

# USING NETWORKS TO UNDERSTAND THE GENOTYPE-PHENOTYPE CONNECTION

---

**John Quackenbush**

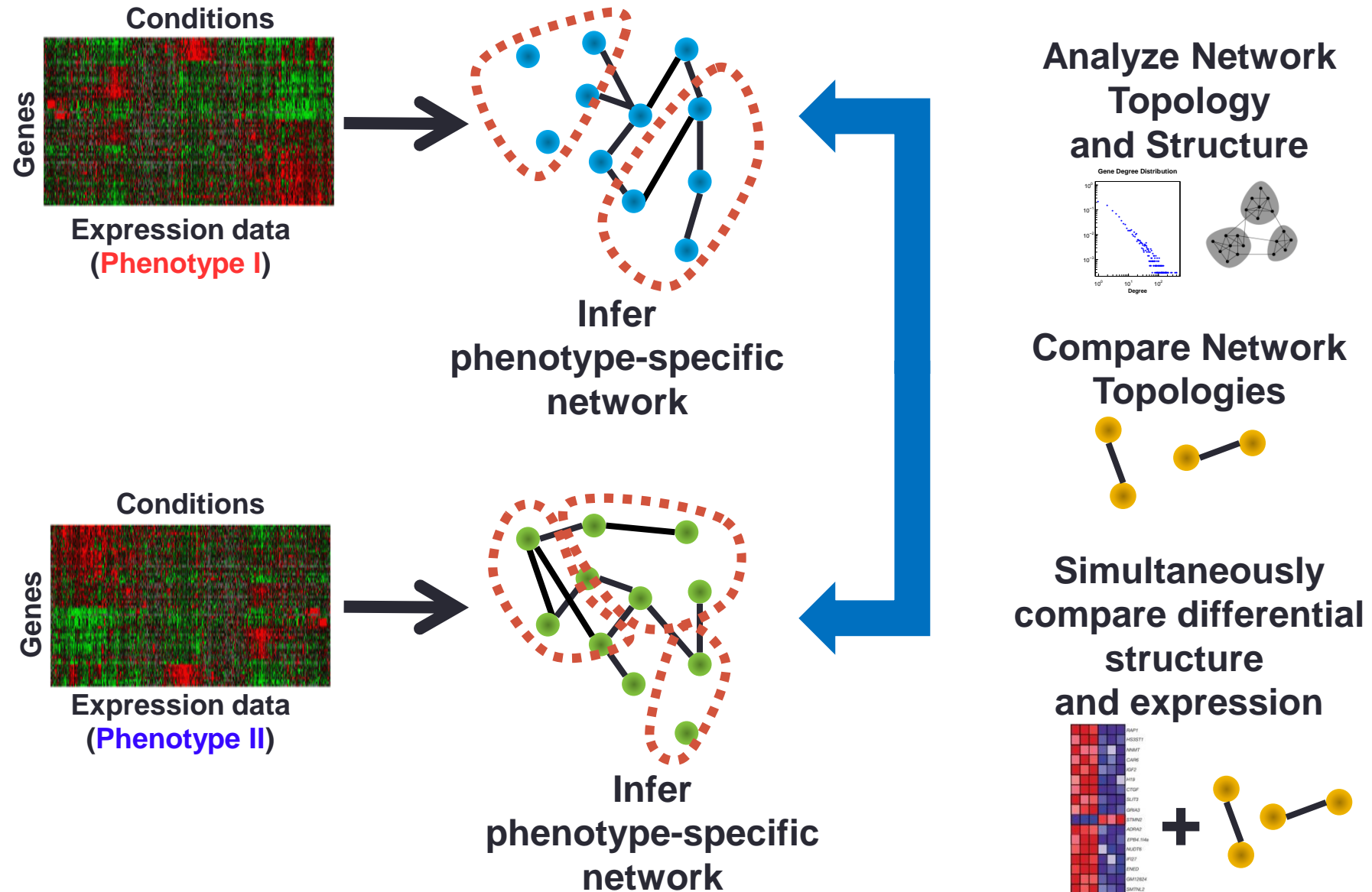
**Dana-Farber Cancer Institute**

**Harvard TH Chan School of Public Health**

**Essentially, all models are wrong,  
but some are useful.**

**– George E. Box**

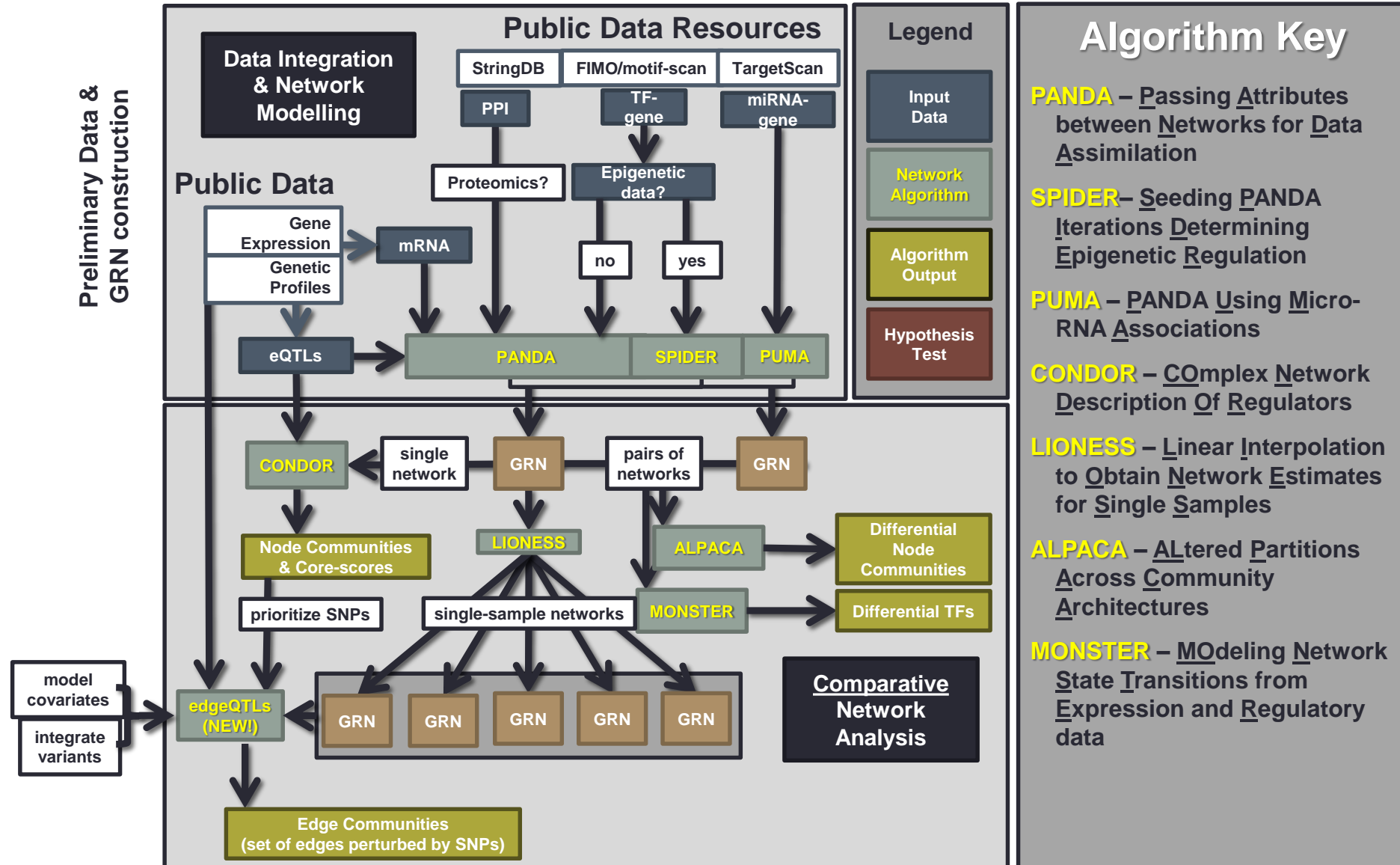
# How we do Network Analysis



# Starting Assumptions

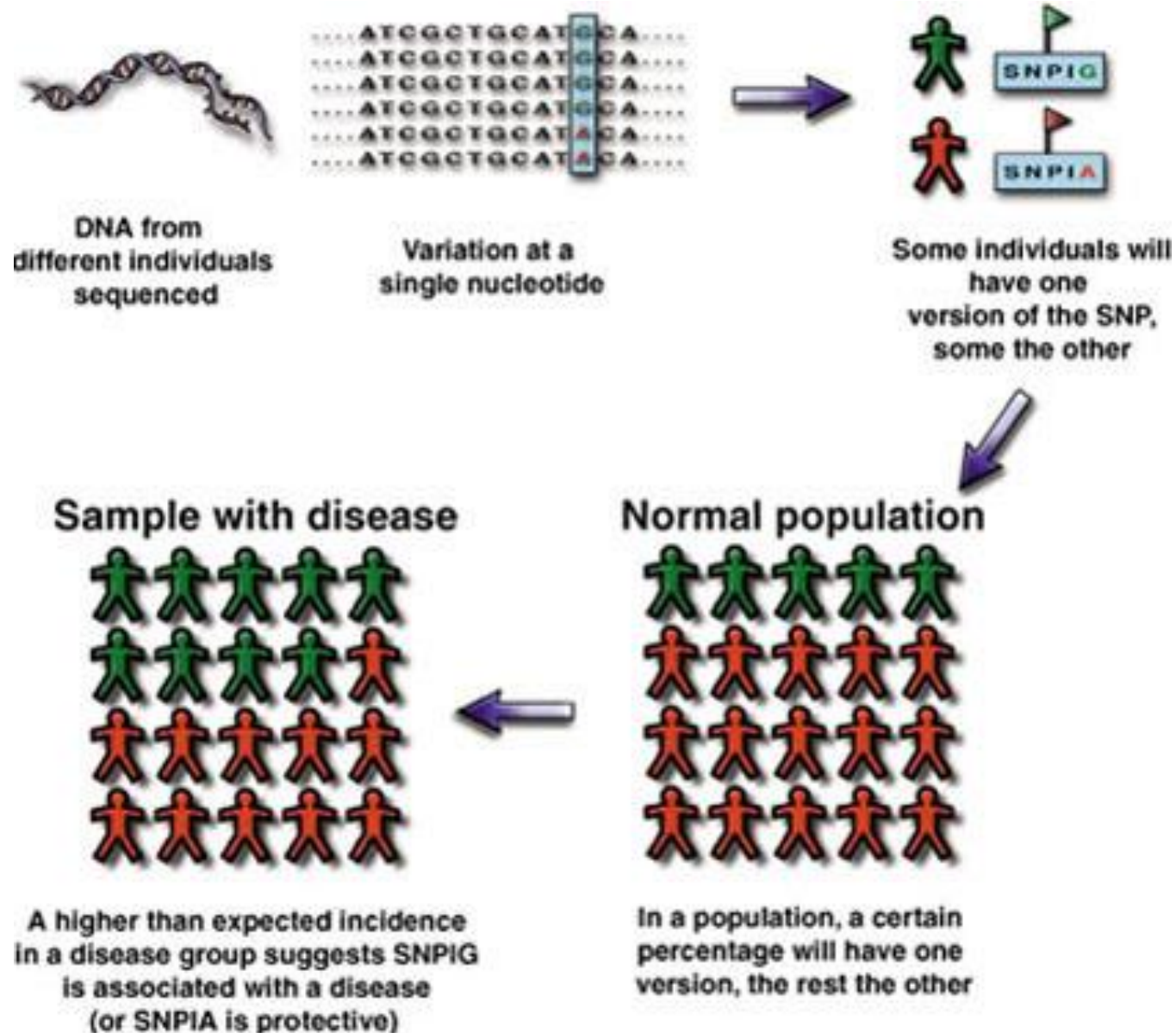
- **There is no single “right” network**
- **The structure of the network matters and network structure often changes between states.**
- **We have to move from asking “Is the network right?” to asking “Is the network useful?”**
- **The real question is “Does a network model inform our understanding of biology?”**

# The Methodological Zoo



**Question 1:**  
**Can we solve the**  
**“GWAS Puzzle”?**

# Genome Wide Association Studies (GWAS)



## Defining the role of common variation in the genomic and biological architecture of adult human height

Using genome-wide data from 253,288 individuals, we identified 697 variants at genome-wide significance that together explained one-fifth of the heritability for adult height. By testing different numbers of variants in independent studies, we show that the most strongly associated ~2,000, ~3,700 and ~9,500 SNPs explained ~21%, ~24% and ~29% of phenotypic variance. Furthermore, all common variants together captured 60% of heritability. The 697 variants clustered in 423 loci were enriched for genes, pathways and tissue types known to be involved in growth and together implicated genes and pathways not highlighted in earlier efforts, such as signaling by fibroblast growth factors, WNT/ $\beta$ -catenin and chondroitin sulfate-related genes. We identified several genes and pathways not previously connected with human skeletal growth, including mTOR, osteoglycin and binding of hyaluronic acid. Our results indicate a genetic architecture for human height that is characterized by a very large but finite number (thousands) of causal variants.

**697 SNPs explain 20% of height**  
**~2,000 SNPs explain 21% of height**  
**~3,700 SNPs explain 24% of height**  
**~9,500 SNPs explain 29% of height**

## Genetic studies of body mass index yield new insights for obesity biology

A list of authors and their affiliations appears at the end of the paper

Obesity is heritable and predisposes to many diseases. To understand the genetic basis of obesity better, here we conduct a genome-wide association study and MetaboChip meta-analysis of body mass index (BMI), a measure commonly used to define obesity and assess adiposity, in up to 339,224 individuals. This analysis identifies 97 BMI-associated loci ( $P < 5 \times 10^{-8}$ ), 56 of which are novel. Five loci demonstrate clear evidence of several independent association signals, and many loci have significant effects on other metabolic phenotypes. The 97 loci account for ~2.7% of BMI variation, and genome-wide estimates suggest that common variation accounts for >20% of BMI variation. Pathway analyses provide strong support for a role of the central nervous system in obesity susceptibility and implicate new genes and pathways, including those related to synaptic function, glutamate signalling, insulin secretion/action, energy metabolism, lipid biology and adipogenesis.

**97 SNPs explain 2.7% of BMI**

**All common SNPs may explain 20% of BMI**

**Do we give up on GWAS, fine map everything,  
or think differently?**

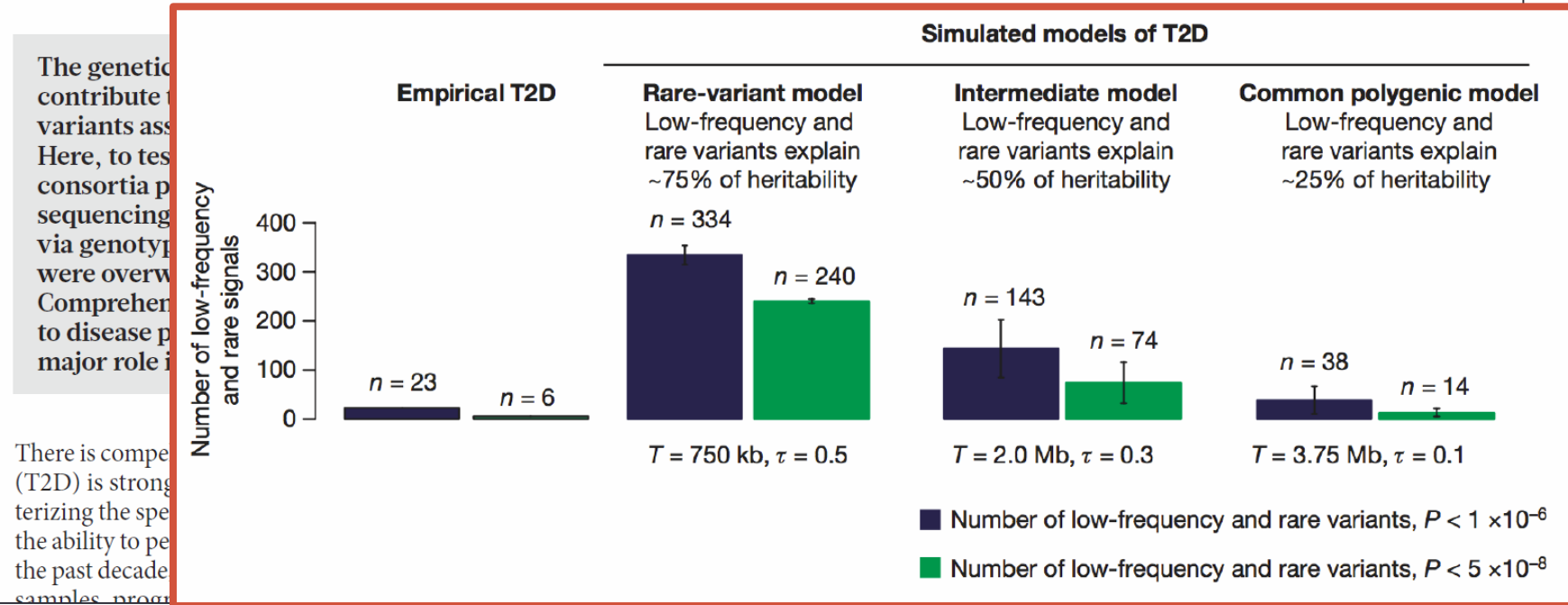
# Rare Variants = Dust

## ARTICLE

doi:10.1038/nature18642

### The genetic architecture of type 2 diabetes

A list of authors and affiliations appears in the online version of the paper



...large-scale sequencing does not support the idea that lower-frequency variants have a major role in predisposition to type 2 diabetes.

# eQTL Analysis

Expression Quantitative Trait Locus Analysis (eQTL Analysis) uses genome-wide data on genetic variants (SNPs) together with gene expression data

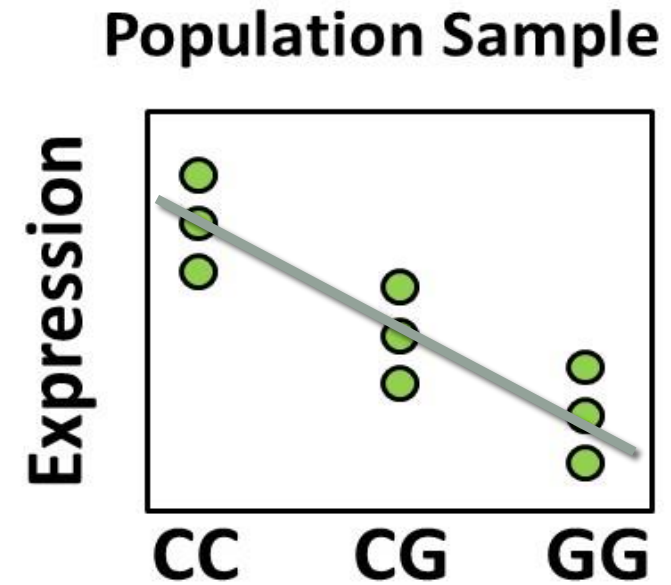
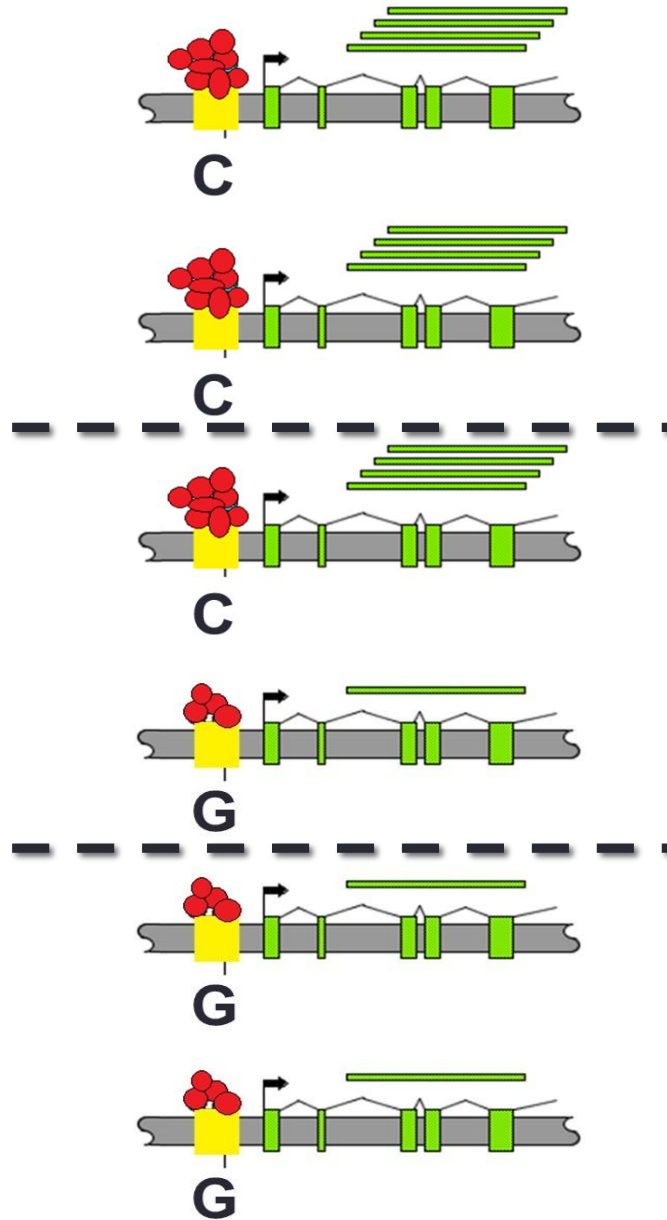
Treat gene expression as a quantitative trait

Ask, “Which SNPs are correlated with the degree of gene expression?”

