USING NETWORKS TO UNDERSTAND THE GENOTYPE-PHENOTYPE CONNECTION

John Quackenbush
Dana-Farber Cancer Institute
Harvard TH Chan School of Public Health

Every revolution in science—from Copernican heliocentric model to the rise of statistical and quantum mechanics, from Darwin's theory of evolution and natural selection to the theory of the gene—has been driven by one and only one thing: access to data.

–John Quackenbush

@johnquackenbush-Every revolution in the history science has been driven by one and only one thing: access to data.

Twitter version, 115 characters with spaces

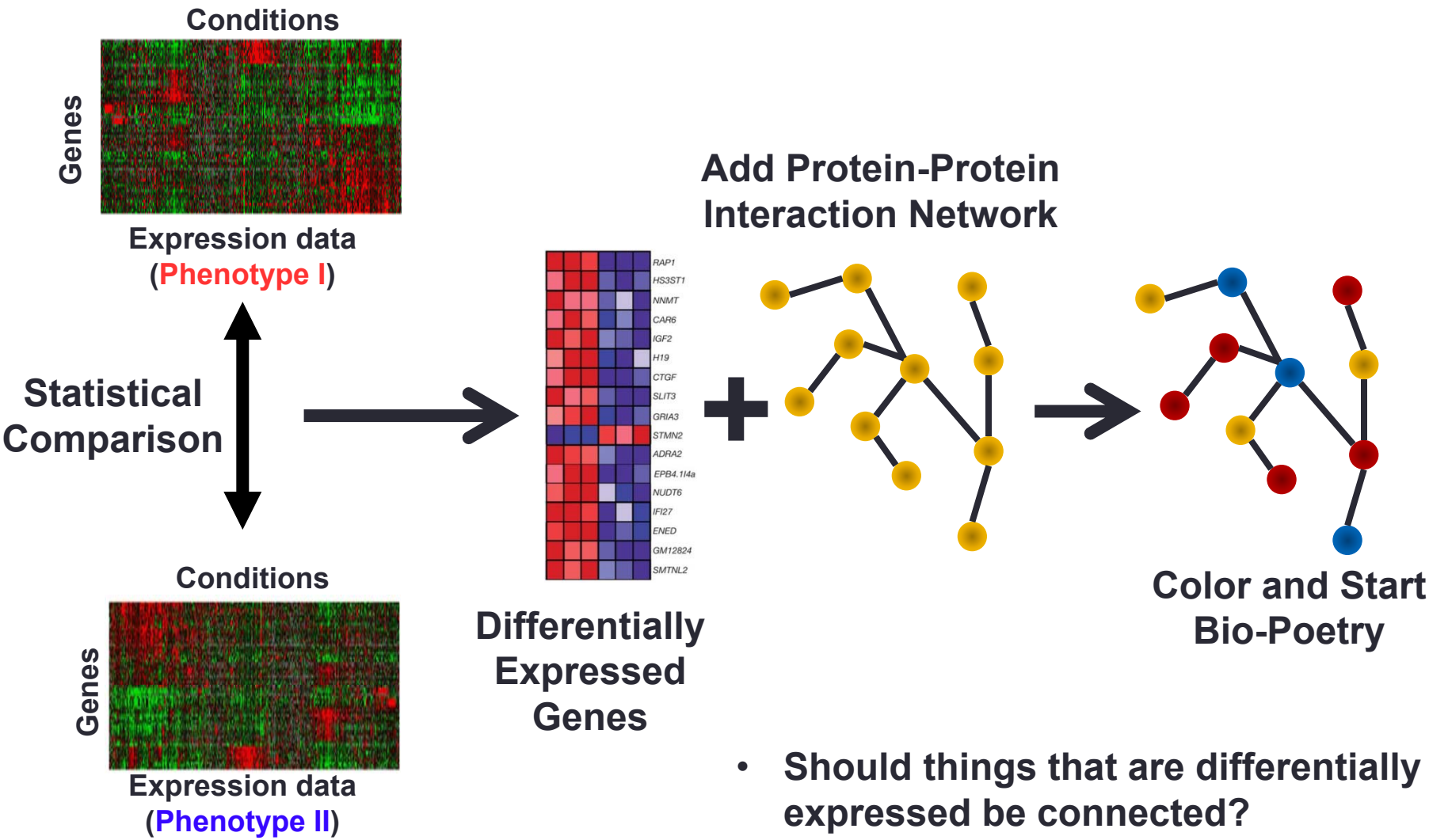
@notrealdonaldtrump-We have the best data. Fantastic Bigly! The other scientists. losers So sad #fakedata #failedscientists

Twitter version, 123 characters with spaces

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

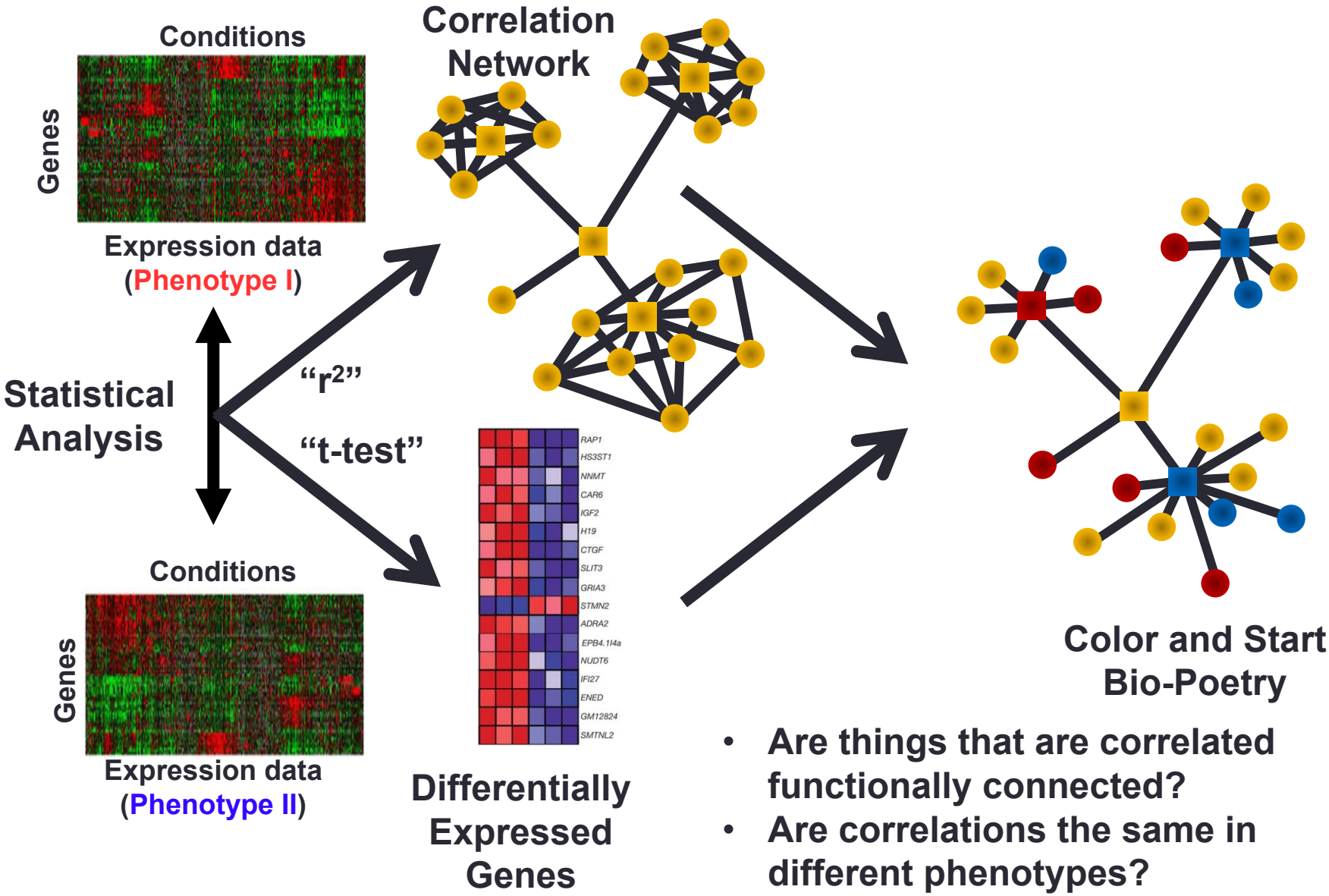
– George E. Box

How *NOT* to do Network Analysis



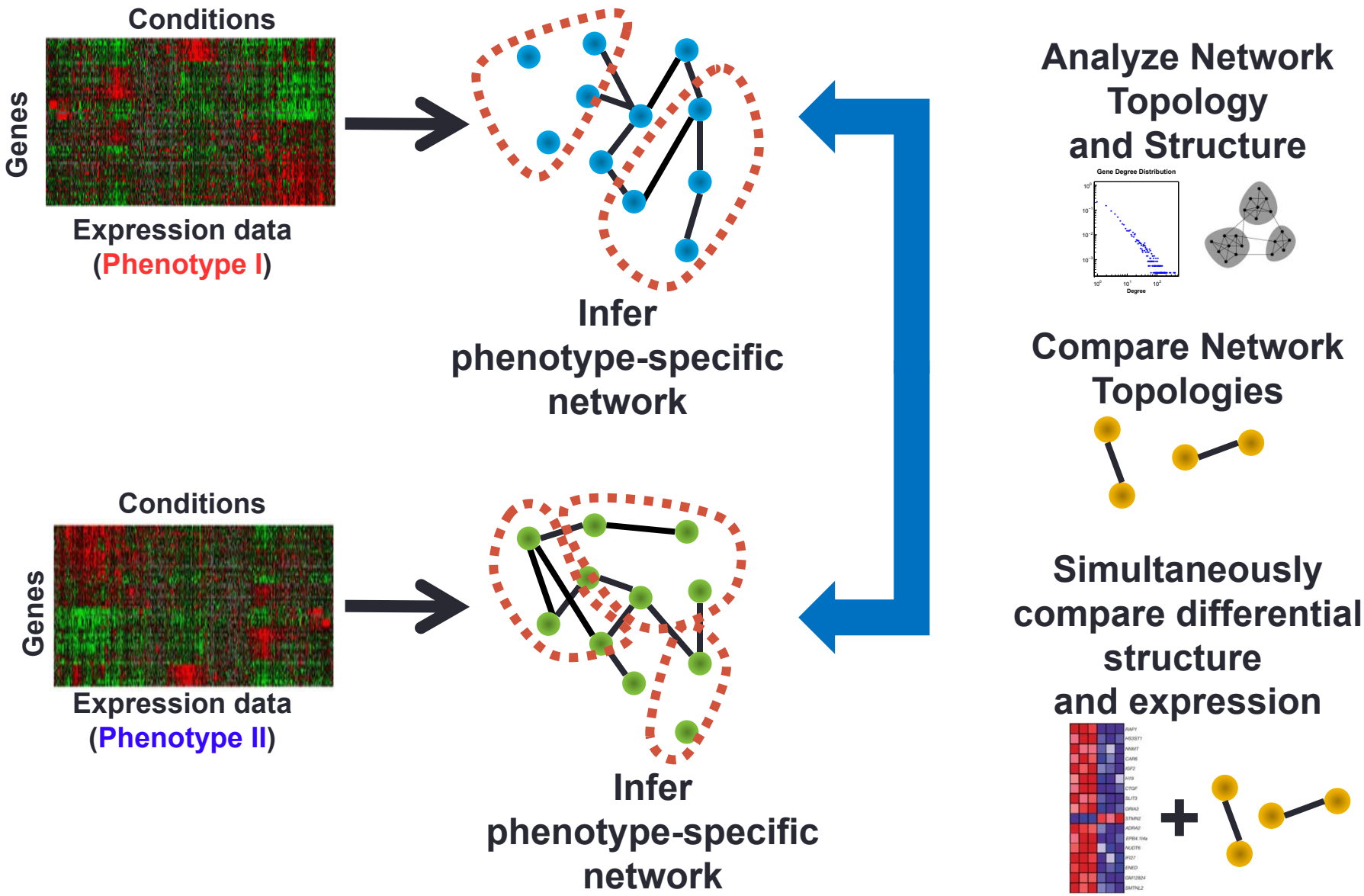
- Should things that are differentially expressed be connected?
- Is the PPI network even “relevant”?

How *NOT* to do Network Analysis (2)



- Are things that are correlated functionally connected?
- Are correlations the same in different phenotypes?

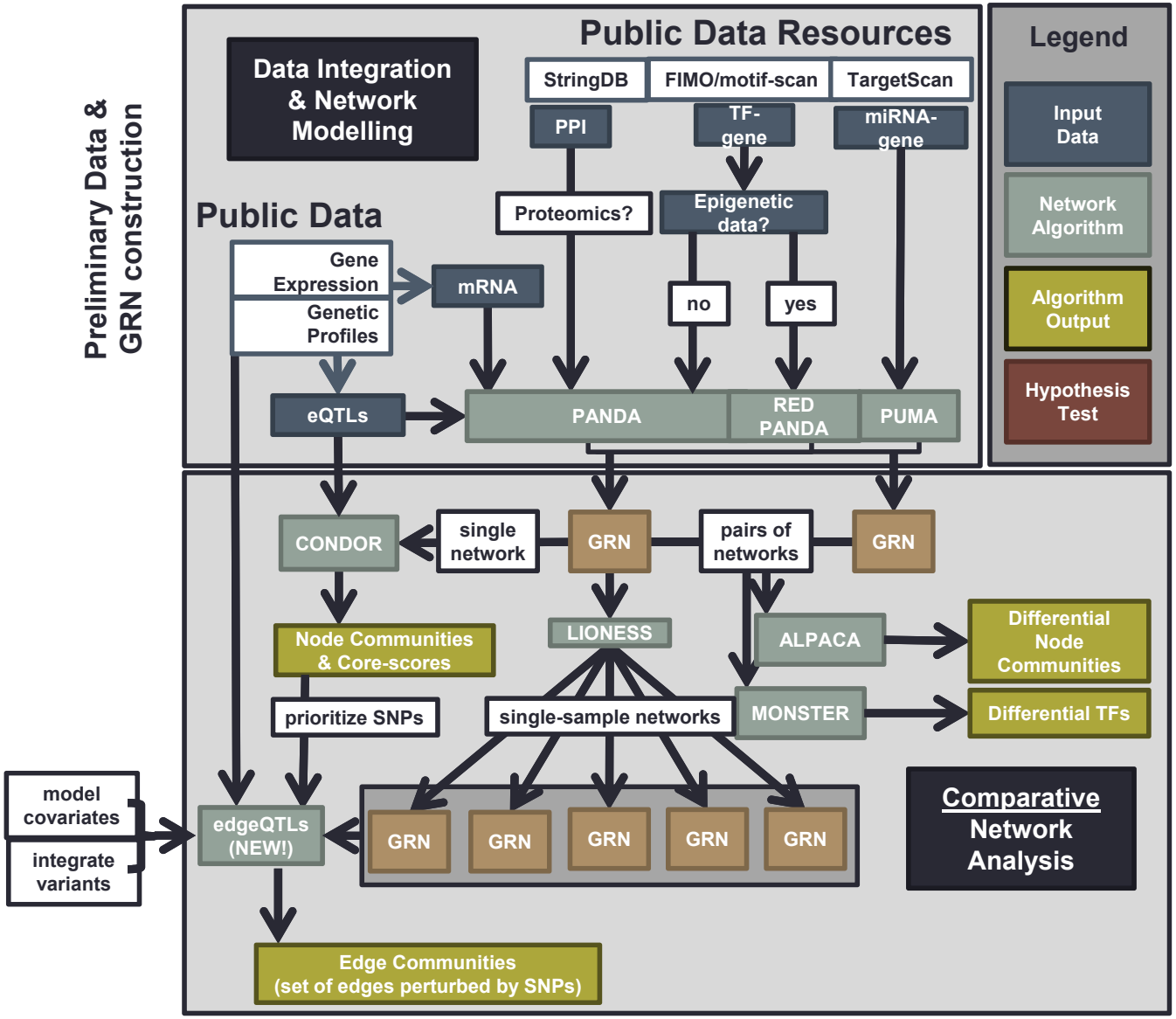
How *we* do Network Analysis



Starting Assumptions

- There is no single “right” network
- The structure of the network matters and network structure often changes between states.
- The “atom” of the network is the edge.
- 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



Algorithm Key

PANDA – Passing Attributes between Networks for Data Assimilation

RED PANDA – Regulation with Epigenetic Data using PANDA

PUMA – PANDA Using Micro-RNA Associations

CONDOR – COmplex Network Description Of Regulators

LIONESS – Linear Interpolation to Obtain Network Estimates for Single Samples

ALPACA – ALtered Partitions Across Community Architectures

MONSTER – MOdeling Network State Transitions from Expression and Regulatory data

Integrative Network Inference: PANDA

OPEN ACCESS Freely available online

PLOS | ONE

Passing Messages between Biological Networks to Refine Predicted Interactions

Kimberly Glass[1,2], Curtis Huttenhower[2], John Quackenbush[1,2], Guo-Cheng Yuan[1,2]*

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.

LIONESS

Cornell University Library

We gratefully acknowledge support from the Simons Foundation and member institutions

arXiv.org > q-bio > arXiv:1505.06440

Search or Article-id (Help | Advanced search) All papers Go!

Quantitative Biology > Molecular Networks

Estimating sample-specific regulatory networks

Marieke Lydia Kuijjer, Matthew Tung, GuoCheng Yuan, John Quackenbush, Kimberly Glass

(Submitted on 24 May 2015)

Biological systems are driven by intricate interactions among the complex array of molecules that comprise the cell. Many methods have been developed to reconstruct network models that attempt to capture those interactions. These methods often draw on large numbers of measured expression samples to tease out subtle signals and infer connections between genes (or gene products). The result is an aggregate network model representing a single estimate for edge likelihoods. While informative, aggregate models fail to capture the heterogeneity that is often represented in a population. Here we propose a method to reverse engineer sample-specific networks from aggregate network models. We demonstrate the accuracy and applicability of our approach in several datasets, including simulated data, microarray expression data from synchronized yeast cells, and RNA-seq data collected from human subjects. We show that these sample-specific networks can be used to study the evolution of network topology across time and to characterize shifts in gene regulation that may not be apparent in the expression data. We believe the ability to generate sample-specific networks will revolutionize the field of network biology and has the potential to usher in an era of precision network medicine.

Subjects: **Molecular Networks (q-bio.MN)**
Cite as: **arXiv:1505.06440 [q-bio.MN]**
(or **arXiv:1505.06440v1 [q-bio.MN]** for this version)

Submission history

From: Kimberly Glass [view email]
[v1] Sun, 24 May 2015 13:45:51 GMT (3558kb,D)

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Link back to: arXiv, form interface, contact.

