From Data to Knowledge:

Extracting Biological Insight from Diverse Data Sources

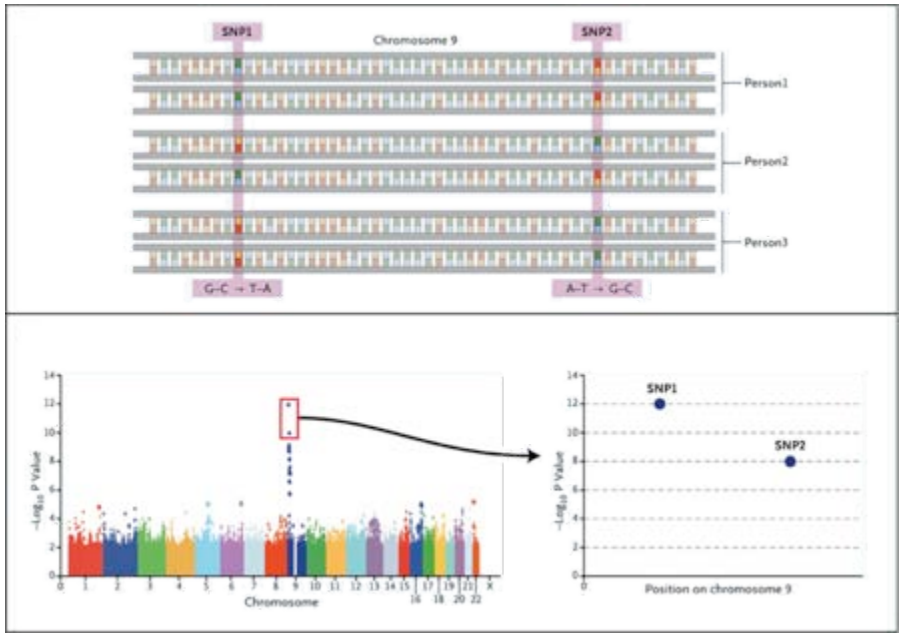
Stephen D. Turner, Ph.D.

Bioinformatics Core Director

bioinformatics@virginia.edu

bioinformatics.virginia.edu

GWAS: One gene, one enzyme, one function?



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

Genome.gov/GWAStudies

*GENETIC CONTROL OF BIOCHEMICAL REACTIONS IN NEUROSPORA**

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

BIOLOGICAL DEPARTMENT, STANFORD UNIVERSITY

Communicated October 8, 1941

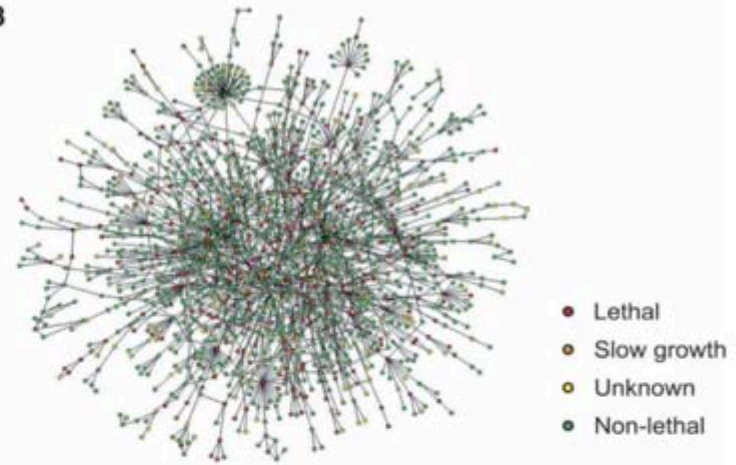
DNA Variation: Limitations

GWAS DOES NOT INFORM:

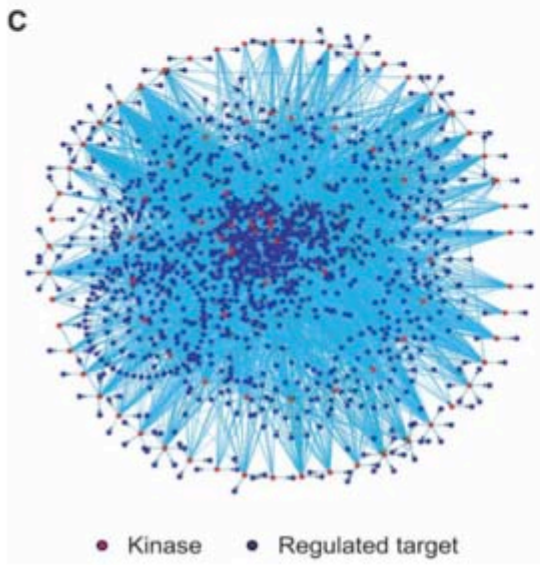
- Which gene is affected
- How gene function is perturbed
- How biological processes are altered

One gene, one enzyme, one function?