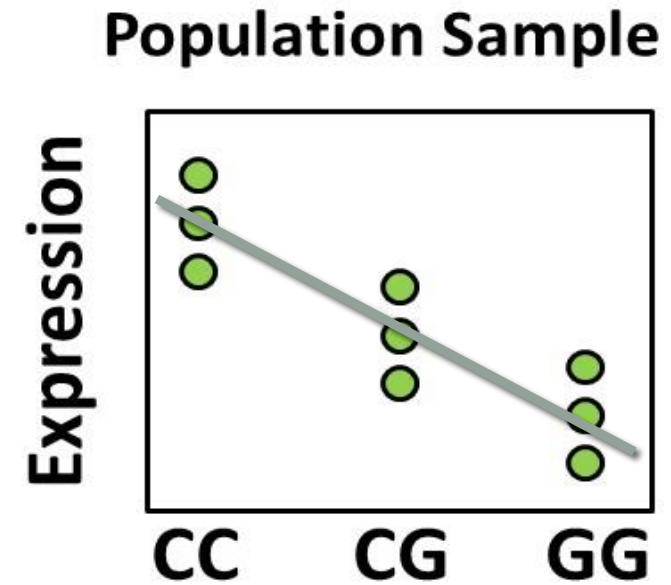
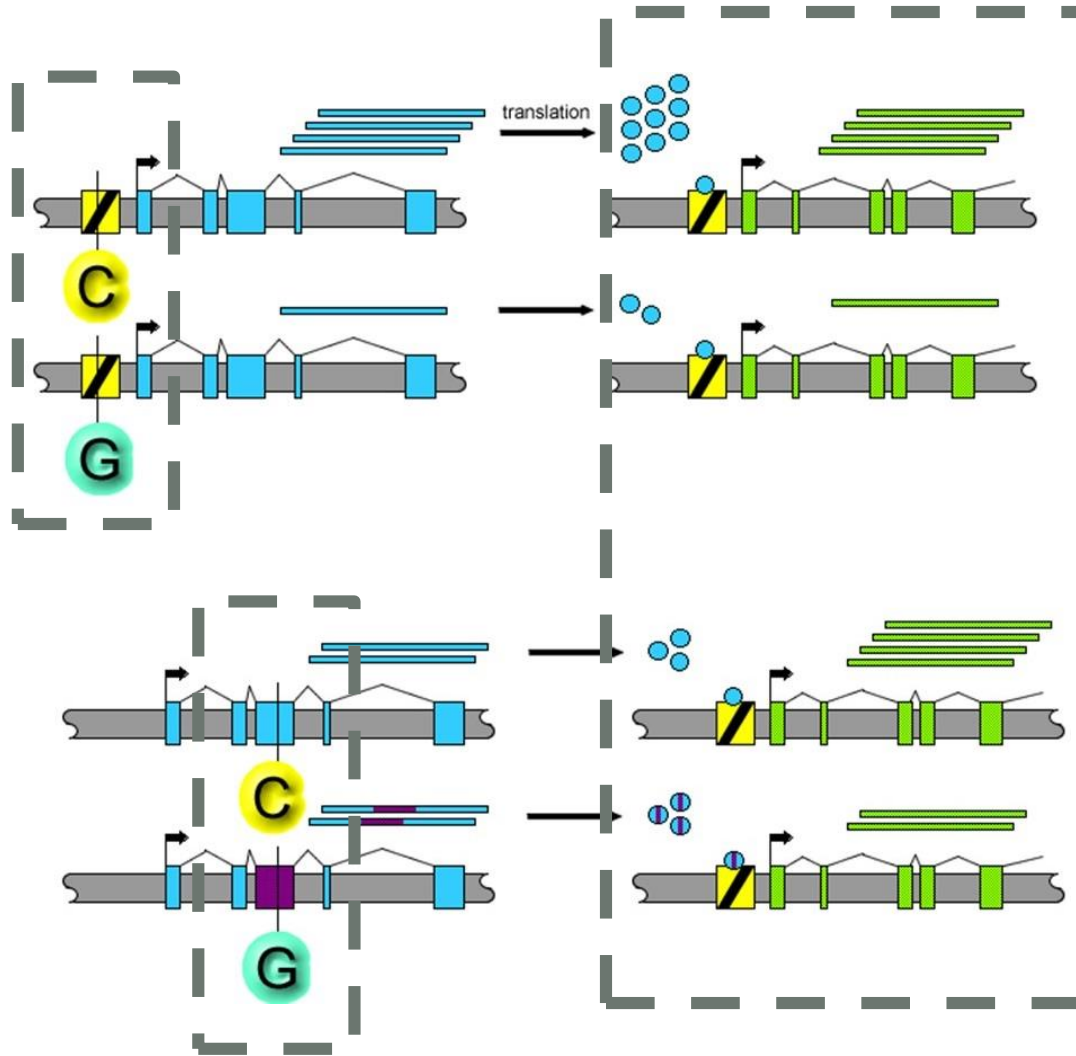
Most people concentrate on cis-acting SNPs

What about trans-acting SNPs?

# *cis*-eQTL Analysis



# *trans*-eQTL Analysis



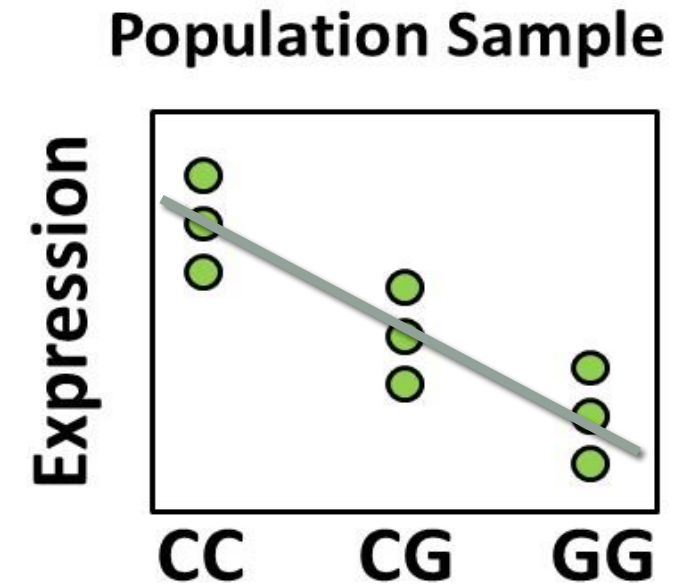
# eQTL Networks: A simple idea

- Perform a “standard eQTL” analysis:

$$Y = \beta_0 + \beta_1 ADD + \varepsilon$$

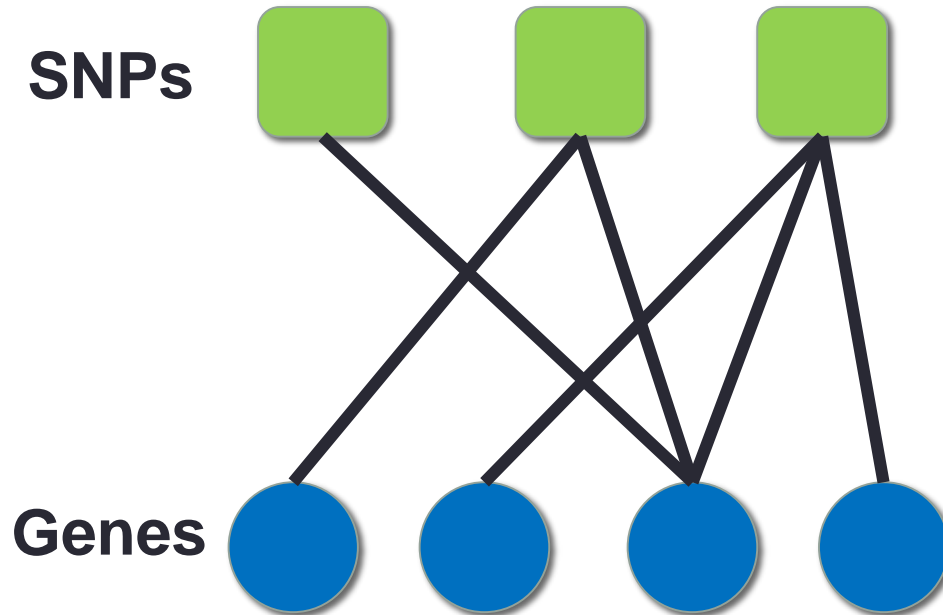
where  $Y$  is the quantitative trait and  $ADD$  is the allele dosage of a genotype.

Representing eQTLs as a network and analyzing its structure should provide insight in the complex interactions that drive disease.



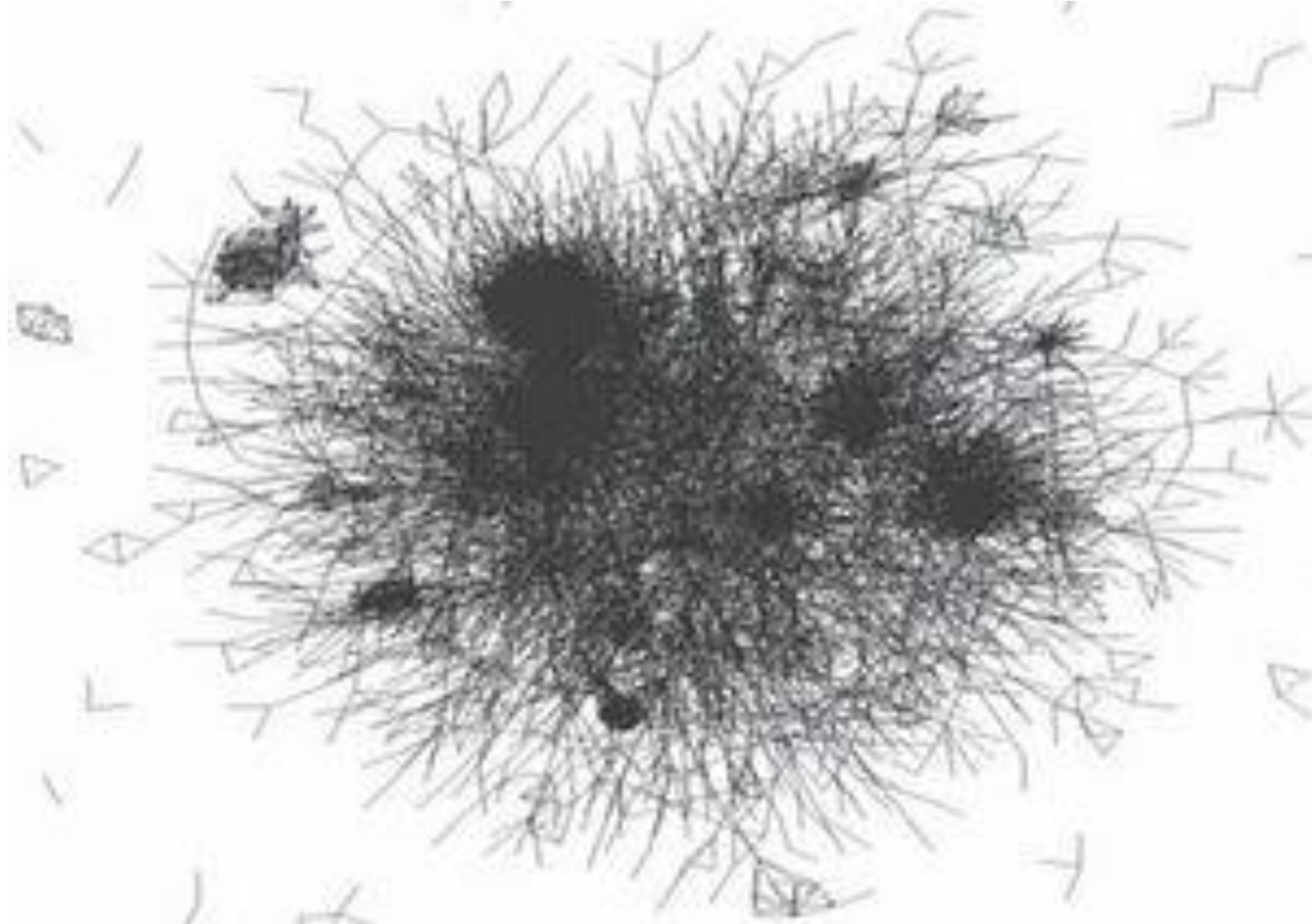
# eQTL Networks: A simple idea

Many strong eQTLs are found near the target gene. But what about multiple SNPs that are correlated with multiple genes?



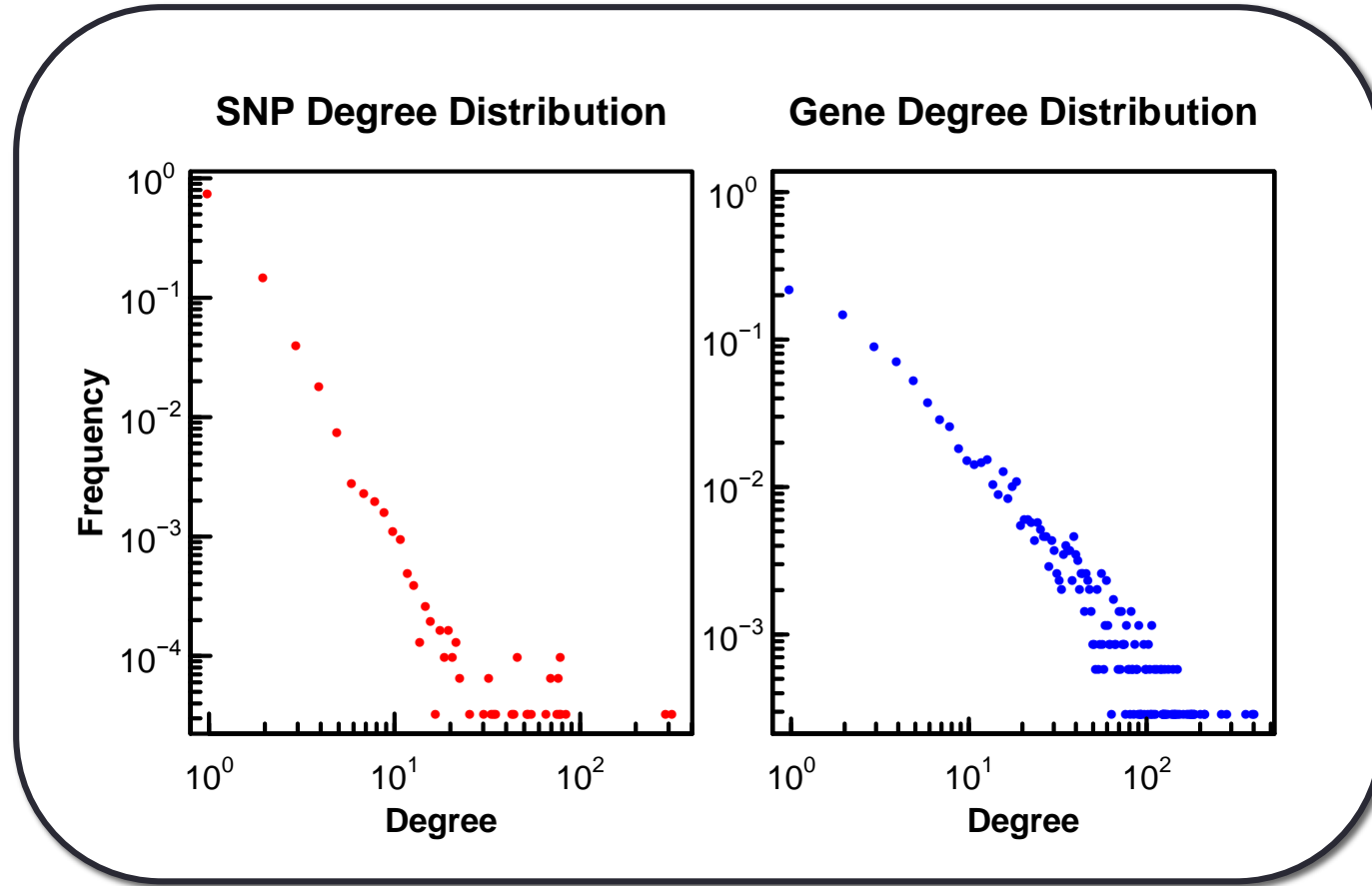
Can a network of SNP-gene associations (*cis* and *trans*) inform the functional roles of these SNPs?

# The Result: A Hairball



Some random hairball I grabbed. I was too lazy to make one.