Download:

- Citation
- PDF (Drag to your desktop to save)
- Other formats

(cc) BY

Current browse context:
q-bio.MN
< prev | next >
new | recent | 1505

Change to browse by:
q-bio

References & Citations
- NASA ADS

Bookmark (what is this?)

http://tinyurl.com/JQ-LIONESS

MONSTER

Cold Spring Harbor Laboratory

bioRχiv beta

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABO

Search

New Results

Estimating Drivers of Cell State Transitions Using Gene Regulatory Network Models

Daniel Schlauch, Kimberly Glass, Craig P Hersh, Edwin K Silverman, John Quackenbush

doi: https://doi.org/10.1101/089003

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract Info/History Metrics Preview PDF

Abstract

Specific cellular states are often associated with distinct gene expression patterns. These states are plastic, changing during development, or in the transition from health to disease. One relatively simple extension of this concept is to recognize that we can classify different cell-types by their active gene regulatory networks and that,

http://tinyurl.com/JQ-MONSTER

CONDOR

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Bipartite Community Structure of eQTLs

John Platig[1,2]*, Peter J. Castaldi[3,4,5], Dawn DeMeo[3,5,6], John Quackenbush[1,2,3‡]

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

CrossMark
click for updates

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

Received: January 8, 2016

Accepted: June 23, 2016

Published: September 12, 2016

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 applying CONDOR to eQTLs in chronic obstructive pulmonary disease (COPD), we found the global network "hub" SNPs were devoid of disease associations through GWAS. However, the network was organized into 52 communities of SNPs and genes, many of which were enriched for genes in specific functional classes. We identified local hubs within each

APLACA

Cold Spring Harbor Laboratory

bioRχiv beta

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS

Search

Advanced Search

New Results

Previous Next

Phenotype-Driven Transitions In Regulatory Network Structure

Posted May 25, 2017.

Megha Padi, John Quackenbush

doi: https://doi.org/10.1101/142281

This article is a preprint and has not been peer-reviewed [what does this mean?].

Download PDF Share

Email Citation Tools

Abstract | Info/History | Metrics | Supplementary material

Tweet Like 19

G+1 0

Preview PDF

Abstract

Complex traits and diseases like human height or cancer are often not caused by a single mutation or genetic variant, but instead

Subject Area

Systems Biology

http://tinyurl.com/JQ-ALPACA

GTEx:
A Sandbox
for Methods Development

We need a big sandbox

RESEARCH ARTICLE

HUMAN GENOMICS

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

The GTEx 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 (GTEx) 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*).

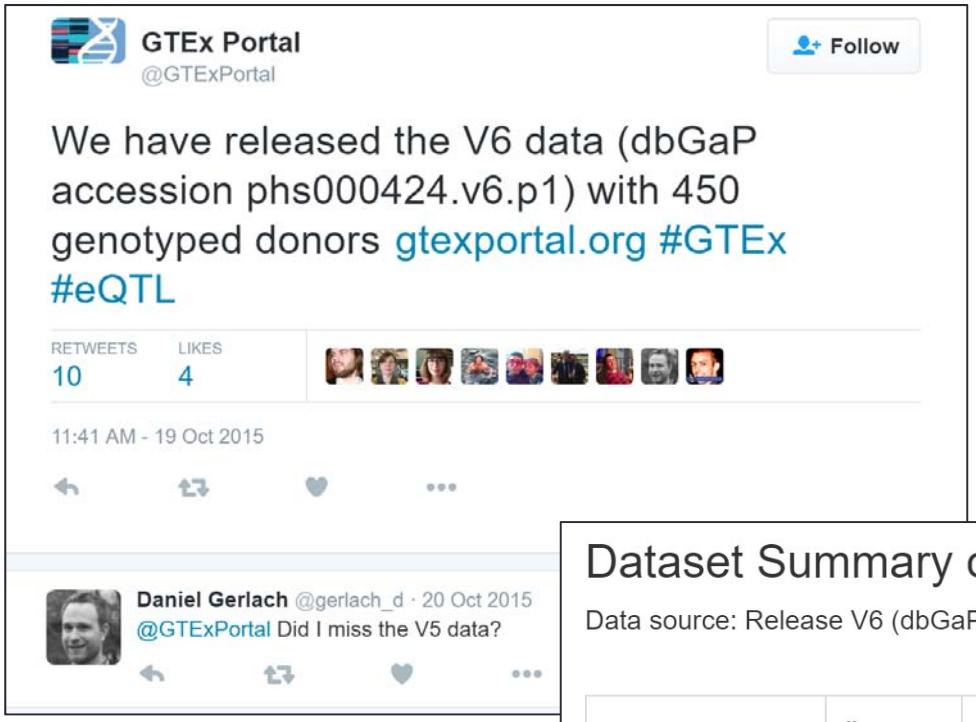
We performed 76–base pair (bp) paired-end mRNA sequencing on a total of 1749 samples, of which 1641 samples from 43 sites, and 175 donors, constituted our final "pilot data freeze" reported on here (*14*). Median sequencing depth was 82.1 million mapped reads per sample (fig. S3A). The final data freeze 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-

ncemag.org/ on May 10, 2016

GTEx: A lot of sand



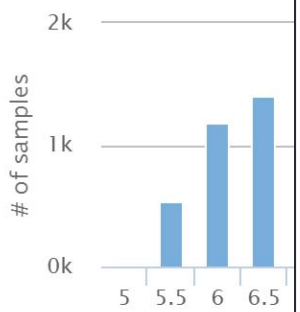
Dataset Summary of Analysis Samples of the V6 Release

Data source: Release V6 (dbGaP Accession phs000424.v6.p1)

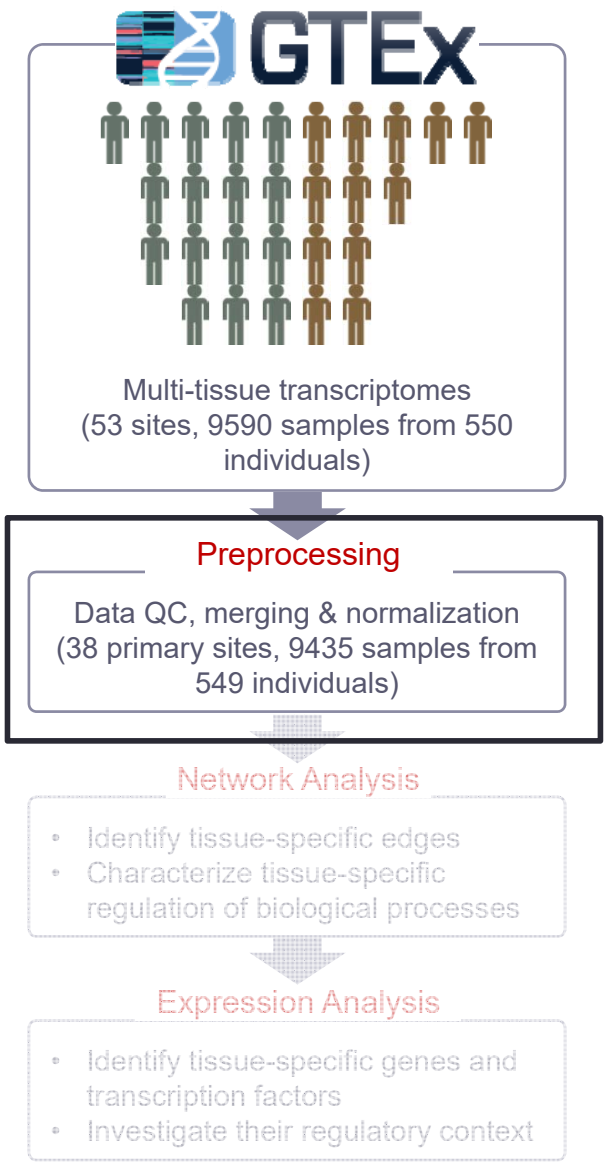
V6 Release	# Tissues	# Donors	# Samples
Total	53	544	8555
With Genotype	53	449	7333
Has eQTL Analysis*	44	449	7051

** Number of samples with genotype >= 70*

Sample RIN Histogram

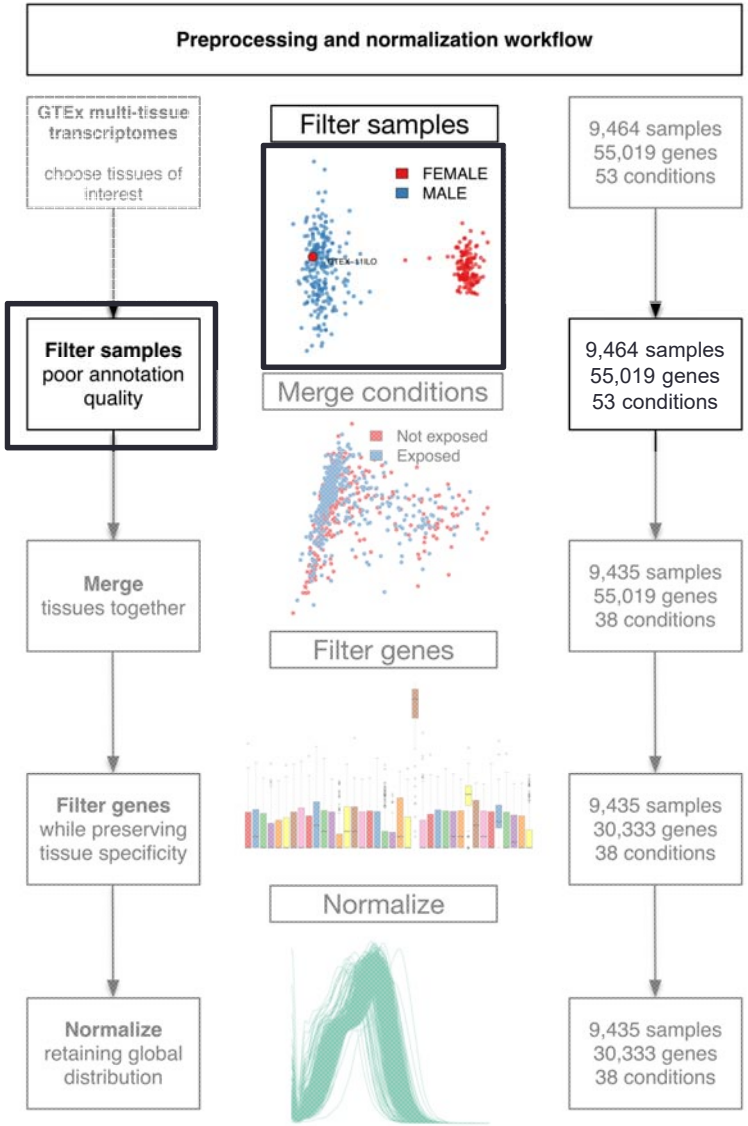


Tissue-Specific Expression Data from GTEx



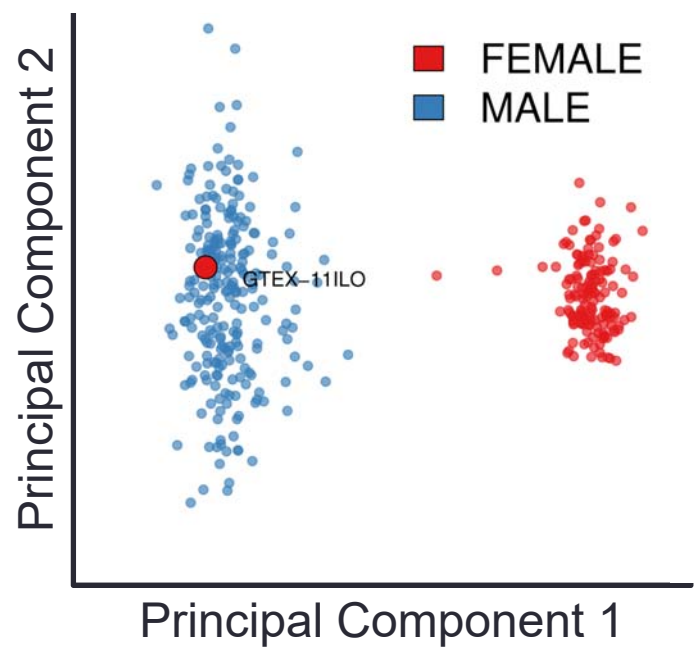
- GTEx: Genotype-Tissue Expression project.
- Version 6 includes ~10,000 expression assays representing postmortem samples collected from ~550 individuals across ~50 different body sites.
- Downloaded, pre-processed and normalized this data.
- Used expression data to explore tissue-specific gene regulation.

Step 1: Filter Samples

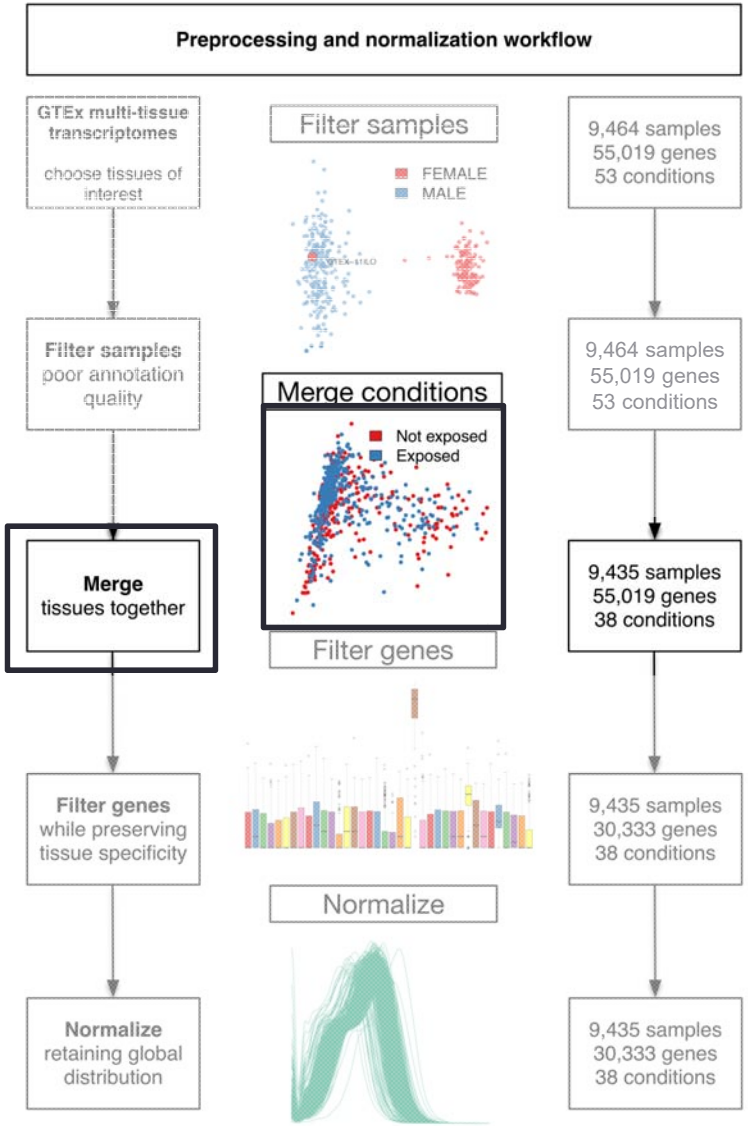


http://tinyurl.com/gtex01

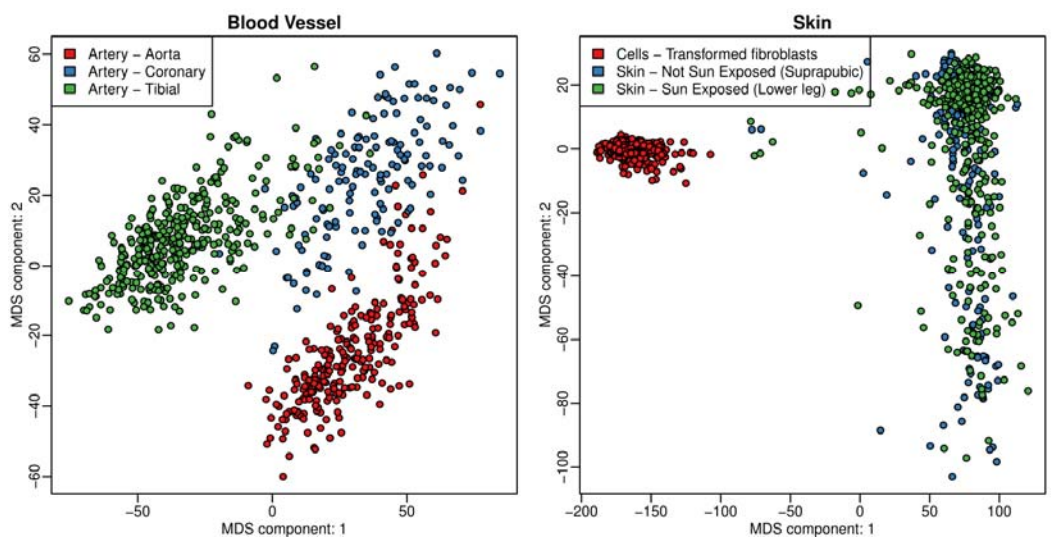
- Ran a principal coordinate analysis on the Y-chromosome expression in each tissue.
- Identified and removed one individual with apparent sex-misidentification.



Step 2: Merge Tissues to Increase Power

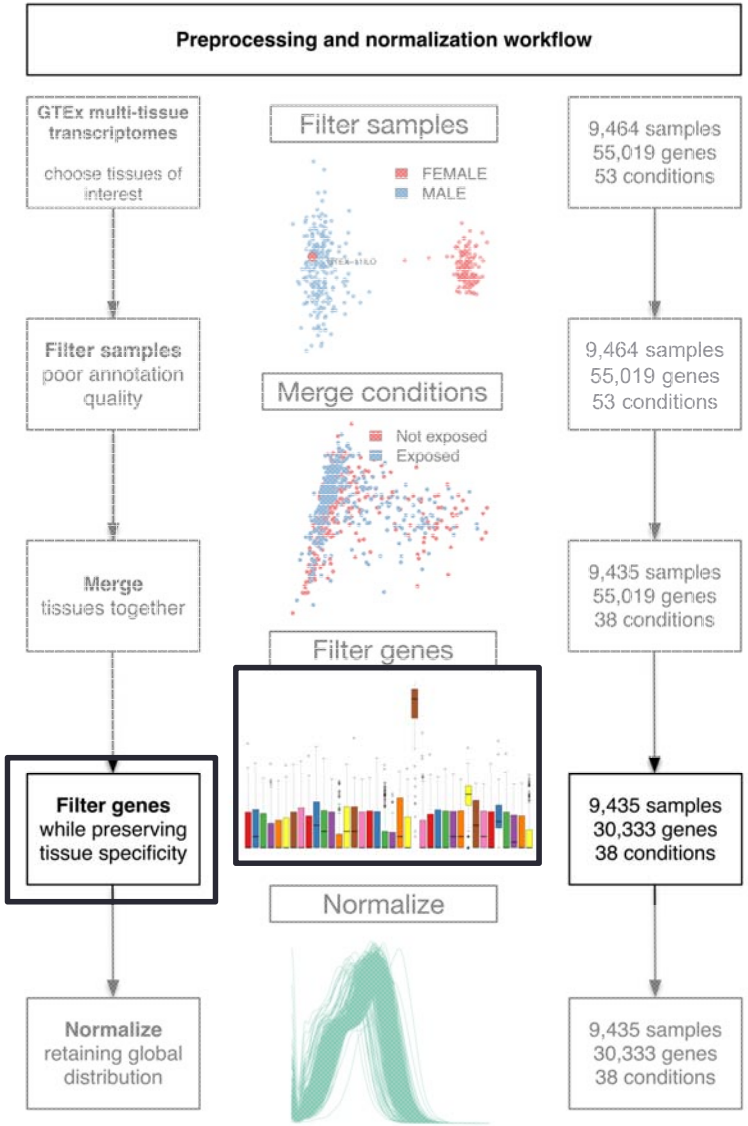


- Ran a principal coordinate analysis on the expression data in each tissue.
- Combined sites from the same tissue that were indistinguishable by PCA.

