2. Because the environment, including disease status, modulates the expression of up to thousands of genes in a coordinated manner
3. The genetic component of most individual transcript abundance is regulated in trans, which also tends to lead to covariance – eg *Stat1* mediates the interferon response

Wikipedia: White Bllood Cells

CIBERSORT

Existing deconvolution methods perform accurately on distinct cell subsets in mixtures with well-defined composition (for example, blood), but are considerably less effective for discriminating closely related cell types (for example, naïve vs. memory B cells).

Input = reference gene expression signatures and unknown profile
Algorithm= linear support vector regression (SVR) – a machine learning approach robust to noise
Output = estimated abundances and p-value for the deconvolution

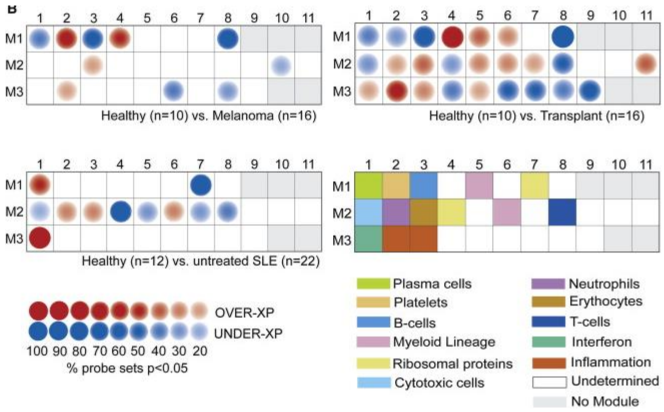


Newman et al (2015) *Nature Methods* **12**: 453-457 “Robust enumeration of cell subsets from tissue expression profiles”

Chaussabel Modules

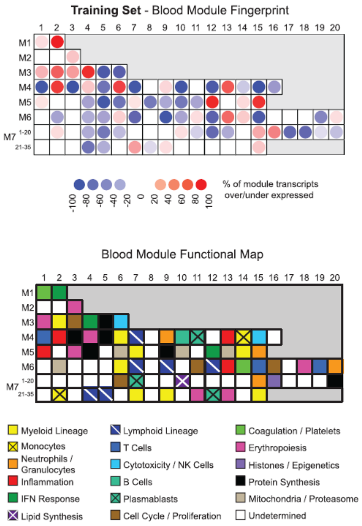
Used k-means clustering to search for conserved modules of genes that are differentially expressed in 8 diseases, namely 239 samples for SLE, JIA, T1D, melanoma, 2 types of bacteremia, influenza, or liver transplantation

Identified 28 modules involving 4742 transcripts (average of 170 per module)



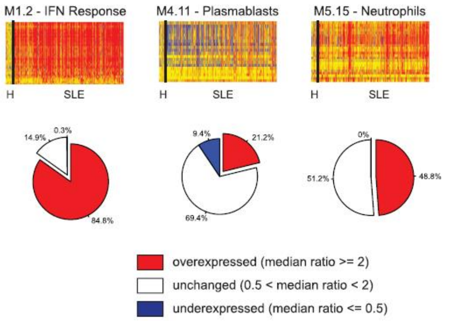
Chaussabel et al (2008) *Immunity* **29**: 150-164 “A modular analysis framework for blood genomics studies: application to SLE”

Update to 95 modules in 2016



158 Pediatric SLE patients
924 longitudinal PB profiles (avg ~ 6 per patient)

First asked how modules correlate with disease, and how many patients show the effect

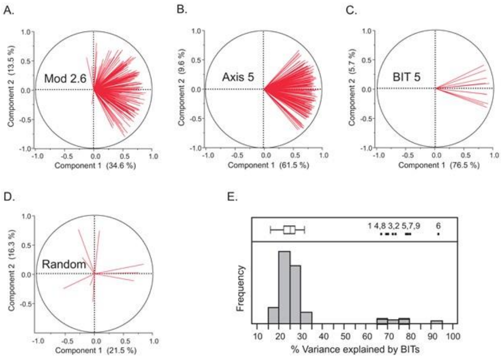


Banchereau et al (2016) *Cell* **165**: 551-565 “Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients”

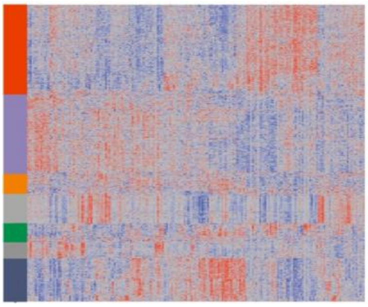
Eight Axes capture basic Immune functions

We define the Axis scores at the first Principal Component of the positively correlated genes

Each Axis corresponds to an identified aspect of immune function, but they explain much more of the variance than the corresponding cell counts. The covariance is due to both cell abundance and transcription.