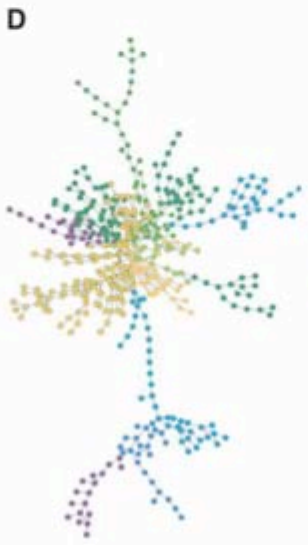
Zhu X. et al. (2007). *Genes & Dev 21:1010-1024.*



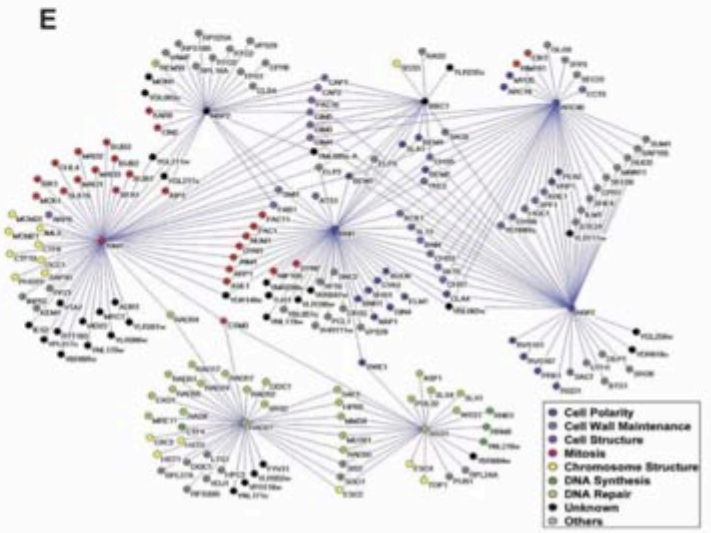
Jeong, H. et al.. *(2001) Nature 411:41–42.*