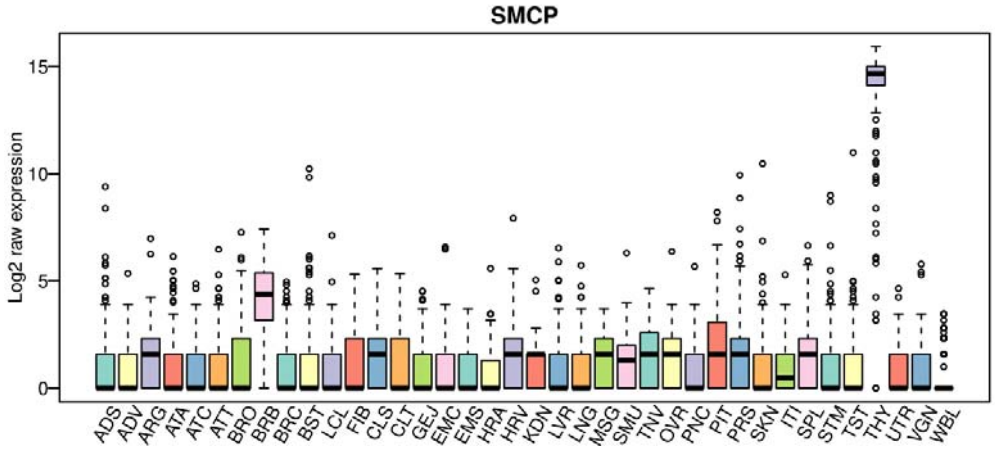


http://tinyurl.com/gtex01

Step 3: Filter Genes with Few Reads

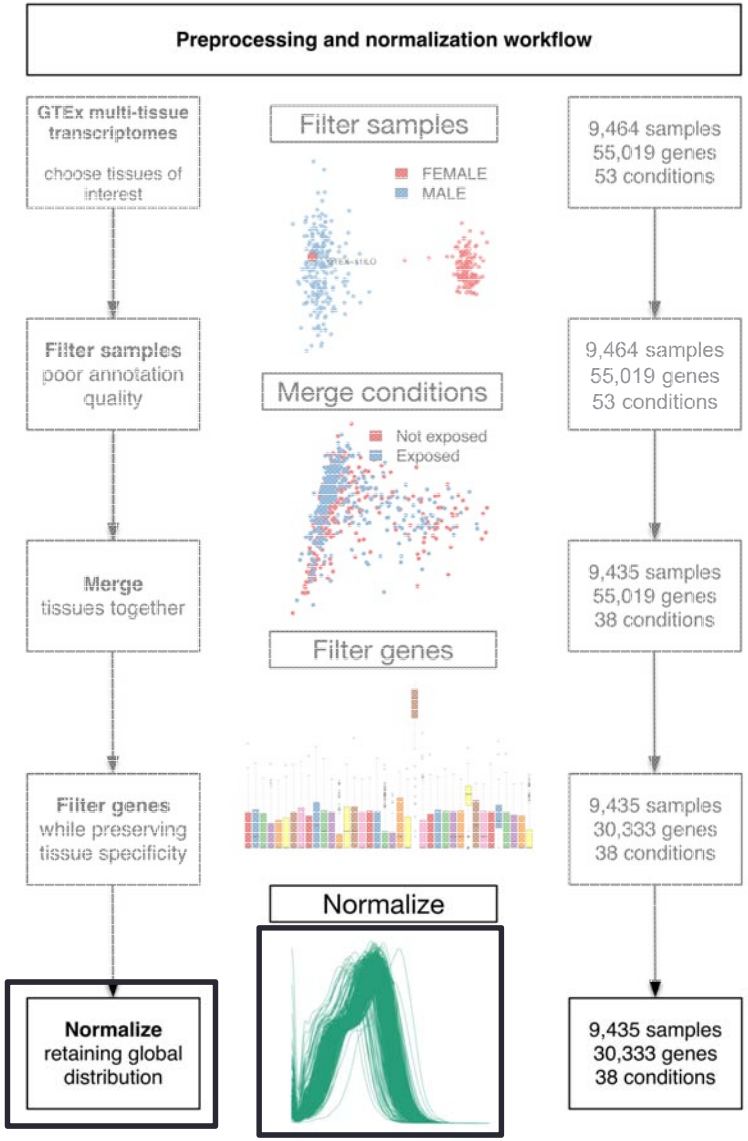


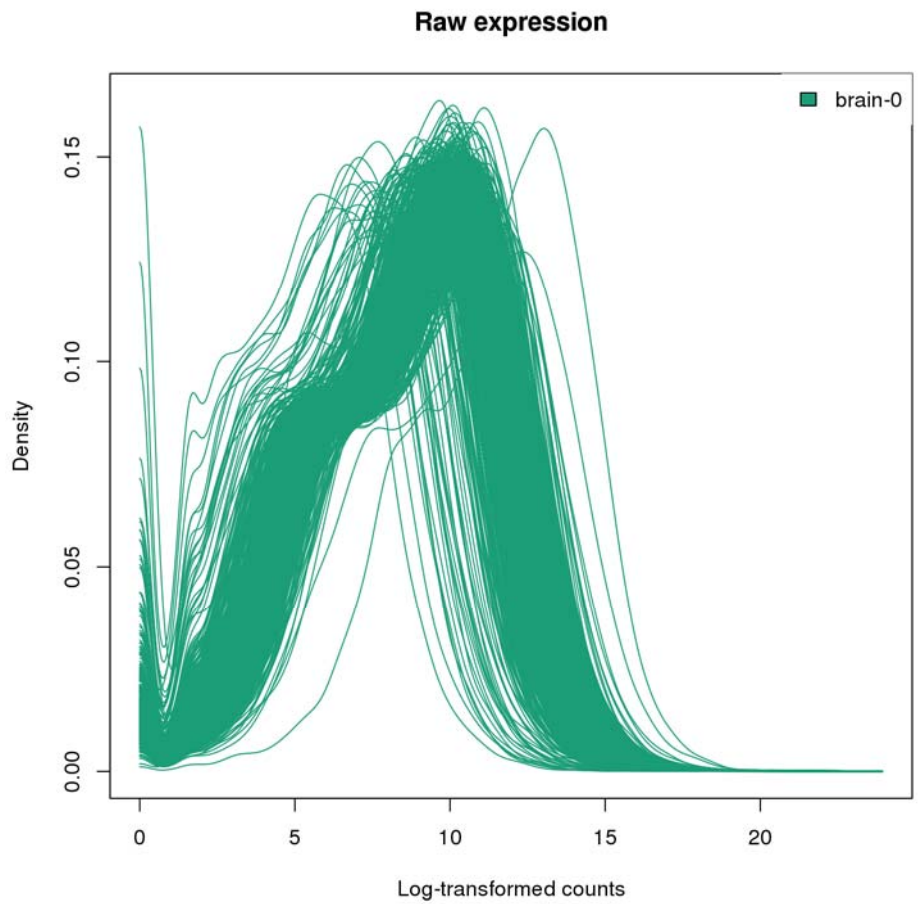
- Performed a sparse-aware filter that allowed us to retain tissue-specific genes.
- Removed genes with <1 CPM in fewer than half of the number of samples of the smallest tissue.



http://tinyurl.com/gtex01

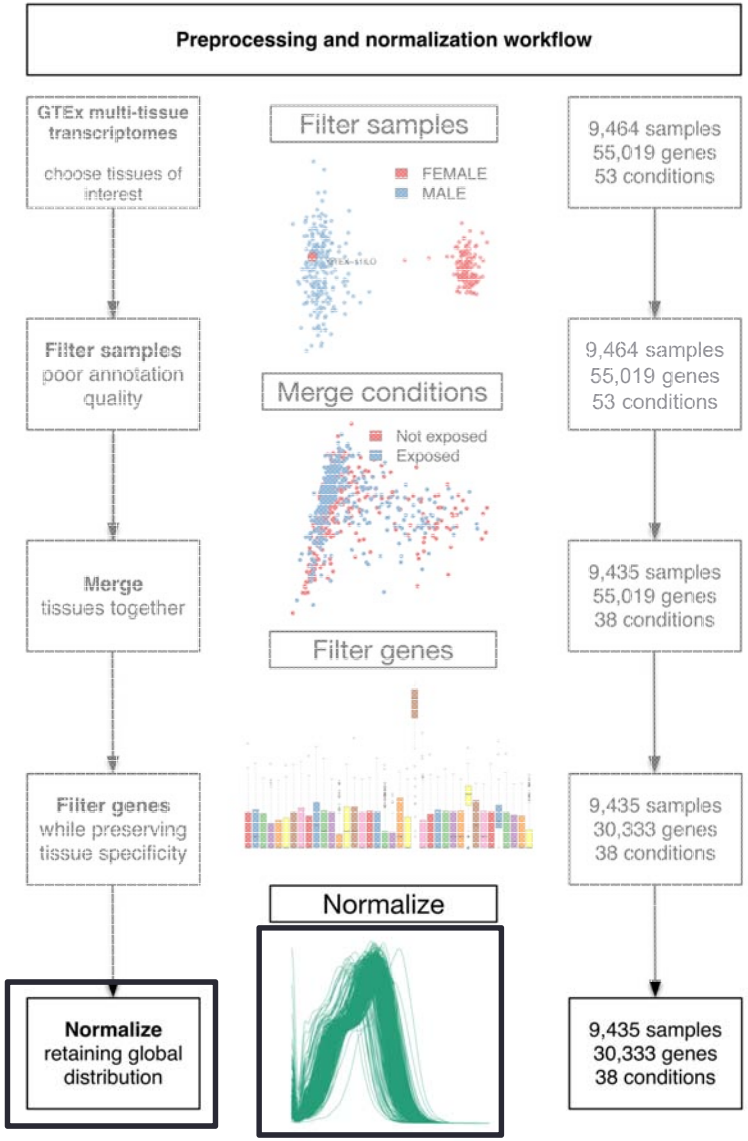
Step 4: Normalize Data





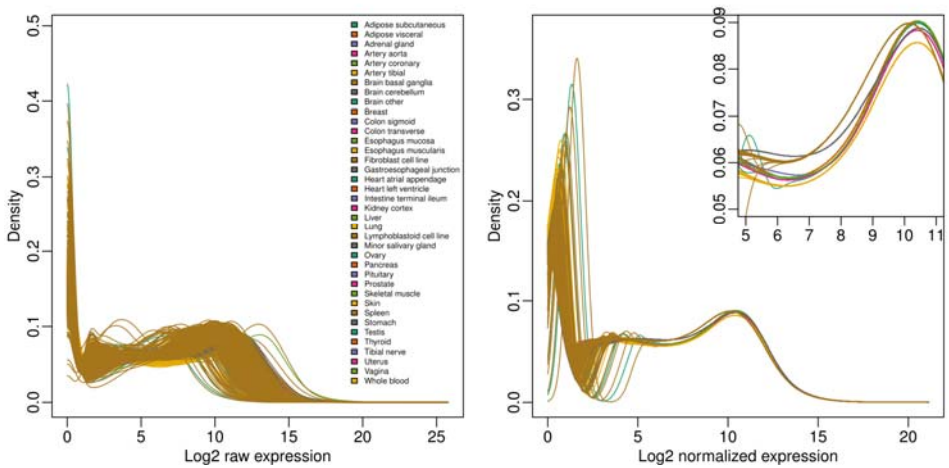
http://tinyurl.com/gtex01

Step 4: Normalize Data



http://tinyurl.com/gtex01

- Ran qsmooth separately on the expression data in each tissue.
- Final data-set included expression information for 30,333 genes across 9,435 samples, representing 38 distinct tissues/body sites.



Question 1:
What is
“Tissue Specificity?”

Integrative Network Inference: PANDA

OPEN ACCESS Freely available online

PLOS | ONE

Passing Messages between Biological Networks to Refine Predicted Interactions

Kimberly Glass[1,2], Curtis Huttenhower[2], John Quackenbush[1,2], Guo-Cheng Yuan[1,2]*

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

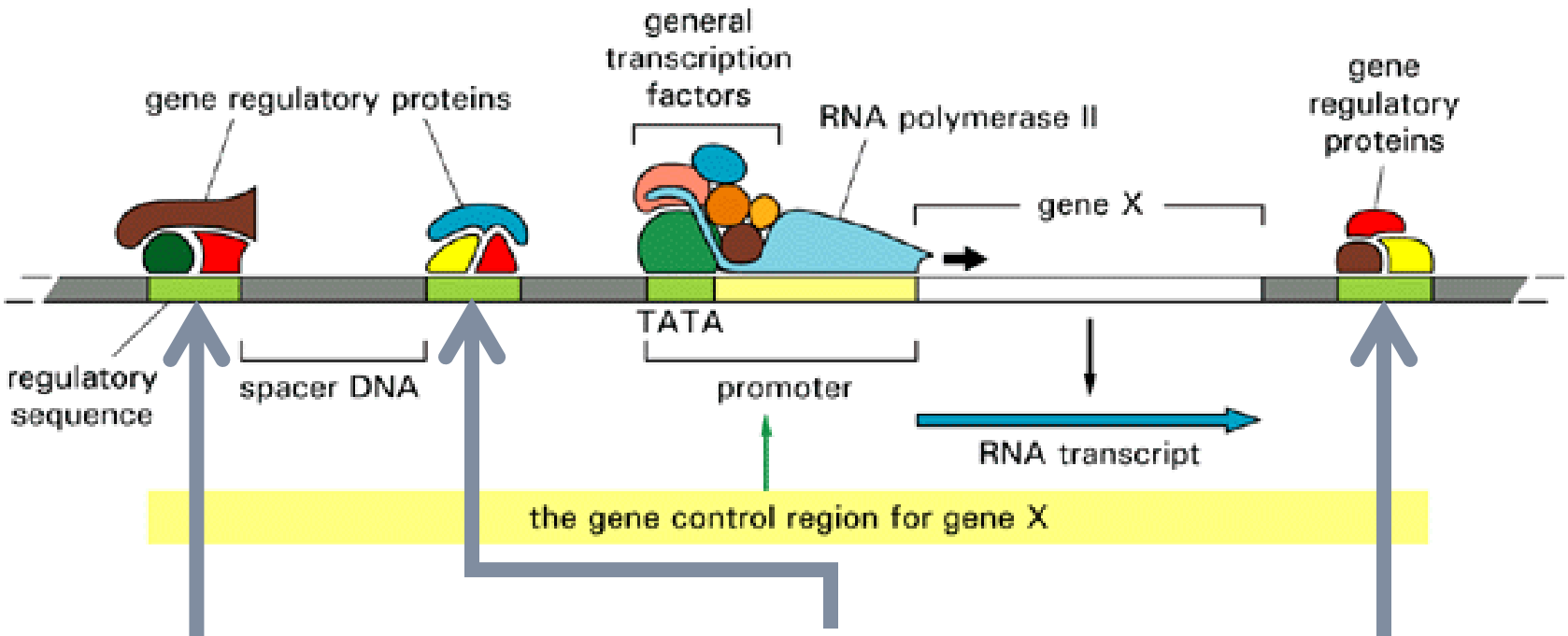
Abstract

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.

Regulation of Transcription

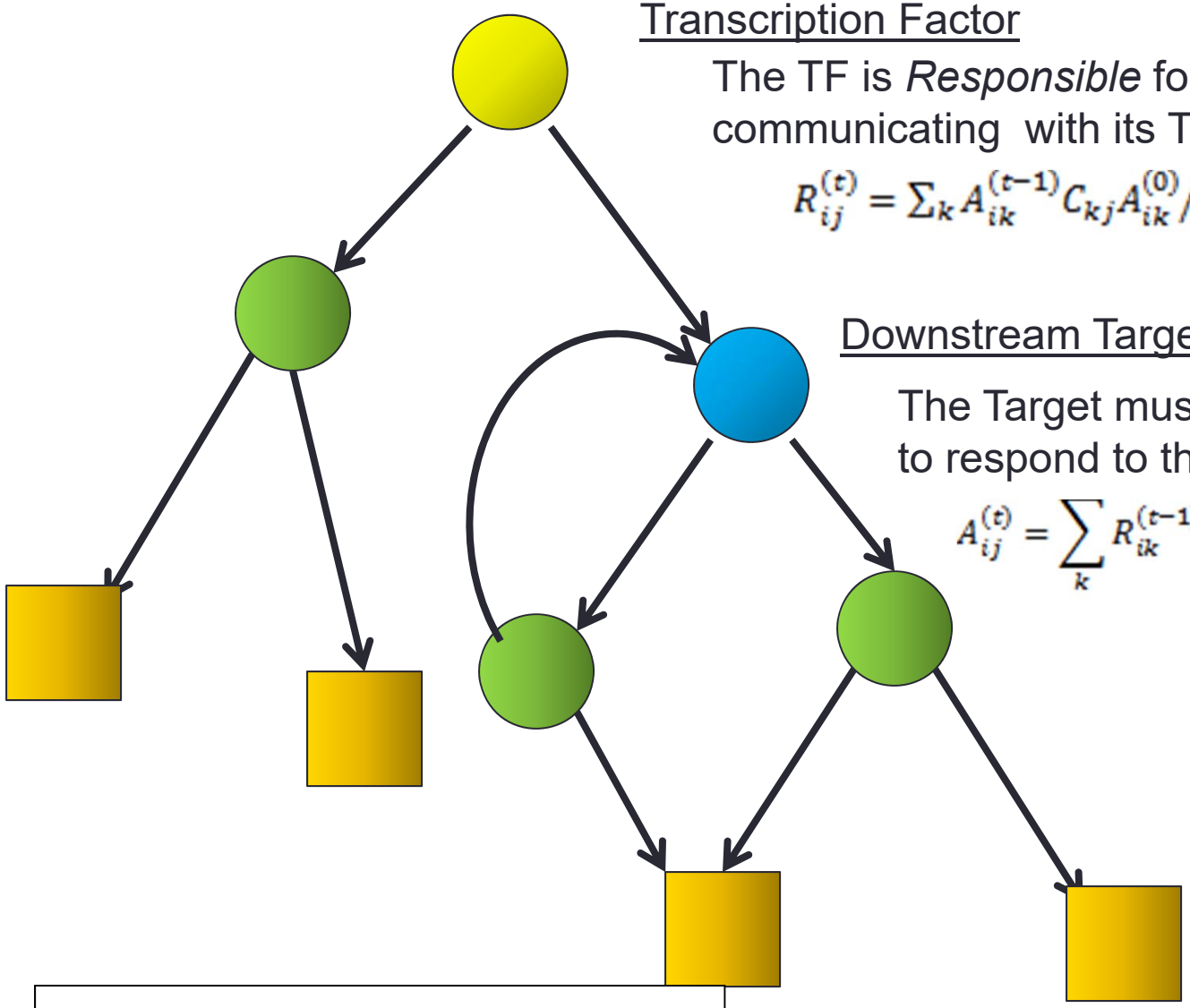
regulatory sequences

promoter



Specific transcription factors

A Simple Idea: Message Passing



Transcription Factor

The TF is *Responsible* for communicating with its Target

$$R_{ij}^{(t)} = \sum_k A_{ik}^{(t-1)} C_{kj} A_{ik}^{(0)} / \sum_k C_{kj} A_{ik}^{(0)}$$