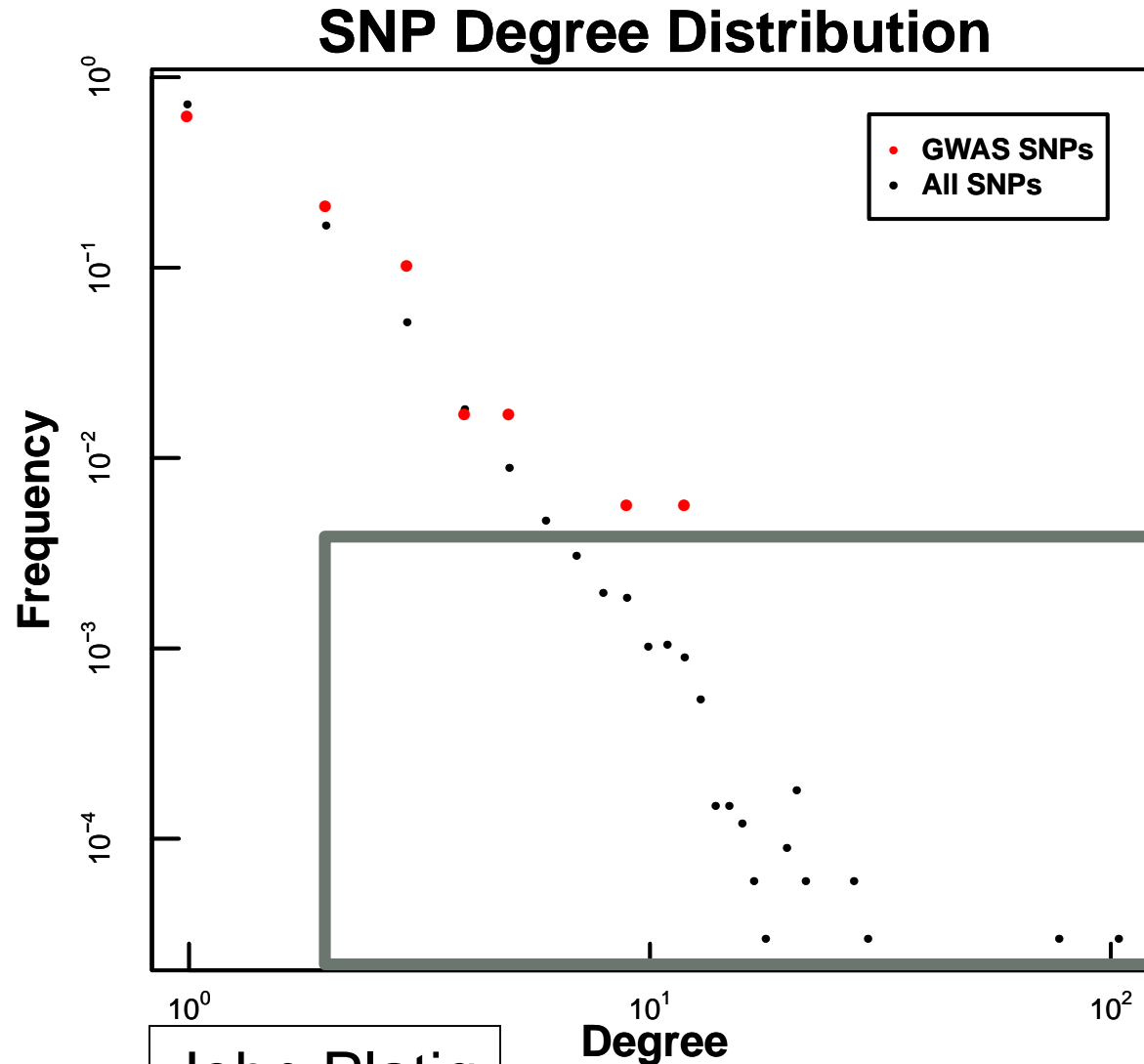
# Results: COPD



~30,000 SNPs and ~3,400 Genes

Degree – number of links per node

# What about GWAS SNPs?



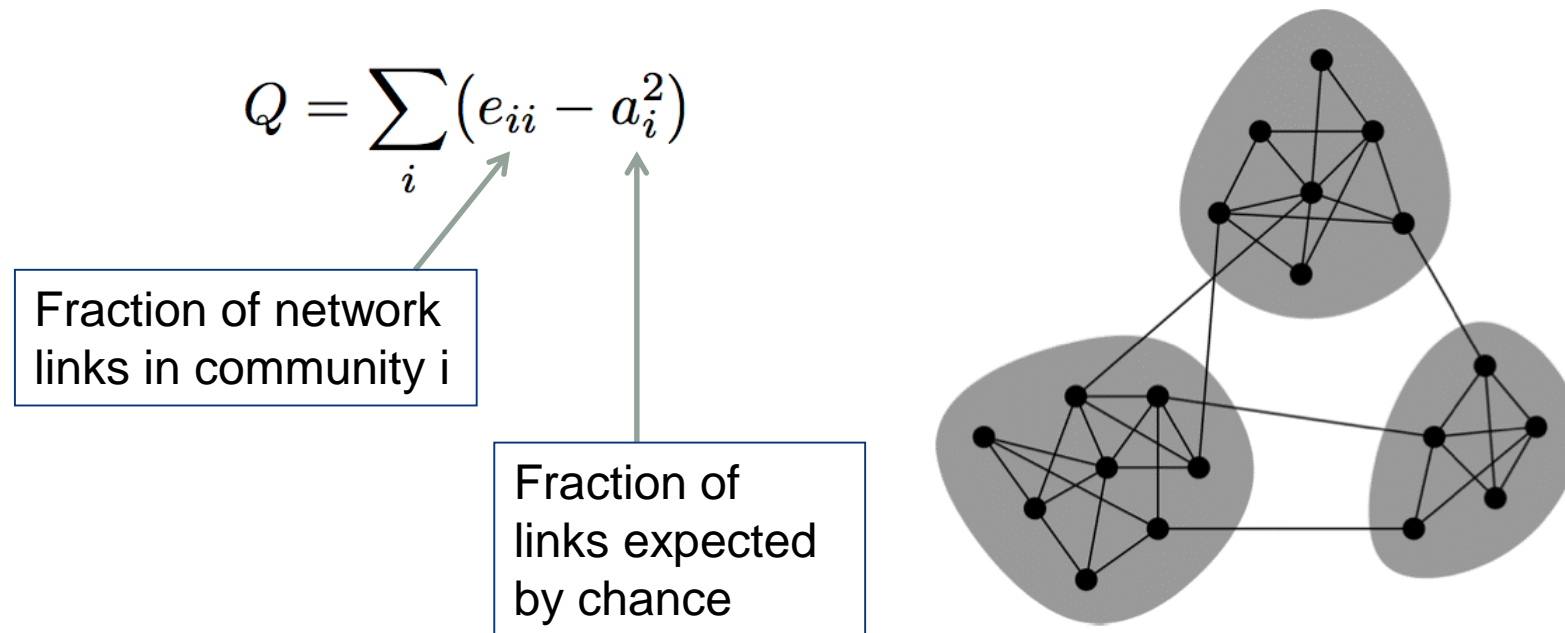
The “hubs”  
are a  
GWAS desert!

**Can we use this network to identify groups of SNPs and genes that play functional roles in the cell?**

Try clustering the nodes into “communities” based on the network structure

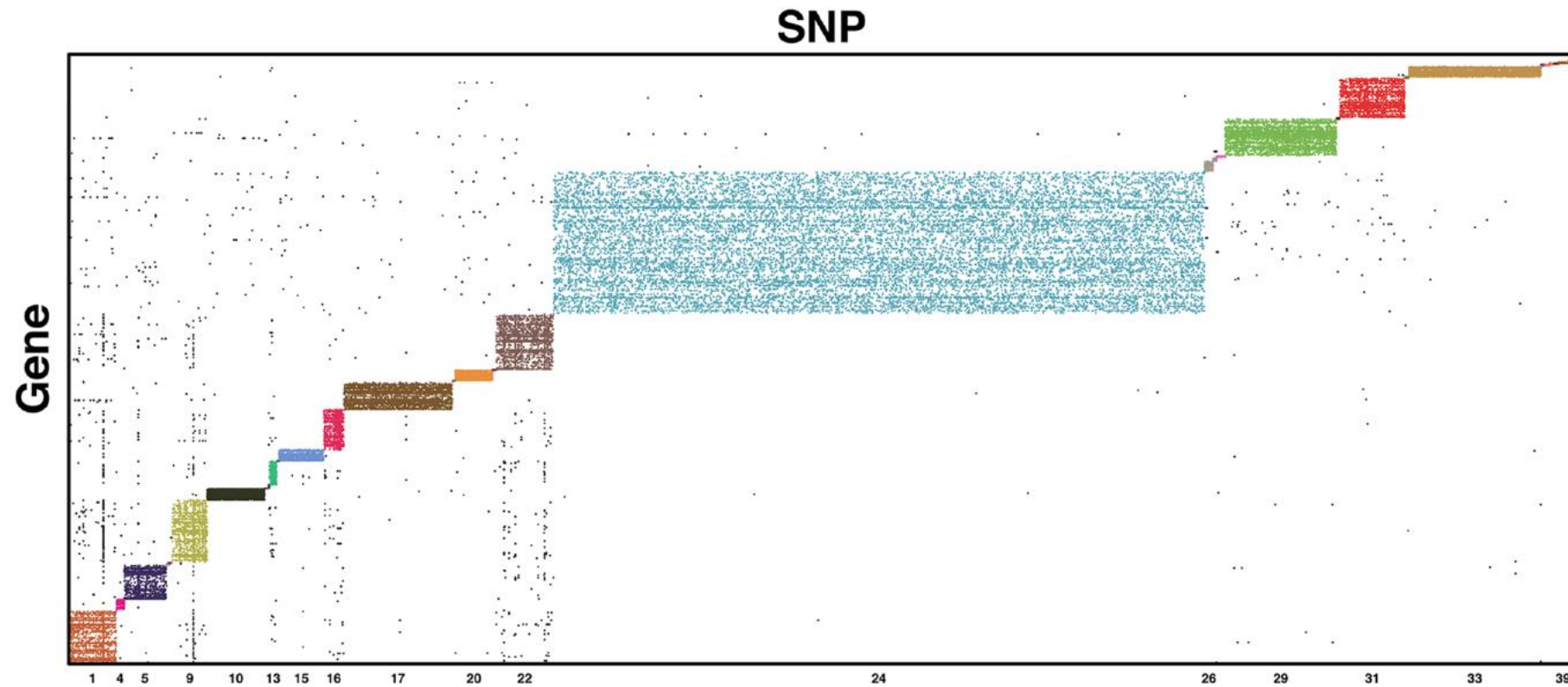
# Communities are groups of highly intra-connected nodes

- Community structure algorithms group nodes such that the number of links within a community is higher than expected by chance
- Formally, they assign nodes to communities such that the modularity,  $Q$ , is optimized



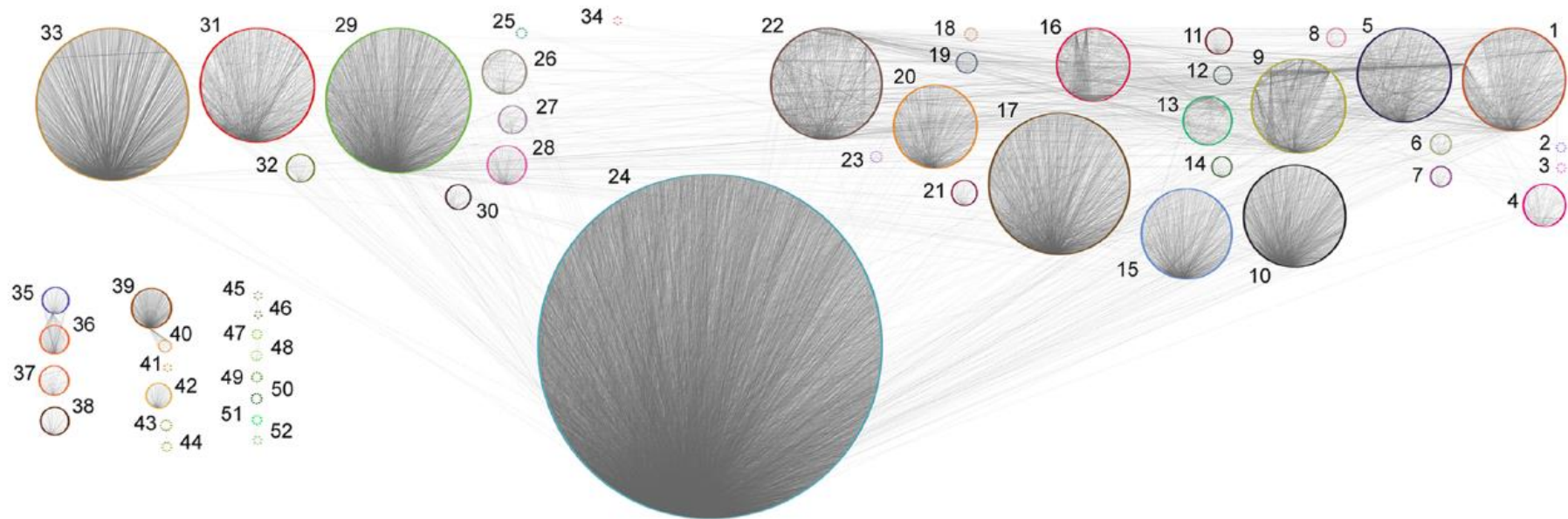
Newman 2006 (PNAS)

# Communities in COPD eQTL networks

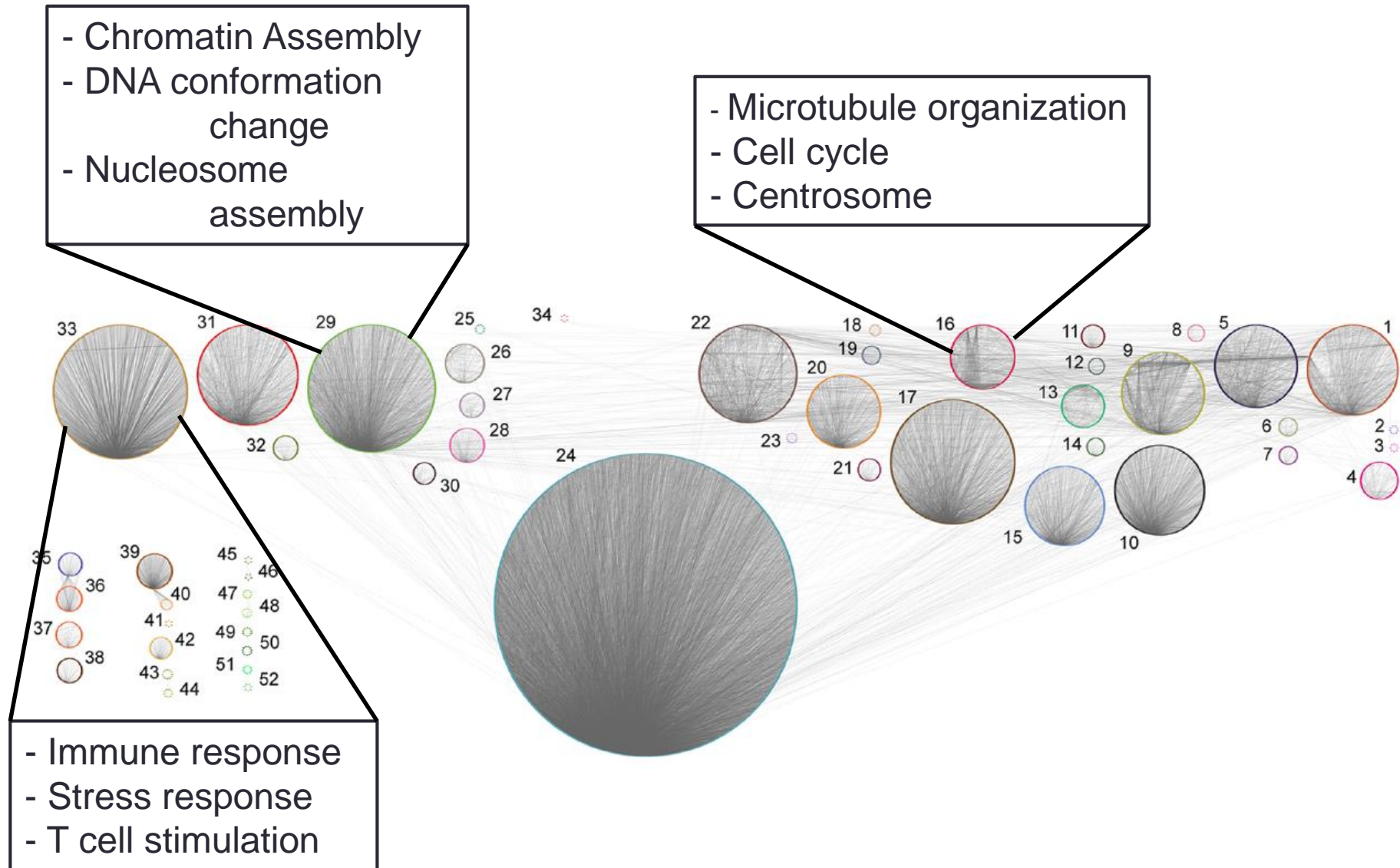


# Communities in COPD eQTL networks

- We identified 52 communities, with  $Q = 0.79$  (out of 1)
- Of 34 communities in the giant connected component, 11 are enriched for genes with coherent functions (GO Terms;  $P < 5 \times 10^{-4}$ )

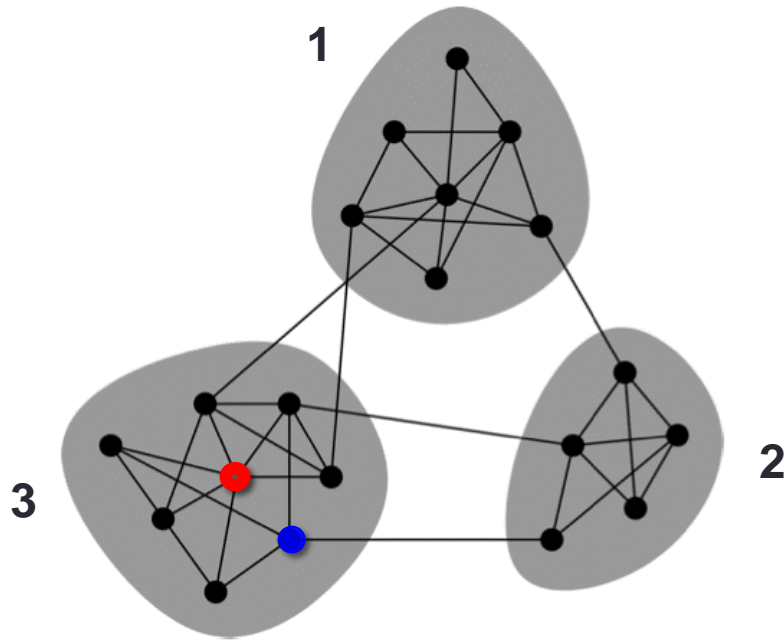


# Communities in COPD eQTL networks



# Identifying community cores

- Score each SNP by its contribution to the modularity of its community
- Do these “core scores” reflect known biology?