Axis 1	T-cell signaling
Axis 2	Reticulocyte development
Axis 3	B-cell signaling
Axis 4	Housekeeping functions
Axis 5	Neutrophils and TLR
Axis 6	Antibody response ?
Axis 7	Interferon response
Axis 10	Mitosis / cell cycle

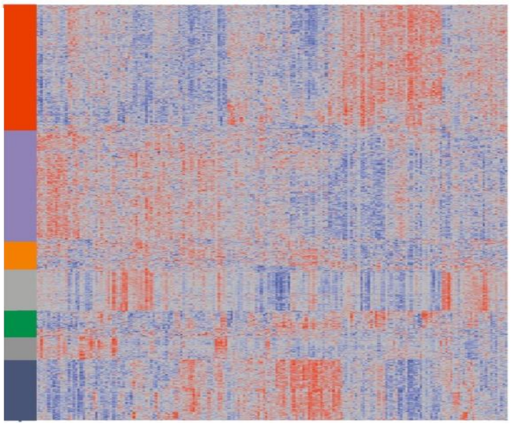


BIT are 10 Blood Informative Transcripts that define each Axis.

Preininger et al (2013) *PLoS Genet.* **9**: e1003362 “Blood Informative Transcripts define nine common axes of peripheral blood gene expression”

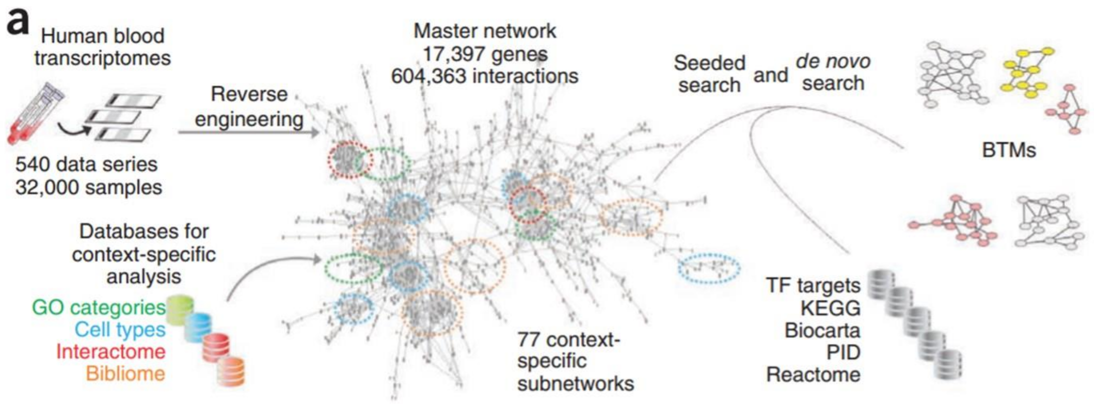
← Individual samples →

T-cell signaling	Axis 1
Neutrophils/Inflammation	Axis 5
Mitotic activity	Axis 10
Interferon-response	Axis 7
Erythropoiesis	Axis 2
B-cell signaling	Axis 3



↑ Individual transcripts ↓

Identifying Blood Transcript Modules



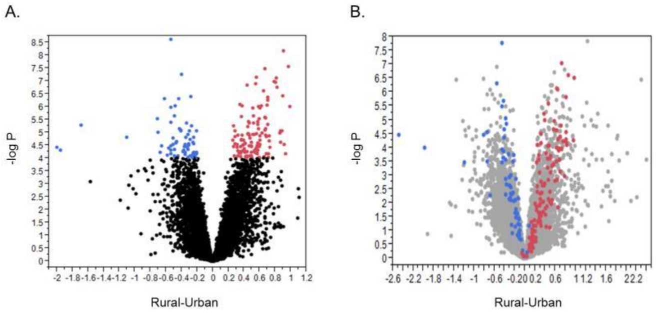
Li et al (2014) *Nat Immunology* **15**: 195-204 "Molecular signatures of antibody responses derived from a systems biological study of 5 human vaccines"

The problem with gene ontology analysis on DE gene sets

1. Although powerful, DE analysis is also intrinsically under-powered, so there is a high false negative rate

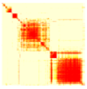
2. Consequently, when you see a gene set annotated as "perturbed by drug x in cell-type y of females with disease z", beware! Most likely a replicate of the experiment would give a completely different list.

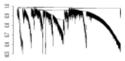
3. Conversely, some annotations, eg "Lupus-associated genes" have multiple completely different lists.



