

# From Data to Knowledge:

Extracting Biological Insight from Diverse Data Sources



Stephen D. Turner, Ph.D.

Bioinformatics Core Director

[bioinformatics@virginia.edu](mailto:bioinformatics@virginia.edu)

[bioinformatics.virginia.edu](http://bioinformatics.virginia.edu)



## Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0)

This is a human-readable summary of the [Legal Code \(the full license\)](#).

[Disclaimer](#)

### You are free:

to **Share** — to copy, distribute and transmit the work

to **Remix** — to adapt the work

### Under the following conditions:



**Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



**Noncommercial** — You may not use this work for commercial purposes.



**Share Alike** — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

### With the understanding that:

**Waiver** — Any of the above conditions can be **waived** if you get permission from the copyright holder.

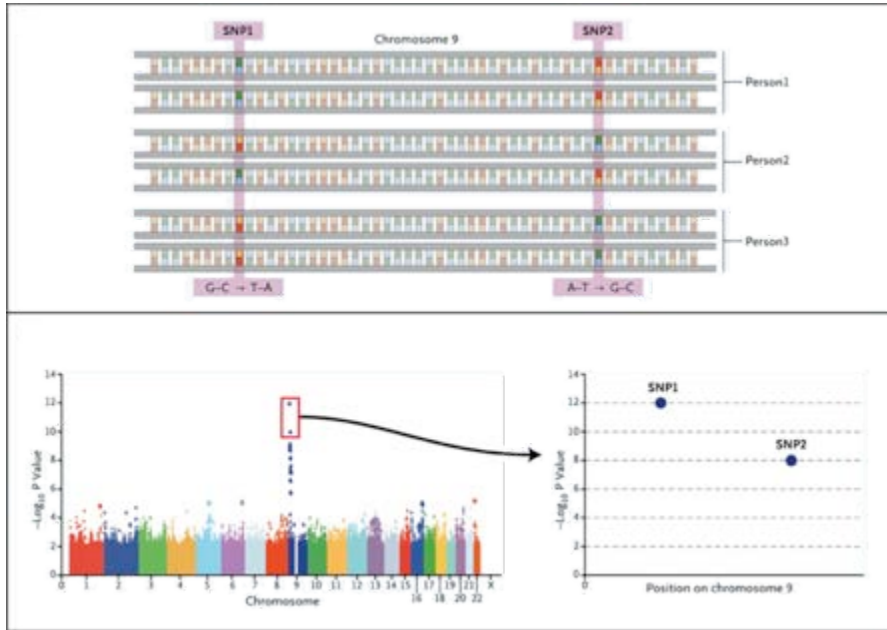
**Public Domain** — Where the work or any of its elements is in the **public domain** under applicable law, that status is in no way affected by the license.

**Other Rights** — In no way are any of the following rights affected by the license:

- Your fair dealing or **fair use** rights, or other applicable copyright exceptions and limitations;
- The author's **moral** rights;
- Rights other persons may have either in the work itself or in how the work is used, such as **publicity** or privacy rights.

**Notice** — For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to this web page.

# GWAS: One gene, one enzyme, one function?



Manolio TA. N Engl J Med 2010;363:166-176.



Genome.gov/GWASudies

## *GENETIC CONTROL OF BIOCHEMICAL REACTIONS IN NEUROSPORA\**

BY G. W. BEADLE AND E. L. TATUM

BIOLOGICAL DEPARTMENT, STANFORD UNIVERSITY

Communicated October 8, 1941

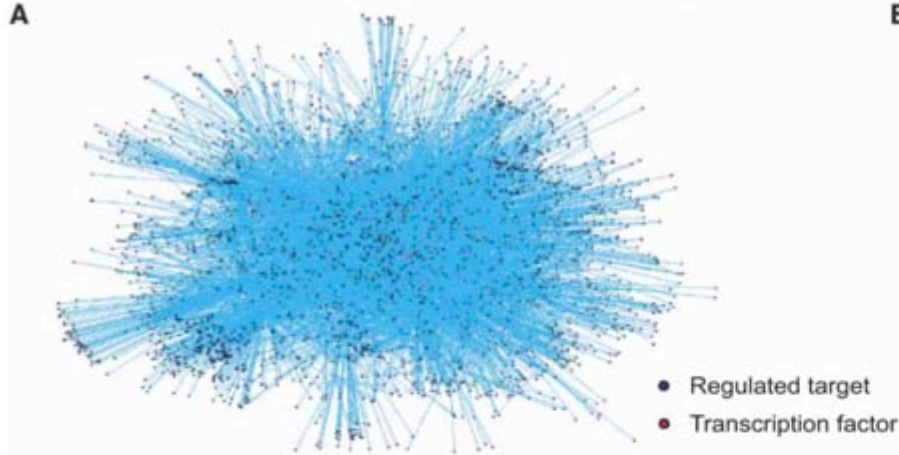
# DNA Variation: Limitations

---

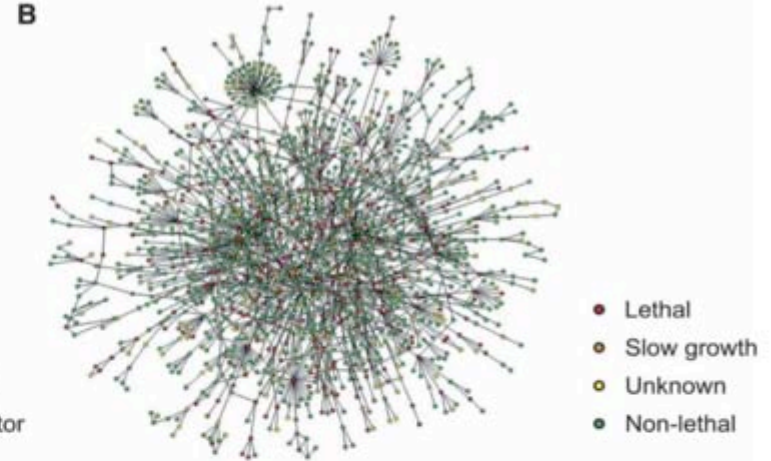
## GWAS DOES NOT INFORM:

- Which gene is affected
- How gene function is perturbed
- How biological processes are altered

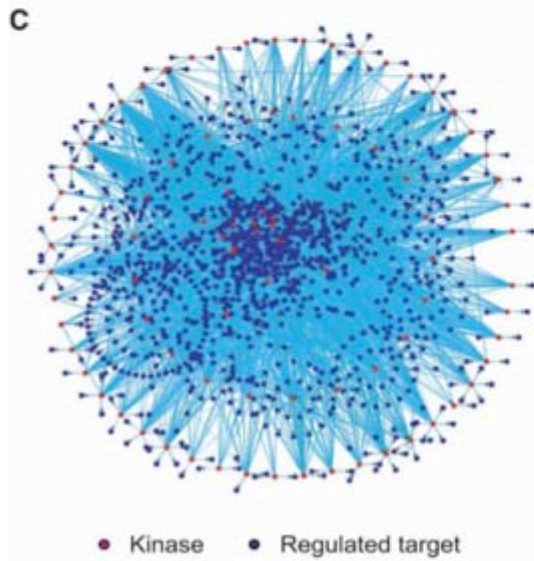
# One gene, one enzyme, one function?