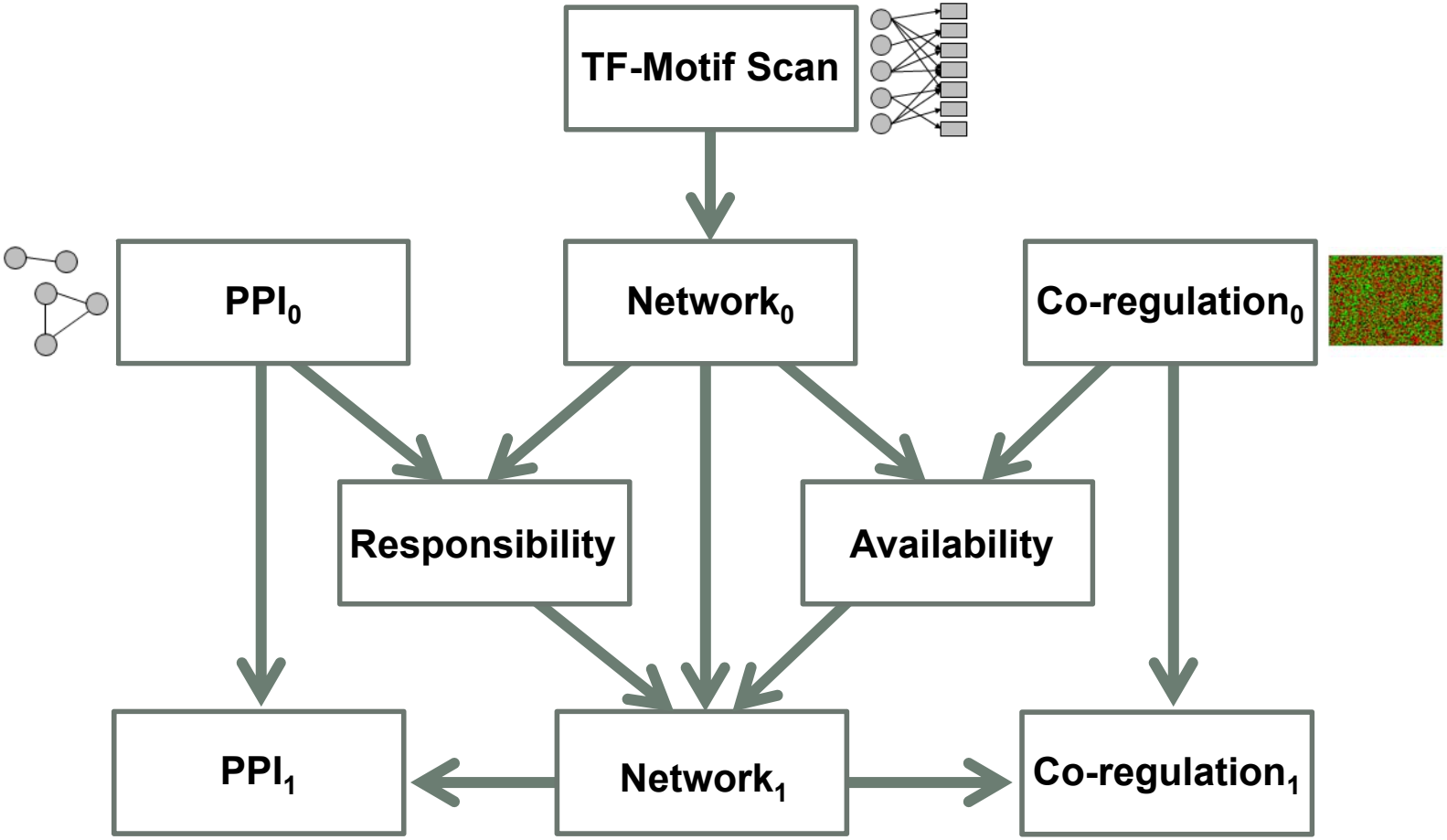
Downstream Target

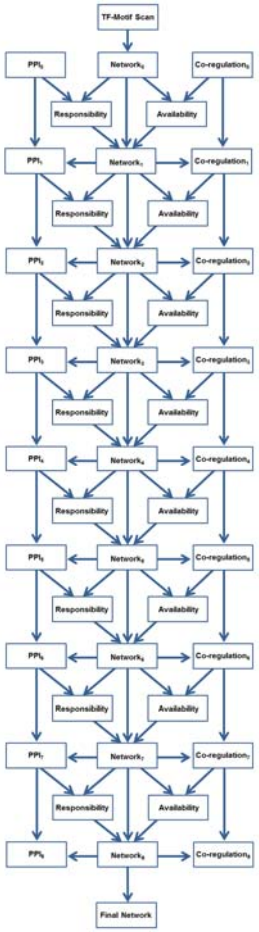
The Target must be *Available* to respond to the TF

$$A_{ij}^{(t)} = \sum_k R_{ik}^{(t-1)} C_{kj} R_{ik}^{(0)} \Big/ \sum_k R_{ik}^{(0)}$$

Kimberly Glass, GC Yuan

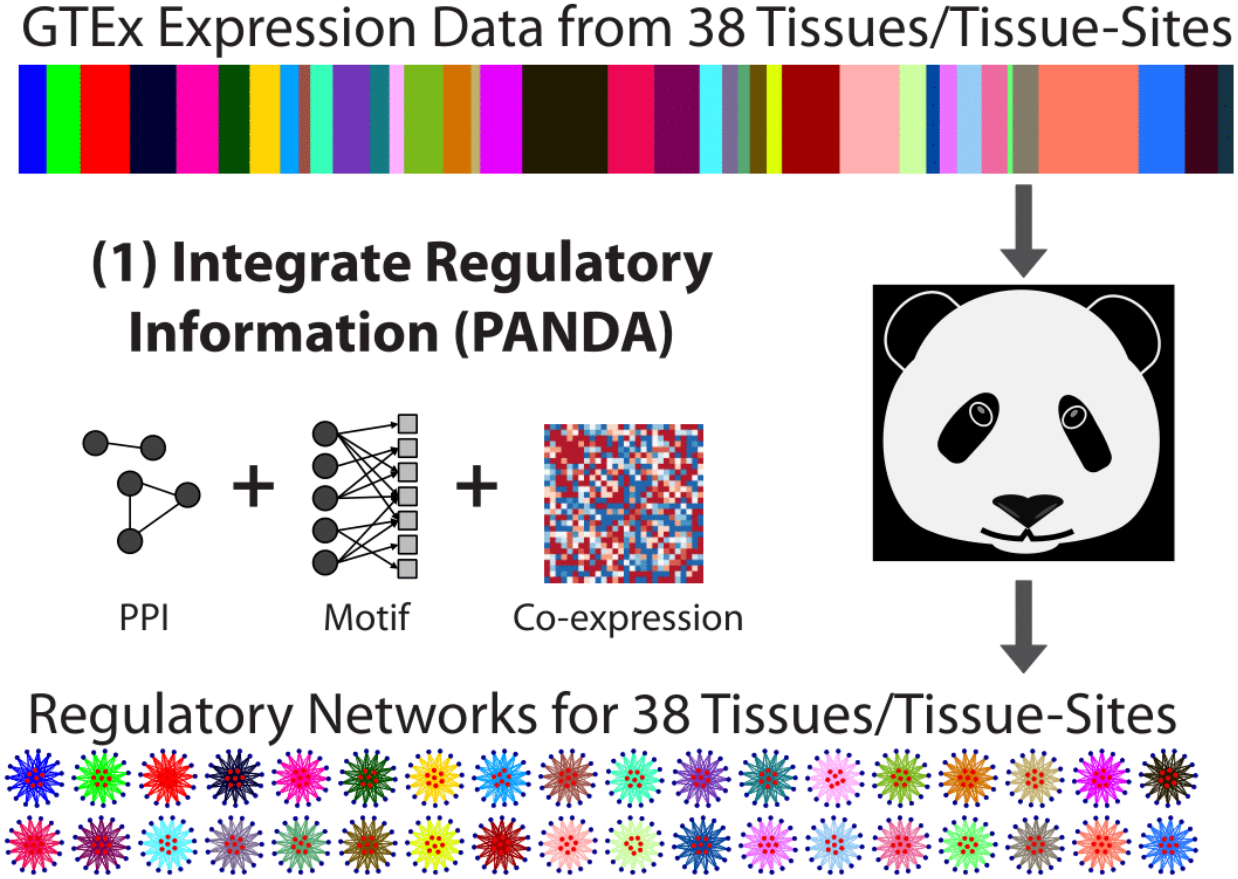
Message-Passing Networks: PANDA





Glass et. al. “Passing Messages Between Biological Networks to Refine Predicted Interactions.” PLoS One. 2013 May 31;8(5):e64832.
Code and related material available on sourceforge: http://sourceforge.net/projects/panda-net/

Reconstructing Tissue-Specific Networks



- **In PANDA, an "edge" is an inferred regulatory association between a TF and a target gene.**
- **The "edge weight" represents the "strength" of the inferred regulatory interaction.**
- **Drawing a network = setting an edge weight threshold.**

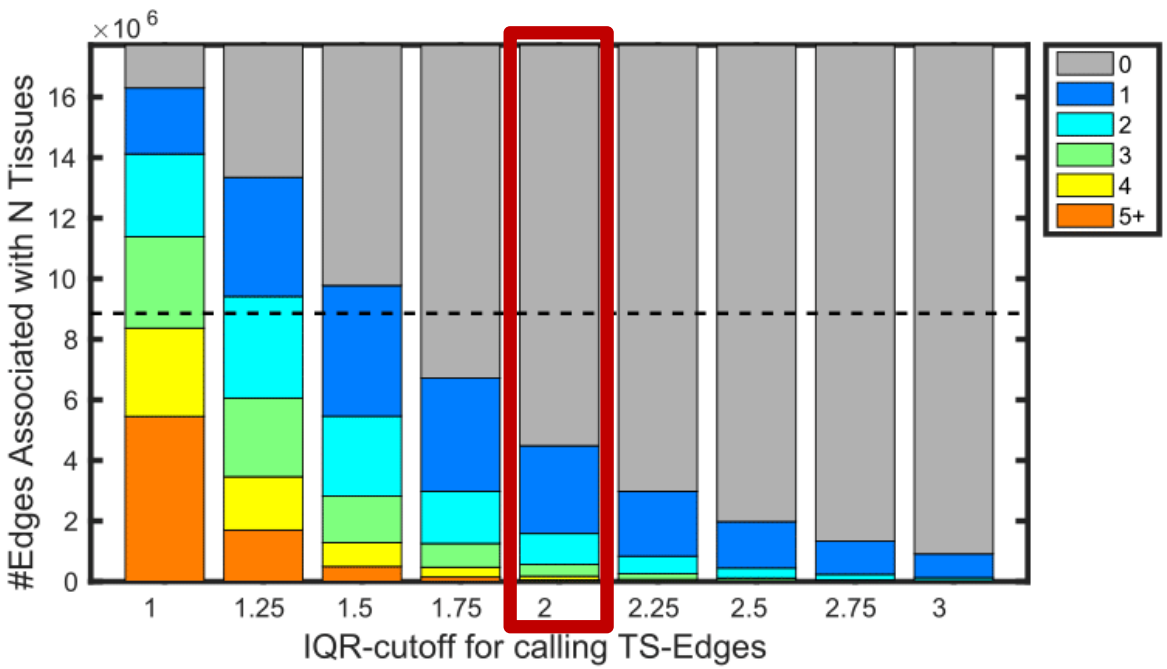
http://tinyurl.com/gtex05

Identifying Tissue-Specific Edges

We compared the weight of an edge between a transcription factor (*i*) and a gene (*j*) in a particular tissue (*t*) to the median and interquartile range (IQR) of its weight across all 38 tissues:

$$s_{ij}^{(t)} = \left(w_{ij}^{(t)} - med\left(w_{ij}^{(all)}\right)\right) / IQR\left(w_{ij}^{(all)}\right)$$

An edge with an edge specificity score $s_{ij}^{(t)} > N$ was identified as specific to tissue *t*.



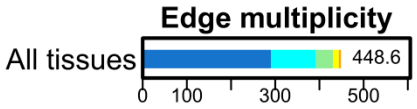
Edge "multiplicity"