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

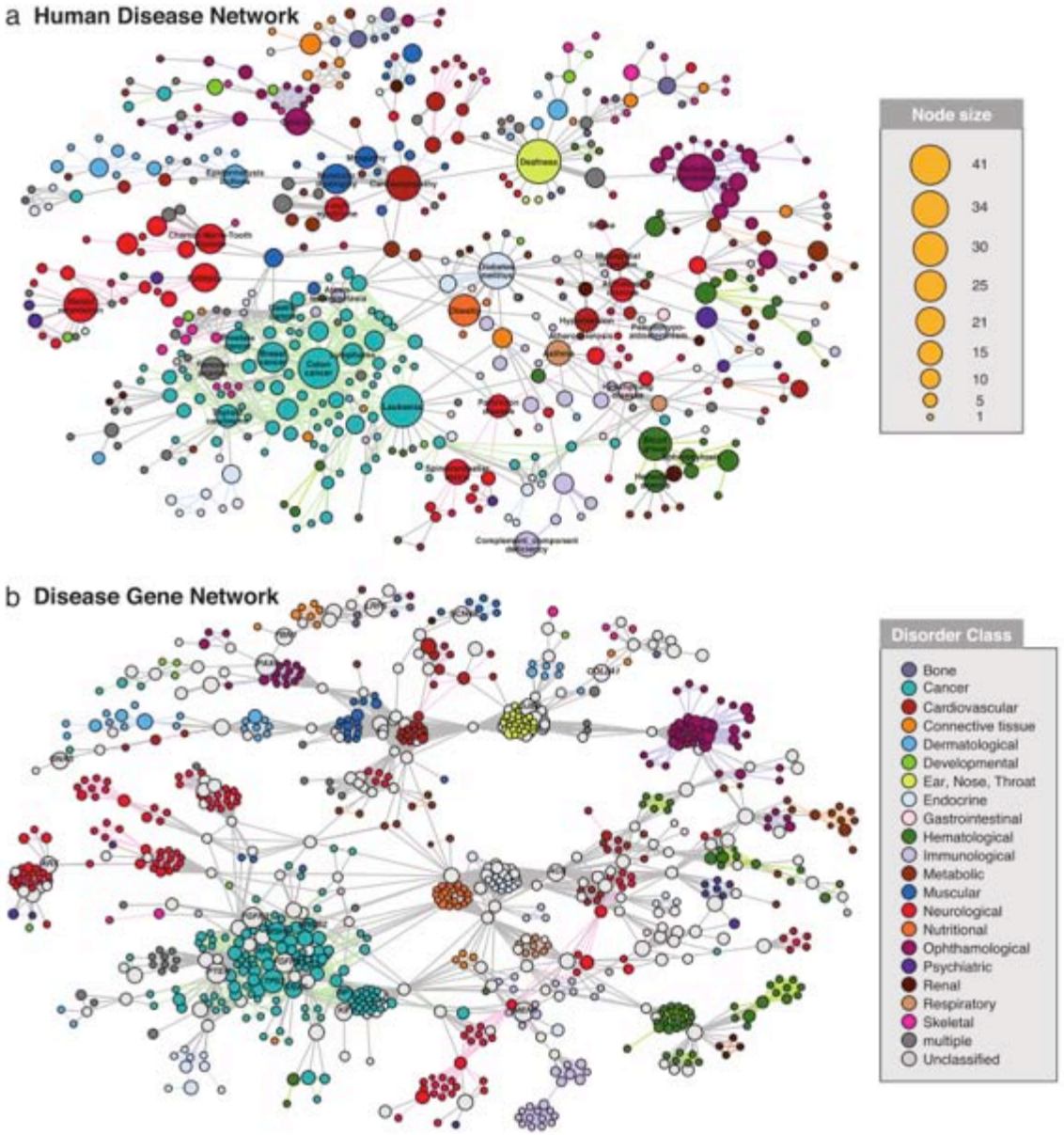


Guimera and Amaral. (2005). *Nature 433:895-900.*



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

Distribution of Disease Genes



Diseases connected if same gene implicated in both.

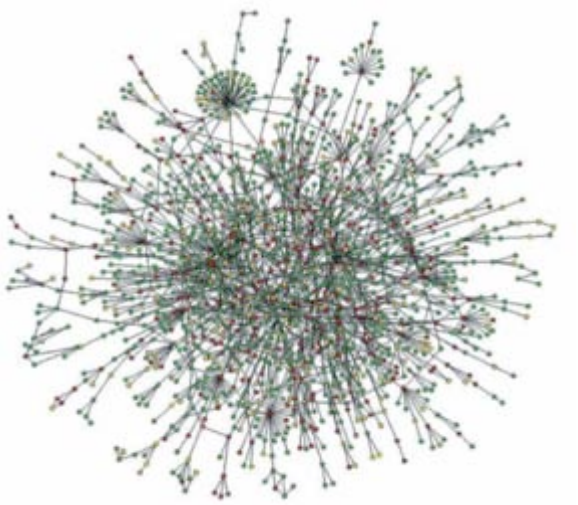
Genes connected if implicated in the same disorder.

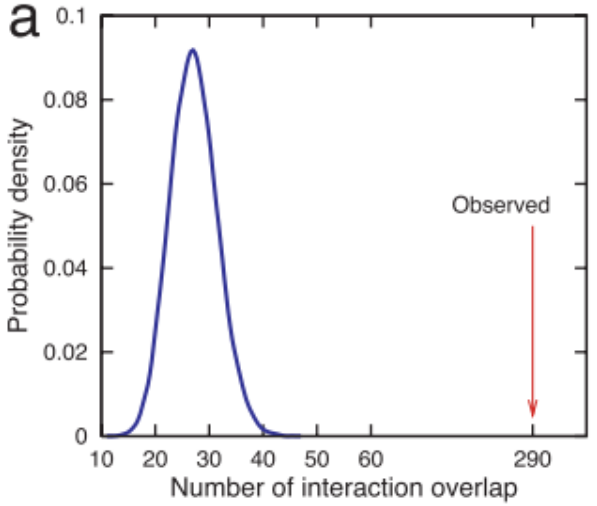
Goh et al. (2007). PNAS 104:8685.

Distribution of Disease Genes

Protein-protein interactions

Overlay with PPI data



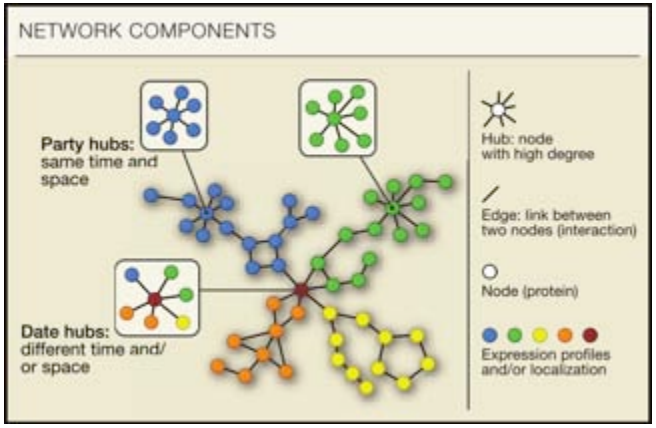


Genes contributing to a common disease interact through protein-protein interactions.