Zhu X. et al. (2007). *Genes & Dev* 21:1010-1024.



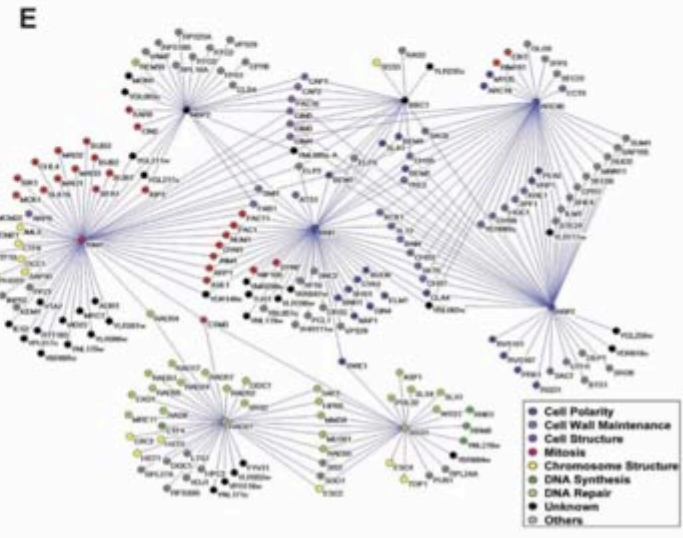
Jeong, H. et al.. (2001) *Nature* 411:41-42.



Ptacek, J. et al. (2005) *Nature* 438:679-684.



Guimera and Amaral. (2005). *Nature* 433:895-900.

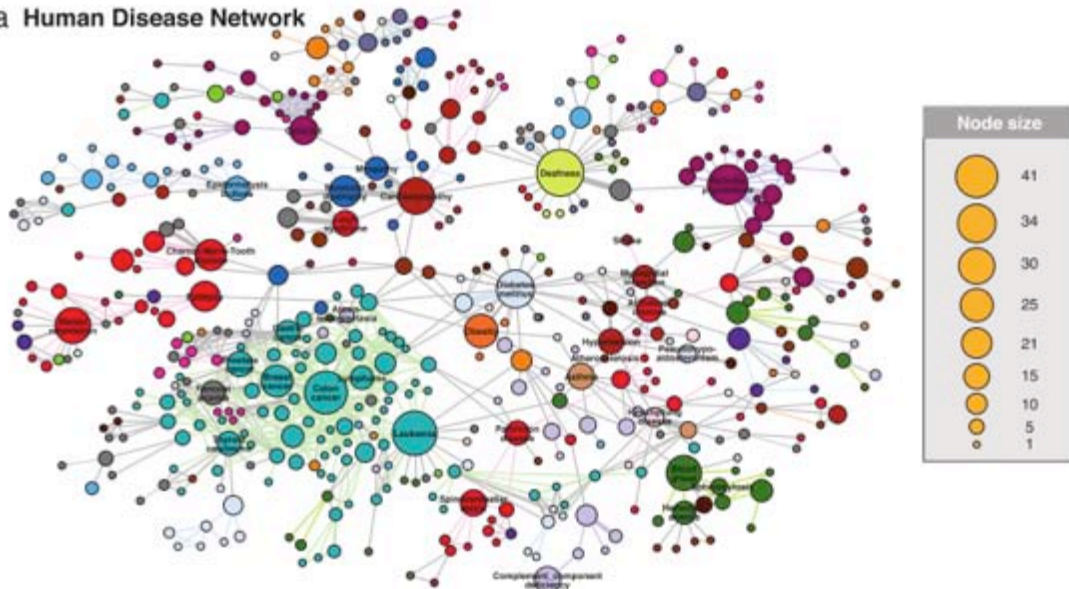


Tong, A.H. et al. (2001). *Science* 294:2364-2368.



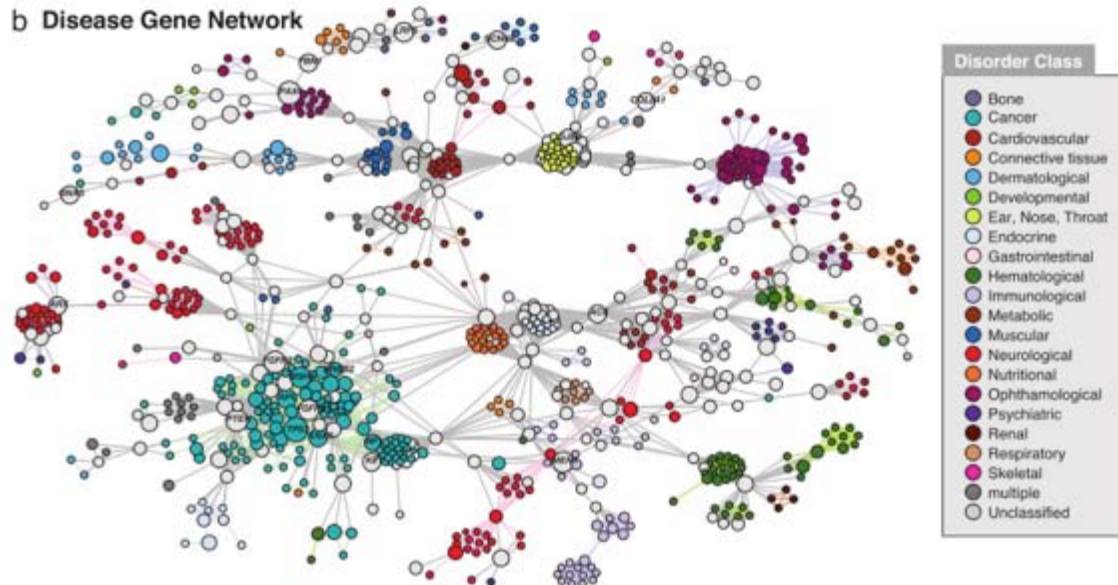
# Distribution of Disease Genes

a Human Disease Network



Diseases connected if same gene implicated in both.