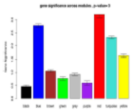
Basic Workflow for Cluster analysis

1. Construct Similarity Matrix of Samples

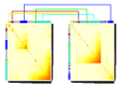




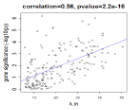
2. Generate Modules with WGCNA (or MMC, or …)



3. Perform Gene Ontology enrichment analysis on the Modules

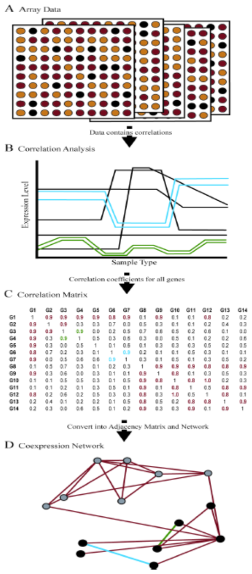


4. Compare Module Preservation across datasets



5. Associate Module Eigenvectors with Traits OR
   search for Molecular Drivers of the Modules

General Framework for Coexpression Network Analysis


A Array Data
Data contains correlations
B Correlation Analysis
Expression Level
Sample Type
Correlation coefficients for all genes
C Correlation Matrix
Convert into Adjacency Matrix and Network
D Coexpression Network

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.9	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

1. Generate gene expression data (Microarray or RNASeq)

2. Measure Pearson correlations between all gene pairs

3. Dichotomize the matrix with some cutoff for the strength of correlation to generate an UNWEIGHTED adjacency matrix

4. OR Weight the correlations to generate a more nuanced network, for example using a power function:

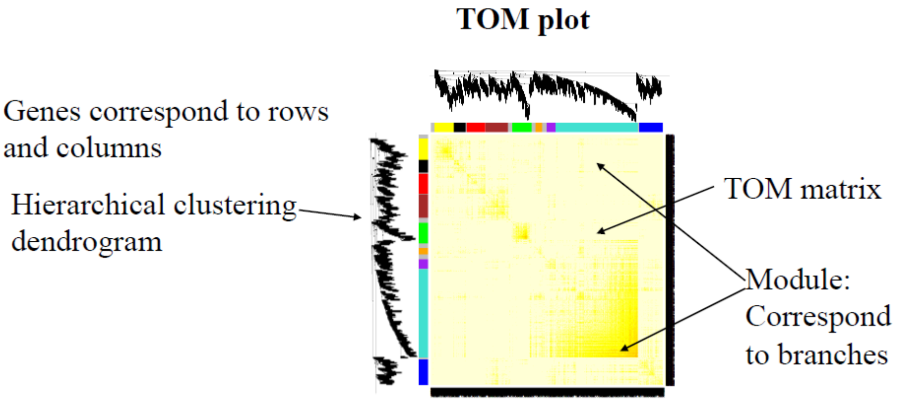
$$a_{ij} = |\, cor(x_i, x_j)\, |^{\beta}$$

Zhang and Horvath (2005) *SAGMB* **4**: 17. A general framework for weighted gene co-expression network analysis.

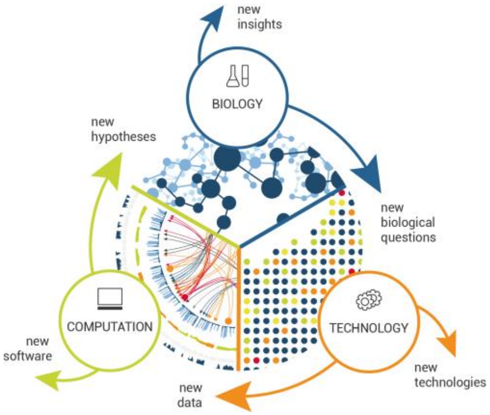
Topological Overlap Matrices

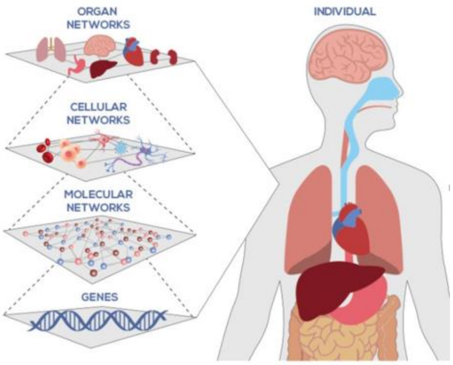
Gene Modules correspond to Branches of the weighted hierarchical tree