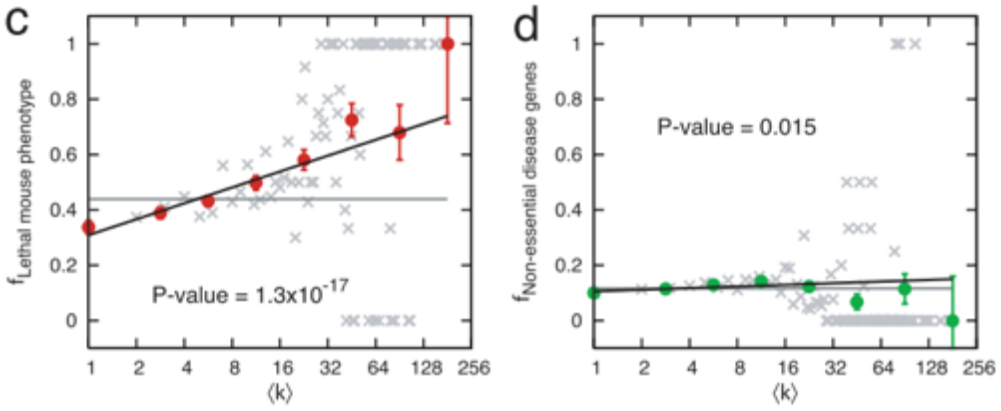
Genes connected if implicated in the same disorder.

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

Distribution of Disease Genes



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

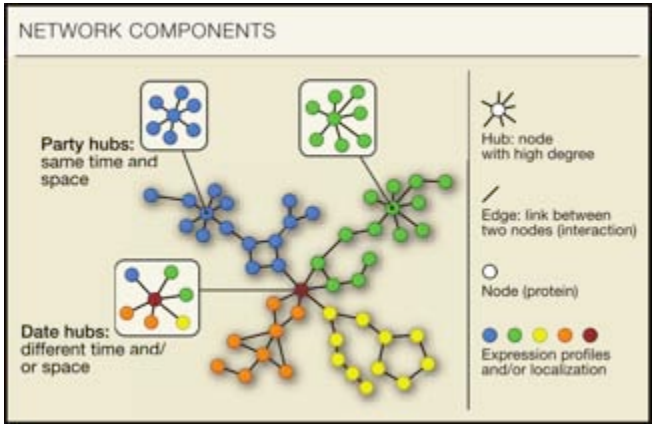


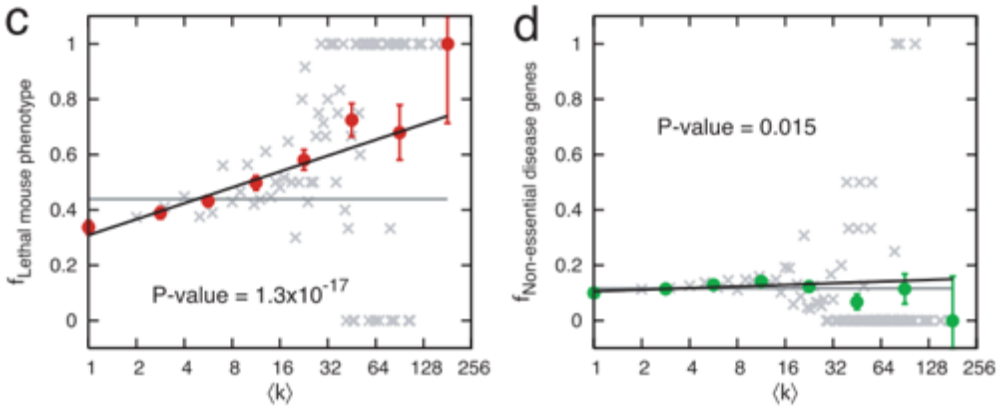
k = degree
= # interaction partners

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

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

Distribution of Disease Genes

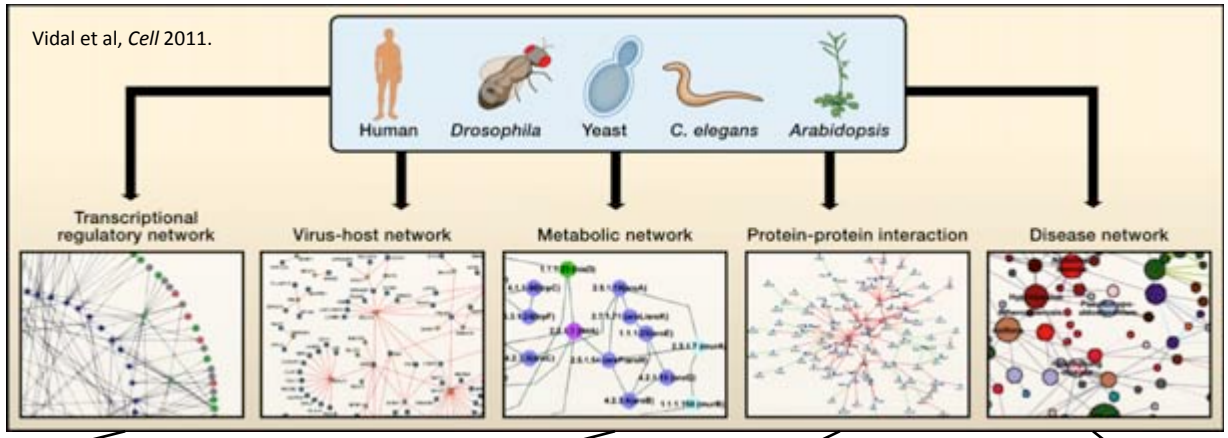




Nonrandom placement of disease genes in interactome!