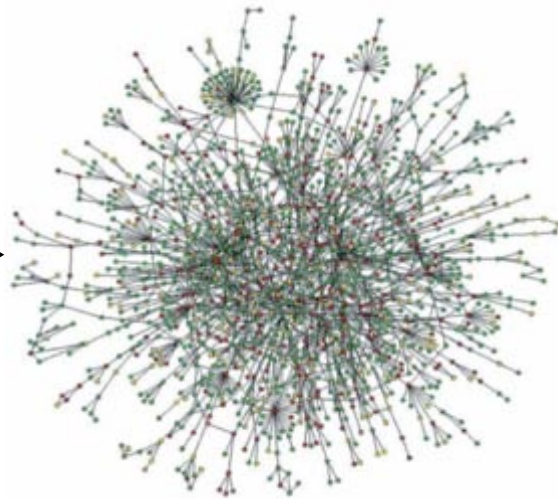
b Disease Gene Network



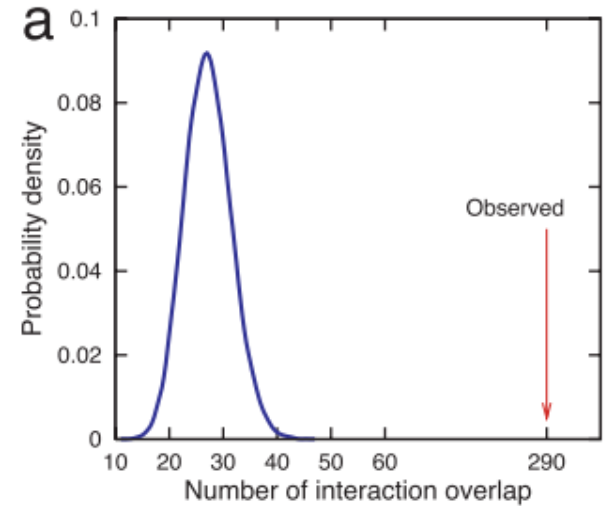
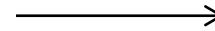
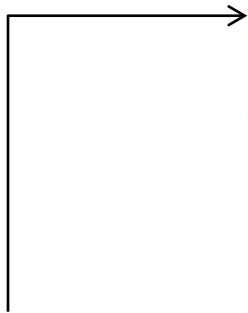
Genes connected if implicated in the same disorder.

# Distribution of Disease Genes

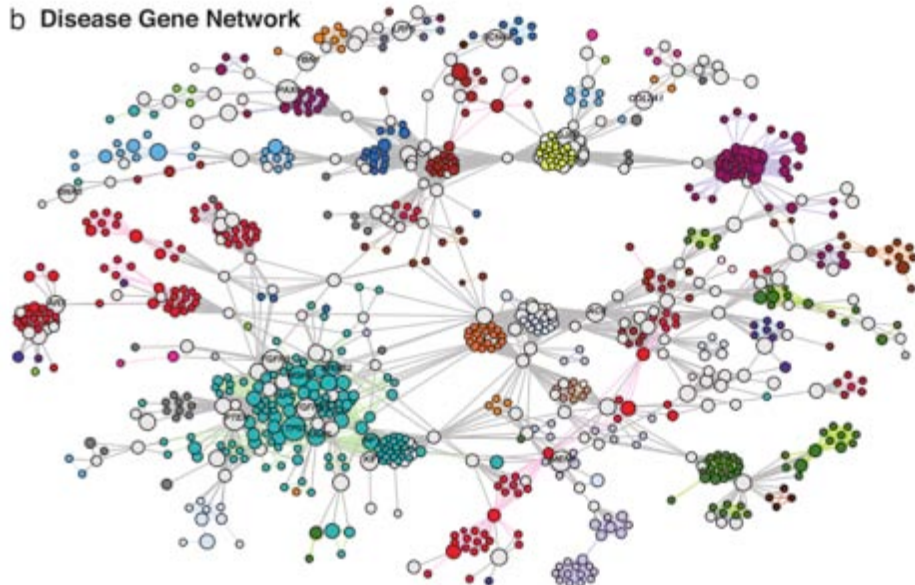
Protein-protein interactions



Overlay with PPI data



**b** Disease Gene Network



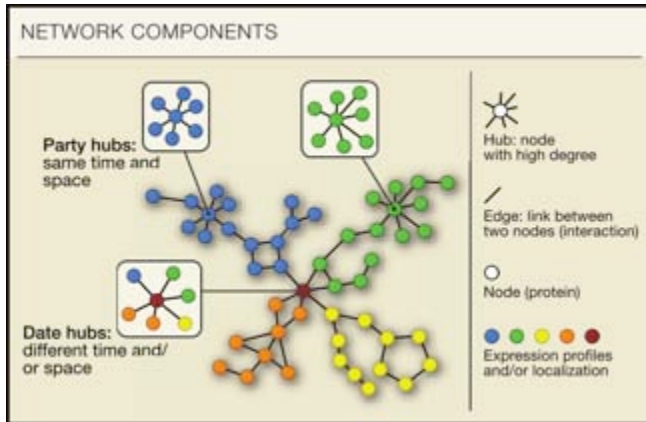
- Disorder Class
- Bone
  - Cancer
  - Cardiovascular
  - Connective tissue
  - Dermatological
  - Developmental
  - Ear, Nose, Throat
  - Endocrine
  - Gastrointestinal
  - Hematological
  - Immunological
  - Metabolic
  - Muscular
  - Neurological
  - Nutritional
  - Ophthalmological
  - Psychiatric
  - Renal
  - Respiratory
  - Skeletal
  - multiple
  - Unclassified

*Genes contributing to a common disease interact through protein-protein interactions.*

Genes connected if implicated in the same disorder.

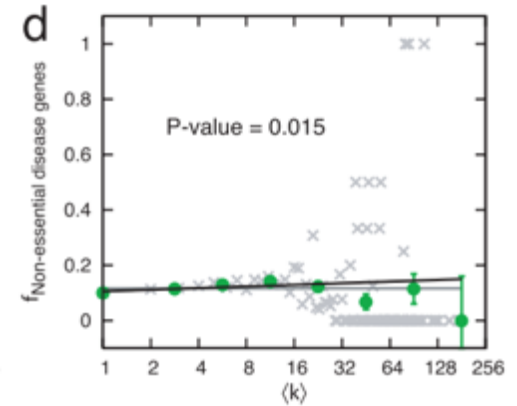
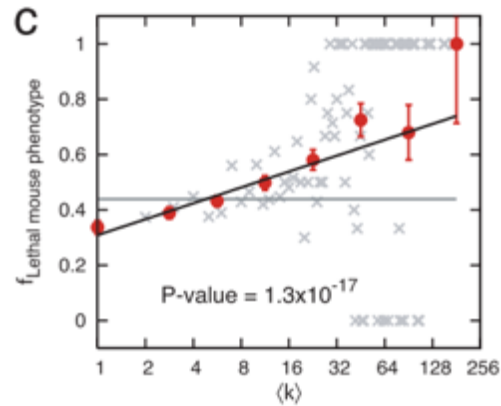
Goh et al. (2007). PNAS 104:8685.