Each Modules is given a color – there may be dozens of them

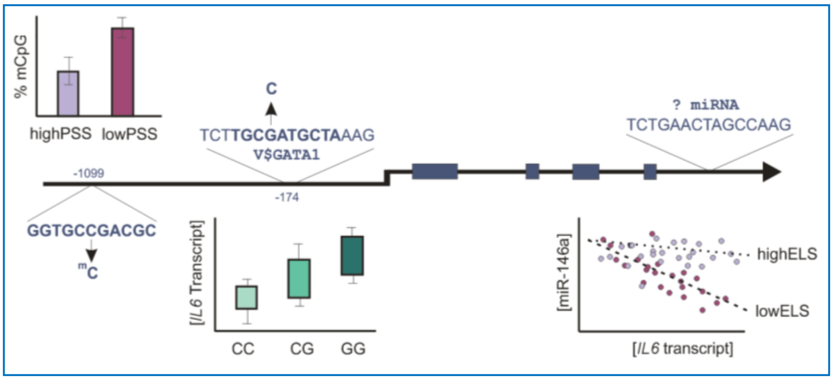


Integrative Systems Biology: big data meets cell biology

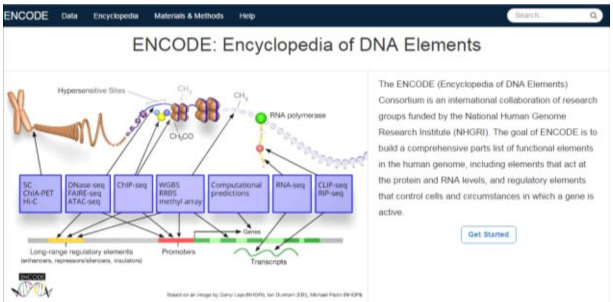




The integrative nature of transcriptional regulation

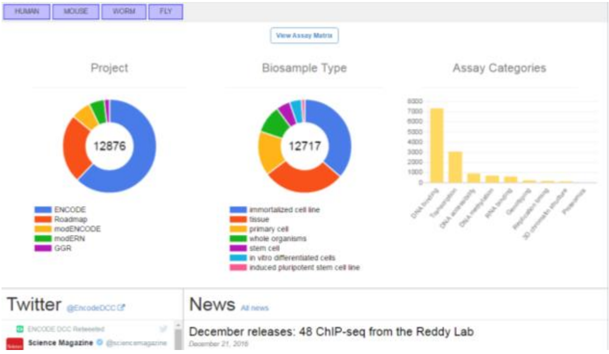


https://www.encodeproject.org/



ENCODE: Encyclopedia of DNA Elements

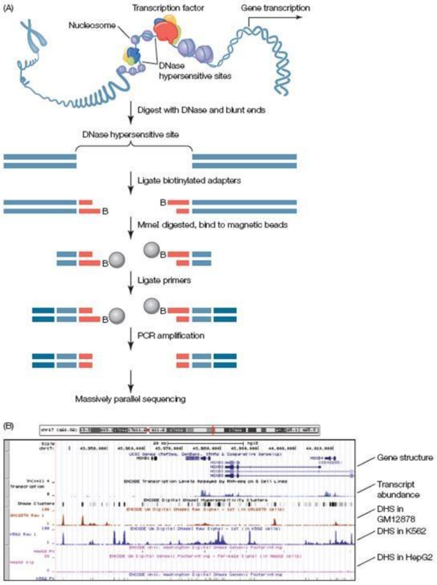
The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at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

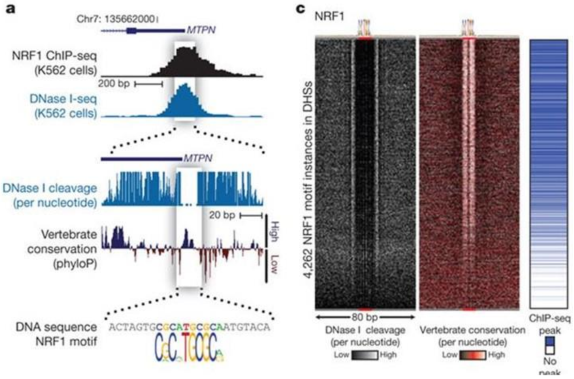


December releases: 48 ChIP-seq from the Reddy Lab
December 21, 2016

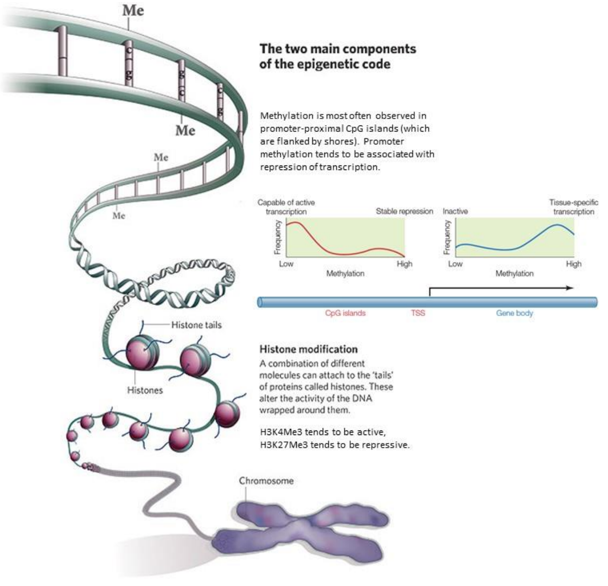
The ENCODE Project Consortium (2011) *PLOS Biology* **9**: 1001046

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

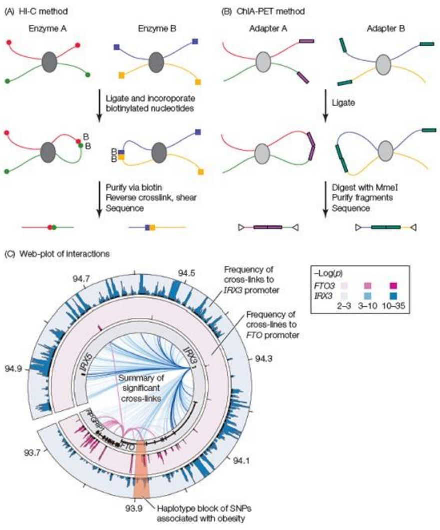
Methylation is most often observed in promoter-proximal CpG islands (which are flanked by shores). Promoter methylation tends to be associated with repression of transcription.