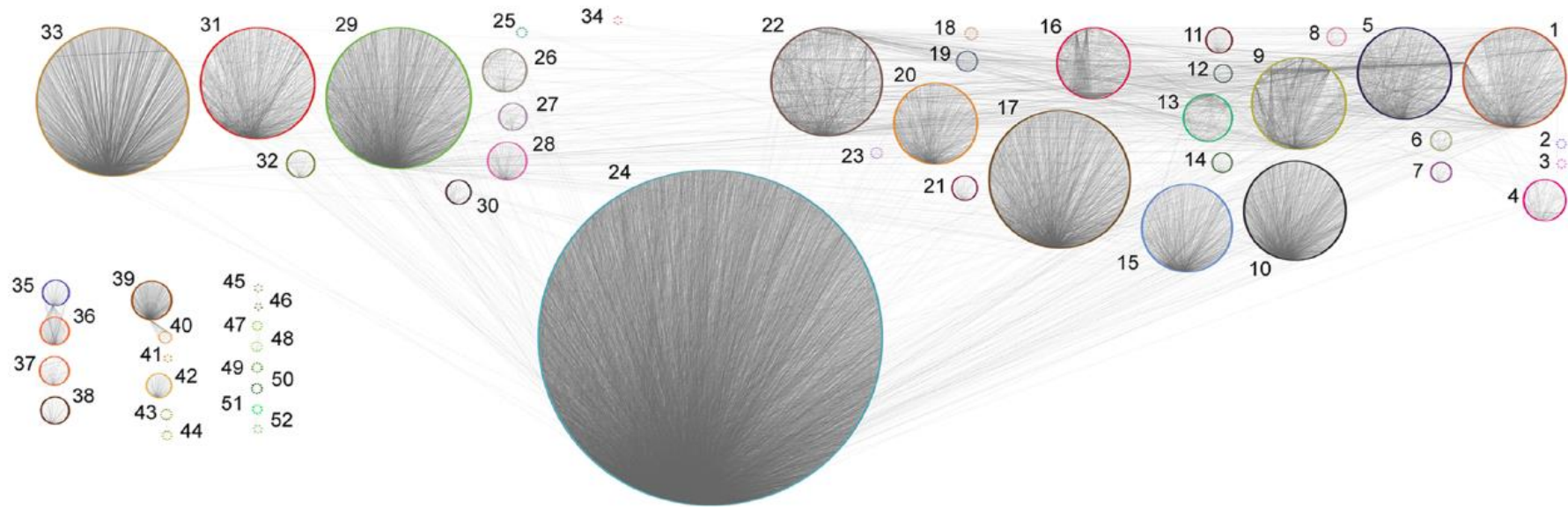


$$Q_{ih} = \frac{Q_i}{Q_h}$$

Newman 2006 (PNAS)

# What about COPD GWAS SNPs?

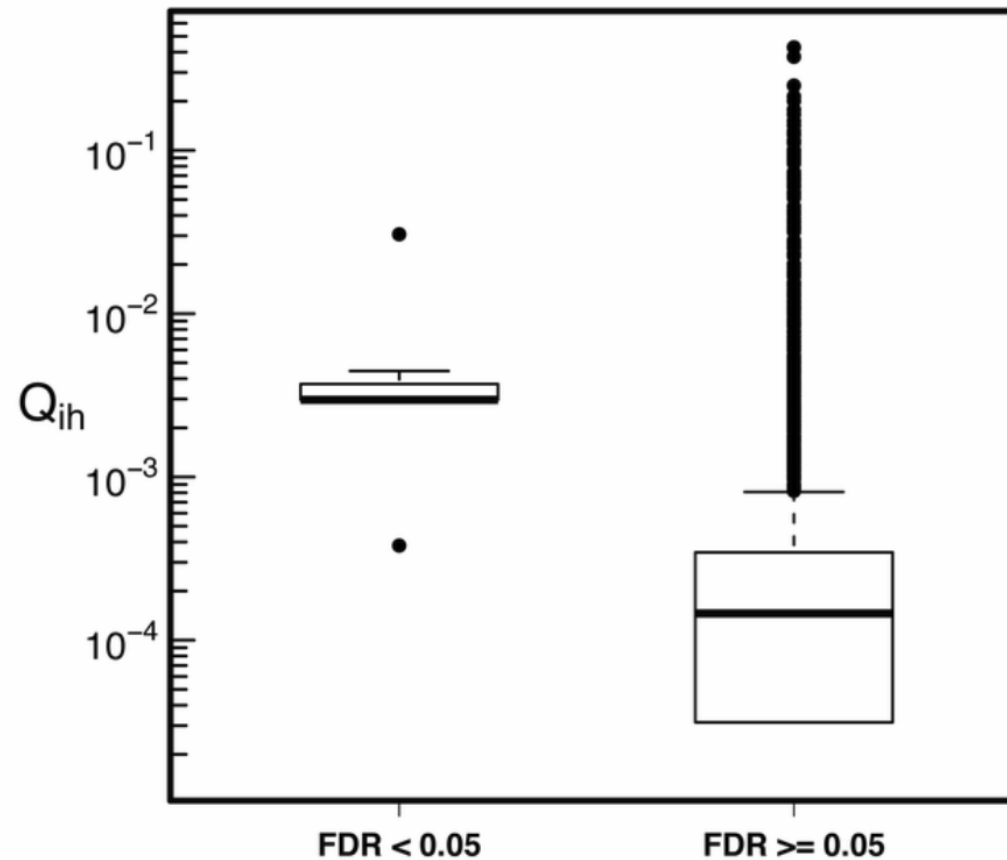
- Use a meta-analysis by Cho et. al. and consider 34 COPD GWAS SNPs (FDR < 0.05)



Cho, Michael H., et al. "Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis." *The Lancet Respiratory Medicine* 2.3 (2014): 214-225.

# Core Scores for COPD GWAS SNPs

The median core score for the 34 FDR-significant GWAS SNPs is **20.3 times higher** than the median for non-significant SNPs



# First in lung tissue/COPD

RESEARCH ARTICLE

## Bipartite Community Structure of eQTLs

John Platig<sup>1,2\*</sup>, Peter J. Castaldi<sup>3,4,5</sup>, Dawn DeMeo<sup>3,5,6</sup>, John Quackenbush<sup>1,2,3‡</sup>

**1** Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, **2** Department of Biostatistics, Harvard Chan School of Public Health, Boston, Massachusetts, United States of America, **3** Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America, **4** Division of General Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America, **5** Harvard Medical School, Boston, Massachusetts, United States of America, **6** Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America

‡This author is the senior author of this work.

\* [jplatig@jimmy.harvard.edu](mailto:jplatig@jimmy.harvard.edu)



### Abstract

Genome Wide Association Studies (GWAS) and expression quantitative trait locus (eQTL) analyses have identified genetic associations with a wide range of human phenotypes. However, many of these variants have weak effects and understanding their combined effect remains a challenge. One hypothesis is that multiple SNPs interact in complex networks to influence functional processes that ultimately lead to complex phenotypes, including disease states. Here we present CONDOR, a method that represents both *cis*- and *trans*-acting SNPs and the genes with which they are associated as a bipartite graph and then uses the modular structure of that graph to place SNPs into a functional context. In

 OPEN ACCESS

**Citation:** Platig J, Castaldi PJ, DeMeo D, Quackenbush J (2016) Bipartite Community Structure of eQTLs. PLoS Comput Biol 12(9): e1005033. doi:10.1371/journal.pcbi.1005033

**Editor:** Florian Markowetz, University of Cambridge, UNITED KINGDOM

**How general is this?**

# Now in thirteen tissues



PNAS PLUS

## Exploring regulation in tissues with eQTL networks

Maud Fagny<sup>a,b</sup>, Joseph N. Paulson<sup>a,b</sup>, Marieke L. Kuijjer<sup>a,b</sup>, Abhijeet R. Sonawane<sup>c</sup>, Cho-Yi Chen<sup>a,b</sup>, Camila M. Lopes-Ramos<sup>a,b</sup>, Kimberly Glass<sup>c</sup>, John Quackenbush<sup>a,b,d,1</sup>, and John Platig<sup>a,b,1</sup>

<sup>a</sup>Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02115; <sup>b</sup>Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA 02115; <sup>c</sup>Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School Boston, MA 02115; and <sup>d</sup>Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115

Edited by Jasper Rine, University of California, Berkeley, CA, and approved August 4, 2017 (received for review May 3, 2017)

**Characterizing the collective regulatory impact of genetic variants on complex phenotypes is a major challenge in developing a genotype to phenotype map. Using expression quantitative trait locus (eQTL) analyses, we constructed bipartite networks in which edges represent significant associations between genetic variants and gene expression levels and found that the network structure informs regulatory function. We show, in 13 tissues, that these eQTL networks are organized into dense, highly modular communities grouping genes often involved in coherent biological processes. We find communities representing shared processes across tissues, as well as communities associated with tissue-specific processes that coalesce around variants in tissue-specific active chromatin regions. Node centrality is also highly informative, with the global and community hubs differing in regulatory potential and likelihood of being disease associated.**

GTEx | expression quantitative trait locus | eQTL | bipartite networks | GWAS

biological pathways. In particular, we find three aspects of the eQTL network topology that inform tissue-level regulatory biology: (i) Communities—which are composed of SNPs and genes with a high density of within-group edges—are enriched for pathways, functionally related genes, and SNPs in tissue-specific active chromatin regions (actively transcribed and open regulatory regions); (ii) community hubs (core SNPs)—which are SNPs highly connected to genes in their community—are enriched for active chromatin regions close to the transcriptional start site and for GWAS association; and (iii) global hubs—which are connected to many genes throughout the network—are enriched for distal elements such as nongenic enhancers and devoid of GWAS association. The picture that emerges from analysis of the eQTL networks is a complex web of associations that reflects the polygenic architecture across tissues and that provides a natural framework for understanding both the shared and tissue-specific effects of genetic variants. These networks, along with SNP and gene network properties across all 13 tissues, are avail-

# GTEEx: A big sandbox

## RESEARCH

### RESEARCH ARTICLE

#### HUMAN GENOMICS

## The Genotype-Tissue Expression (GTEEx) pilot analysis: Multitissue gene regulation in humans

The GTEEx Consortium\*†

Understanding the functional consequences of genetic variation, and how it affects complex human disease and quantitative traits, remains a critical challenge for biomedicine. We present an analysis of RNA sequencing data from 1641 samples across 43 tissues from 175 individuals, generated as part of the pilot phase of the Genotype-Tissue Expression (GTEEx) project. We describe the landscape of gene expression across tissues, catalog thousands of tissue-specific and shared regulatory expression quantitative trait loci (eQTL) variants, describe complex network relationships, and identify signals from genome-wide association studies explained by eQTLs. These findings provide a systematic understanding of the cellular and biological consequences of human genetic variation and of the heterogeneity of such effects among a diverse set of human tissues.

Over the past decade, there has been a marked increase in our understanding of the role

are for any given GWAS locus or disease. Hence, understanding the role of regulatory variants,

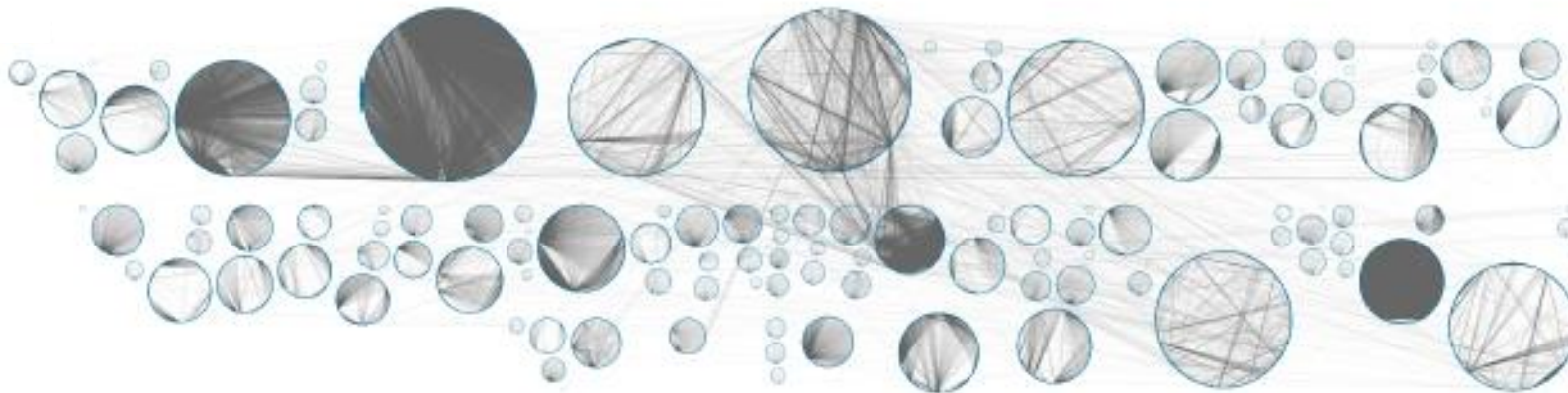
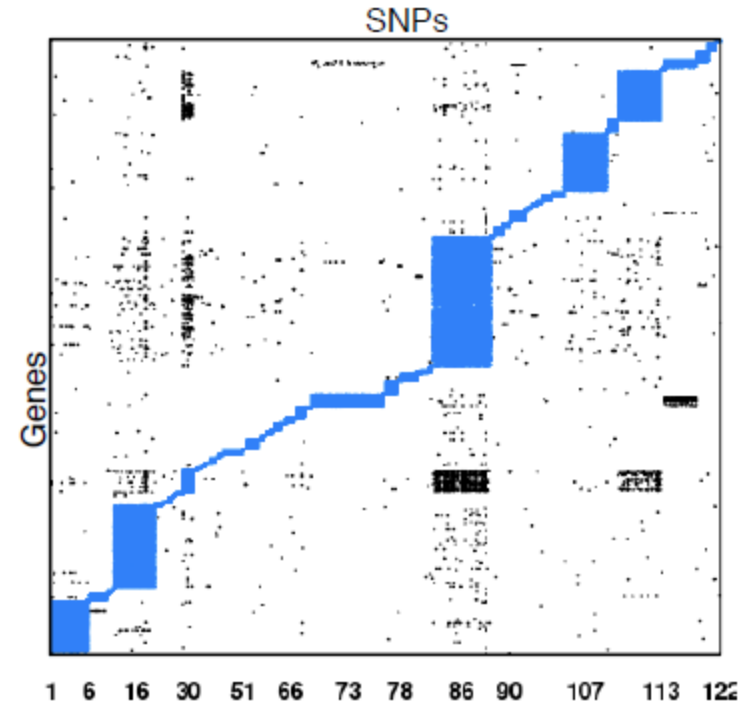
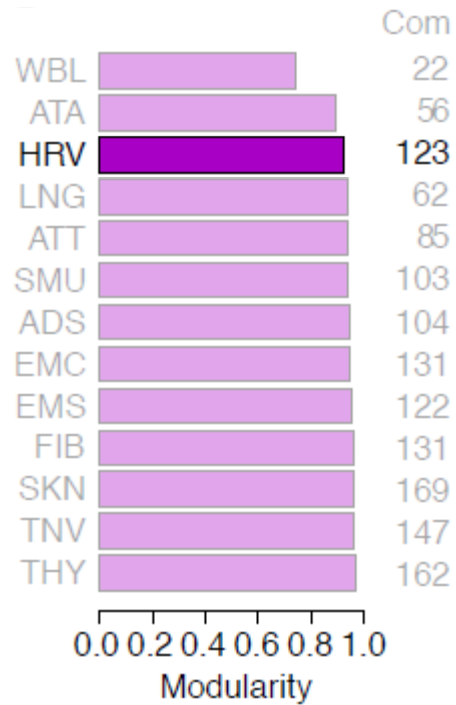
statistical power, we prioritized RNA sequencing of samples from nine tissues that were most frequently collected and that routinely met minimum RNA quality criteria: adipose (subcutaneous), tibial artery, heart (left ventricle), lung, muscle (skeletal), tibial nerve, skin (Sun-exposed), thyroid, and whole blood (Table 1) (14).

We performed 76-base pair (bp) paired-end mRNA sequencing on a total of 1749 samples, of which 1641 samples from 43 sites, and 175 donors, constituted our final “pilot data freeze” reported on here (14). Median sequencing depth was 82.1 million mapped reads per sample (fig. S3A). The final data freeze included samples from 43 body sites: 29 solid-organ tissues, 11 brain subregions (with two duplicated regions), a whole-blood sample, and two cell lines derived from donor blood [EBV-transformed lymphoblastoid cell lines (LCLs)] and skin samples (cultured fibroblasts) (Table 1 and tables S1 and S2). Median sample size for the nine high-priority tissues was 105; median sample size for the other 34 sampled sites was 18.5.

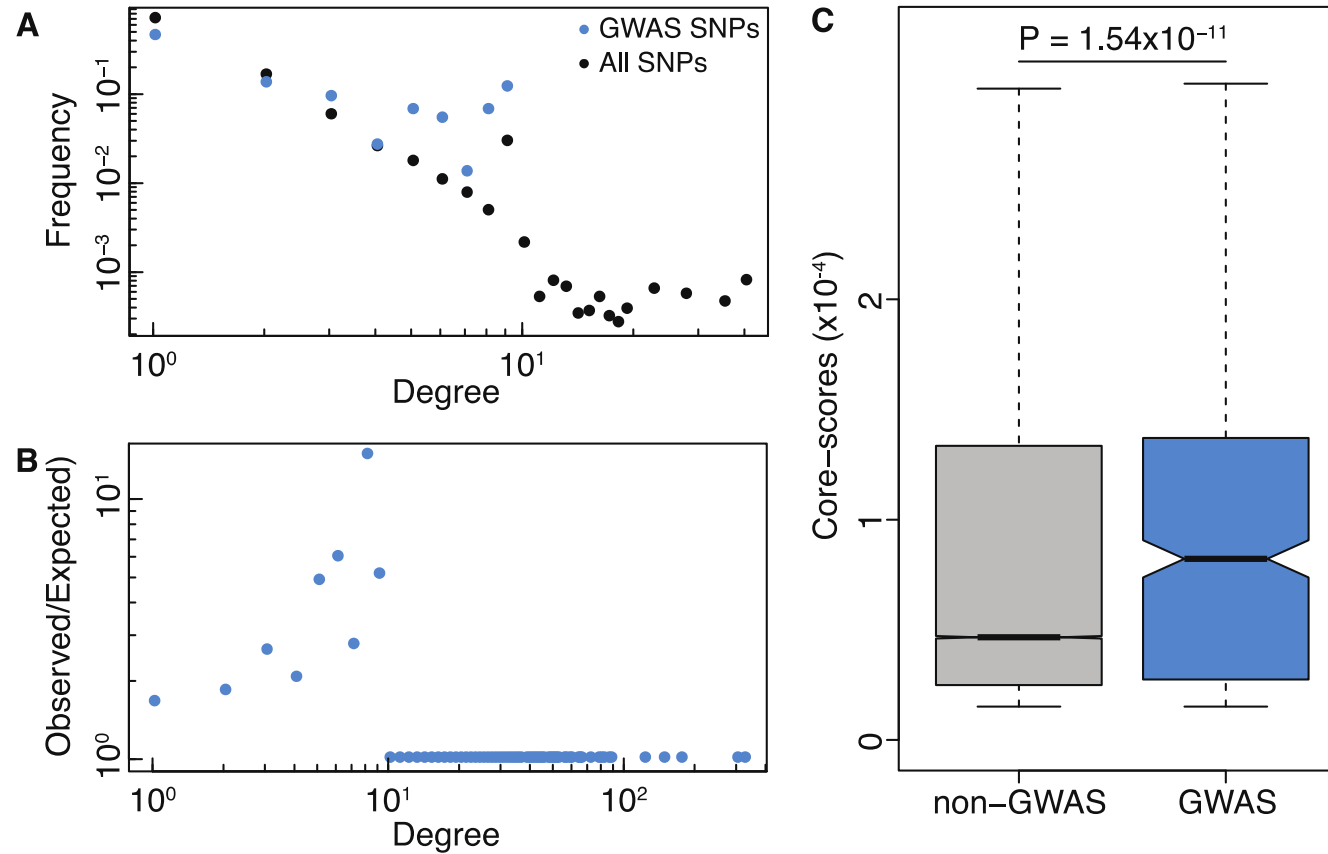
### Gene expression across tissues

We examined the patterns of expression of 53,934 transcribed genes across tissues [on the basis of Gencode V12 annotations] (14, 15). The number of biotypes [protein-coding genes, pseudogenes, and long noncoding RNAs (lncRNAs)] that were transcribed above a minimal threshold [reads per kilo-