# Distribution of Disease Genes



Seebacher and Gavin (2011). *Cell* 144:1000-1001

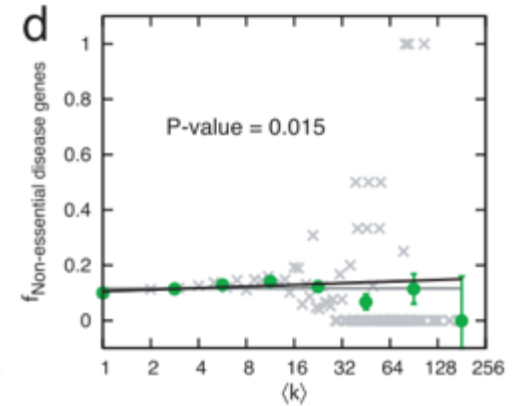
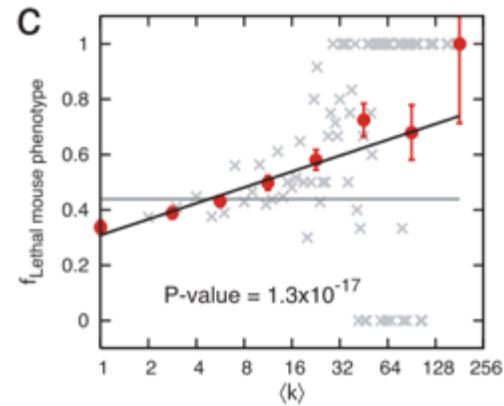
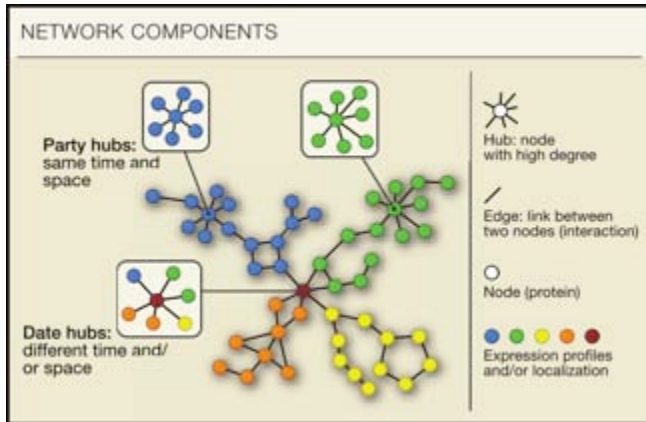
$k$  = degree  
= # interaction partners



- “Essential” genes
  - Encode hubs
  - Are expressed globally
- “Non-essential” disease genes
  - Do not encode hubs
  - Tissue specific expression

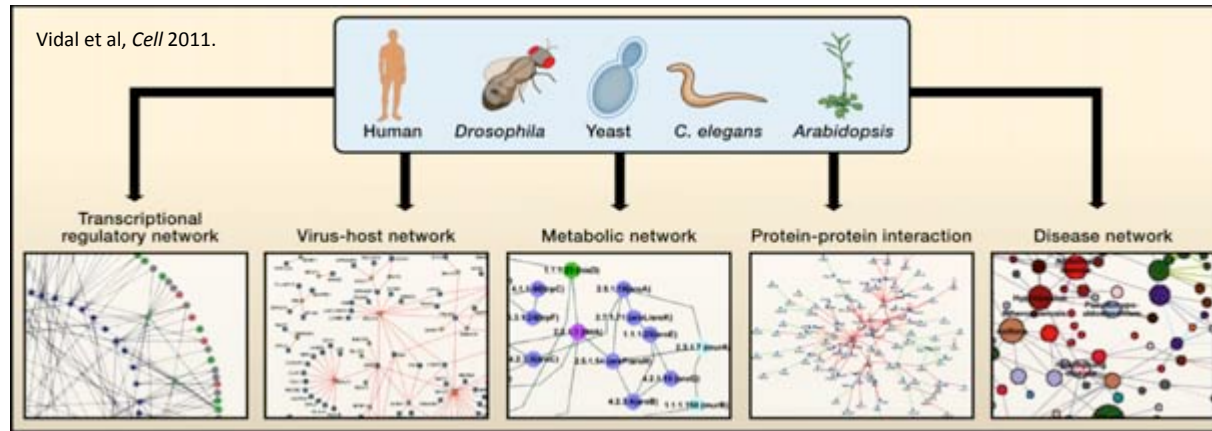


# Distribution of Disease Genes



**Nonrandom placement of  
disease genes in interactome!**

# Interactome Mapping & Data Integration



microRNA.org



PHOSIDA  
Posttranslational Modification Database

# Data Integration: Genetic Variation & Gene Expression

## Mapping the Genetic Architecture of Gene Expression in Human Liver

Wei Li, Zhongyuan Zhang, Jiyun Zhang, et al. *Nature Genetics* 2012, 44:1181-1190

**Abstract:** We have conducted a genome-wide association study (GWAS) to identify genetic variants associated with gene expression levels in human liver. We identified 1,128 independent SNPs associated with the expression of 1,128 genes. These SNPs are enriched in regulatory regions and are associated with chromatin accessibility and transcription factor binding. Our findings provide insights into the genetic architecture of gene expression and its relationship to disease risk.

**Key words:** GWAS, gene expression, liver, genetic variants, chromatin accessibility, transcription factors.

**Introduction:** Gene expression is a complex trait influenced by both genetic and environmental factors. Understanding the genetic architecture of gene expression is crucial for identifying disease risk and developing targeted therapies. This study aims to map the genetic architecture of gene expression in human liver using GWAS.

**Methods:** We performed a GWAS on gene expression data from 1,000 individuals. We identified 1,128 independent SNPs associated with the expression of 1,128 genes. We then performed functional annotations and pathway analysis to understand the biological significance of these SNPs.

**Results:** We identified 1,128 independent SNPs associated with the expression of 1,128 genes. These SNPs are enriched in regulatory regions and are associated with chromatin accessibility and transcription factor binding. Our findings provide insights into the genetic architecture of gene expression and its relationship to disease risk.

## ARTICLES

### A genome-wide association study of global gene expression

James L. Hwang, et al. *Nature Genetics* 2012, 44:1191-1200

**Abstract:** We have conducted a genome-wide association study (GWAS) to identify genetic variants associated with global gene expression levels. We identified 1,128 independent SNPs associated with the expression of 1,128 genes. These SNPs are enriched in regulatory regions and are associated with chromatin accessibility and transcription factor binding.

**Key words:** GWAS, gene expression, global, genetic variants, chromatin accessibility, transcription factors.

**Introduction:** Gene expression is a complex trait influenced by both genetic and environmental factors. Understanding the genetic architecture of gene expression is crucial for identifying disease risk and developing targeted therapies. This study aims to map the genetic architecture of global gene expression using GWAS.

**Methods:** We performed a GWAS on global gene expression data from 1,000 individuals. We identified 1,128 independent SNPs associated with the expression of 1,128 genes.