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

Interactome Mapping & Data Integration



JASPAR

KEGG Kyoto Encyclopedia of Genes and Genomes

IMEx

BioGRID 3.1

HuDiNe Search

miRBase

microRNA.org

Human Protein Reference Database

OMIM Online Mendelian Inheritance in Man

PHOSIDA Posttranslational Modification Database

Data Integration: Genetic Variation & Gene Expression


Mapping the Genetic Architecture of Gene Expression in Human Liver

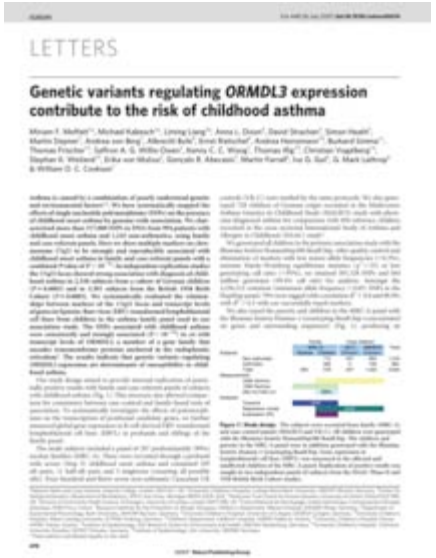

Genetic Inheritance of Gene Expression in Human Cell Lines


Genetic analysis of genome-wide variation in human gene expression

+


ARTICLES
nature genetics
A genome-wide association study of global gene expression

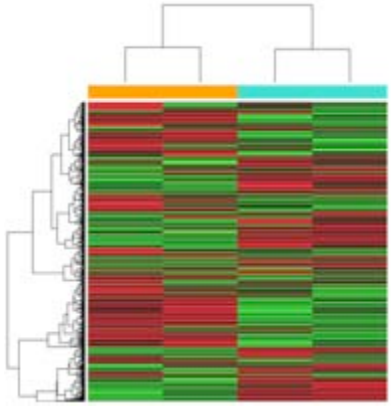

nature genetics
Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function


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

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

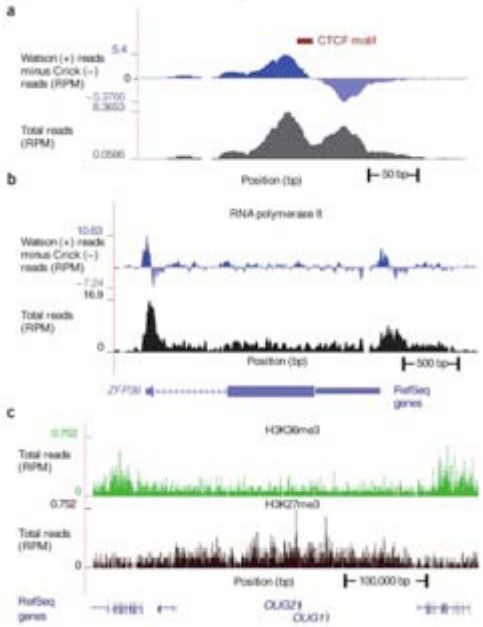
Data Integration: Gene expression + DNA Binding