the number of tissues in which a edge is identified as specific

$$m_{ij} = \sum_{t} \left[s_{ij}^{(t)} > N\right]$$

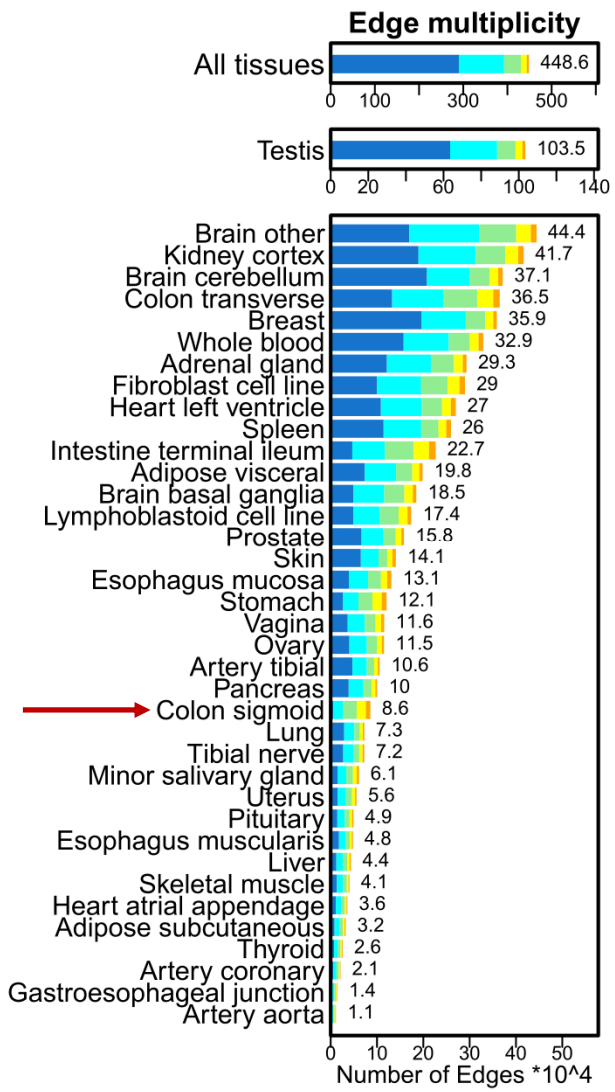
http://tinyurl.com/gtex05

Interpreting Edge Multiplicity

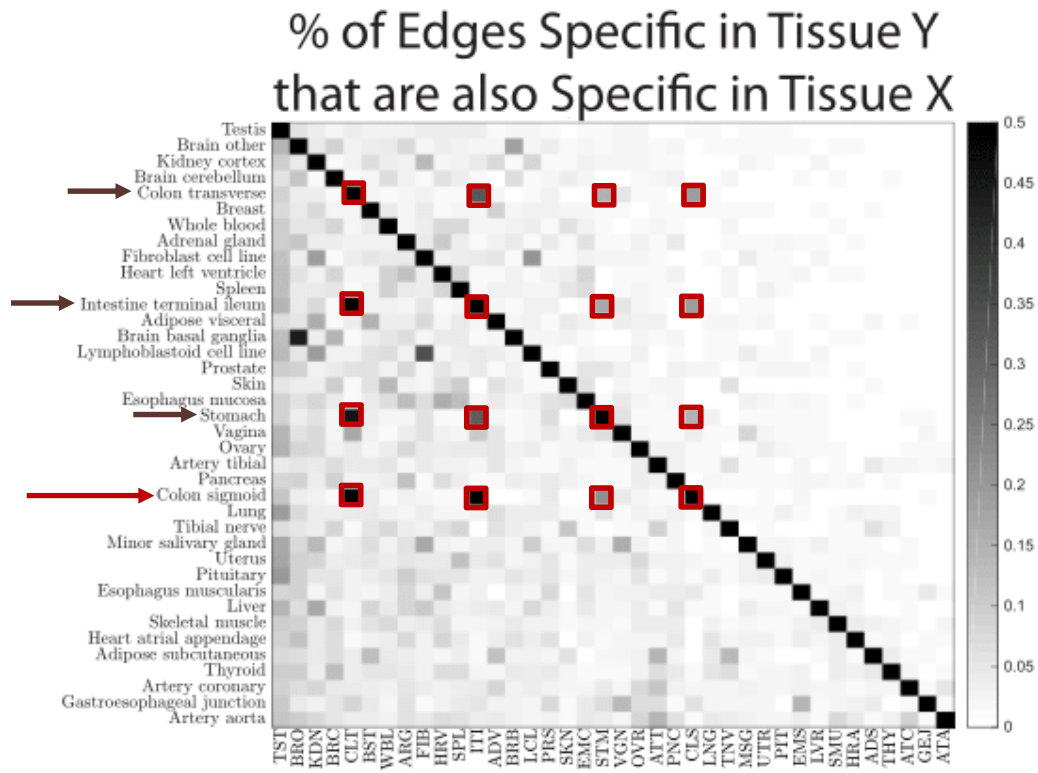


http://tinyurl.com/gtex05

Interpreting Edge Multiplicity



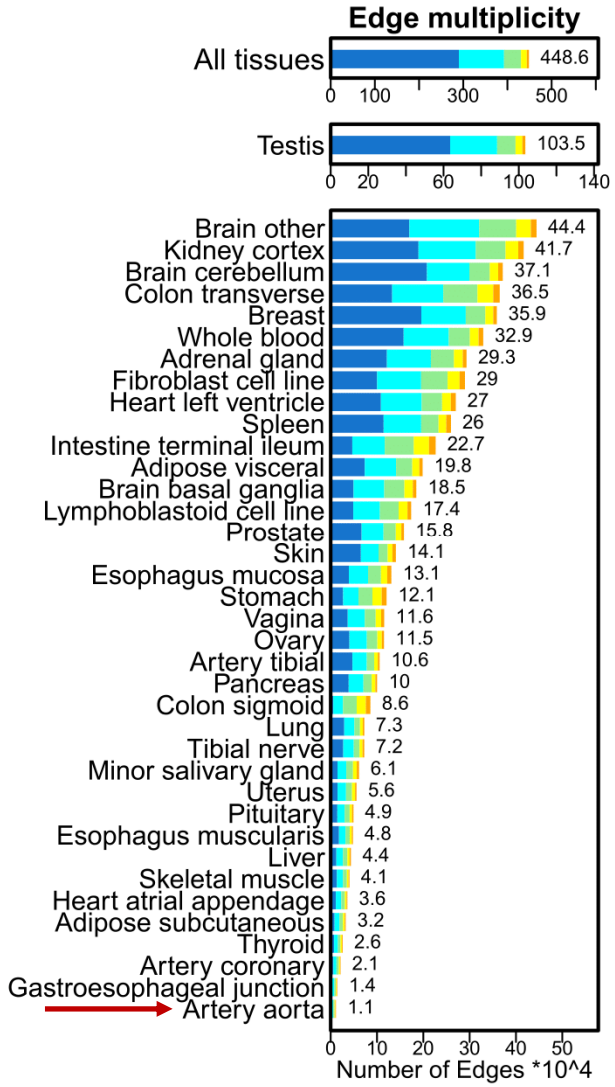
http://tinyurl.com/gtex05

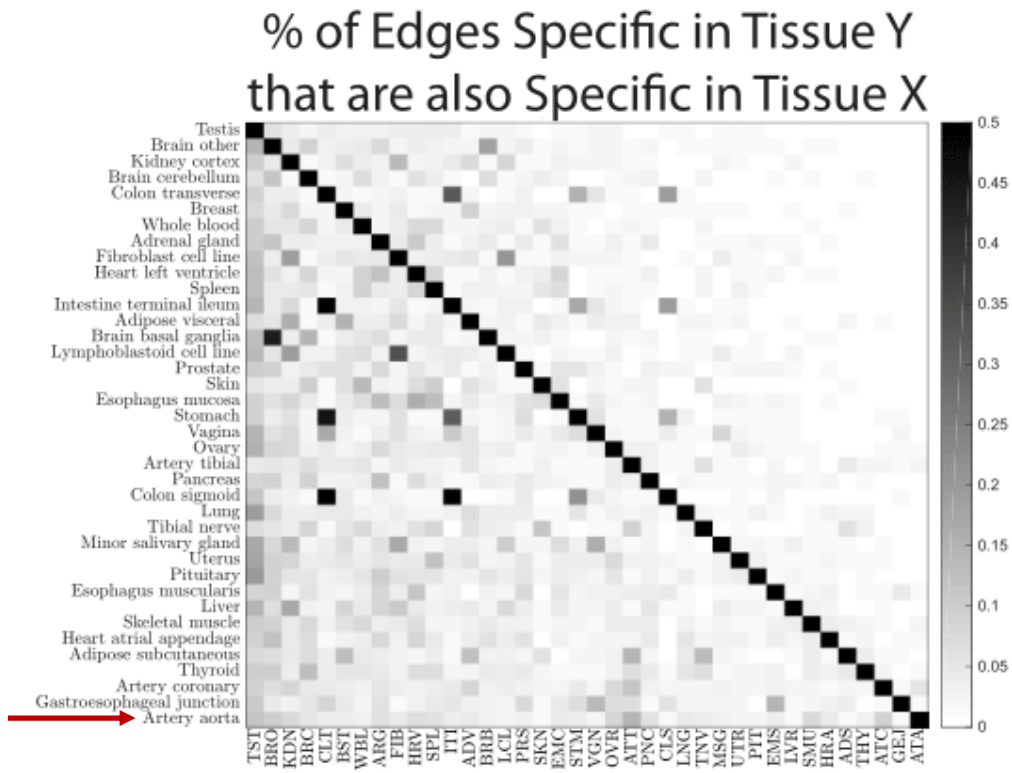


93.4% of sigmoid colon edges have multiplicity>1

These edges are shared with the transverse colon (82.1%), small intestine (51%), and stomach (21.4%)

Interpreting Edge Multiplicity





78.7% of aorta edges have multiplicity>1

These edges are shared across <u>many</u> tissues, including the tibial artery (14.8%), testis (11.2%), brain (9.8%), coronary artery (9.3%), and ovary (8.5%)

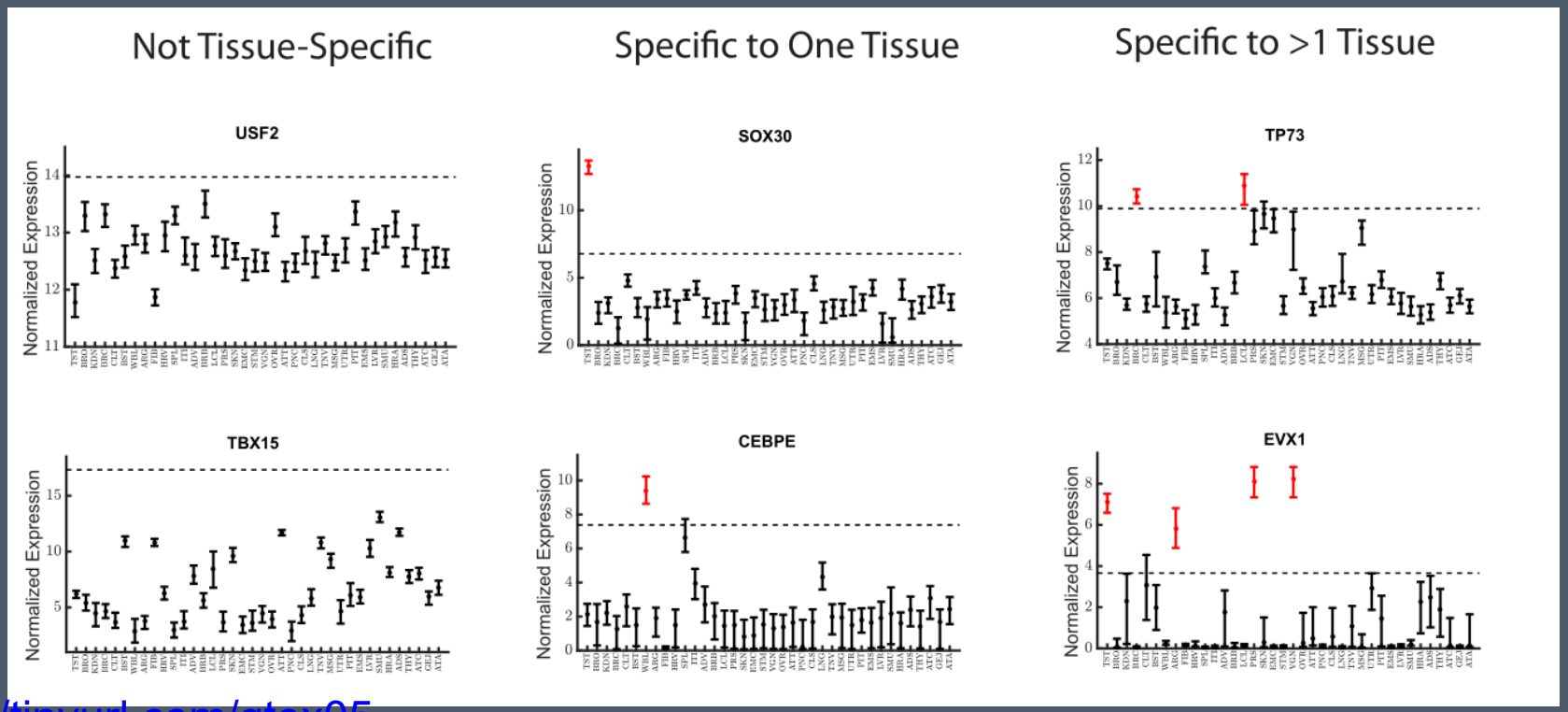
http://tinyurl.com/gtex05

Defining Node Tissue-Specificity

We compared the median expression level ($e_j^{(t)}$) of a gene (*j*) in a particular tissue (*t*), to the median and interquartile range of its expression across all samples:

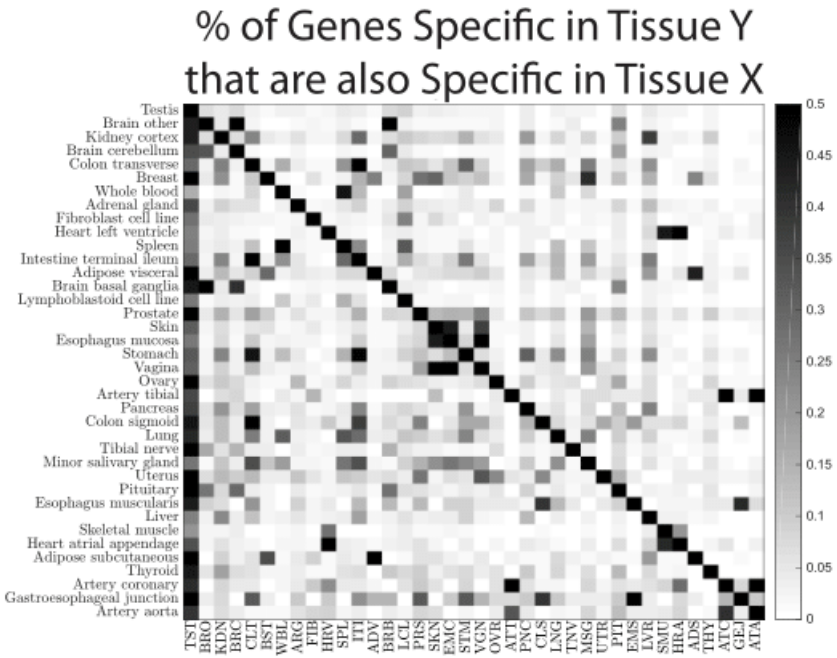
$$s_j^{(t)} = \left(med\left(e_j^{(t)}\right) - med\left(e_j^{(all)}\right)\right) / IQR\left(e_j^{(all)}\right)$$

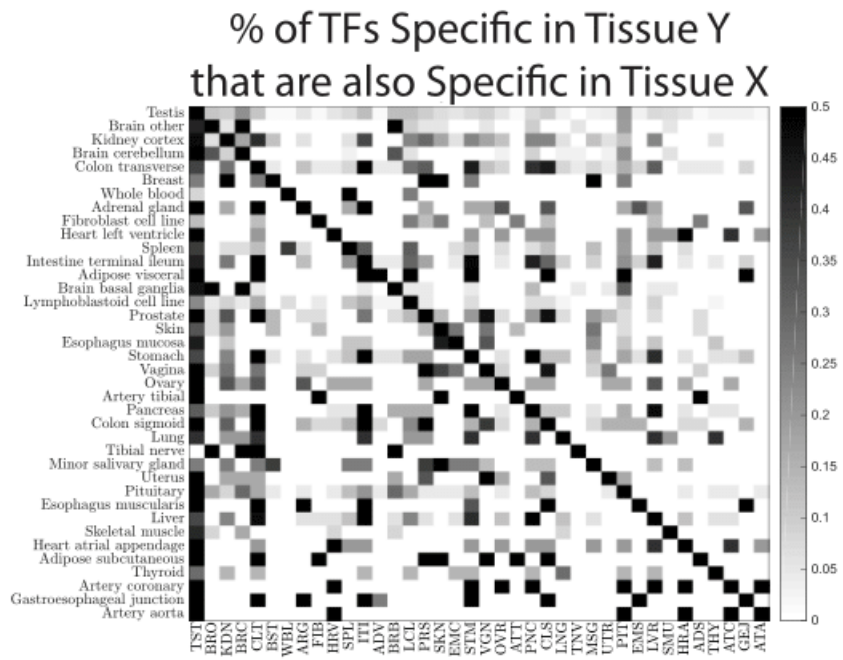
A gene with a specificity score $s_j^{(t)} > N$ was identified as specific to tissue *t*.



http://tinyurl.com/gtex05

Complex Patterns of Shared-Specificity





The amount of shared specificity among genes and transcription factors is much more complex than that of the edges.

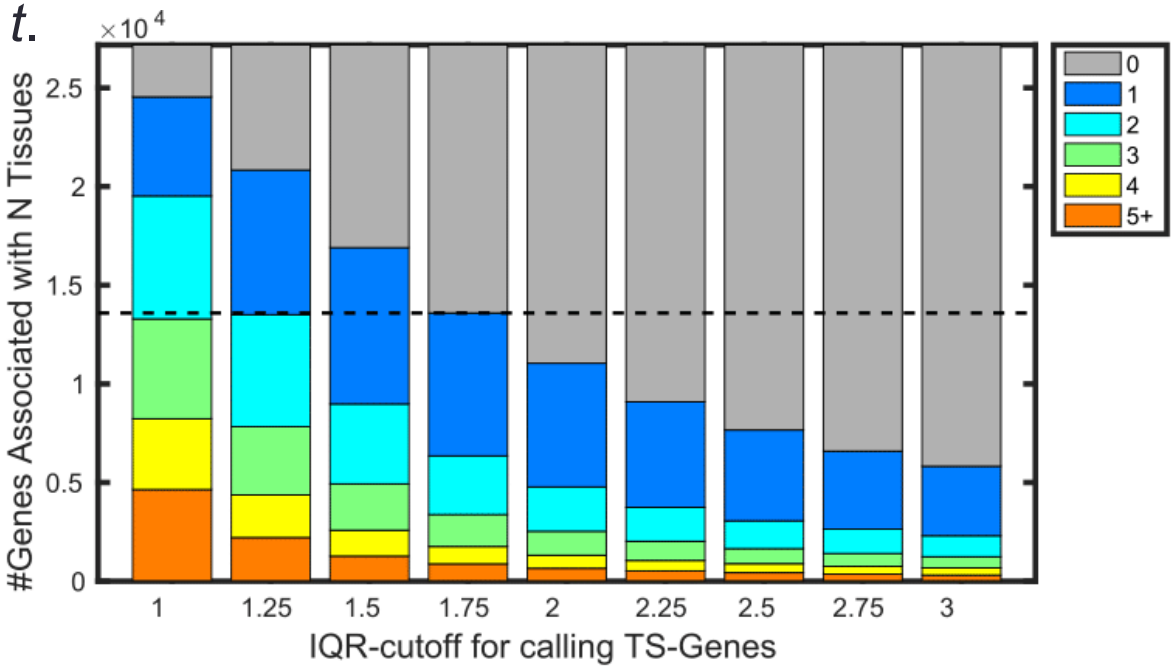
http://tinyurl.com/gtex05

Defining Tissue-Specificity

We compared the median expression level ($e_j^{(t)}$) of a gene (*j*) in a particular tissue (*t*), to the median and interquartile range of its expression across all samples:

$$s_j^{(t)} = \left(med\left(e_j^{(t)}\right) - med\left(e_j^{(all)}\right)\right) / IQR\left(e_j^{(all)}\right)$$

A gene with a specificity score $s_j^{(t)} > N$ was identified as specific to tissue *t*.

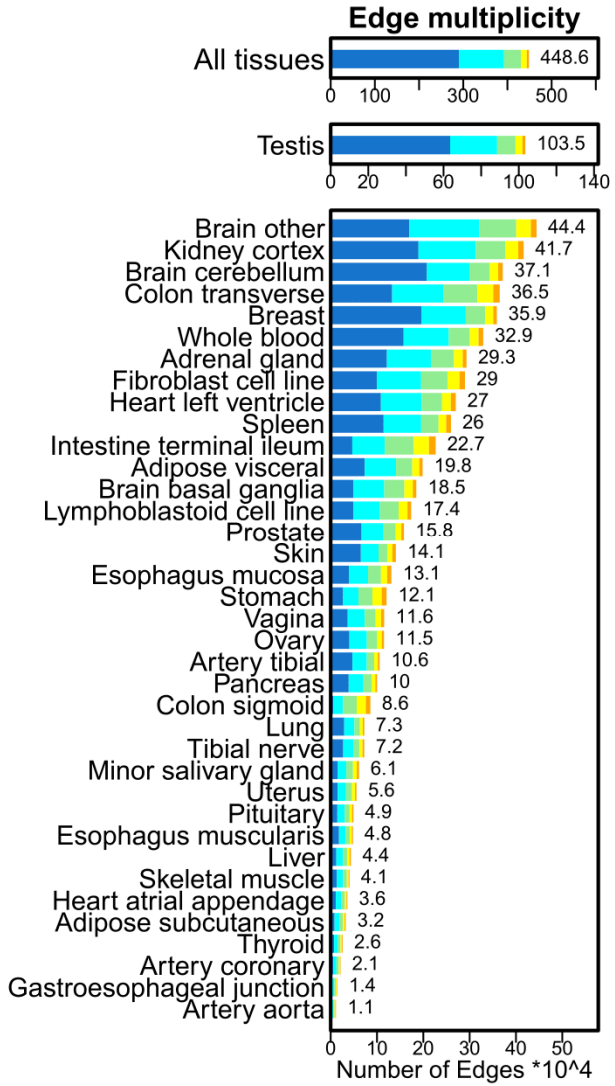


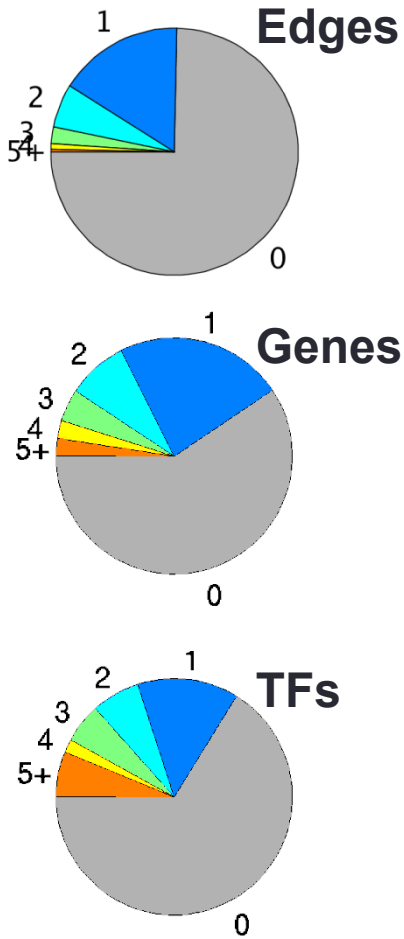
Gene "multiplicity"
the number of tissues in which a gene is identified as specific