**Results:** We identified 1,128 independent SNPs associated with the expression of 1,128 genes. These SNPs are enriched in regulatory regions and are associated with chromatin accessibility and transcription factor binding.

**Conclusion:** Our findings provide insights into the genetic architecture of global gene expression and its relationship to disease risk.

## Genetic Inheritance of Gene Expression in Human Cell Lines

S. A. Madhukar, et al. *Nature Genetics* 2012, 44:1201-1210

**Abstract:** We have conducted a study to understand the genetic inheritance of gene expression in human cell lines. We identified 1,128 independent SNPs associated with the expression of 1,128 genes. These SNPs are enriched in regulatory regions and are associated with chromatin accessibility and transcription factor binding.

**Key words:** Genetic inheritance, gene expression, human cell lines, genetic variants, chromatin accessibility, transcription factors.

**Introduction:** Gene expression is a complex trait influenced by both genetic and environmental factors. Understanding the genetic inheritance of gene expression is crucial for identifying disease risk and developing targeted therapies. This study aims to map the genetic inheritance of gene expression in human cell lines using GWAS.

**Methods:** We performed a GWAS on gene expression data from human cell lines. We identified 1,128 independent SNPs associated with the expression of 1,128 genes.

**Results:** We identified 1,128 independent SNPs associated with the expression of 1,128 genes. These SNPs are enriched in regulatory regions and are associated with chromatin accessibility and transcription factor binding.

**Conclusion:** Our findings provide insights into the genetic inheritance of gene expression in human cell lines.

## ARTICLES

### Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function

Oliver J. Hellmich, et al. *Nature Genetics* 2012, 44:1211-1220

**Abstract:** We have conducted a complex trait analysis of gene expression to uncover polygenic and pleiotropic networks that modulate nervous system function. We identified 1,128 independent SNPs associated with the expression of 1,128 genes. These SNPs are enriched in regulatory regions and are associated with chromatin accessibility and transcription factor binding.

**Key words:** Complex trait analysis, gene expression, polygenic, pleiotropic networks, nervous system function, genetic variants, chromatin accessibility, transcription factors.

**Introduction:** Gene expression is a complex trait influenced by both genetic and environmental factors. Understanding the genetic architecture of gene expression is crucial for identifying disease risk and developing targeted therapies. This study aims to map the genetic architecture of gene expression in the nervous system using complex trait analysis.

**Methods:** We performed a complex trait analysis of gene expression data from 1,000 individuals. We identified 1,128 independent SNPs associated with the expression of 1,128 genes.

**Results:** We identified 1,128 independent SNPs associated with the expression of 1,128 genes. These SNPs are enriched in regulatory regions and are associated with chromatin accessibility and transcription factor binding.

**Conclusion:** Our findings provide insights into the genetic architecture of gene expression in the nervous system.

## Genetic analysis of genome-wide variation in human gene expression

Shantanu Datta, et al. *Nature Genetics* 2012, 44:1221-1230

**Abstract:** We have conducted a genetic analysis of genome-wide variation in human gene expression. We identified 1,128 independent SNPs associated with the expression of 1,128 genes. These SNPs are enriched in regulatory regions and are associated with chromatin accessibility and transcription factor binding.

**Key words:** Genetic analysis, genome-wide variation, human gene expression, genetic variants, chromatin accessibility, transcription factors.

**Introduction:** Gene expression is a complex trait influenced by both genetic and environmental factors. Understanding the genetic architecture of gene expression is crucial for identifying disease risk and developing targeted therapies. This study aims to map the genetic architecture of genome-wide variation in human gene expression using GWAS.

**Methods:** We performed a GWAS on genome-wide variation in human gene expression data from 1,000 individuals. We identified 1,128 independent SNPs associated with the expression of 1,128 genes.

**Results:** We identified 1,128 independent SNPs associated with the expression of 1,128 genes. These SNPs are enriched in regulatory regions and are associated with chromatin accessibility and transcription factor binding.

**Conclusion:** Our findings provide insights into the genetic architecture of genome-wide variation in human gene expression.

## LETTERS

### Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma

Miriam F. Feenstra, et al. *Nature Genetics* 2012, 44:1231-1240

**Abstract:** We have conducted a study to understand the genetic variants regulating ORMDL3 expression and their contribution to the risk of childhood asthma. We identified 1,128 independent SNPs associated with the expression of 1,128 genes. These SNPs are enriched in regulatory regions and are associated with chromatin accessibility and transcription factor binding.

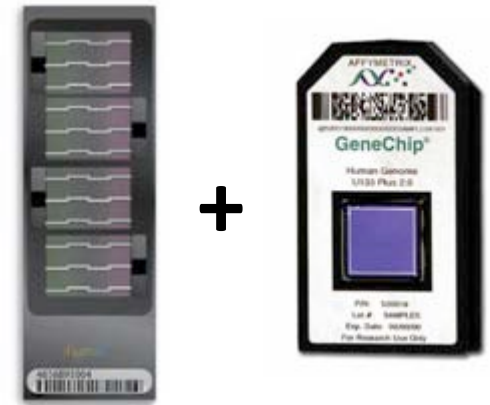
**Key words:** Genetic variants, ORMDL3 expression, childhood asthma, genetic variants, chromatin accessibility, transcription factors.

**Introduction:** Gene expression is a complex trait influenced by both genetic and environmental factors. Understanding the genetic architecture of gene expression is crucial for identifying disease risk and developing targeted therapies. This study aims to map the genetic architecture of ORMDL3 expression and its relationship to childhood asthma risk.

**Methods:** We performed a study to identify genetic variants regulating ORMDL3 expression. We identified 1,128 independent SNPs associated with the expression of 1,128 genes.

**Results:** We identified 1,128 independent SNPs associated with the expression of 1,128 genes. These SNPs are enriched in regulatory regions and are associated with chromatin accessibility and transcription factor binding.

**Conclusion:** Our findings provide insights into the genetic architecture of ORMDL3 expression and its relationship to childhood asthma risk.



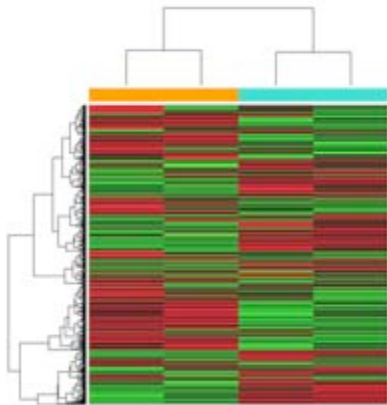
Are DNA variants that are associated with disease also associated with gene expression levels?