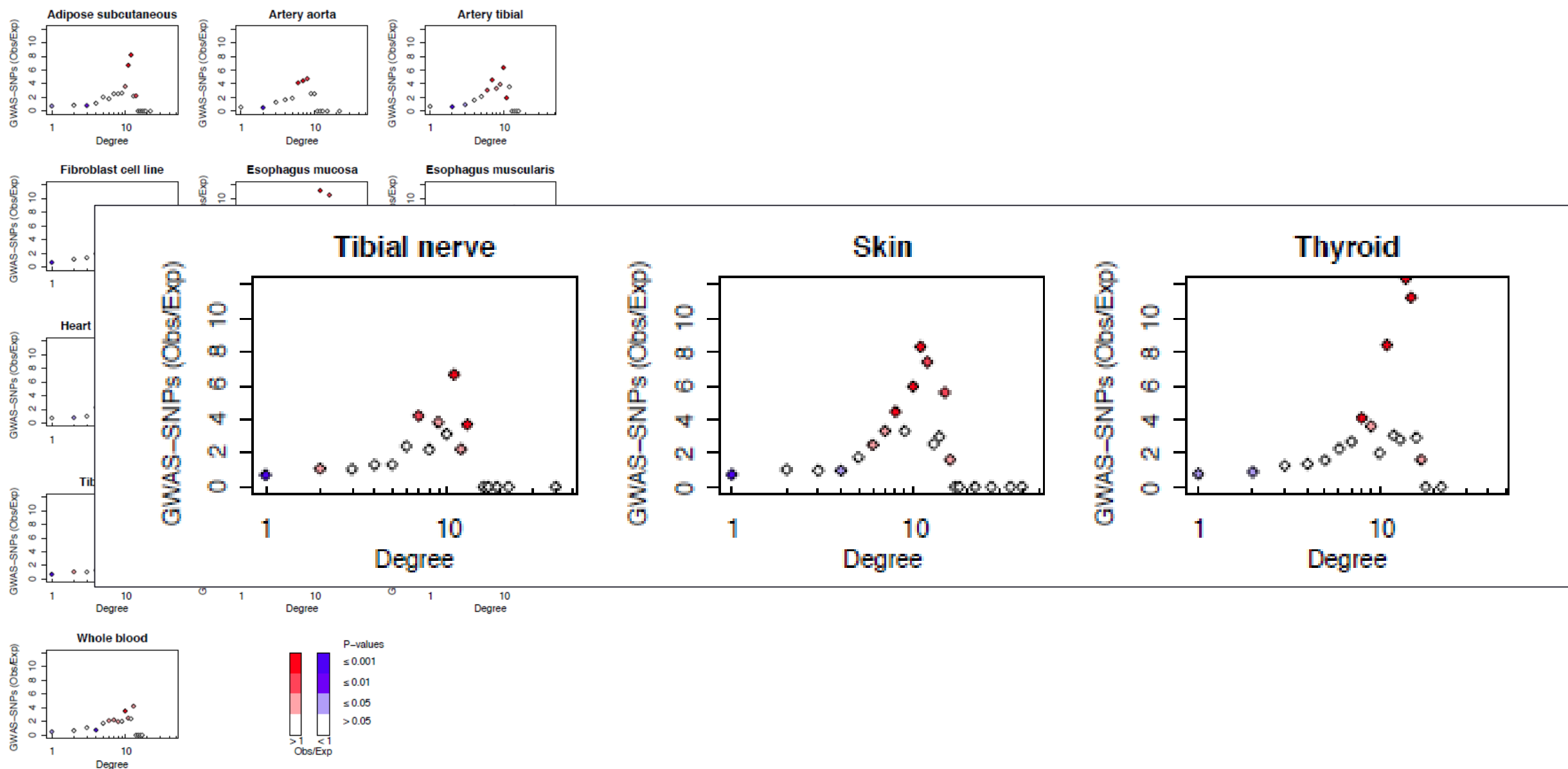
# eQTL networks are highly modular



# GWAS SNPs are cores, but not hubs

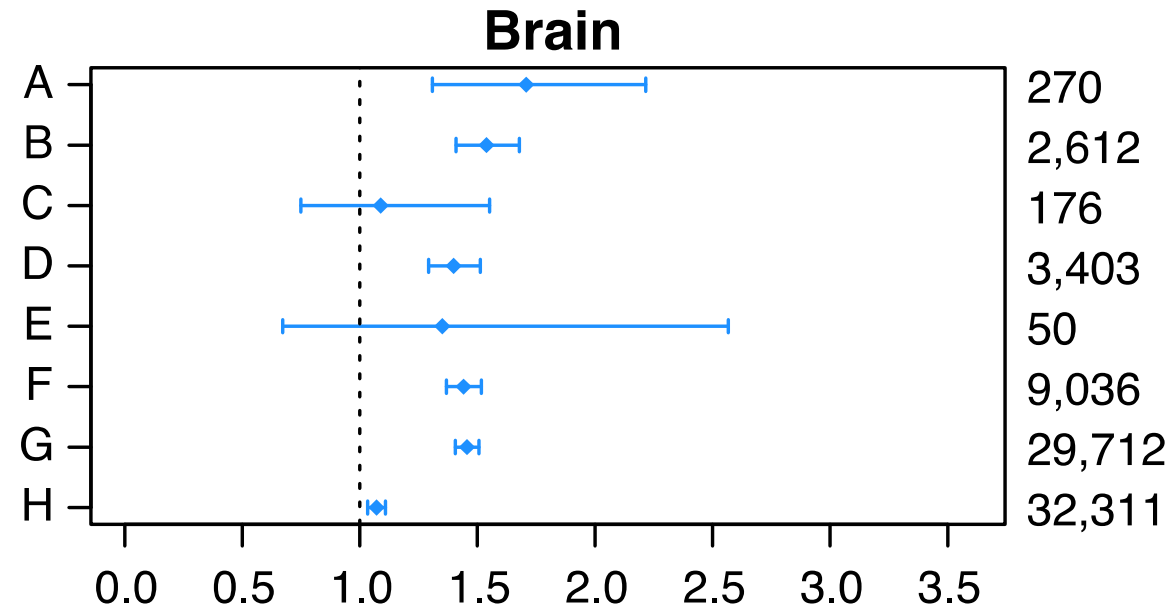


# GWAS SNPs not Hubs—in every tissue

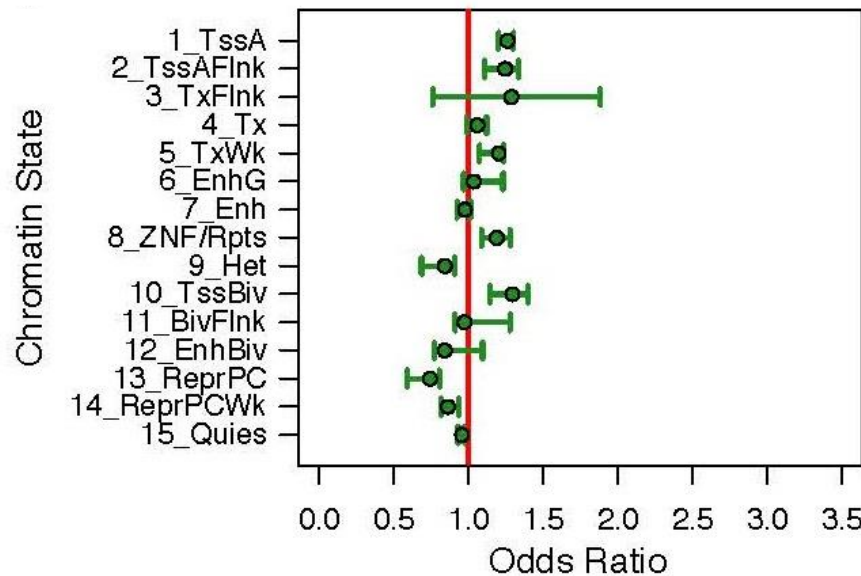


# Core SNPs are more likely to be functionally annotated

Category	Annotation
A	TF binding + matched TF motif + matched DNase footprint + DNase peak
B	TF binding + any motif + DNase footprint + DNase peak
C	TF binding + matched TF motif + DNase peak
D	TF binding + any motif + DNase peak
E	TF binding + matched TF motif
F	TF binding + DNase peak
G	TF binding or DNase peak
H	Motif hit
I	No Information

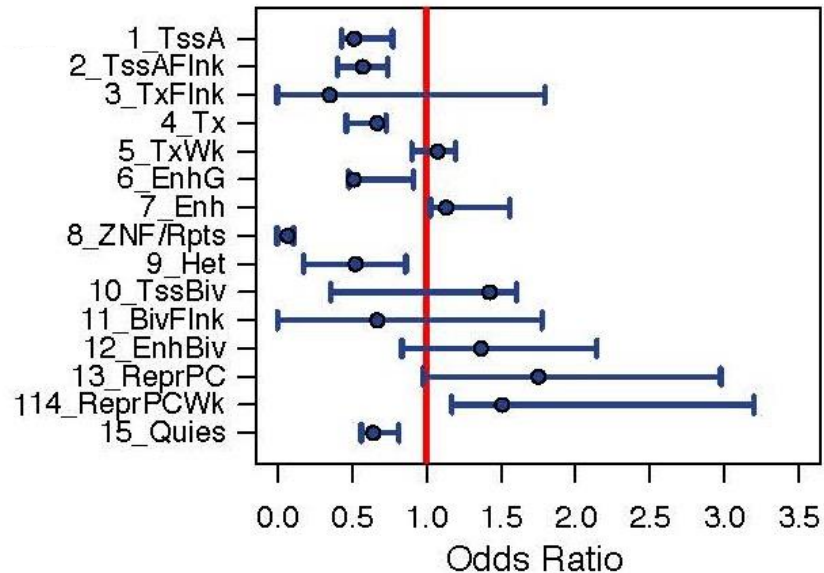


# Core SNPs are different from Hubs: Roadmap Epigenomics Project



## Local Hubs (Core SNPs)

Tissue-specific active  
chromatin



## Global Hubs

Nongenic Enhancers  
Polycomb Repressed  
Regions

# What does this tell us?

- The SNPs that are global hubs are not GWAS hits—meaning that they are not linked to diseases or traits.
- The SNPs and genes group into communities that share function—a family of SNPs regulate a function.
- Disease-associated (GWAS) SNPs map to communities whose genes have functions that make biological sense.
- “Core” SNPs are far more likely to be disease SNPs.
- Tissue-specific functions are in tissue-specific communities with tissue-specific genes organized around SNPs in tissue-specific open chromatin.

**Question 2:**  
**Can we model**  
**gene regulatory processes?**

# Integrative Network Inference: PANDA

OPEN ACCESS Freely available online



## Passing Messages between Biological Networks to Refine Predicted Interactions

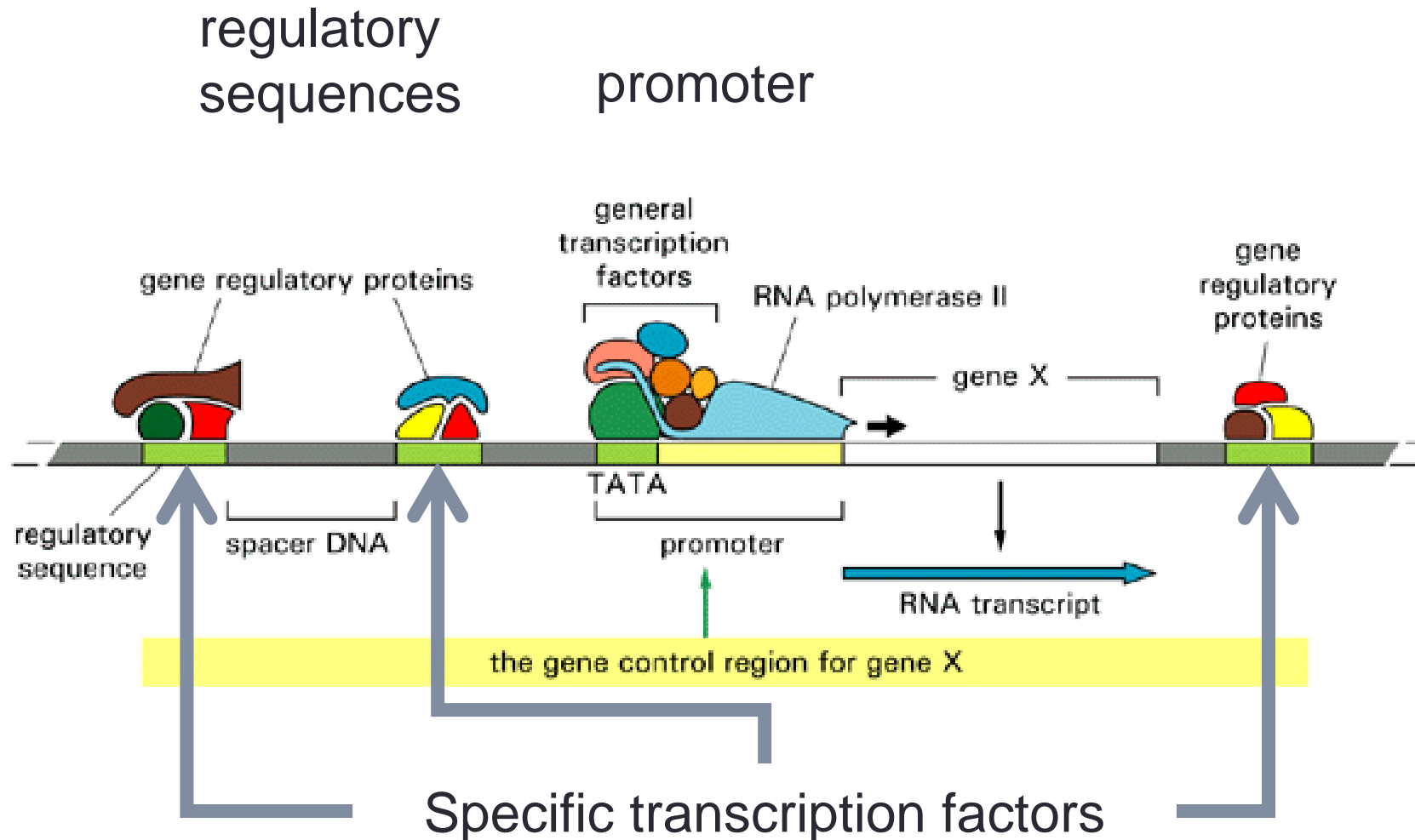
Kimberly Glass<sup>1,2</sup>, Curtis Huttenhower<sup>2</sup>, John Quackenbush<sup>1,2</sup>, Guo-Cheng Yuan<sup>1,2\*</sup>

**1** Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, **2** Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of America

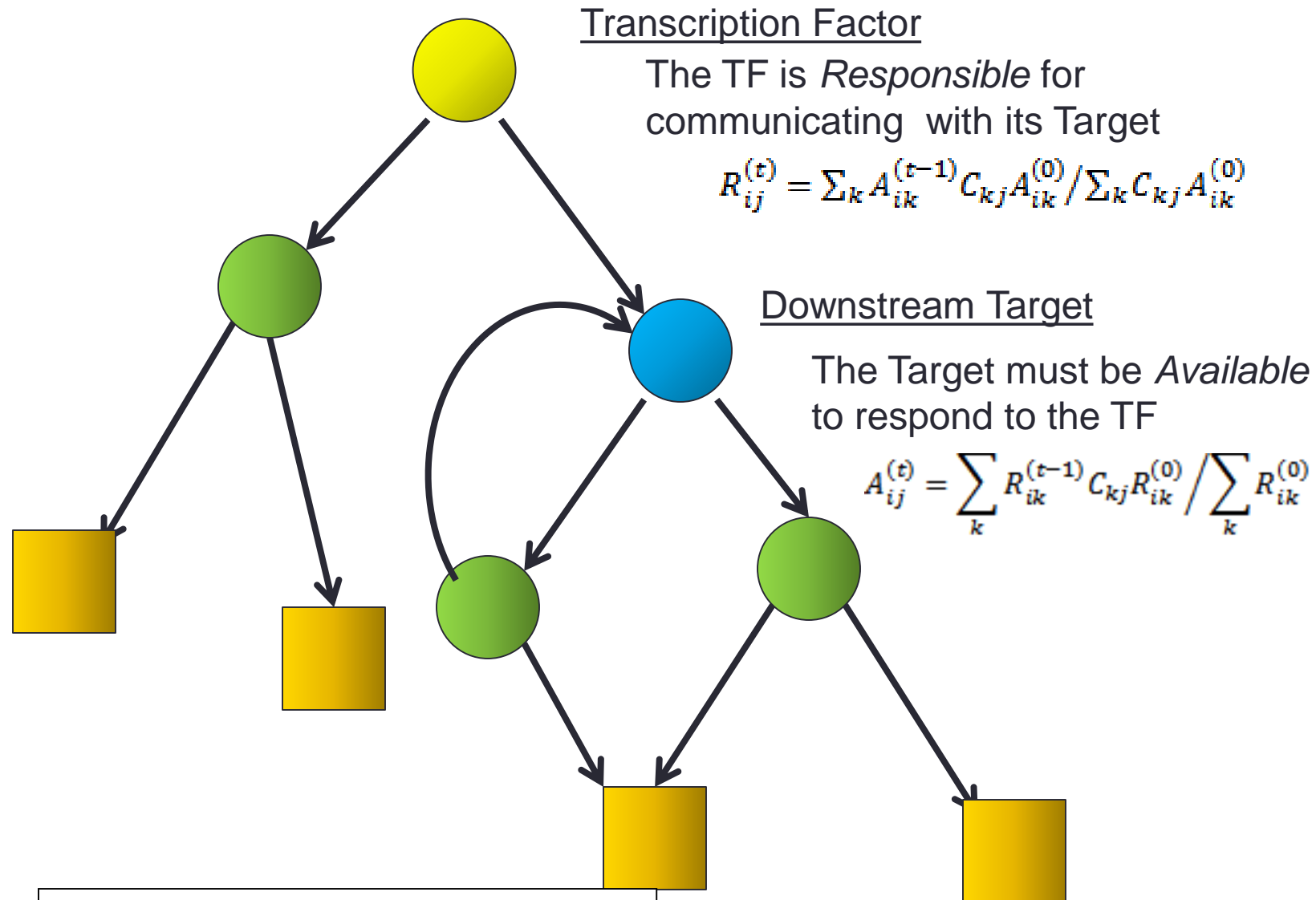
### Abstract

Regulatory network reconstruction is a fundamental problem in computational biology. There are significant limitations to such reconstruction using individual datasets, and increasingly people attempt to construct networks using multiple, independent datasets obtained from complementary sources, but methods for this integration are lacking. We developed PANDA (Passing Attributes between Networks for Data Assimilation), a message-passing model using multiple sources of information to predict regulatory relationships, and used it to integrate protein-protein interaction, gene expression, and sequence motif data to reconstruct genome-wide, condition-specific regulatory networks in yeast as a model. The resulting networks were not only more accurate than those produced using individual data sets and other existing methods, but they also captured information regarding specific biological mechanisms and pathways that were missed using other methodologies. PANDA is scalable to higher eukaryotes, applicable to specific tissue or cell type data and conceptually generalizable to include a variety of regulatory, interaction, expression, and other genome-scale data. An implementation of the PANDA algorithm is available at [www.sourceforge.net/projects/panda-net](http://www.sourceforge.net/projects/panda-net).