Gene expression arrays + ChIP-Seq

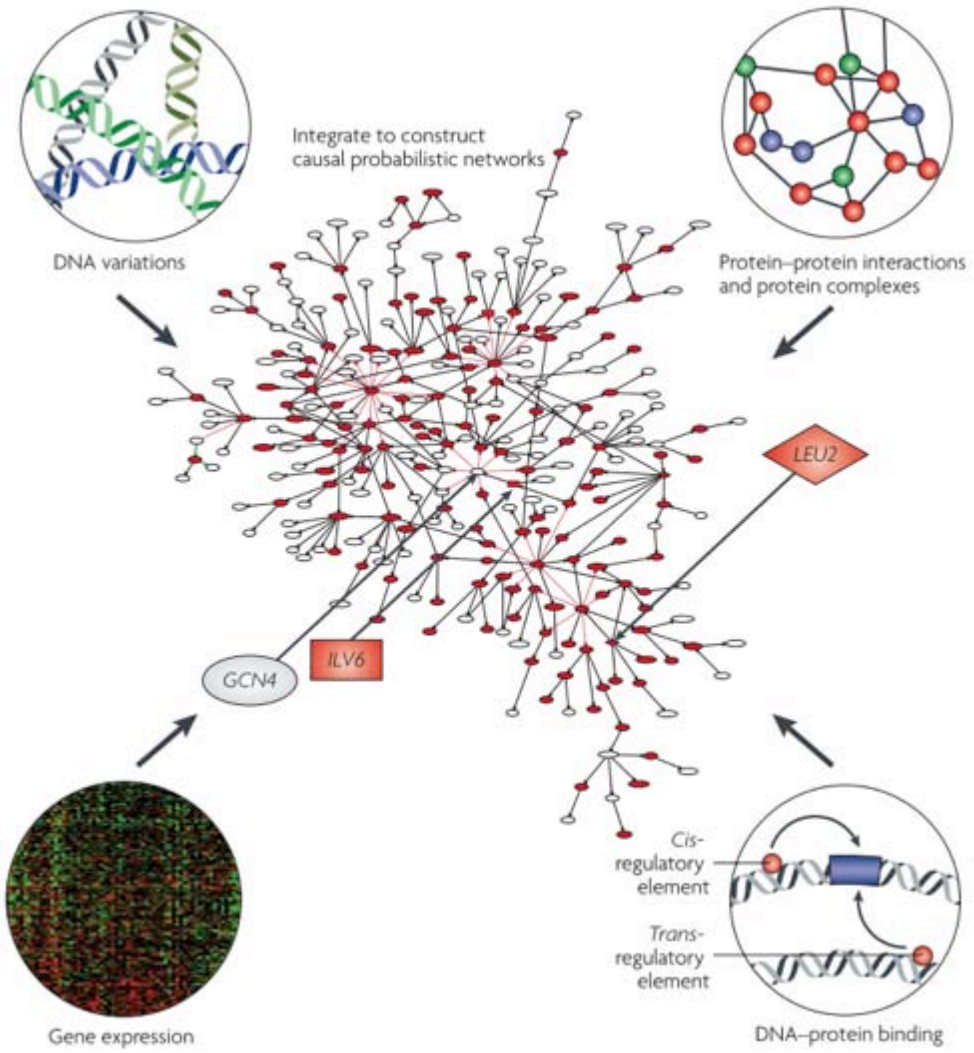


SKA3	spindle and kinetochore associated comp	221150	-3.6929178	0.0006329
KRT6A	keratin 6A	3853	-3.6046662	0.0006329
CDK1	cyclin-dependent kinase 1	983	-4.3548774	0.0006329
CA6	carbonic anhydrase VI	765	-4.1299515	0.0006329
SCRG1	stimulator of chondrogenesis 1	11341	-3.3564348	0.0006329
KIAA0101	KIAA0101	9768	-4.2823439	0.0006329
DHFR	dihydrofolate reductase	1719	-3.3352926	0.0006573
HIST1H1B	histone cluster 1, H1b	3009	-3.7588129	0.0006573
CDC20	cell division cycle 20 homolog (S. cerevisi	991	-3.8883015	0.0006573
NCAPG	non-SMC condensin I complex, subunit G	64151	-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, mem	374393	-3.388453	0.0006573
CASP1	caspase 1, apoptosis-related cysteine pep	834	-3.099429	0.0006573
TTK	TTK protein kinase	7272	-3.4101424	0.0006573
CLCA4	chloride channel accessory 4	22802	-3.2658635	0.0007305
GINS2	GINS complex subunit 2 (Psf2 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



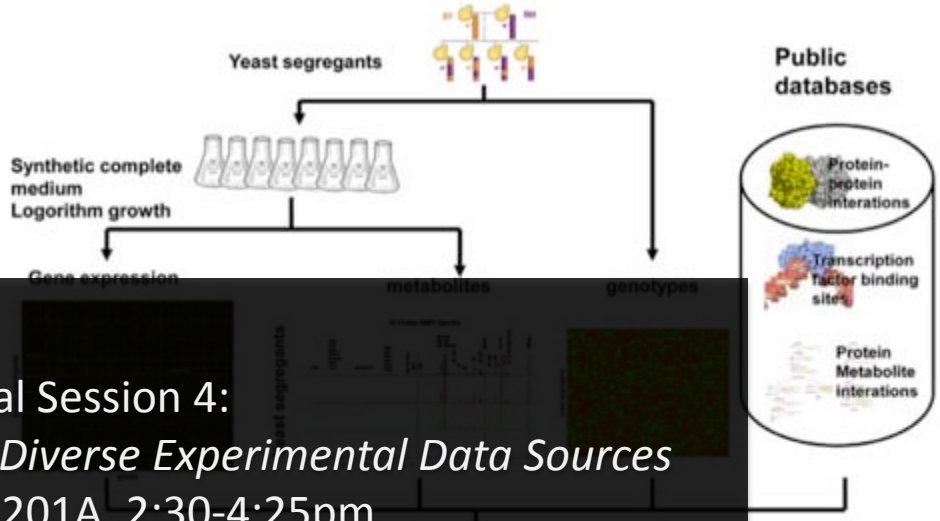
Schadt et al. 2009. Network view of disease and compound screening. *Nat Rev Drug Discovery 8:286*.