# Data Integration: Gene expression + DNA Binding

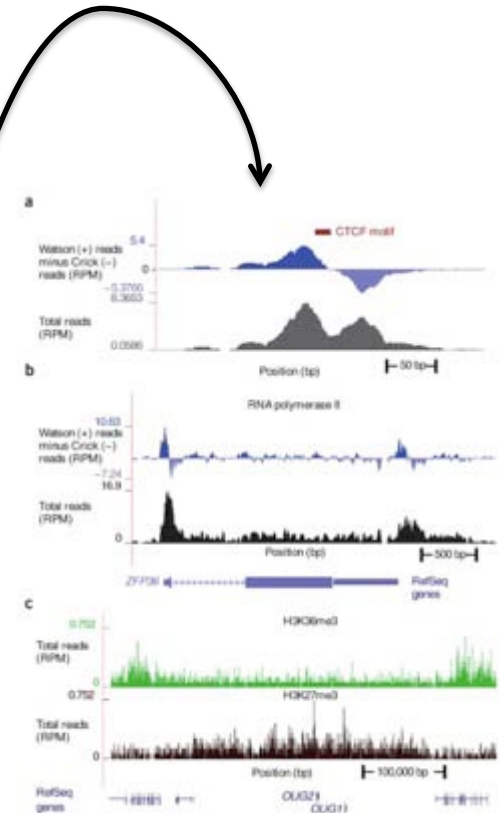
Gene expression arrays + ChIP-Seq



SKA3	spindle and kinetochore associated comp	221150	-3.6929178	0.0006329
KRT6A	keratin 6A	3853	-3.6046662	0.0006329
CDK1	cyclin-dependent kinase 1	983	-4.3548774	0.0006329
CA6	carbonic anhydrase VI	765	-4.1299515	0.0006329
SCRG1	stimulator of chondrogenesis 1	11341	-3.3564348	0.0006329
KIAA0101	KIAA0101	9768	-4.2823439	0.0006329
DHFR	dihydrofolate reductase	1719	-3.3352926	0.0006573
HST1H18	histone cluster 1, H1b	3009	-3.7588129	0.0006573
CDC20	cell division cycle 20 homolog (S. cerevisiae)	991	-3.8883015	0.0006573
NCAPG	non-SMC condensin I complex, subunit G	64153	-3.8576129	0.0006573
CCNB2	cyclin B2	9133	-3.882789	0.0006573
PRR11	proline rich 11	55771	-3.7027197	0.0006573
BUB1	budding uninhibited by benzimidazoles 1	699	-3.8005975	0.0006573
FAM111B	family with sequence similarity 111, member	374393	-3.388453	0.0006573
CASP1	caspase 1, apoptosis-related cysteine peptidase	834	-3.099429	0.0006573
TTK	TTK protein kinase	7272	-3.4101424	0.0006573
CLCA4	chloride channel accessory 4	22802	-3.2658635	0.0007305
GNS2	GINS complex subunit 2 (Pif2 homolog)	51659	-3.2227793	0.0007305
PSG5	pregnancy specific beta-1-glycoprotein 5	5673	3.3272996	0.0007631

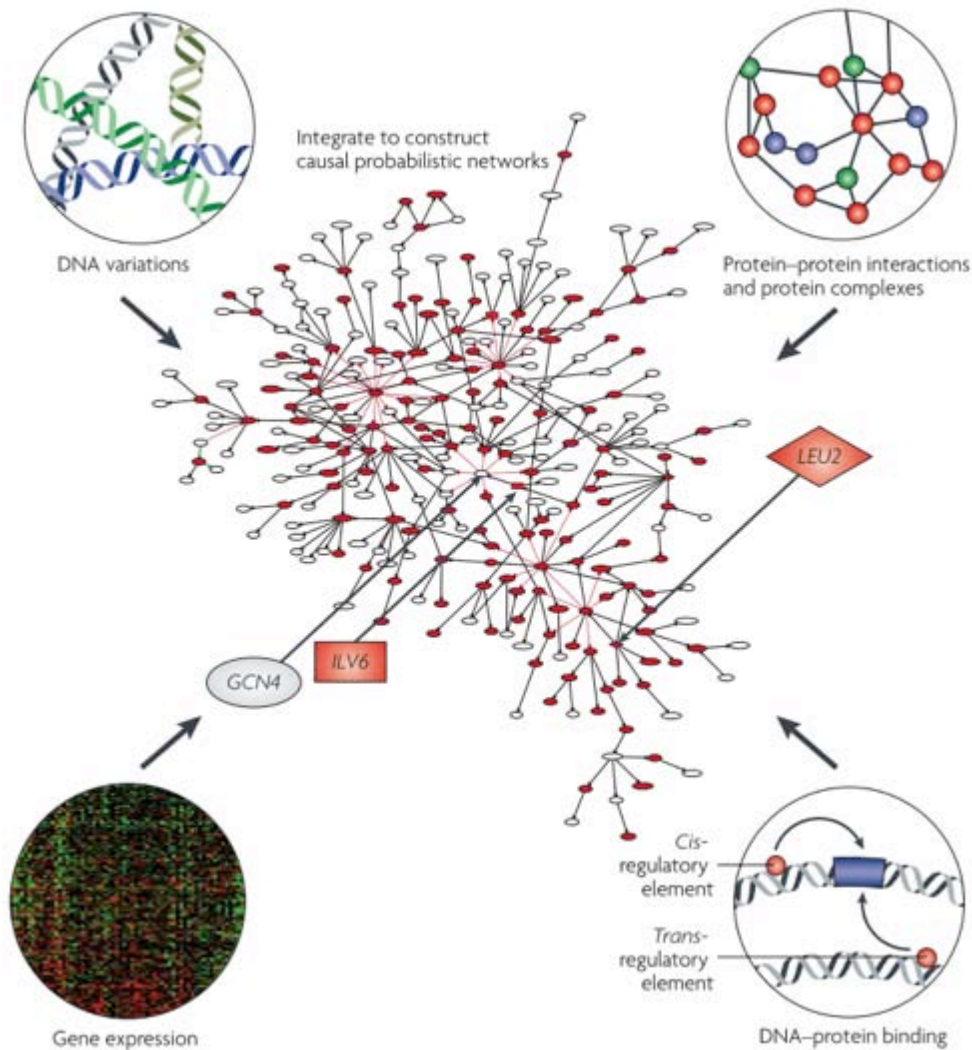


**BIOBASE**  
BIOLOGICAL DATABASES





# Data Integration: 4 Dimensions



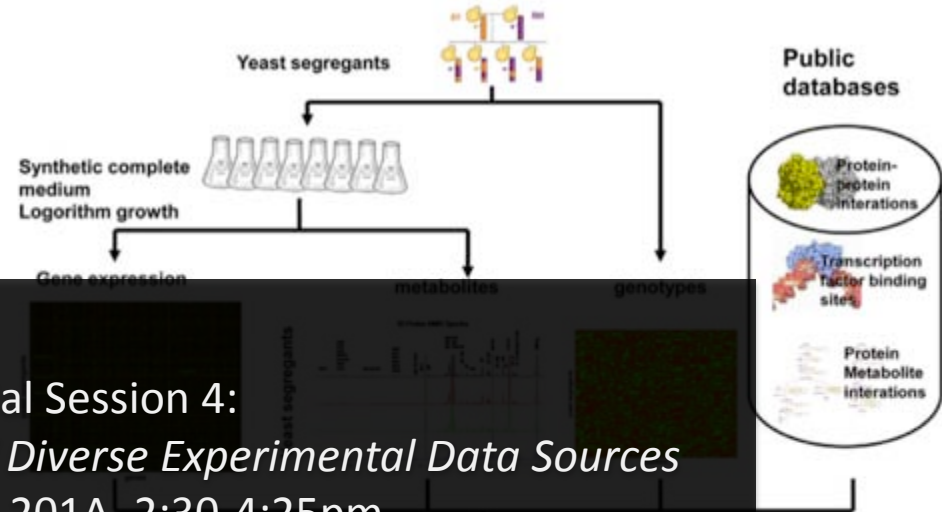
Schadt et al. 2009. Network view of disease and compound screening. *Nat Rev Drug Discovery* 8:286.