$$m_j = \sum_t \left[s_j^{(t)} > N\right]$$

http://tinyurl.com/gtex05

Edge vs Node Multiplicity

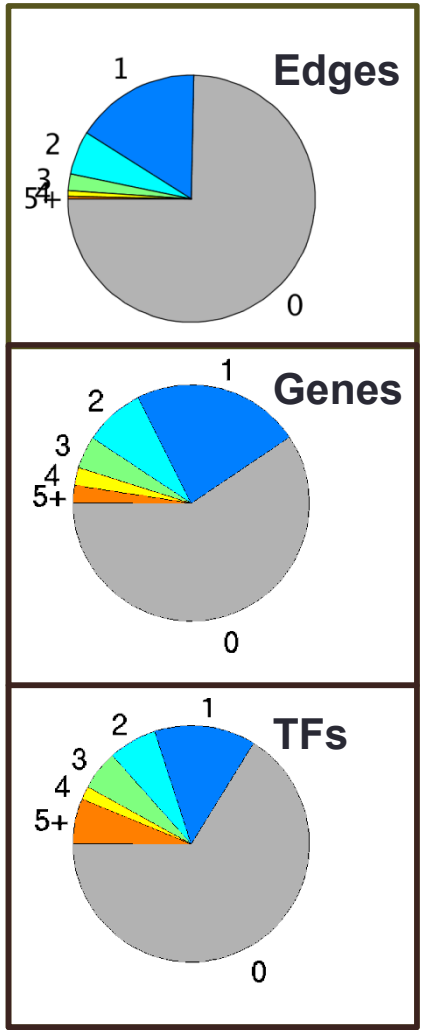




http://tinyurl.com/gtex05

Some Take-Home Messages

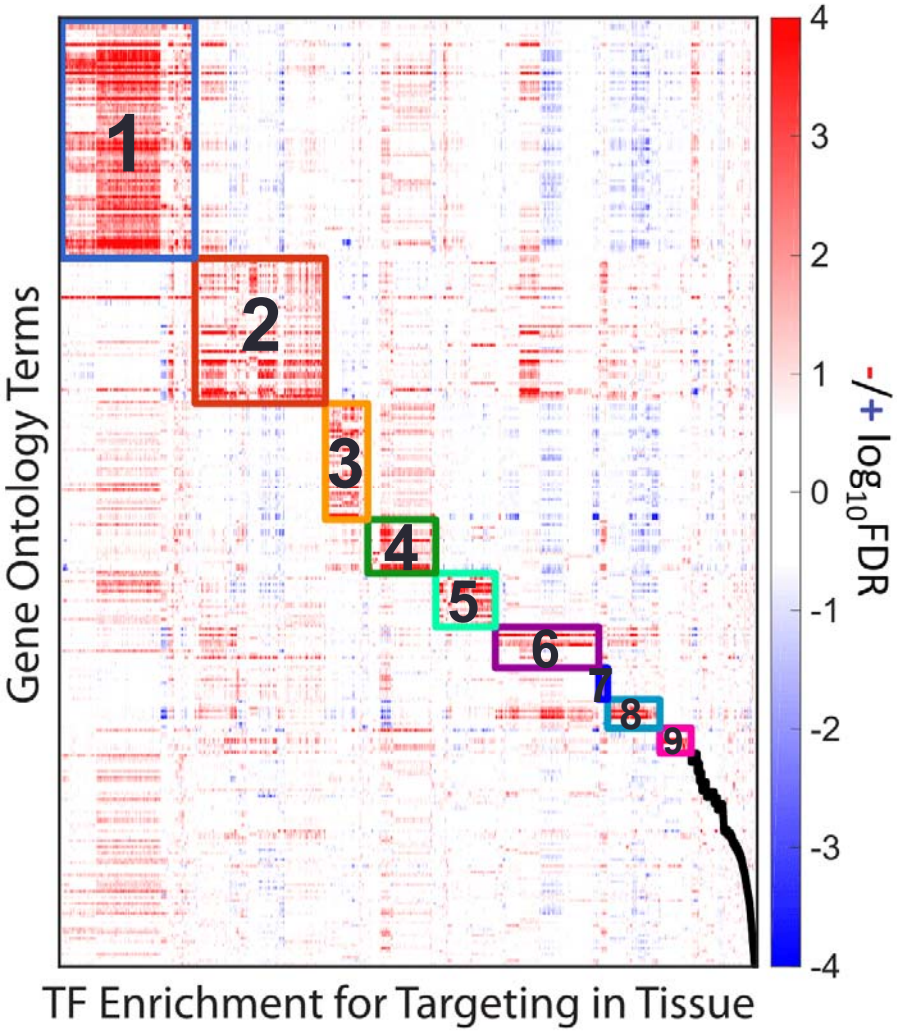
- Network edges are more likely to be identified as specific to only a single tissue (multiplicity=1) compared to genes.
- Transcription factors are more likely to be identified as “specific” in multiple different tissues (multiplicity>1) compared to genes.
- A smaller percentage of transcription factors are identified as tissue-specific than genes.



Tissue-specific regulation is likely caused by changes in regulation and regulatory edges

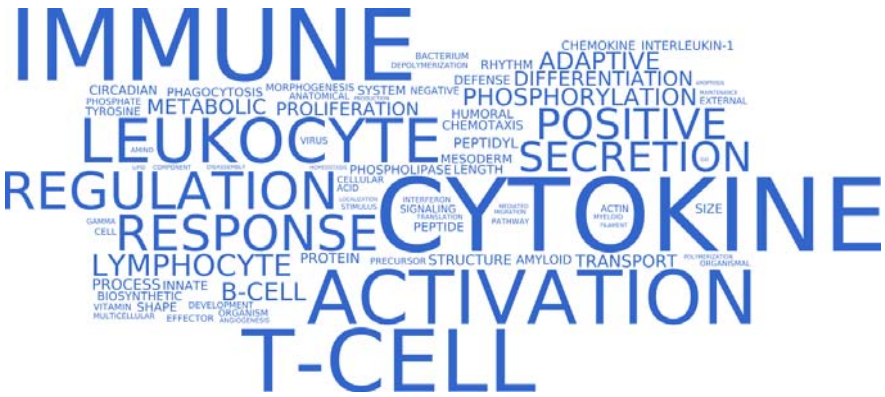
Structure of Tissue-Specific Regulation

Look at the targets of TS edges and map these gene targets to function.

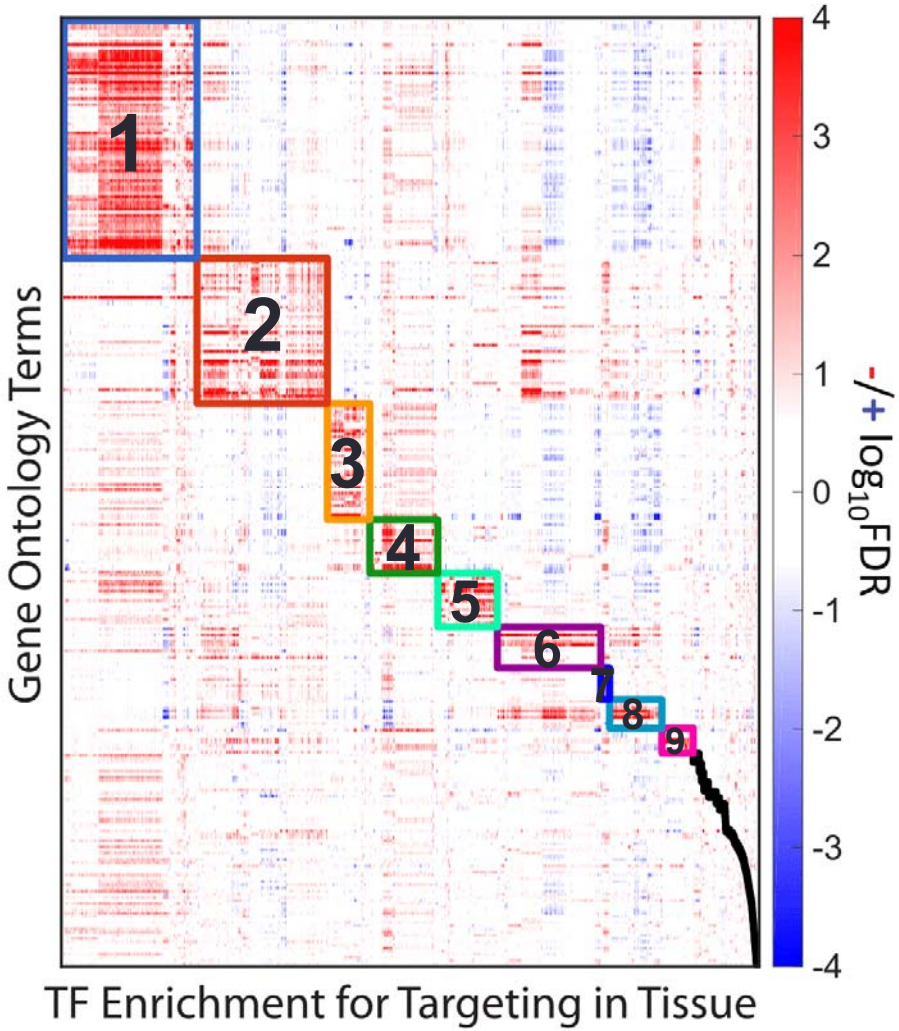


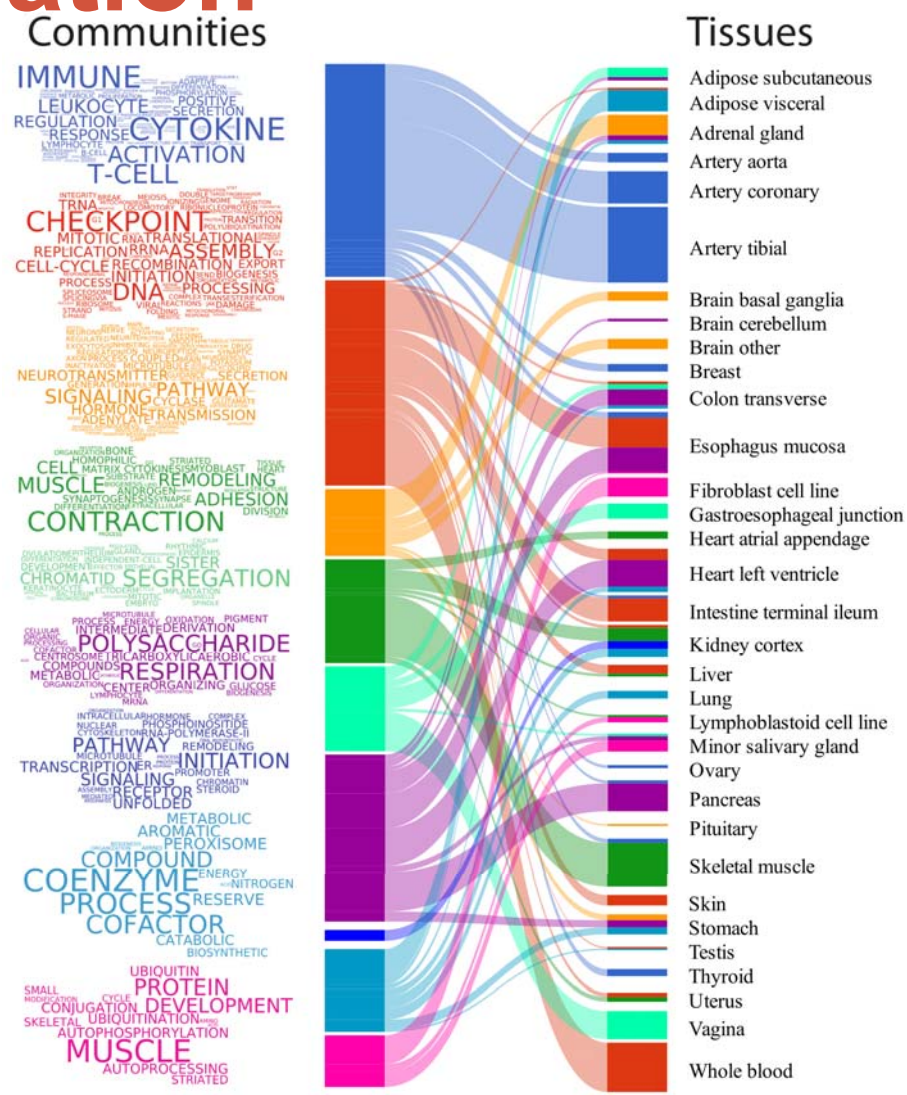
ID	Pathways	Tissues	Top Category (most significant associations in community)
1	75	22	IMMUNE RESPONSE
2	46	27	RIBONUCLEOPROTEIN COMPLEX BIOGENESIS AND ASSEMBLY
3	37	13	SYNAPTIC TRANSMISSION
4	17	17	SYNAPTOGENESIS
5	17	19	EPIDERMIS DEVELOPMENT
6	13	19	CELLULAR RESPIRATION
7	10	6	CHROMATIN REMODELING
8	9	16	COFACTOR METABOLIC_PROCESS
9	8	7	PROTEIN MODIFICATION BY SMALL PROTEIN CONJUGATION

Word Cloud of Pathways in Community 1



Structure of Tissue-Specific Regulation

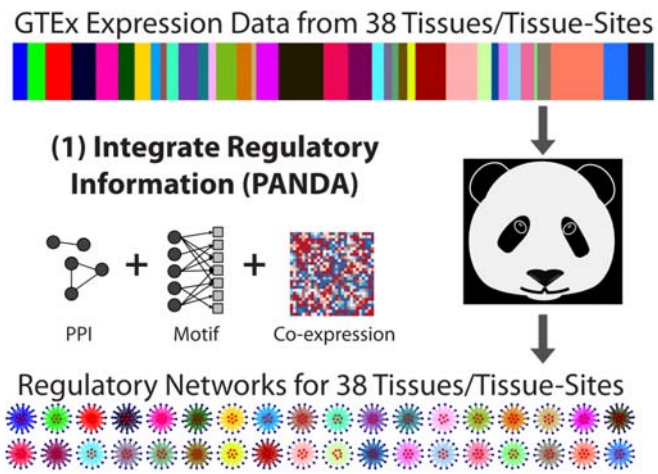


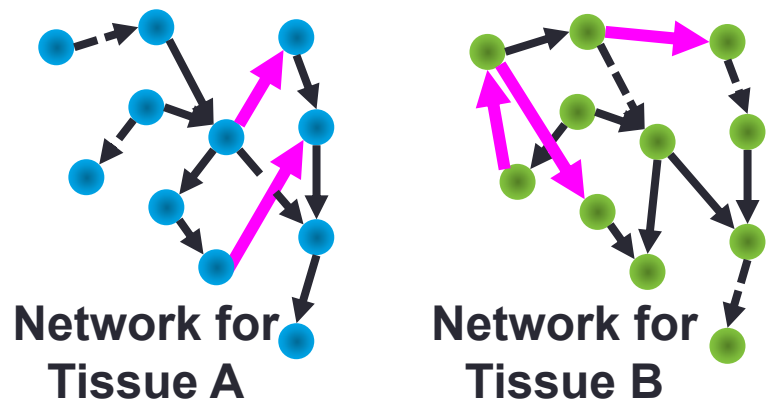


width = number of TFs

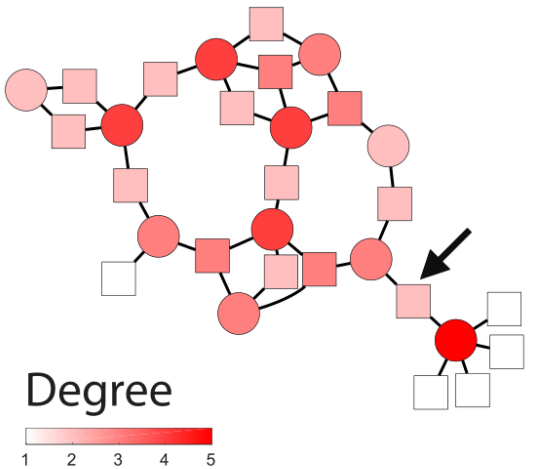
How is Tissue-Specificity Maintained?

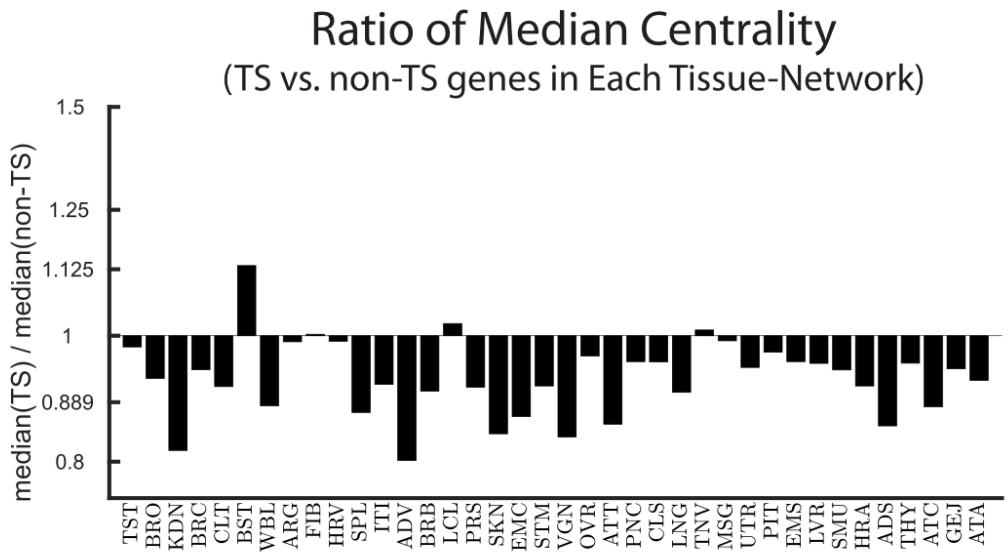
- Our network analysis so-far has focused on tissue-specific edges (or edges that appear in one or a few tissues).
- These edges are part of a larger network that also includes common edges, which exist across all tissues.
- We wanted to know where tissue-specific genes are located within this larger network context.





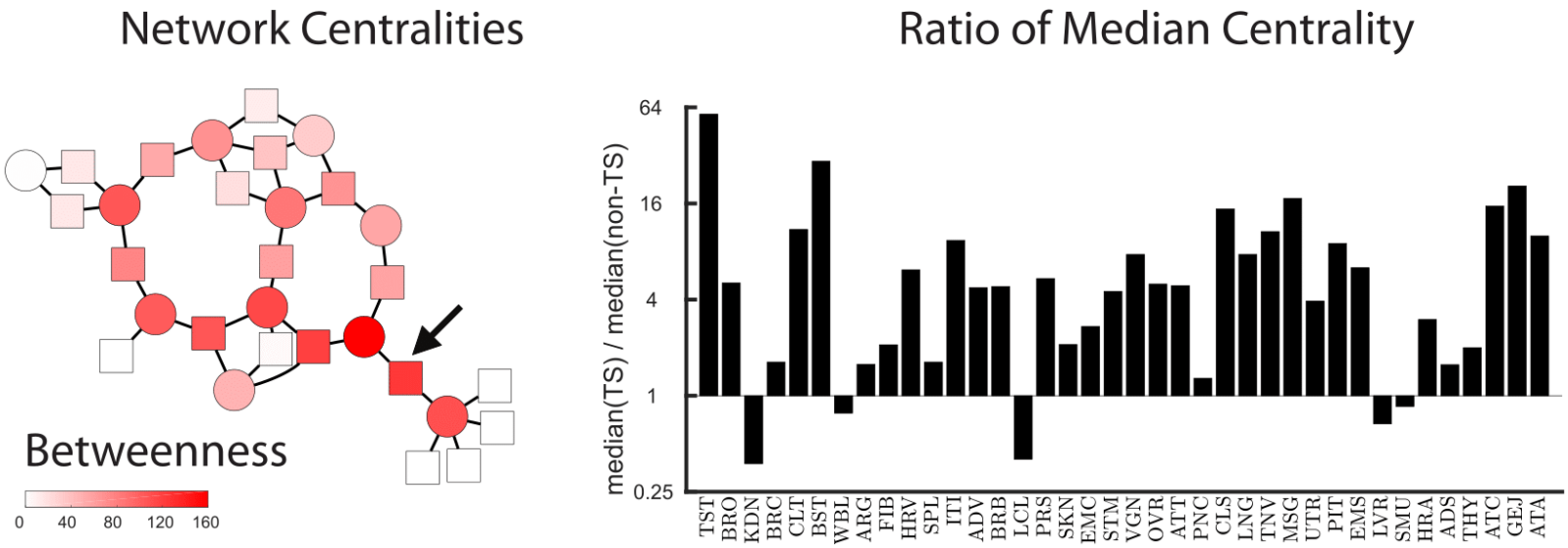
Degree of Tissue-Specific Genes





- Calculated the degree of genes (number of edges connected to a gene) in each of the 38 PANDA-predicted networks.
- Compared the degree of tissue-specific genes to the degree of non-tissue-specific genes in each of these networks.
- Tissue-specific genes are depleted for regulatory edges.