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):

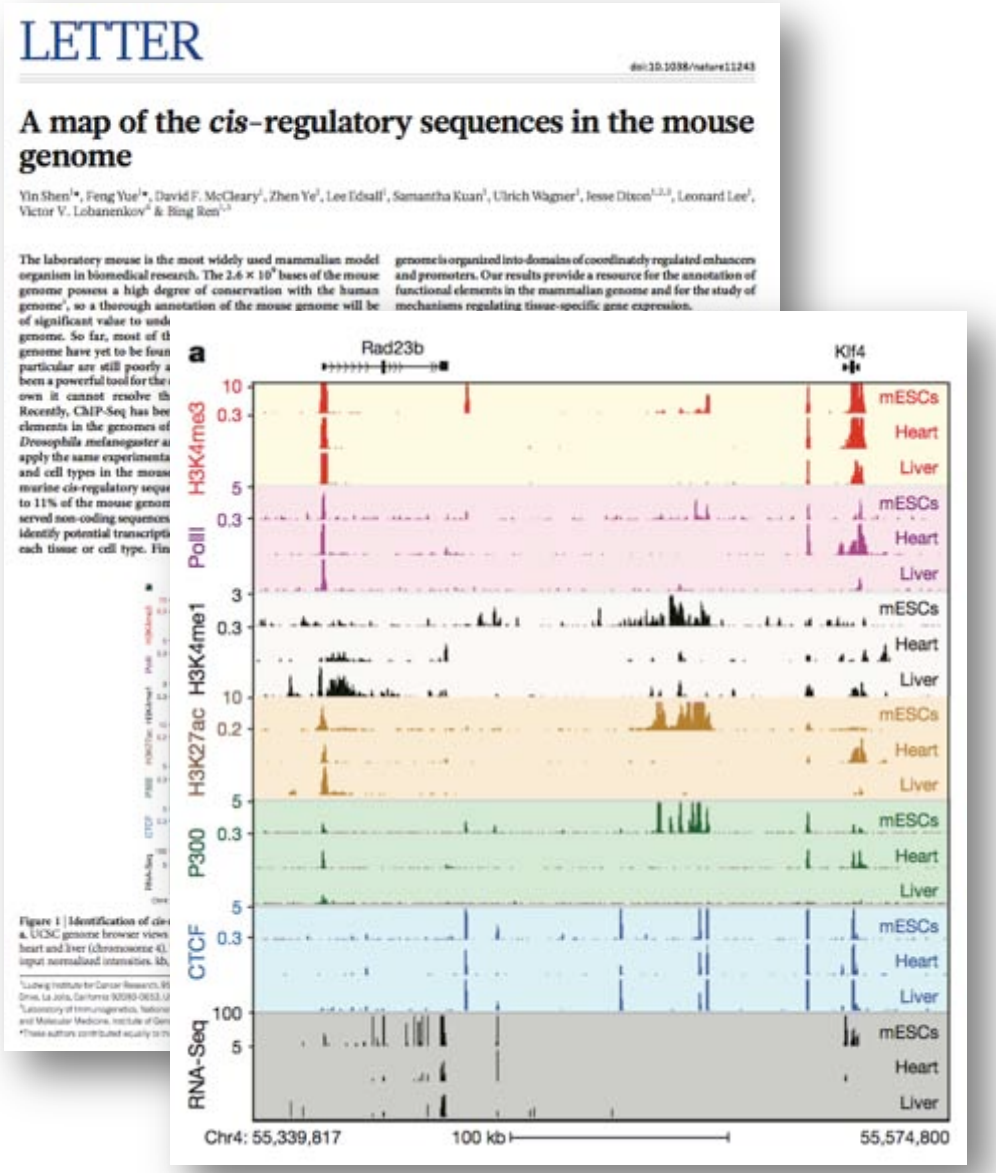
Stitching together multiple data dimensions…

- Metabolites linked to DNA variants (MetQTLs)
- MetQTLs co-localize with eQTLs
- Using a Bayesian network

Infer causality

  - Nodes: DNA variation, gene expresion, metabolite concentration
  - Priors: Protein-DNA binding, protein-protein interaction, metabolite-protein interaction
  - Edges: Inferred relationships → mechanism