Probabilistic Bayesian Network Integrating:

1. Genetic variation
2. Gene expression
3. Protein-protein interactions
4. Transcript factor binding

# Data Integration: 6 Dimensions

Zhu J, ... Schadt EE. 2012. Stitching together Multiple Data Dimensions Reveals Interacting Metabolomic and Transcriptomic Networks that Modulate Cell Regulation. *PLoS Biol.*



1. Metabolite concentrations
2. RNA expression
3. DNA Variation
4. DNA-protein binding
5. Protein-protein interaction
6. Protein-metabolite interaction

Special Session 4:  
*Bioinformatic Integration of Diverse Experimental Data Sources*

**Today**, room 201A, 2:30-4:25pm

Part D (4:00-4:25): network

*Stitching together multiple data dimensions...*

- Metabolites linked to DNA variants (MetQTLs)
- MetQTLs co-localize with eQTLs
- Using a Bayesian network
  - Nodes: DNA variation, gene expresion, metabolite concentration
  - Priors: Protein-DNA binding, protein-protein interaction, metabolite-protein interaction
  - Edges: Inferred relationships → mechanism

Infer causality

# Data Integration: Mouse *Cis*-Regulatory Map

LETTER

doi:10.1038/nature11243

## A map of the *cis*-regulatory sequences in the mouse genome

Yin Shen<sup>1\*</sup>, Feng Yue<sup>1\*</sup>, David F. McCleary<sup>1</sup>, Zhen Ye<sup>1</sup>, Lee Edsall<sup>1</sup>, Samantha Kuan<sup>1</sup>, Ulrich Wagner<sup>1,2,3</sup>, Jesse Dixon<sup>1,2,3</sup>, Leonard Lee<sup>1</sup>, Victor V. Lobanov<sup>4</sup> & Bing Ren<sup>1,5</sup>

The laboratory mouse is the most widely used mammalian model organism in biomedical research. The  $2.6 \times 10^9$  bases of the mouse genome possess a high degree of conservation with the human genome<sup>6</sup>, so a thorough annotation of the mouse genome will be of significant value to understand the mouse genome. So far, most of the genome has yet to be fully annotated. In particular, cis-regulatory sequences are still poorly understood. Recently, ChIP-Seq has been used to identify cis-regulatory elements in the genomes of *Drosophila melanogaster* and other model organisms. We apply the same experimental design and cell types in the mouse genome to identify cis-regulatory sequences. We find that 11% of the mouse genome is covered by cis-regulatory sequences. We identify potential transcription factor binding sites for each tissue or cell type. The mouse genome is organized into domains of coordinately regulated enhancers and promoters. Our results provide a resource for the annotation of functional elements in the mammalian genome and for the study of mechanisms regulating tissue-specific gene expression.

of significant value to understand the mouse genome. So far, most of the genome has yet to be fully annotated. In particular, cis-regulatory sequences are still poorly understood. Recently, ChIP-Seq has been used to identify cis-regulatory elements in the genomes of *Drosophila melanogaster* and other model organisms. We apply the same experimental design and cell types in the mouse genome to identify cis-regulatory sequences. We find that 11% of the mouse genome is covered by cis-regulatory sequences. We identify potential transcription factor binding sites for each tissue or cell type. The mouse genome is organized into domains of coordinately regulated enhancers and promoters. Our results provide a resource for the annotation of functional elements in the mammalian genome and for the study of mechanisms regulating tissue-specific gene expression.

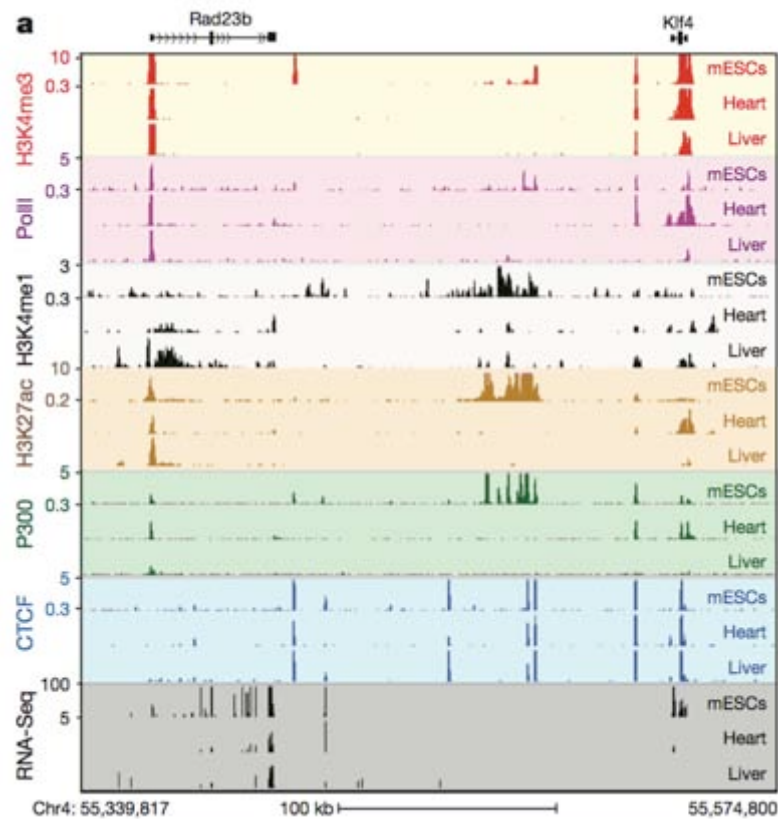


Figure 1 | Identification of cis-regulatory sequences in the mouse genome. UCSC genome browser view of heart and liver (chromosome 4), input normalized intensities. kb.