Histone modification

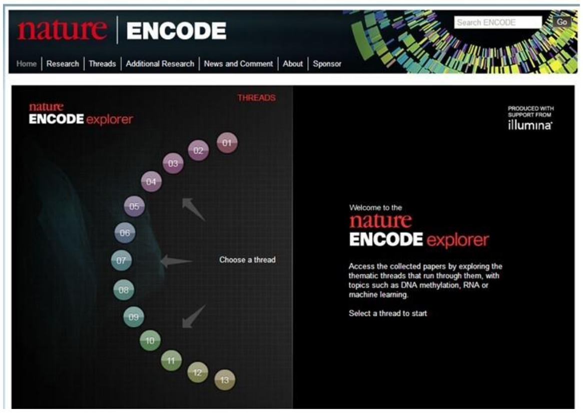
A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.

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

Chromatin conformation



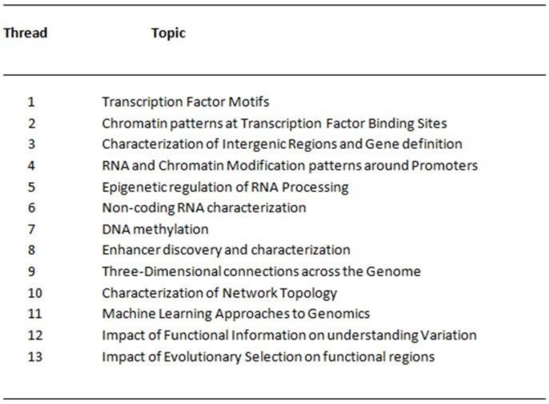
ENCODE *Nature* threads 2012


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.

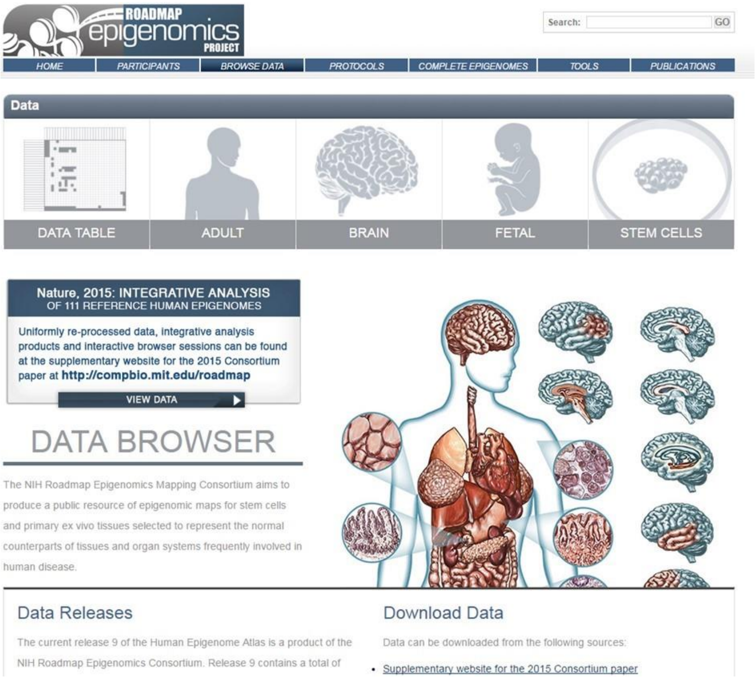
Select a thread to start



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

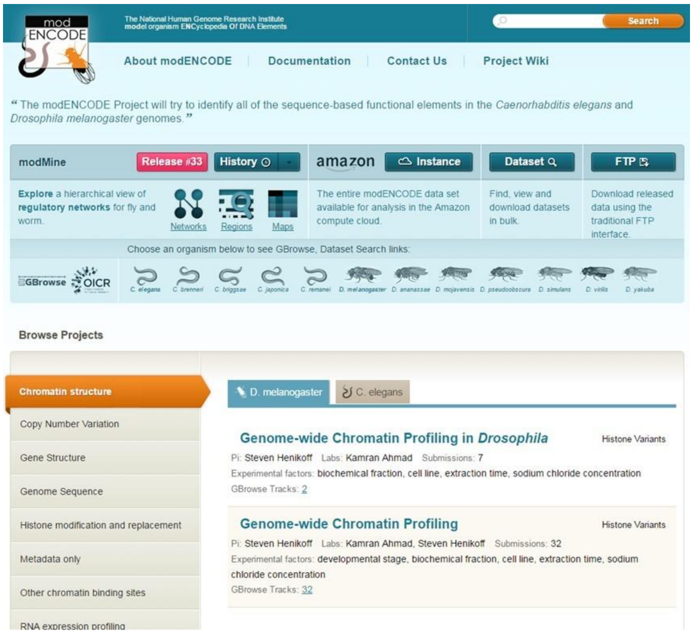
Download Data

Data can be downloaded from the following sources:

- Supplementary website for the 2015 Consortium paper

http://www.roadmapepigenomics.org/

Model Organism ENCODE



“ 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

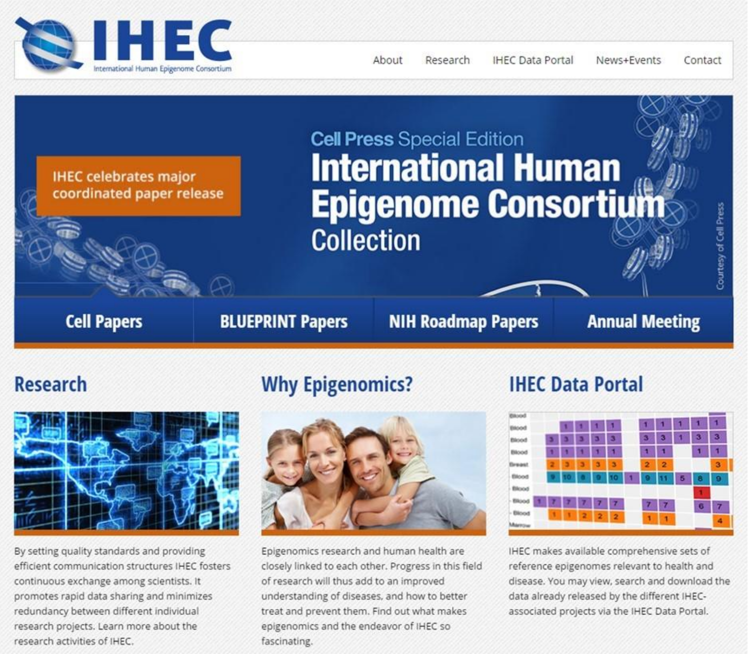
Pi: Steven Henikoff Labs: Kamran Ahmad, Steven Henikoff Submissions: 32

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

GBrowse Tracks: 32

http://www.modencode.org/

International Human Epigenome Consortium


IHEC
International Human Epigenome Consortium
About
Research
IHEC Data Portal
News+Events
Contact
IHEC celebrates major coordinated paper release
Cell Press Special Edition
International Human Epigenome Consortium Collection
Courtesy of Cell Press
Cell Papers
BLUEPRINT Papers
NIH Roadmap Papers
Annual Meeting

Research

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.

Why Epigenomics?

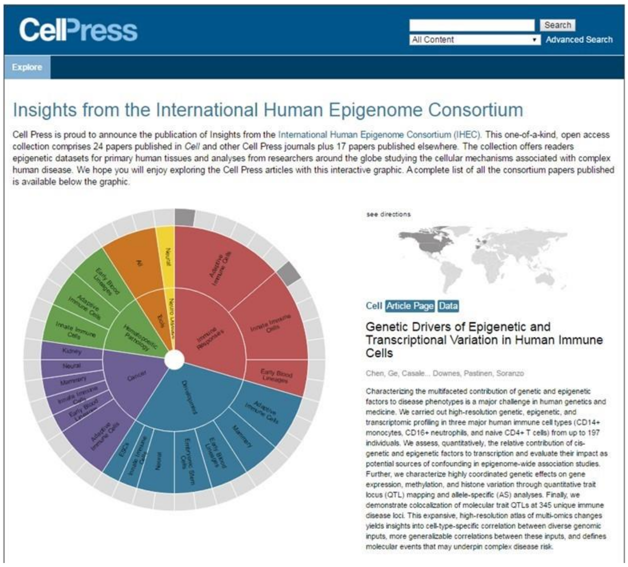
Epigenomics research and human health are closely linked to each other. Progress in this field of research will thus add to an improved understanding of diseases, and how to better treat and prevent them. Find out what makes epigenomics and the endeavor of IHEC so fascinating.

IHEC Data Portal

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.

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

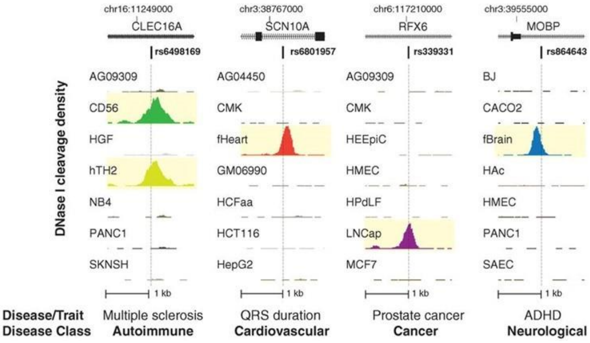
93% of GWAS peak SNPs are located in regulatory regions rather than affecting the protein sequence

