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

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.

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

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).


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.

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

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:

\[ a_{ij} = |\text{cor}(x_i, x_j)|^\beta \]
Topological Overlap Matrices

Gene Modules correspond to Branches of the weighted hierarchical tree

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

TOM plot

Genes correspond to rows and columns
Hierarchical clustering dendrogram

Integrative Systems Biology: big data meets cell biology
The integrative nature of transcriptional regulation

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

https://www.encodeproject.org/
DHS and TFBS: DNAse hypersensitive sites and TF Binding

Three modes of epigenetic regulation
ENCODE Nature threads 2012

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/
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

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

CADD score annotation of likely deleteriousness

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 $-10 \times \log_{10}(\text{rank/total})$, 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


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 $-10 \times \log_{10}(\text{rank/total})$, 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

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 β: 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


---

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

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)


---

Single Cell Genetics

**Parallel eQTL**

- AA
- AG
- GG

- Monocyte
- Lymphocyte
- Neutrophil

**Perturb-Seq**


Datlinger *et al* (2017) *Nat Methods* 14: 297-301