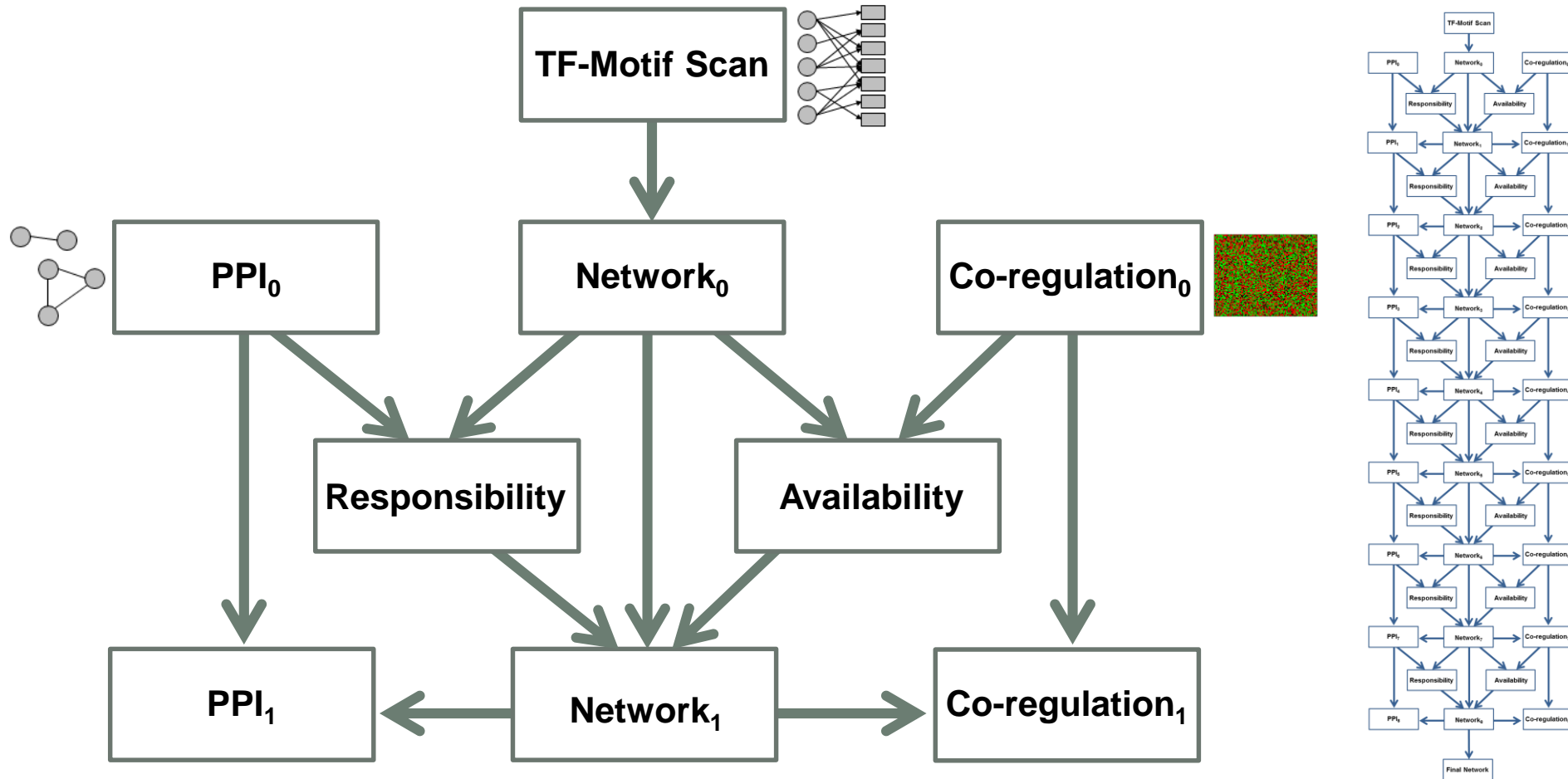
# Regulation of Transcription



# A Simple Idea: Message Passing



# Message-Passing Networks: PANDA



# Subtypes of Ovarian Cancer

OPEN ACCESS Freely available online



## Angiogenic mRNA and microRNA Gene Expression Signature Predicts a Novel Subtype of Serous Ovarian Cancer

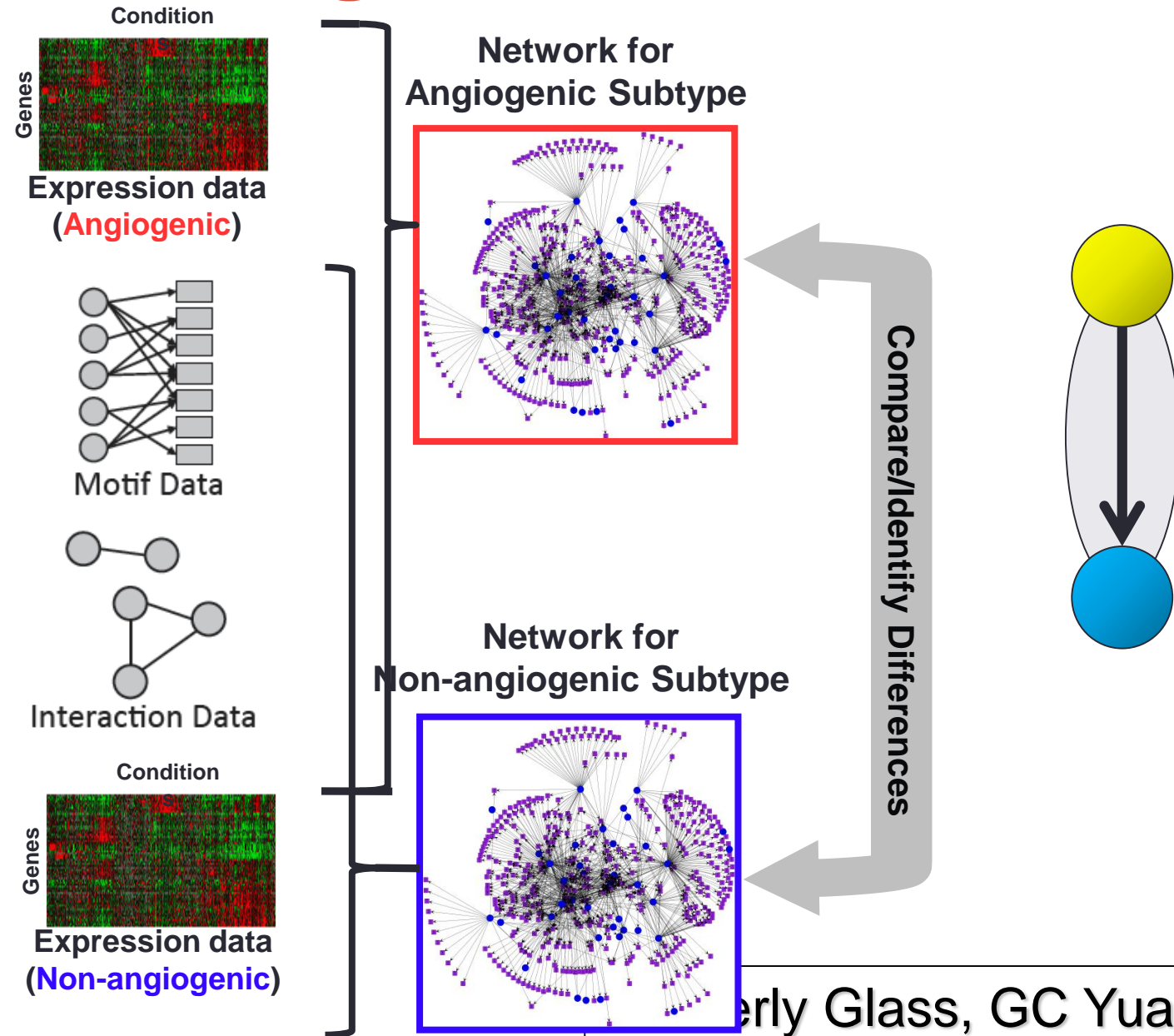
**Stefan Bentink<sup>1,6\*</sup>, Benjamin Haibe-Kains<sup>1,6\*</sup>, Thomas Risch<sup>1</sup>, Jian-Bing Fan<sup>3</sup>, Michelle S. Hirsch<sup>4,7</sup>, Kristina Holton<sup>1</sup>, Renee Rubio<sup>1</sup>, Craig April<sup>3</sup>, Jing Chen<sup>3</sup>, Eliza Wickham-Garcia<sup>3</sup>, Joyce Liu<sup>2,7</sup>, Aedin Culhane<sup>1,6</sup>, Ronny Drapkin<sup>4,5,7</sup>, John Quackenbush<sup>1,2,6\*</sup>†, Ursula A. Matulonis<sup>5,7</sup>†**

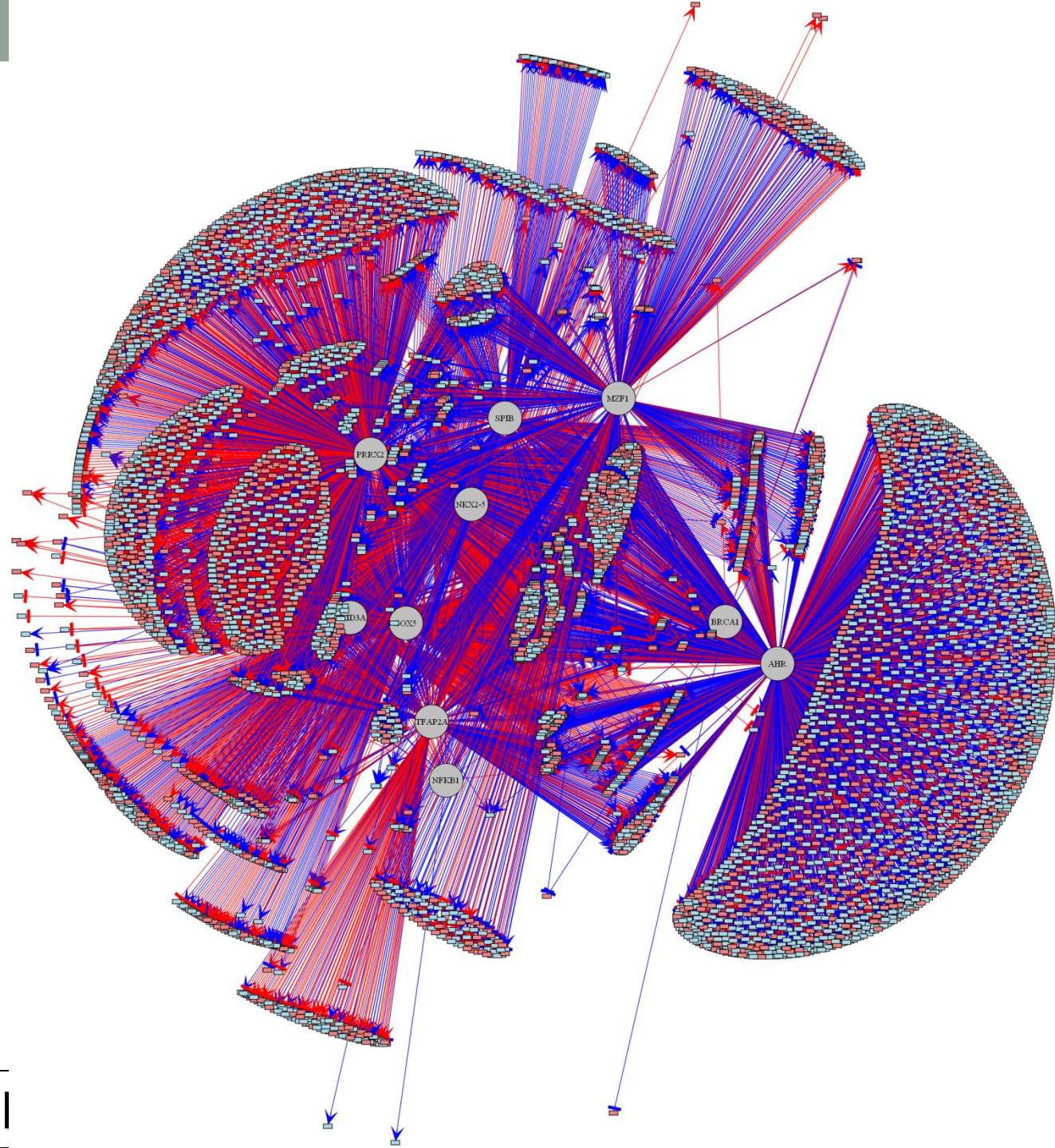
**1** Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, **2** Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, **3** Illumina, Inc., San Diego, California, United States of America, **4** Department of Pathology, Division of Woman's and Perinatal Pathology, Brigham and Women's Hospital, Boston, Massachusetts, United States of America, **5** Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, **6** Harvard School of Public Health, Boston, Massachusetts, United States of America, **7** Harvard Medical School, Boston, Massachusetts, United States of America

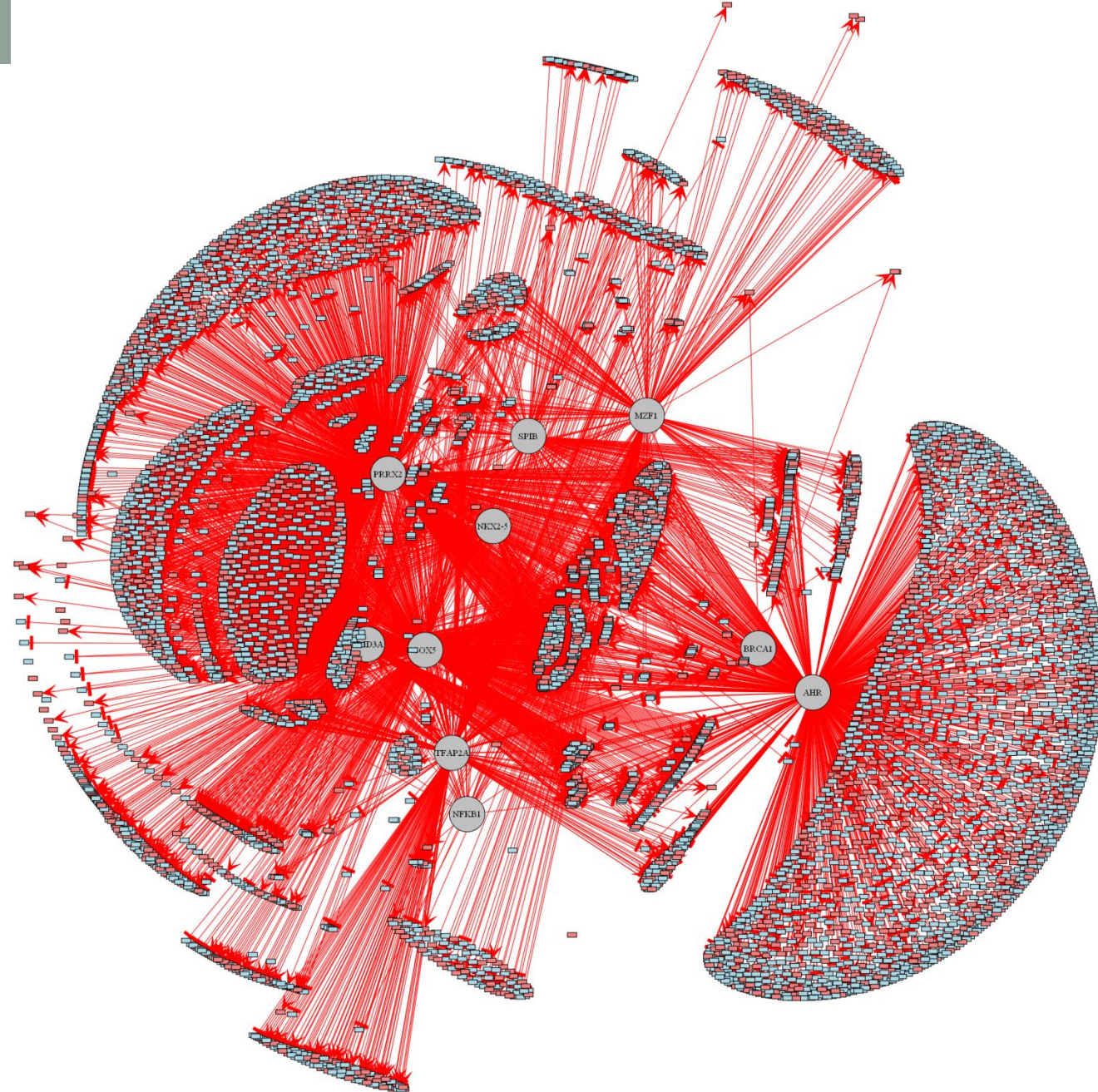
### Abstract

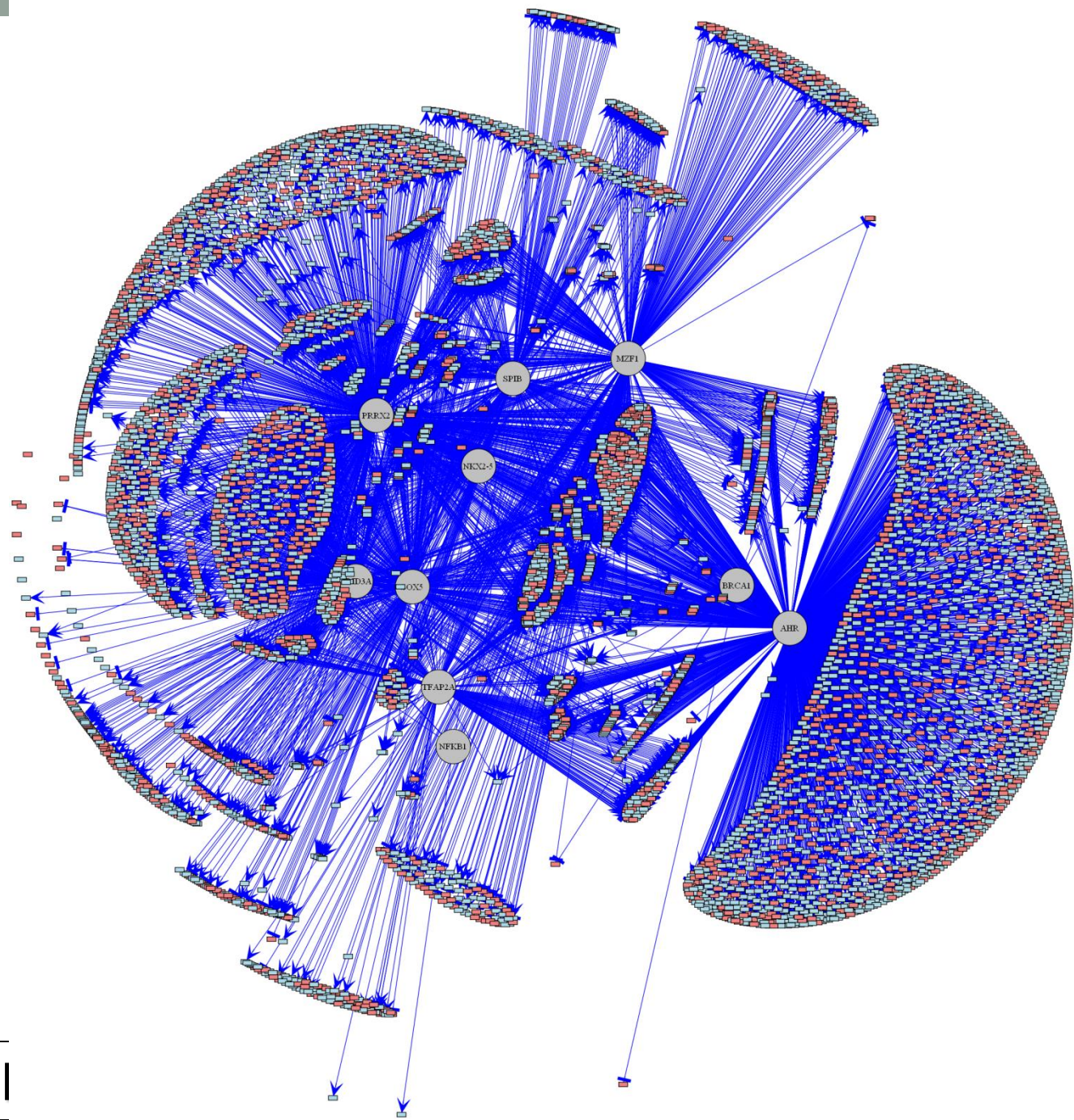
Ovarian cancer is the fifth leading cause of cancer death for women in the U.S. and the seventh most fatal worldwide. Although ovarian cancer is notable for its initial sensitivity to platinum-based therapies, the vast majority of patients eventually develop recurrent cancer and succumb to increasingly platinum-resistant disease. Modern, targeted cancer drugs intervene in cell signaling, and identifying key disease mechanisms and pathways would greatly advance our treatment abilities. In order to shed light on the molecular diversity of ovarian cancer, we performed comprehensive transcriptional profiling on 129 advanced stage, high grade serous ovarian cancers. We implemented a re-sampling based version of the