Data Integration: Mouse *Cis*-Regulatory Map



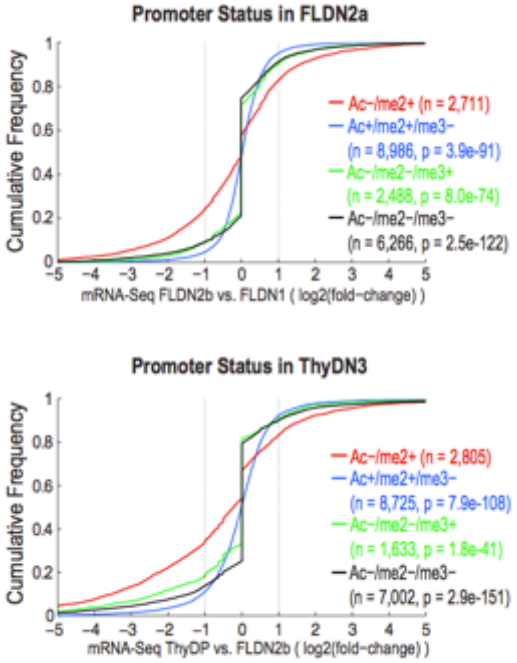
Shen et al. A map of the cis-regulatory sequences in the mouse genome. *Nature*, July 2012.

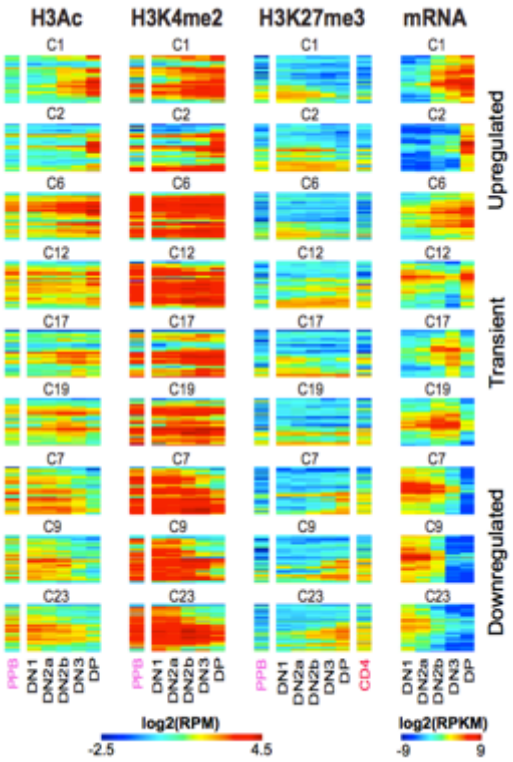
- RNA-Seq and ChIP-Seq for 6 DNA-binding factors * 19 cell types
  - **ChIP**: PolII, H3K4me3, H3K4me1, H3K27ac, P300, CTCF
  - **Adult Tissues**: bone marrow, cerebellum, cortex, heart, intestine, kidney, liver, lung, olfactory bulb, placenta, spleen, testis, thymus
  - **Embryonic Tissues**: brain,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

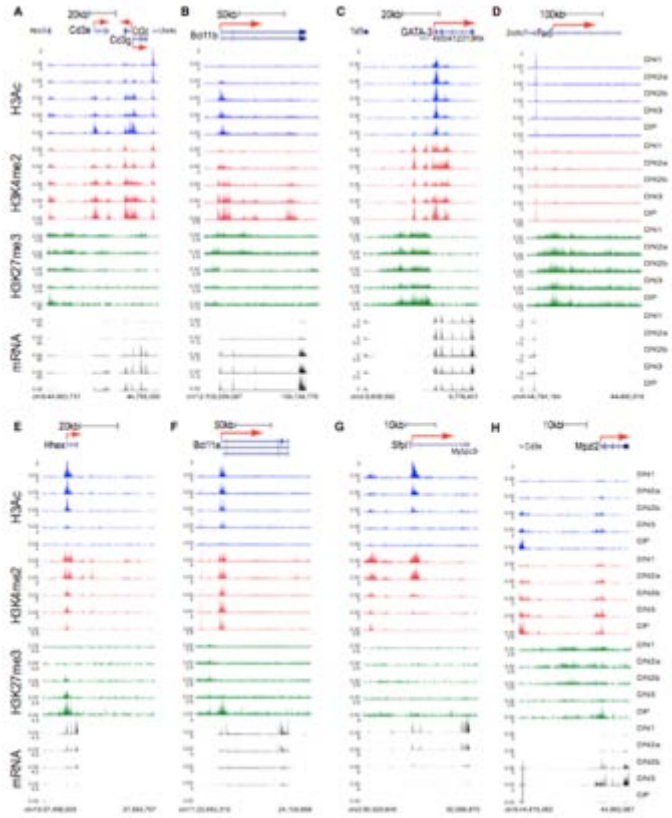
University of Virginia School of Medicine bioinformatics core

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

Web: bioinformatics.virginia.edu

E-mail: bioinformatics@virginia.edu

Blog: www.GettingGeneticsDone.com

Twitter: twitter.com/genetics_blog