Betweenness of Tissue-Specific Genes



- Calculated the <u>betweenness</u> of genes (number of paths through a gene) in each of the 38 PANDA-predicted network.
- Compared the betweenness of tissue-specific genes to non-tissue-specific genes in each of these networks.
- Tissue-specific genes are <u>enriched</u> for regulatory paths.

What we discover about Tissue-Specific Regulation

- “Universal” (high multiplicity) genes are concentrated in the center of the network and have high degree and high betweenness.
- “Tissue-Specific” (low multiplicity, TS) genes have intermediate betweenness
- This means that TS genes exist toward the edge of the network
- However, TS genes increase their betweenness in their “target” tissues, meaning that as they increase their expression, there are more regulatory paths running through them.

Question 2:
Can we solve the
“GWAS Puzzle?”

ARTICLES

nature genetics

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/β-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

ARTICLE

doi:10.1038/nature14177

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

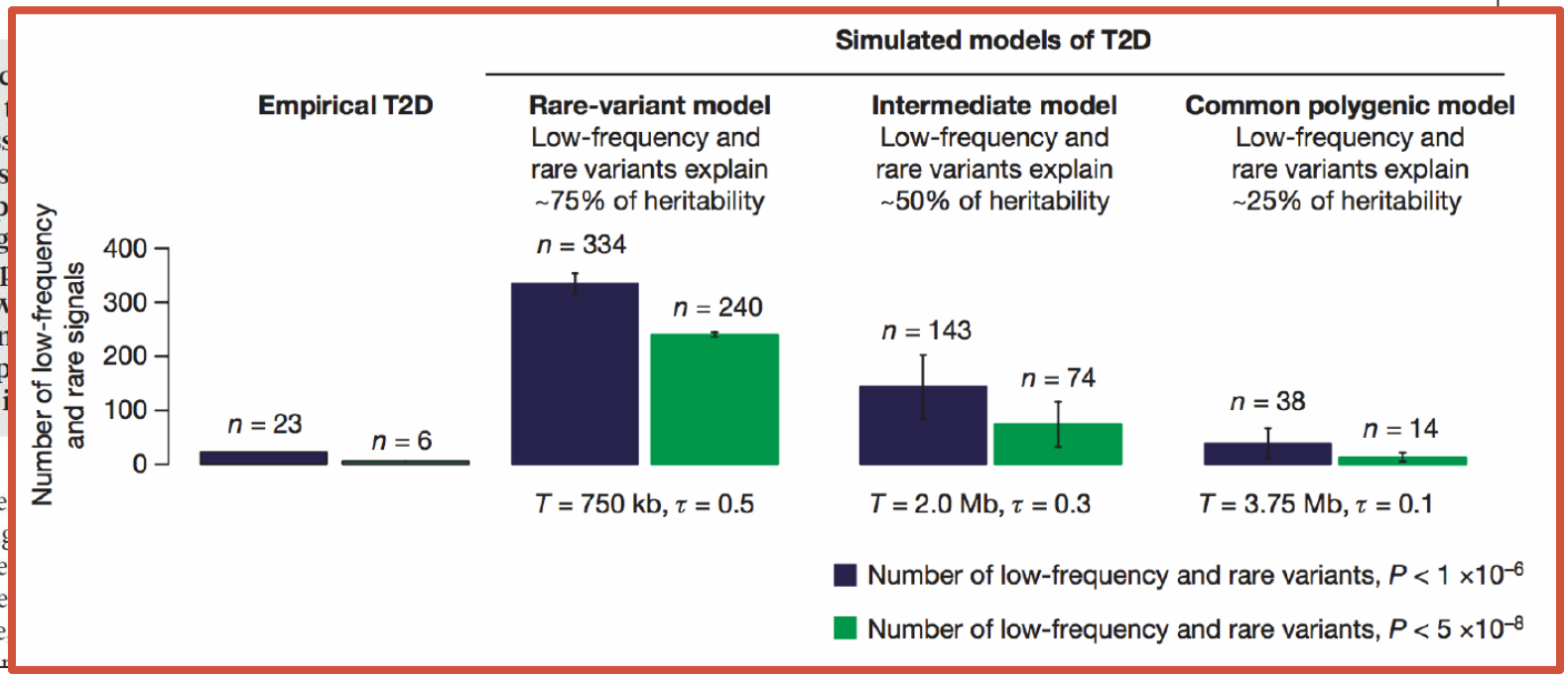
doi:10.1038/nature18642

The genetic architecture of type 2 diabetes

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

The genetic
contribute t
variants ass
Here, to tes
consortia p
sequencing
via genotyp
were overw
Comprehen
to disease p
major role i

There is compe
(T2D) is strong
terizing the spe
the ability to pe
the past decade
samples, progr



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

eQTL Analysis

Use genome-wide data on genetic variants (SNPs = Single Nucleotide Polymorphisms) and gene expression data together

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?

John Platig

eQTL Networks: A simple idea

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

- Perform a “standard eQTL” analysis:

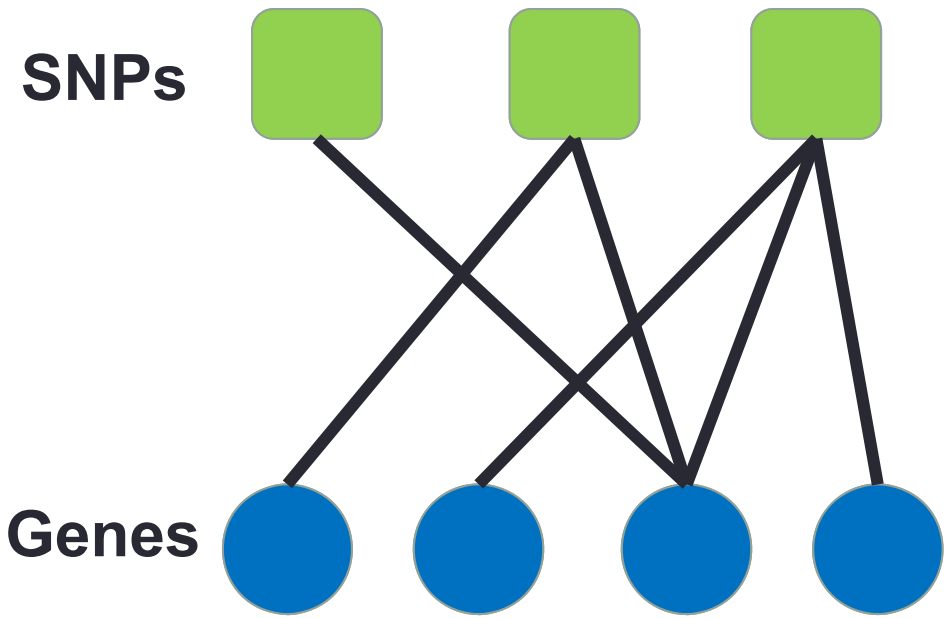
$$Y = \beta_0 + \beta_1 ADD + \varepsilon$$

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

John Platig

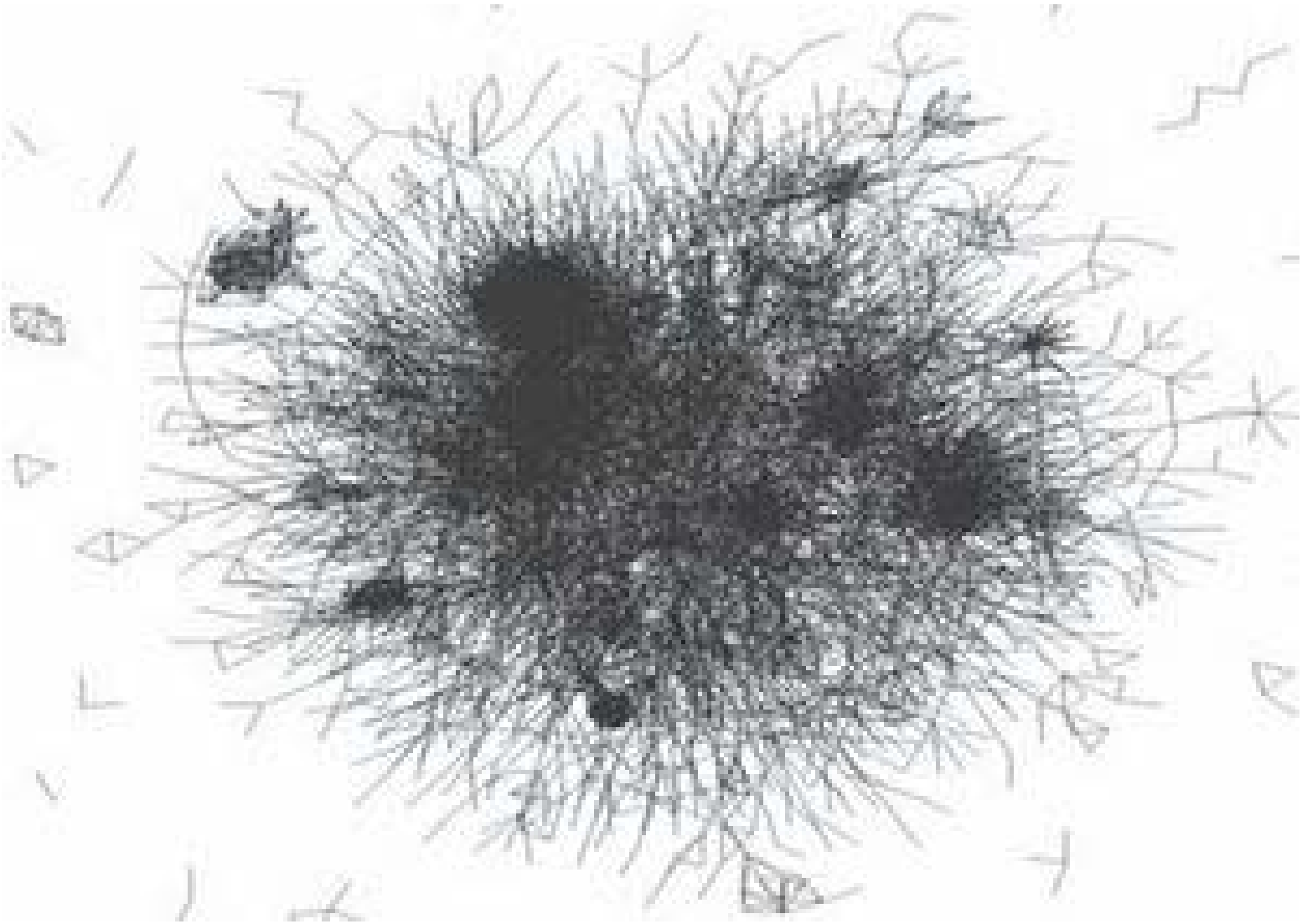
eQTL Networks: A simple idea

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



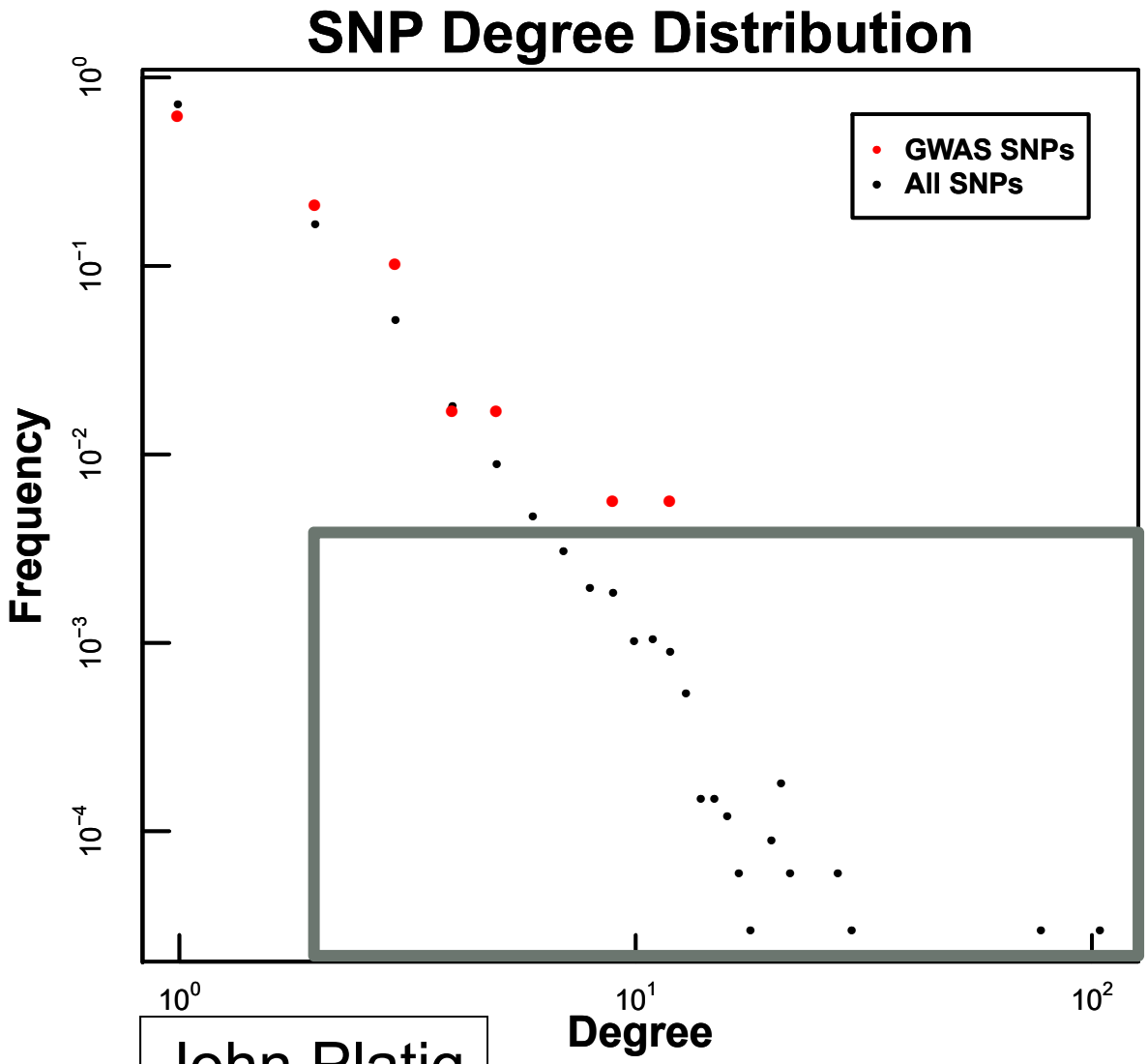
Can a network of SNP-gene associations (*cis* and *trans*) inform the functional roles of these SNPs?

The Result: A Hairball



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

What about GWAS SNPs?



The “hubs” are a GWAS desert!

John Platig

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

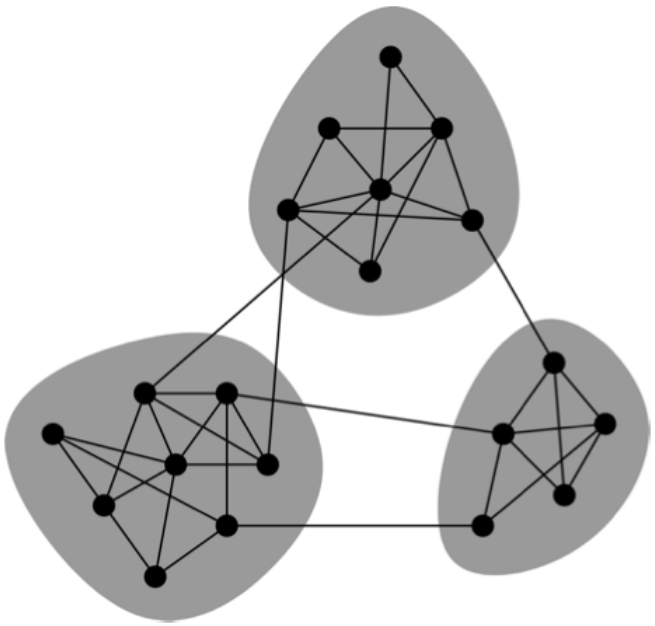
Communities are groups of highly intra-connected nodes

- 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

$$Q = \sum_i (e_{ii} - a_i^2)$$

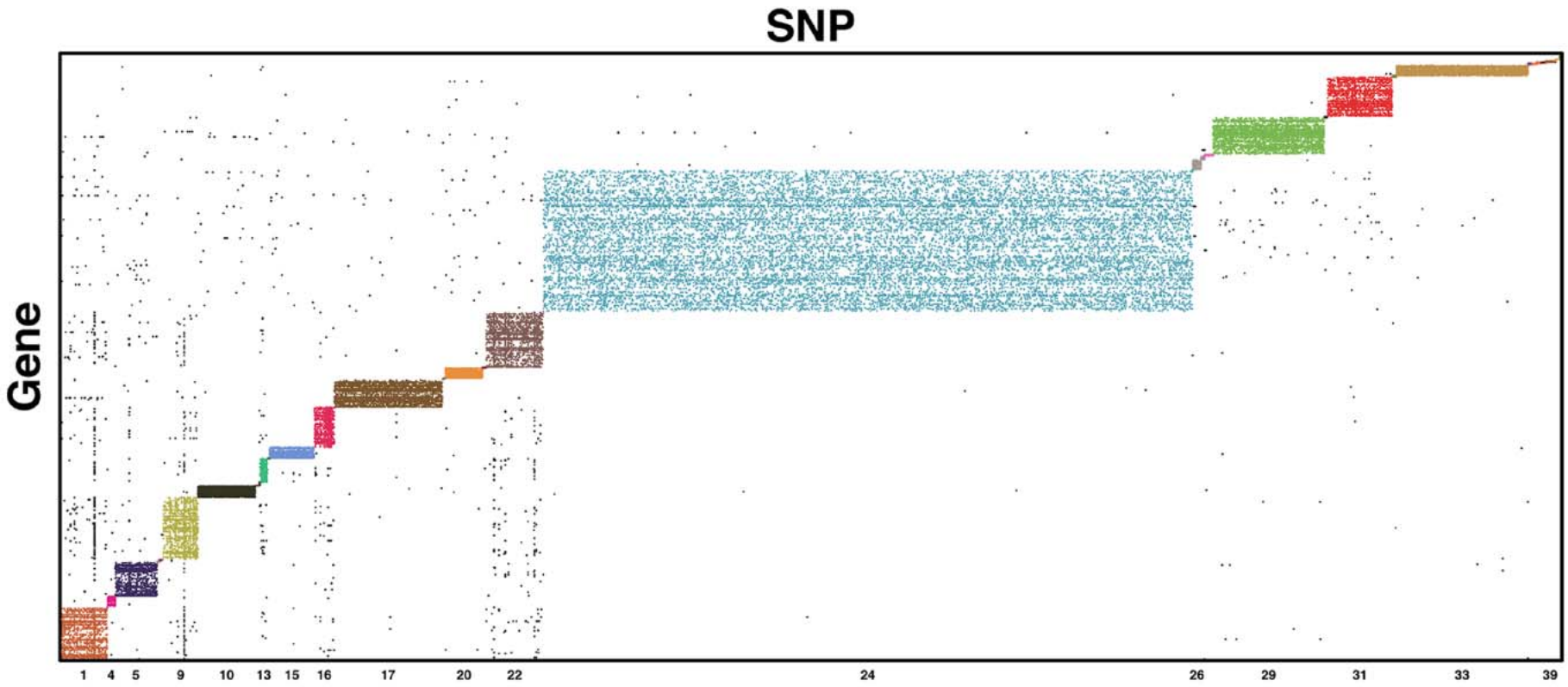
Fraction of network links in community i