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



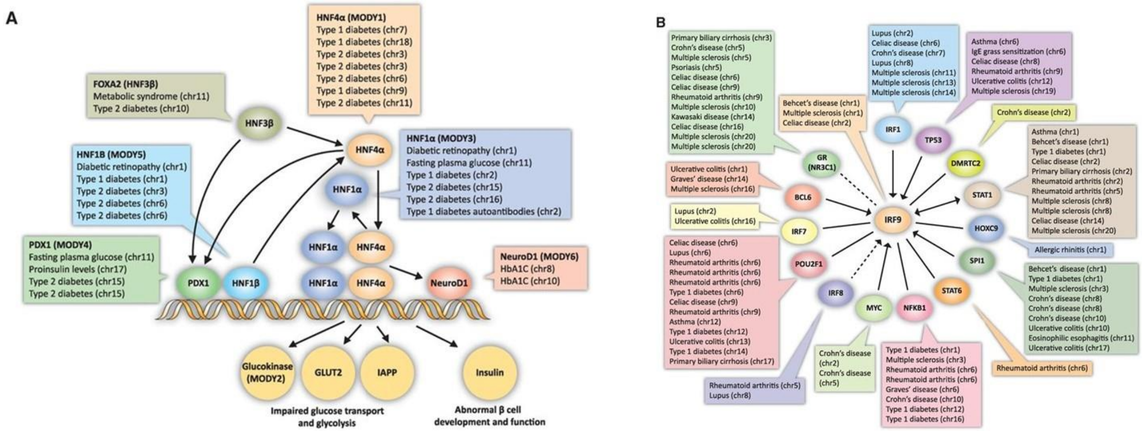
Maurano et al (2012) *Science* **337**: 1190-1195

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

CADD score annotation of likely deleteriousness

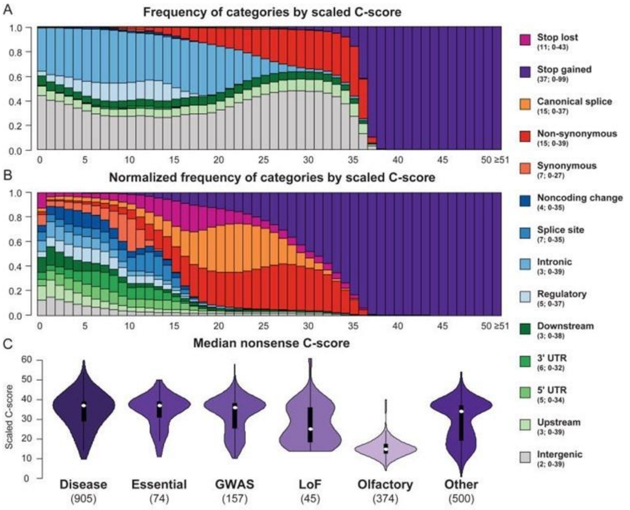
http://cadd.gs.washington.edu/

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%

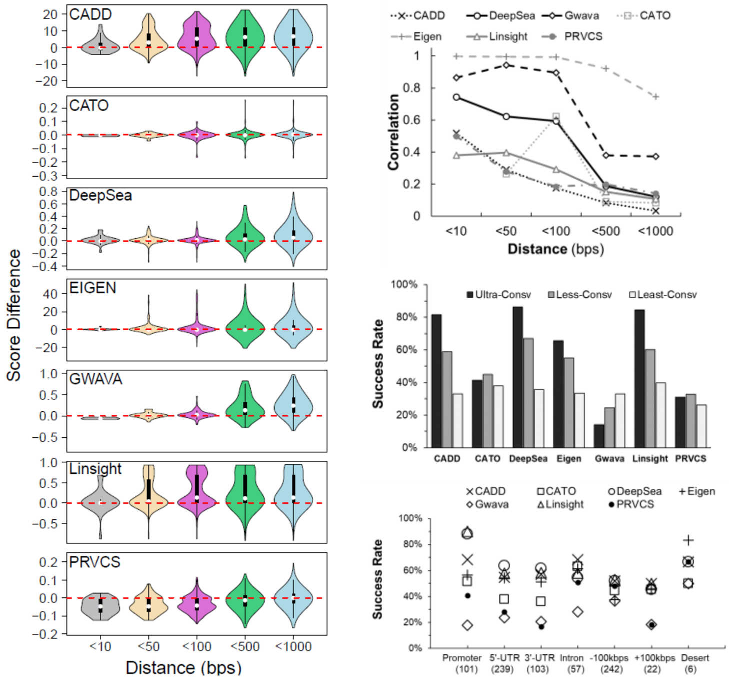
The score attempts unbiased prediction of "deleteriousness", based on machine learning comparison of 15M observed and simulated human variants



Kircher et al (2014) *Nature Genetics* **46**: 310-315

Beware Regulatory Annotation

Li Liu, Max Sanderford, Sudhir Kumar, GG
Under review



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)

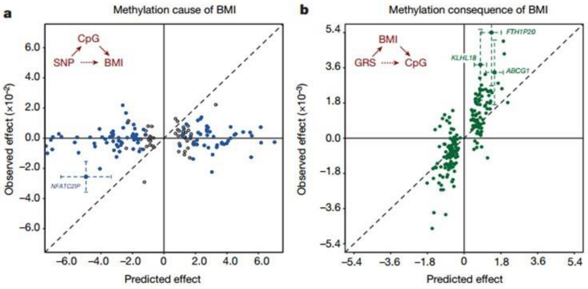
ccQTL: a genotype that influences the level of chromatin conformation / cross-linking