# PANDA: Integrative Network Models







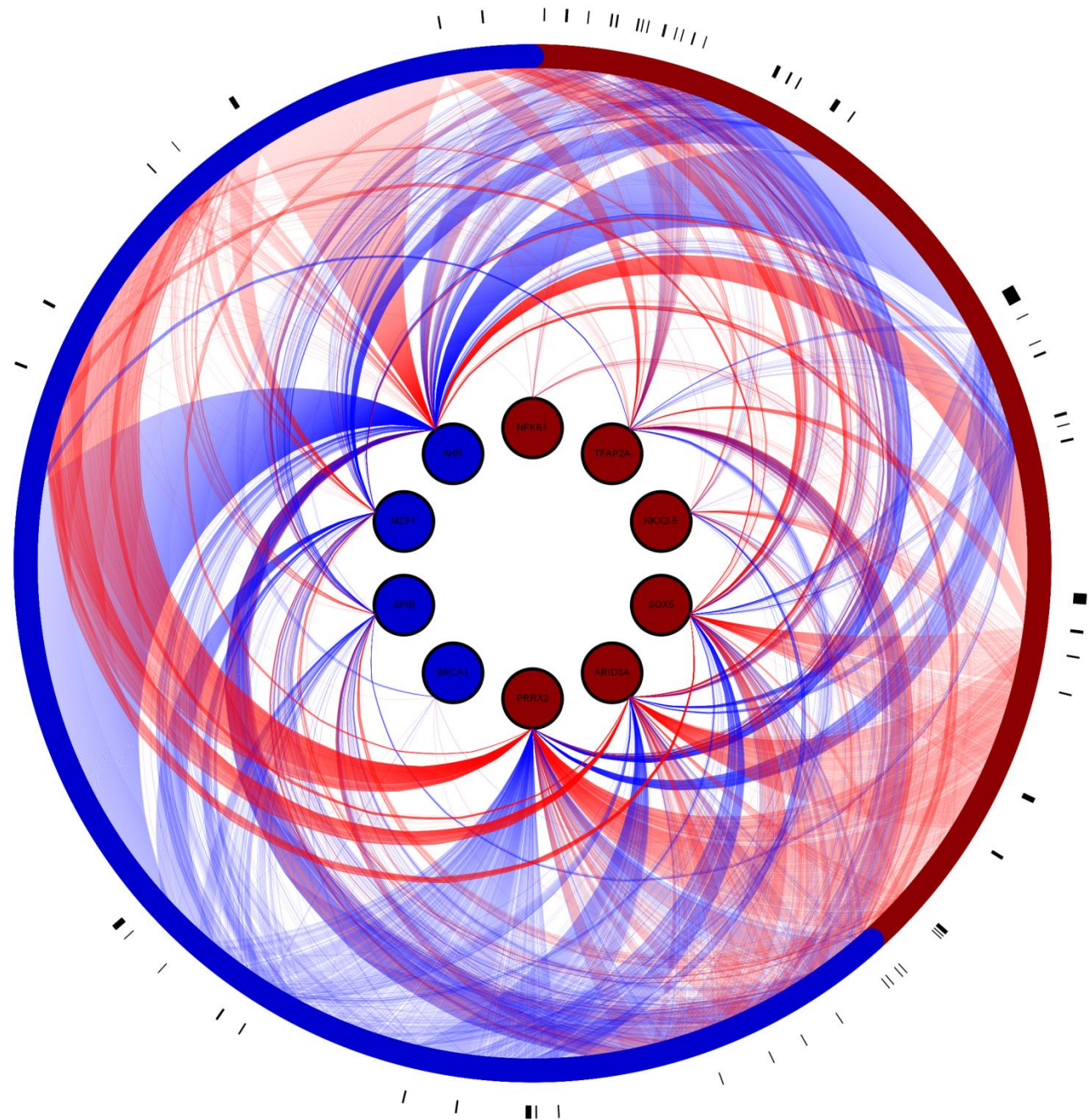


**Inner ring: key TFs**  
**Colored by Edge**  
**Enrichment (A or N)**

**Outer ring: genes**  
**Colored by Differential**  
**Expression (A or N)**

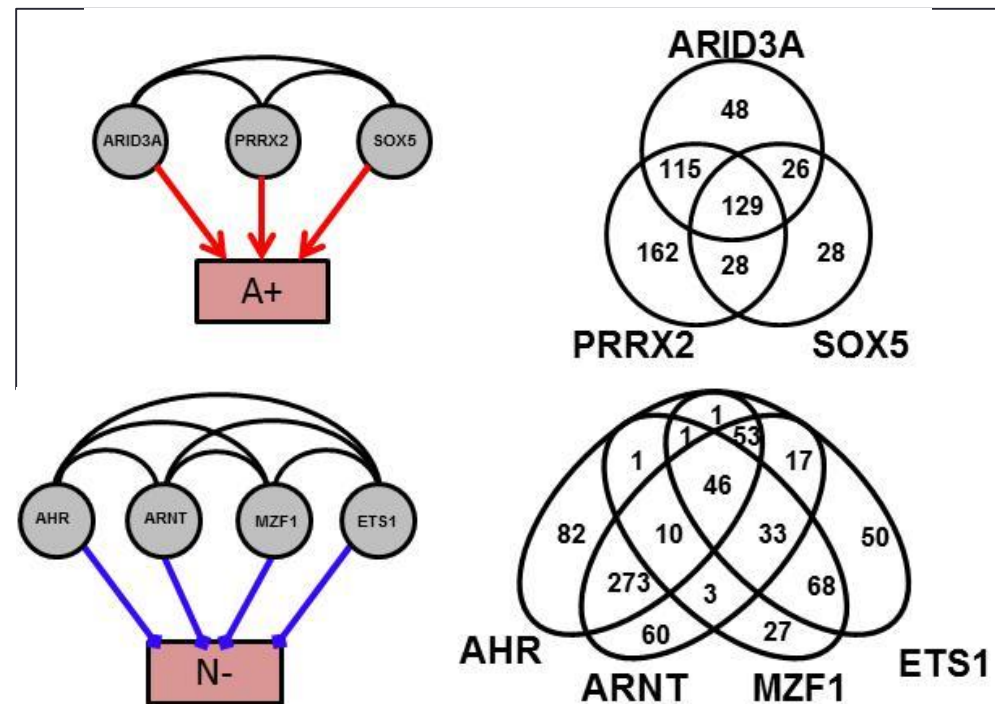
**Interring Connections**  
**Colored by**  
**Subnetwork (A or N)**

**Ticks – genes**  
**annotated to**  
**“angiogenesis” in GO,**



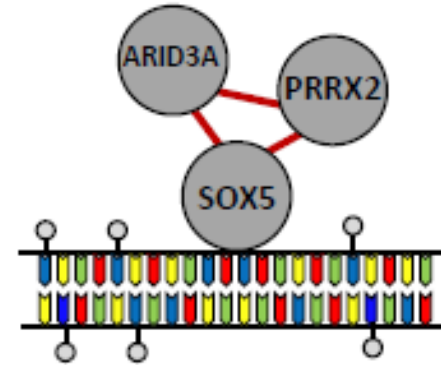
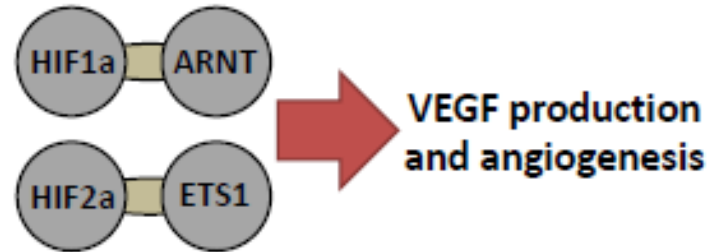
# Complex Regulatory Patterns Emerge

TF1	TF2	sig.	#	Class	Co-regulatory TF Pairs
ARID3A	PRRX2	1.16E-23	244	A+	
ARID3A	SOX5	1.01E-14	155	A+	
PRRX2	SOX5	3.83E-12	157	A+	
ARNT	MZF1	5.83E-23	92	N-	
AHR	ARNT	6.13E-16	382	N-	
ETS1	MZF1	9.08E-16	148	N-	



# Regulatory Patterns suggest Therapies

## ANGIOGENIC BEHAVIOR



High levels of CpG methylation

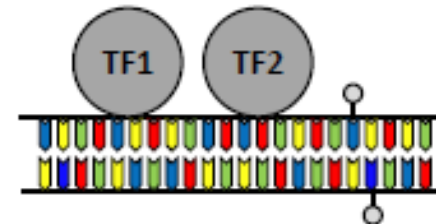
## TREATMENT MODEL

(1) Prevent ARNT/HIF1a and ETS1/HIF2a dimerization



(2) Promote ARNT/AHR and ETS1/AHR dimerization

(3) Decrease genome-wide methylation





## RESEARCH

# A network model of ovarian cancer

Kimberly Glass<sup>1,2</sup>

### Abstract

**Background:** We are interested in the mechanisms that influence the model information.

**Results:** We find that the subtypes, large-scale angiogenesis, and these factors are differentially expressed previously unreported of combinatorial.

**Conclusions:** The network models away from between subtypes.

**Keywords:** Network Angiogenesis



## RESEARCH

# Sexually related

Kimberly Glass<sup>1,2</sup>  
Guo-Cheng Yuan

### Abstract

**Background:** The network models in men and women that influence the network structure.

**Results:** Here we find that the Data Assimilation Pulmonary Disease adapting statistical targeting pattern function and the transcriptional.

**Conclusions:** The network models were not evident provides a primary in gene regulation.

**Keywords:** Network Disease



Contents lists available at ScienceDirect

## Genomics

journal homepage: [www.elsevier.com/locate/ygeno](http://www.elsevier.com/locate/ygeno)



# Diet-induced weight loss leads to a switch in gene regulatory network control in the rectal mucosa

Ashley J. Vargas<sup>a,b</sup>, John Quackenbush<sup>a,c</sup>, Kimberly Glass<sup>c,d,\*</sup>

<sup>a</sup> Harvard School of Public Health, Harvard University, Boston, MA, USA

<sup>b</sup> Cancer Prevention Fellowship Program, National Cancer Institute, Rockville, MD, USA

<sup>c</sup> Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

<sup>d</sup> Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA

## ARTICLE INFO

### Article history:

Received 7 April 2016

Received in revised form 9 August 2016

Accepted 10 August 2016

Available online xxxx

### Keywords:

Weight loss  
Colorectal cancer  
Colorectum  
Diet  
Gene network  
Gene regulation  
Diet

## ABSTRACT

**Background:** Weight loss may decrease risk of colorectal cancer in obese individuals, yet its effect in the colorectum is not well understood. We used integrative network modeling, Passing Attributes between Networks for Data Assimilation, to estimate transcriptional regulatory network models from mRNA expression levels from rectal mucosa biopsies measured pre- and post-weight loss in 10 obese, pre-menopausal women.

**Results:** We identified significantly greater regulatory targeting of glucose transport pathways in the post-weight loss regulatory network, including "regulation of glucose transport" (FDR = 0.02), "hexose transport" (FDR = 0.06), "glucose transport" (FDR = 0.06) and "monosaccharide transport" (FDR = 0.08). These findings were not evident by gene expression analysis alone. Network analysis also suggested a regulatory switch from NFkB1 to MAX control of MYC post-weight loss.

**Conclusions:** These network-based results expand upon standard gene expression analysis by providing evidence for a potential mechanistic alteration caused by weight loss.

© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

More application papers coming....

# Understanding Tissue-Specific Gene Regulation

Abhijeet Rajendra Sonawane,<sup>1,2</sup> John Platig,<sup>3,4</sup> Maud Fagny,<sup>3,4</sup> Cho-Yi Chen,<sup>3,4</sup> Joseph Nathaniel Paulson,<sup>3,4</sup> Camila Miranda Lopes-Ramos,<sup>3,4</sup> Dawn Lisa DeMeo,<sup>1,2,5</sup> John Quackenbush,<sup>1,2,3,4,6</sup> Kimberly Glass,<sup>1,2,7,8</sup> and Marieke Lydia Kuijjer<sup>3,4,7,\*</sup>

<sup>1</sup>Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA

<sup>2</sup>Department of Medicine, Harvard Medical School, Boston, MA 02115, USA

<sup>3</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

<sup>4</sup>Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

<sup>5</sup>Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA

<sup>6</sup>Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

<sup>7</sup>These authors contributed equally

<sup>8</sup>Lead Contact

\*Correspondence: [kimberly.glass@channing.harvard.edu](mailto:kimberly.glass@channing.harvard.edu) (K.G.), [mkuijjer@jimmy.harvard.edu](mailto:mkuijjer@jimmy.harvard.edu) (M.L.K.)

<https://doi.org/10.1016/j.celrep.2017.10.001>

## SUMMARY

Although all human tissues carry out common processes, tissues are distinguished by gene expression patterns, implying that distinct regulatory programs control tissue specificity. In this study, we investigate gene expression and regulation across 38 tissues profiled in the Genotype-Tissue Expression project. We find that network edges (transcription factor to target gene connections) have higher

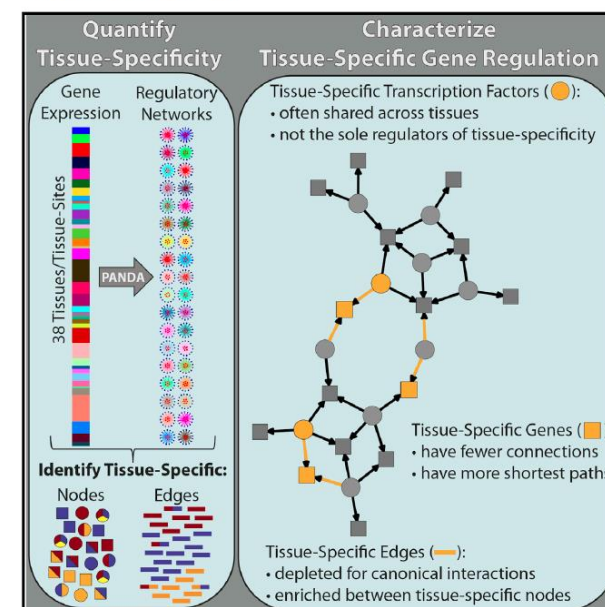
biological function requires the combinatorial multiple regulatory elements, primarily transcription factors, that work together with other genetic and environmental factors to mediate the transcription of genes and their products (Vaquerizas et al., 2009).

Gene regulatory network modeling provides a framework that can summarize the complex interactions between transcription factors, genes, and gene products (Oltvai, 2004; Gerstein et al., 2012). Despite the importance of the regulatory process, the most widely used network methods are based on pairwise gene co-expression

## Cell Reports

# Understanding Tissue-Specific Gene Regulation

## Graphical Abstract



## Highlights

- Regulatory network connections are more tissue specific than nodes (genes and transcription factors)

## Article

## Authors

Abhijeet Rajendra Sonawane, John Platig, Maud Fagny, ..., John Quackenbush, Kimberly Glass, Marieke Lydia Kuijjer

## Correspondence

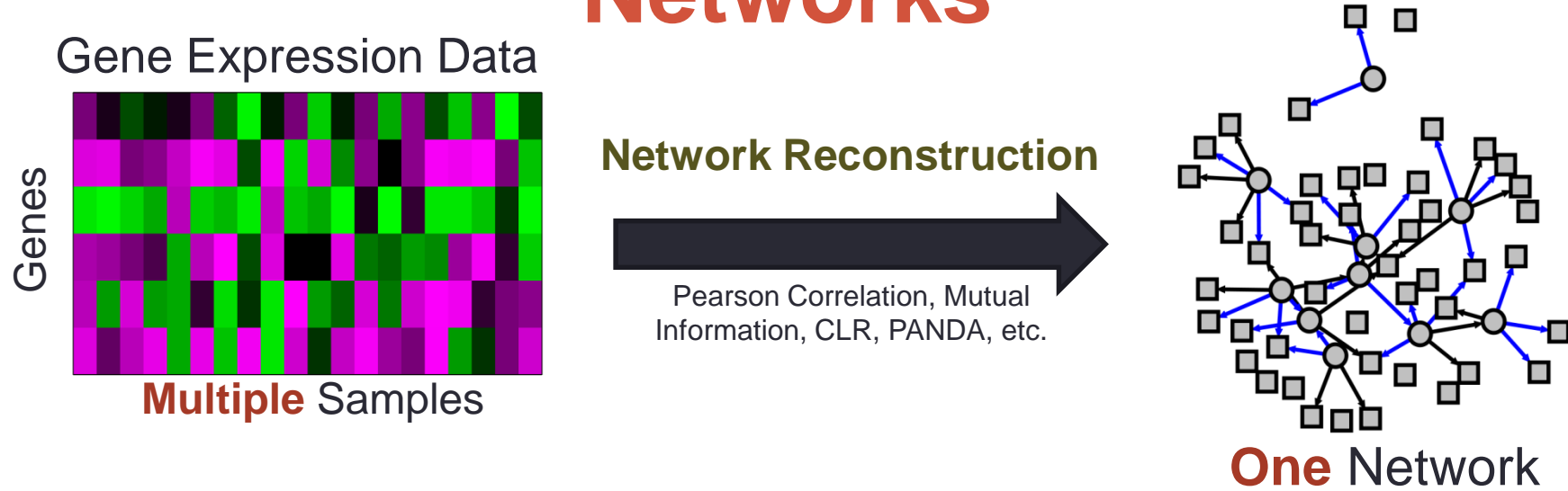
[kimberly.glass@channing.harvard.edu](mailto:kimberly.glass@channing.harvard.edu) (K.G.), [mkuijjer@jimmy.harvard.edu](mailto:mkuijjer@jimmy.harvard.edu) (M.L.K.)

## In Brief

Understanding gene regulation is important for many fields in biology and medicine. Sonawane et al. reconstruct and investigate regulatory networks for 38 human tissues. They find that regulation of tissue-specific function is largely *independent* of transcription factor expression and that tissue specificity appears to be mediated by tissue-specific regulatory network paths.