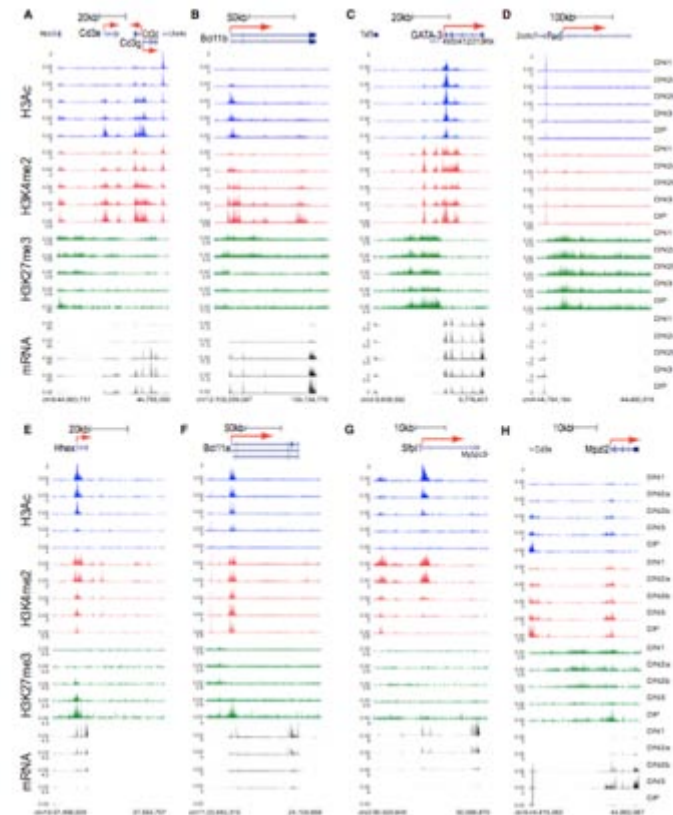
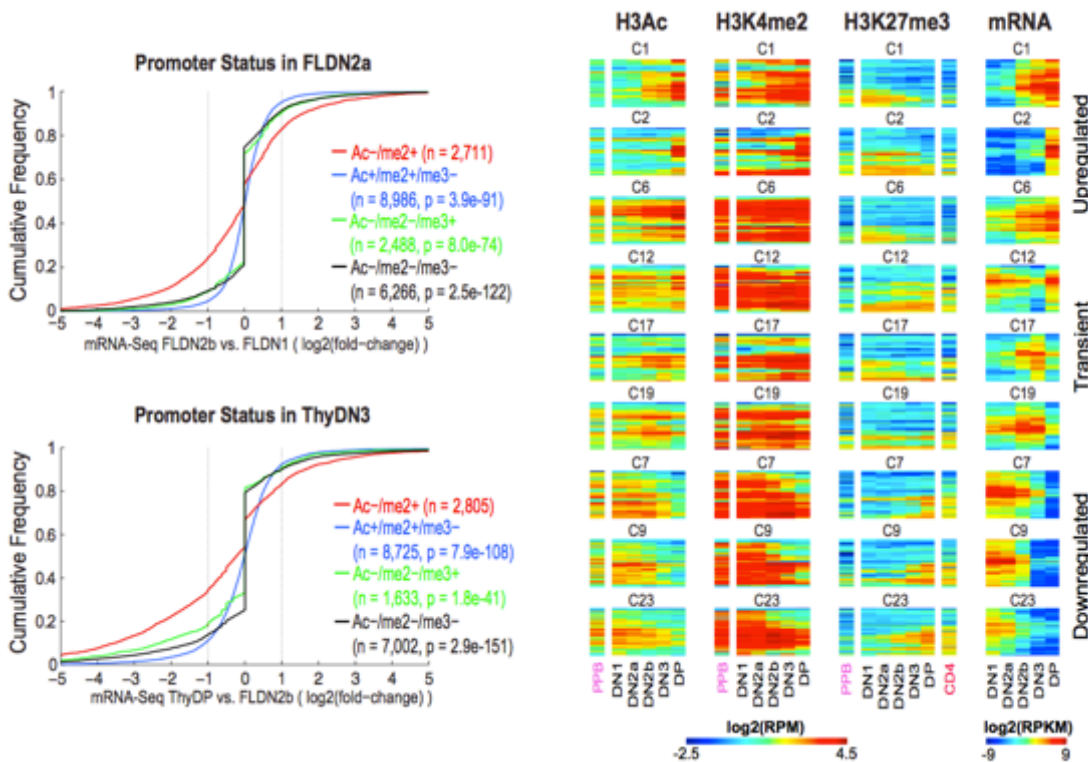
<sup>1</sup> Ludwig Institute for Cancer Research, 3601 La Jolla Village Drive, La Jolla, California 92037-0632, USA  
<sup>2</sup> Laboratory of Immunogenetics, National Institute of Health, Bethesda, Maryland 20892, USA  
<sup>3</sup> Department of Cell Biology and Biophysics, University of California, San Diego, La Jolla, California 92037, USA  
<sup>4</sup> These authors contributed equally to this work.

Shen et al. A map of the cis-regulatory sequences in the mouse genome. *Nature*, July 2012.

- RNA-Seq and ChIP-Seq for 6 DNA-binding factors \* 19 cell types
  - **ChIP:** PolII, H3K4me3, H3K4me1, H3K27ac, P300, CTCF
  - **Adult Tissues:** bone marrow, cerebellum, cortex, heart, intestine, kidney, liver, lung, olfactory bulb, placenta, spleen, testis, thymus
  - **Embryonic Tissues:** brain, heart, limb, liver
  - **Cell lines:** mESCs, MEFs
- Found 300,000 *cis*-reg features
  - 11% mouse genome
  - 70% conserved non-coding sequence

# Data Integration: Epigenome & Transcriptome

- Zhang JA, Mortazavi A, Williams BA, Wold BJ, Rothenberg EV. Dynamic Transformations of Genome-wide Epigenetic Marking and Transcriptional Control Establish T Cell Identity. *Cell* 2012.
- ChIP-Seq + RNA-Seq in sequential T-cell developmental stages
- Changes in gene expression co-occur w/ histone modification at *cis*-regulatory sites.





# Summary

- Data is cheap and diverse.
  - Genetic variation: GWAS, next-gen sequencing
  - Gene expression: Microarray, RNA-seq
  - Proteomics: Y2H, CoAP/MS
- Cellular components interact in a network with other cellular components.
- Disease is the result of an abnormality in that network.
- Integrate multiple data types, understand network, understand disease.



# Thank you



UNIVERSITY  
*of* VIRGINIA  
SCHOOL *of* MEDICINE



bioinformatics core

[bioinformatics.virginia.edu](http://bioinformatics.virginia.edu)

Web: [bioinformatics.virginia.edu](http://bioinformatics.virginia.edu)

E-mail: [bioinformatics@virginia.edu](mailto:bioinformatics@virginia.edu)

Blog: [www.GettingGeneticsDone.com](http://www.GettingGeneticsDone.com)

Twitter: [twitter.com/genetics\\_blog](https://twitter.com/genetics_blog)