Epigenome-Wide Association Studies (EpiWAS) for Metabolic Disease

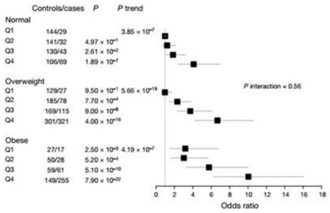
Methyl450 array study of whole blood DNA for 5,387 Europeans and Asians
Identified 278 CpG sites in 207 genes associated with BMI at $p<10^{-7}$: consistent across ethnicities, 90% replicated

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



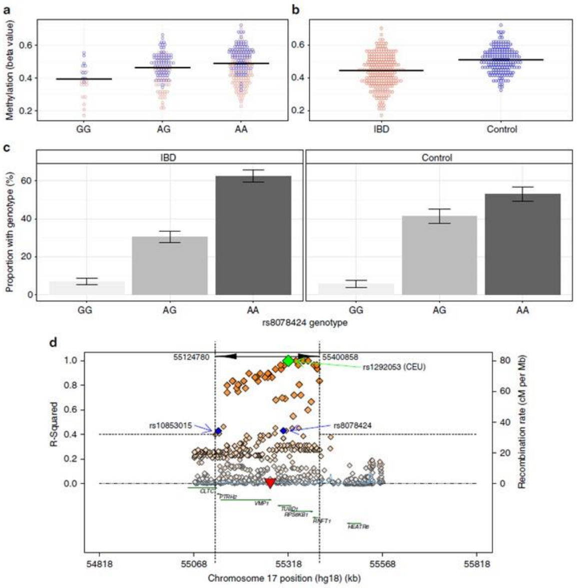
Wahl et al (2016) *Nature* **541**: 81-85

meQTL for Inflammatory Bowel Disease

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



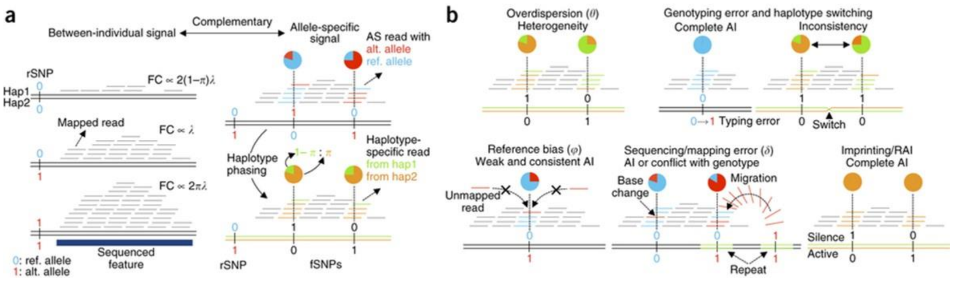
Ventham et al (2016) *Nature Communications* **7**: 13507

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)

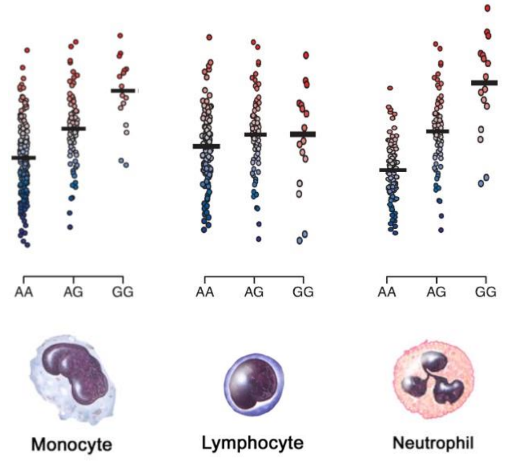
An emerging software for allele-specific ATAC-Seq (and RNASeq) analysis is RASQUAL
(Robust Allele-Specific Quantitation and Quality Control)



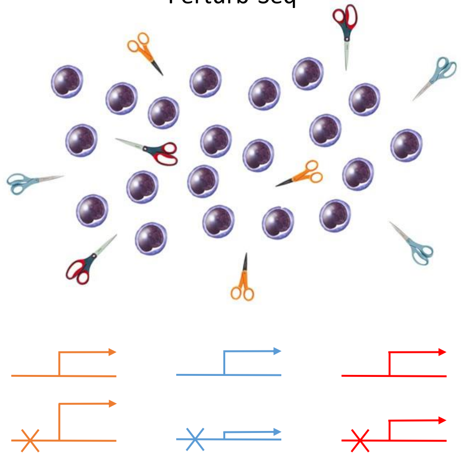
Kumasaka, Knights and Gaffney (2015) *Nature Genetics* **48**: 206-13

Single Cell Genetics

Parallel eQTL



Perturb-Seq



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