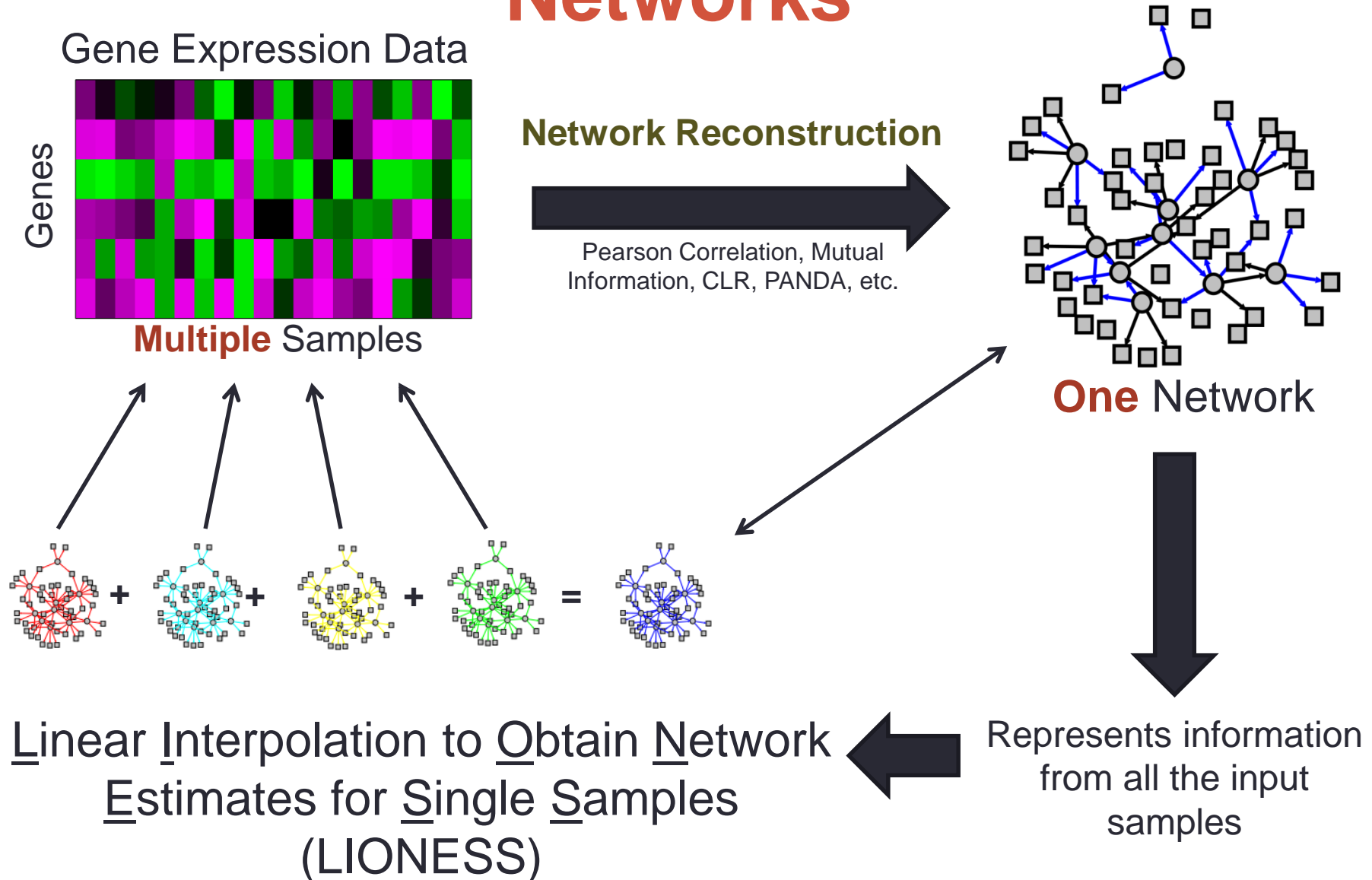
**Question 3:**  
**Can we move beyond**  
**THE Network?**

# Reconstructing Gene Regulatory Networks

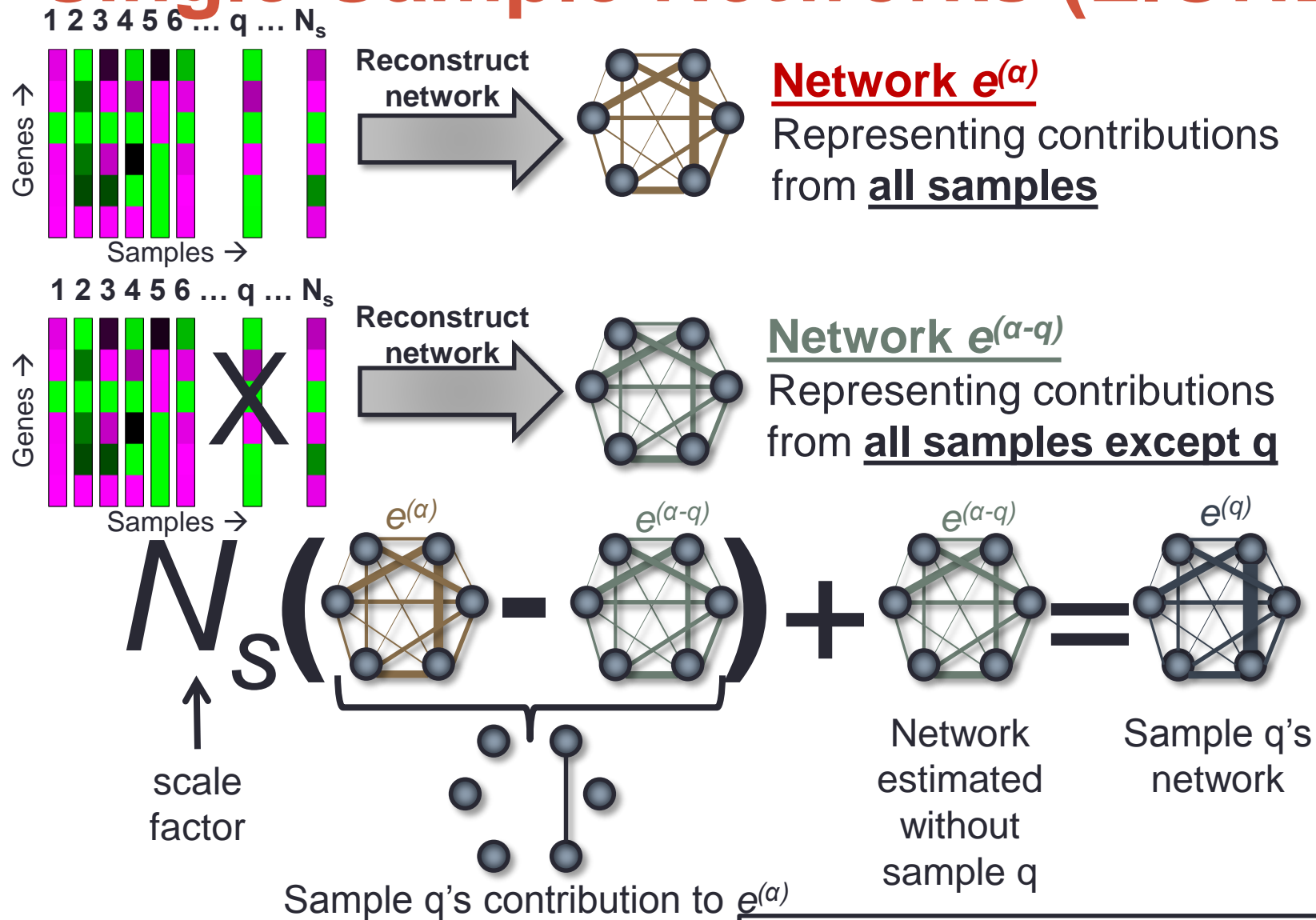


**We generally estimate “Aggregate” Networks.**

# Reconstructing Gene Regulatory Networks



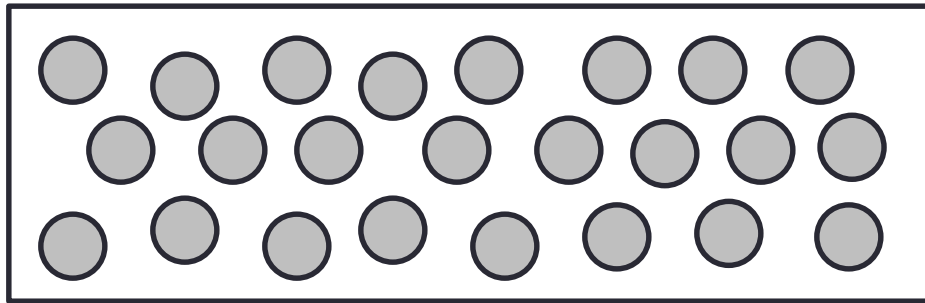
# Single-Sample Networks (LIONESS)



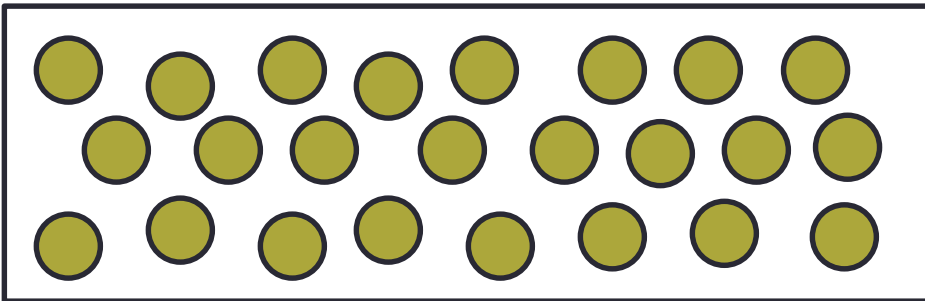
$$N_s \left( e_{ij}^{(\alpha)} - e_{ij}^{(\alpha-q)} \right) + e_{ij}^{(\alpha-q)} = e_{ij}^{(q)}$$

# A Quick Test Using Yeast Cell Cycle Data

- Data includes 48 total expression arrays taken over a time-course (every 5 min) on synchronized yeast cells (~2 cell cycles)
- Includes technical replicates (Cy3/Cy5 and Cy5/Cy3)
- Estimate single-sample networks using data for each replicate.

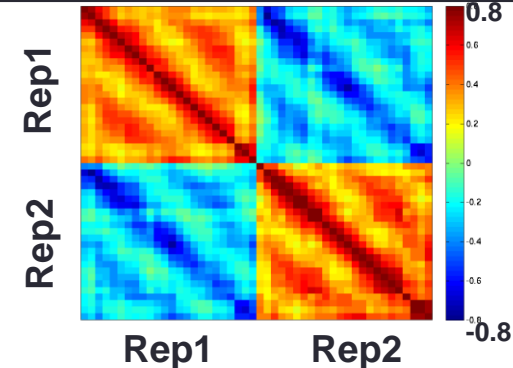


Replicate 1 (24 Samples → 24 Networks)

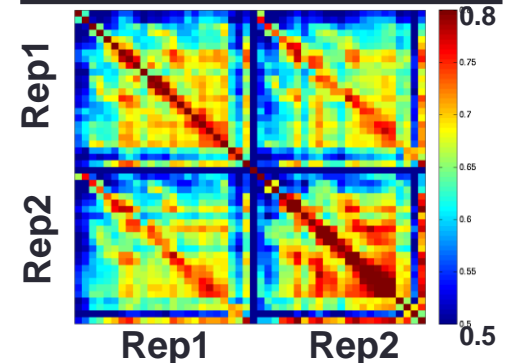


Replicate 2 (24 Samples → 24 Networks)

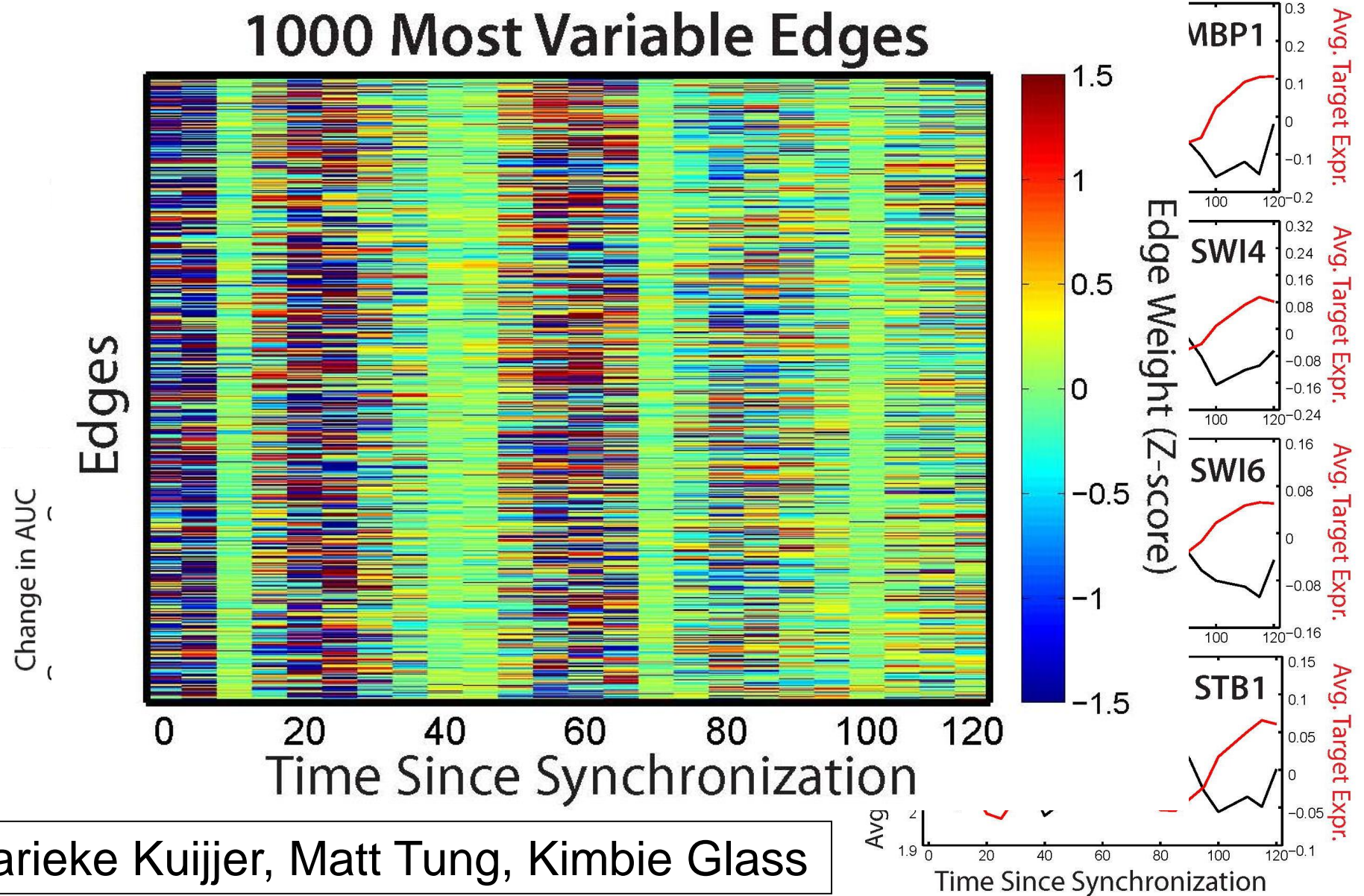
Correlation in Expression



Correlation in Networks

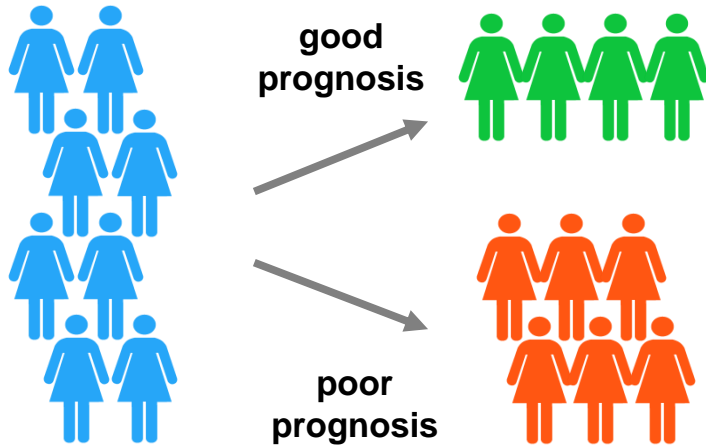


# Validation and Insight

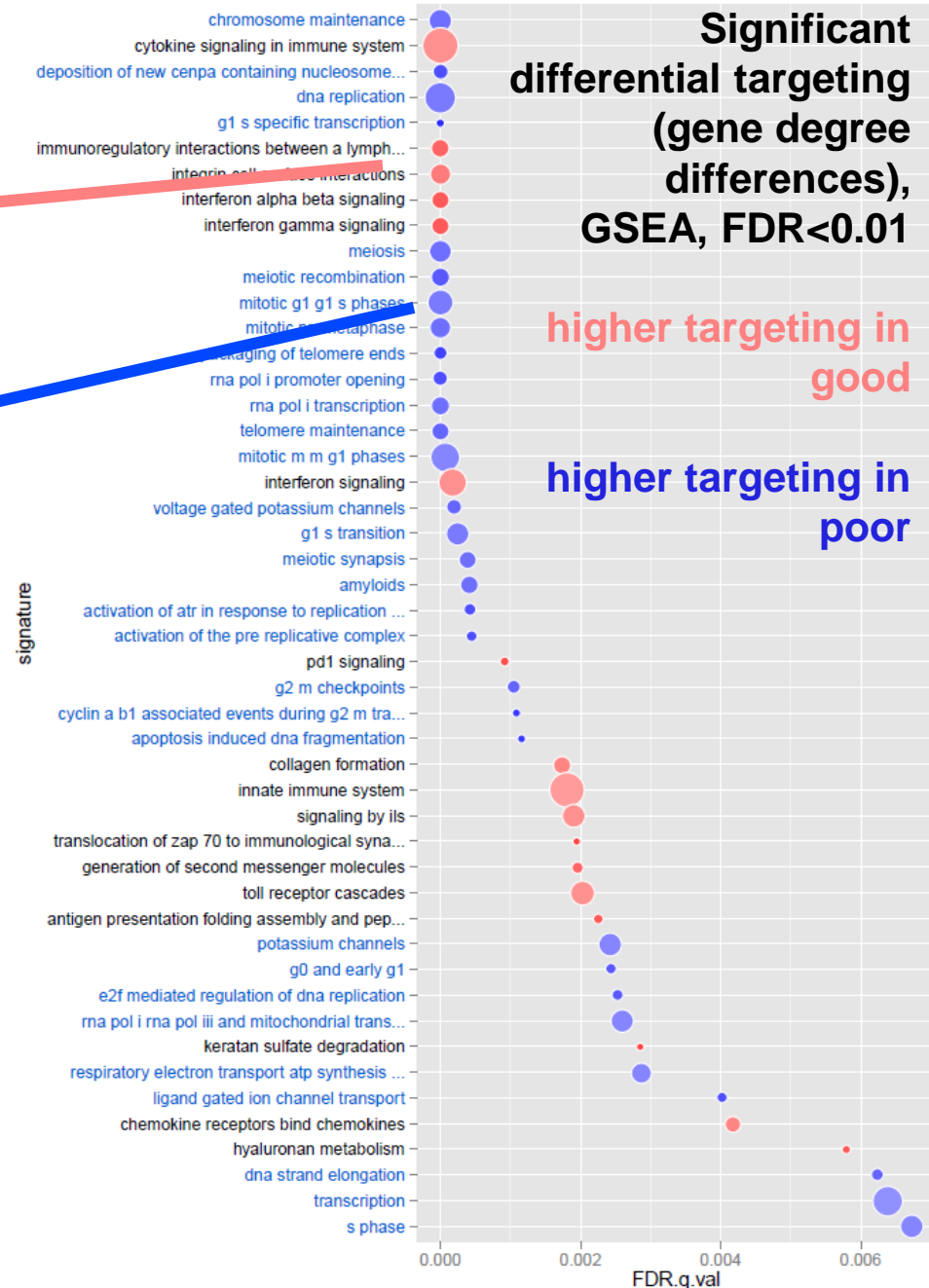


Marieke Kuijjer, Matt Tung, Kimbie Glass

# Glioblastoma network signatures



- Single-sample networks for TCGA glioblastoma patients
- 3yr survival to define good and poor prognosis
- *LIMMA* analysis using network “edge Z-scores”
- Analysis points to important roles for *FOS-JUN* and *NFkB*
- Gene degree differences identify **mitosis** and **immune**-related genes



**Before I came here I was confused  
about this subject.**

**After listening to your lecture,  
I am still confused but at a higher level.**

**- Enrico Fermi, (1901-1954)**

# Acknowledgments

<http://compbio.dfci.harvard.edu>

## Gene Expression Team

Fieda Abderazzaq  
Aedin Culhane  
Jessica Mar  
Renee Rubio

## DFCI Radiology

Hugo Aerts

## University of Queensland

Christine Wells  
Lizzy Mason

## CDNM, Brigham and Women's Hospital

Peter Castaldi  
Michael Cho  
Dawn DeMeo  
Kimberly Glass  
Ed Silverman  
Xiaobo Zhou

## Center for Cancer Computational Biology

Fieda Abderazzaq  
Stas Alekseev  
Nicole Flanagan  
Ed Harms  
Lev Kuznetsov  
Brian Lawney  
Antony Partensky  
John Quackenbush  
Renee Rubio  
Yaoyu E. Wang

<http://cccb.dfci.harvard.edu>



## Administrative Support

Nicole Trotman



## Students and Postdocs

Joseph Barry  
Joey (Cho-Yi) Chen  
Marieke Kuijjer  
Camila Lopes-Ramos  
Zachary McCaw  
Megha Padi  
Joseph Paulson  
John Platig  
Daniel Schlauch  
Daphne Tsoucas

## Alumni

Martin Aryee  
Stefan Bentink  
Kimberly Glass  
Benjamin Haibe-Kains  
Kaveh Maghsoudi  
Jess Mar  
Melissa Merritt  
Alejandro Quiroz  
J. Fah Sathirapongsasuti