Fraction of links expected by chance



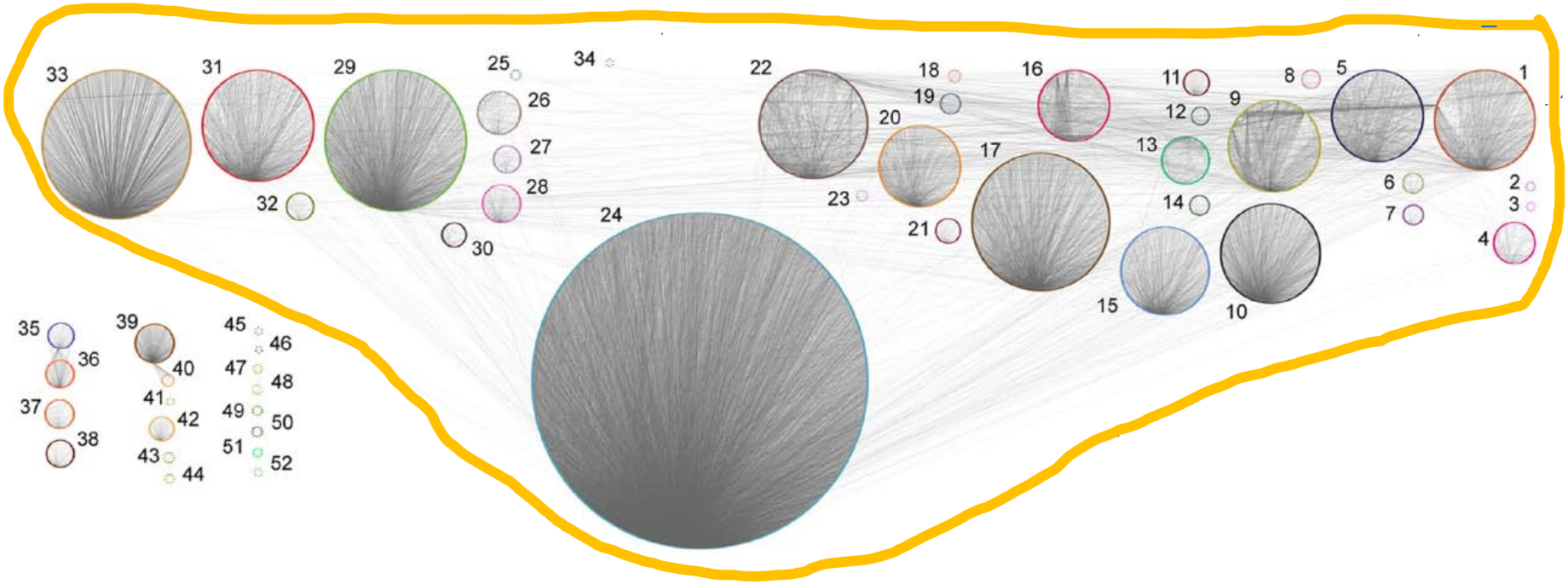
Newman 2006 (PNAS)

Communities in COPD eQTL networks



Communities in COPD eQTL networks

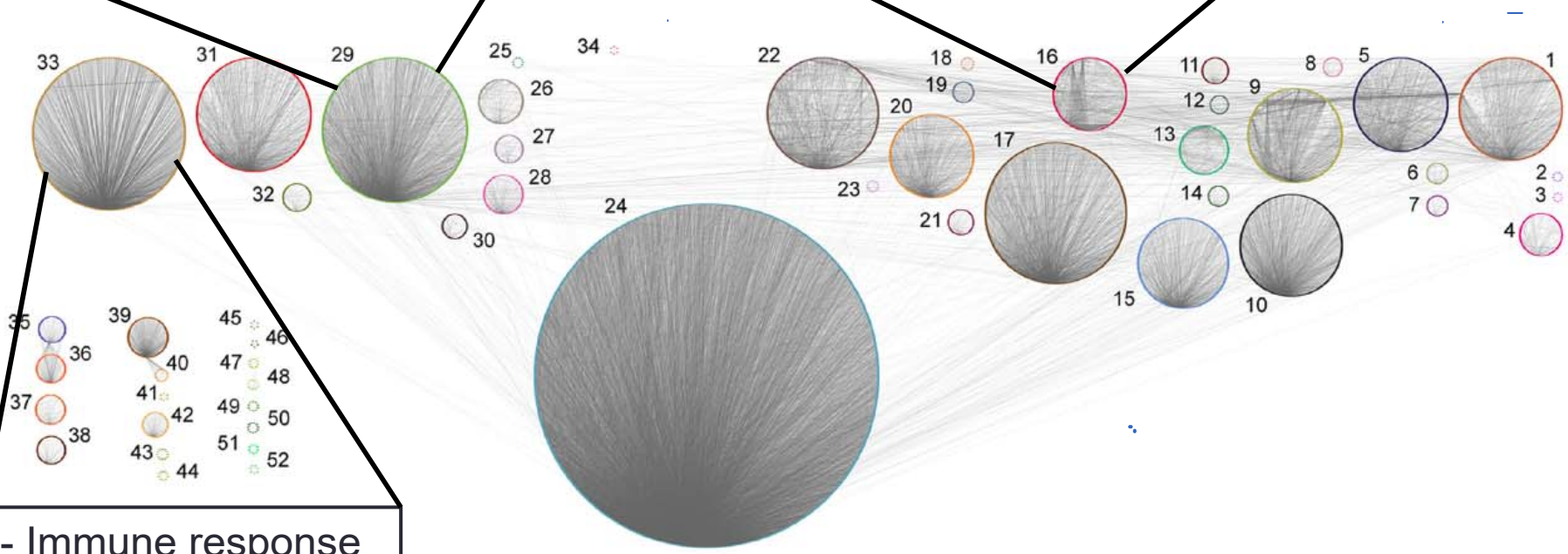
- We identified 52 communities, with Q = 0.79 (out of 1)
- Of 34 communities in the giant connected component, 11 are enriched for GO terms ($P<5x10^{-4}$)



Communities in COPD eQTL networks

- Chromatin Assembly
- DNA conformation change
- Nucleosome assembly

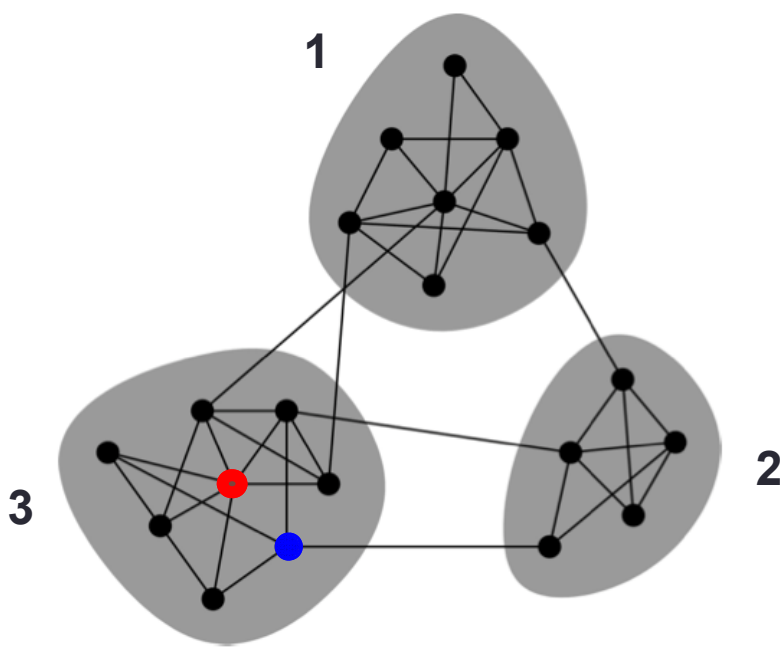
- Microtubule organization
- Cell cycle
- Centrosome



- Immune response
- Stress response
- T cell stimulation

Identifying community cores

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

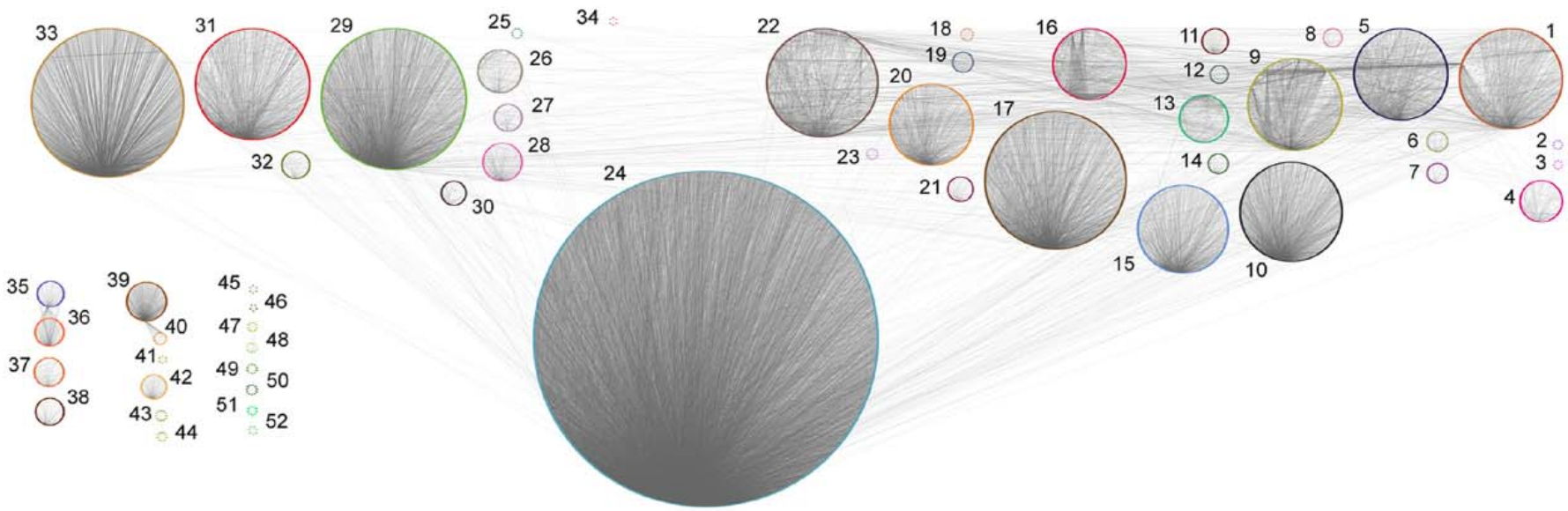


Newman 2006 (PNAS)

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

What about COPD GWAS SNPs?

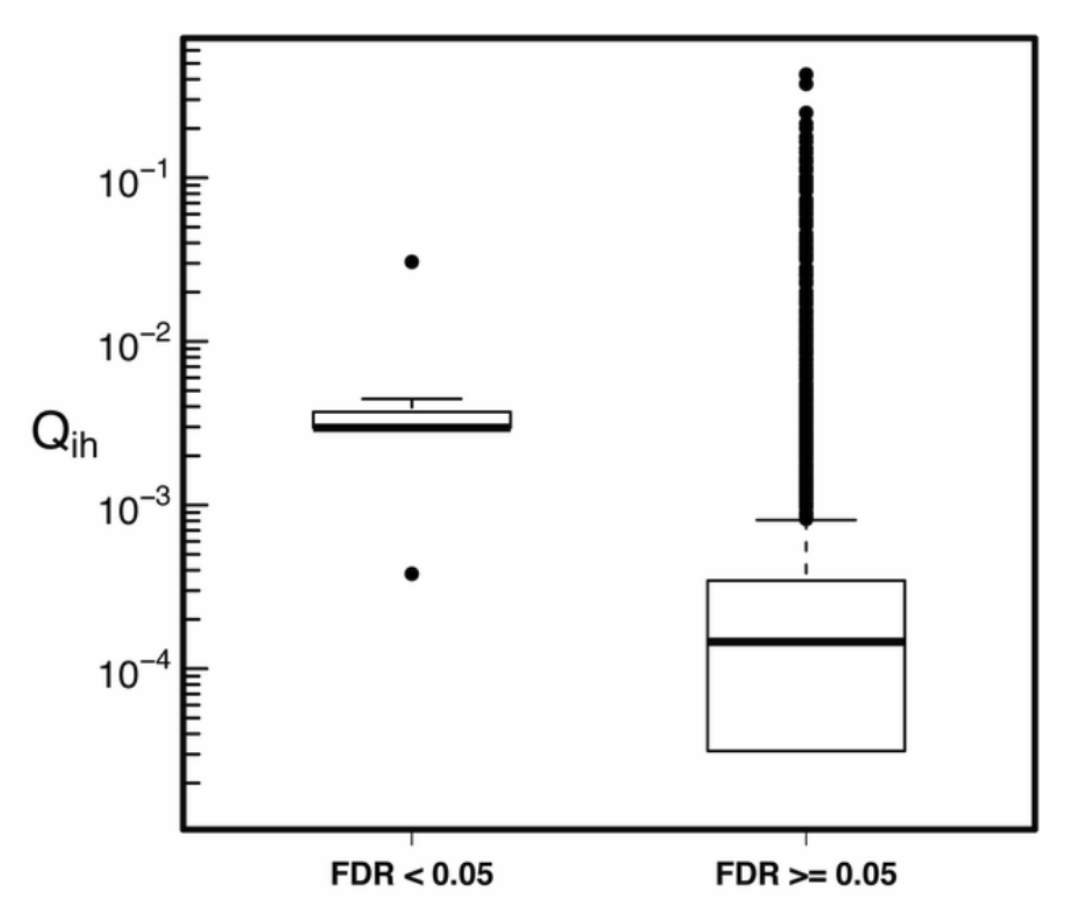
- Use a meta-analysis by Cho et. al. and consider <u>34 COPD GWAS SNPs</u> (FDR < 0.05)



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**



What does this tell us

- The SNPs that are global hubs are not GWAS hits.
- The SNPs and genes group into communities that make sense—a family of SNPs regulate a function.
- GWAS SNPs map to communities whose genes have functions that make biological sense.
- “Core” SNPs are far more likely to be disease SNPs.
- The structure of the network is informative in ways that individual SNP-gene eQTL associations are not.

How general is this?

GTEx eQTL Workflow

GTEx

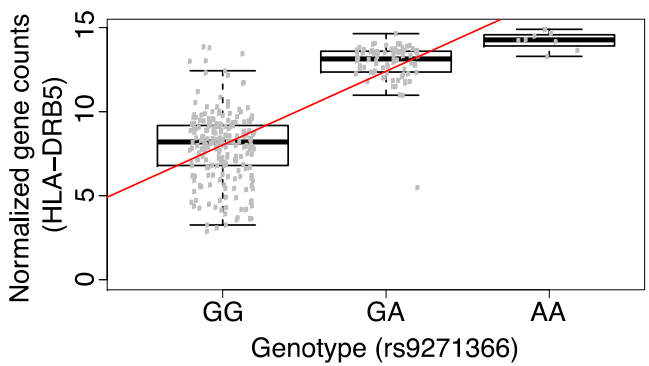
12 tissues, 450 individuals, 4904 samples
Imputed Genotypes (5,640,985 SNPs)
RNAseq data (29,242 genes)

Data Preprocessing

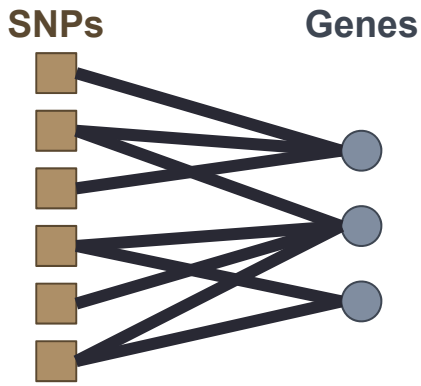
Genotype filtering (*plink*)
MAF ≥ 0.5 & SNP calling > 0.9
RNAseq data QC & normalization (*qsmooth*)

eQTL mapping

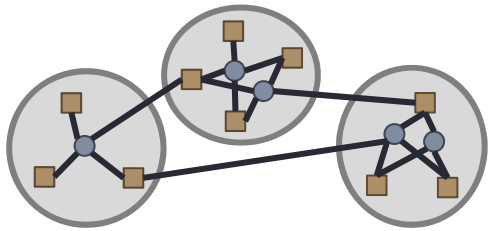
cis- and trans-eQTLs mapping (*MatrixEQTL*)



Bipartite Networks



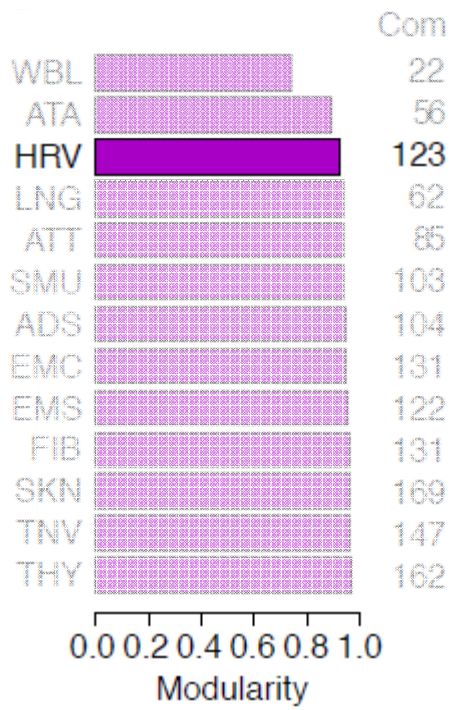
Communities detection
Find community cores
Bipartite Modularity Maximization (*condor*[2])

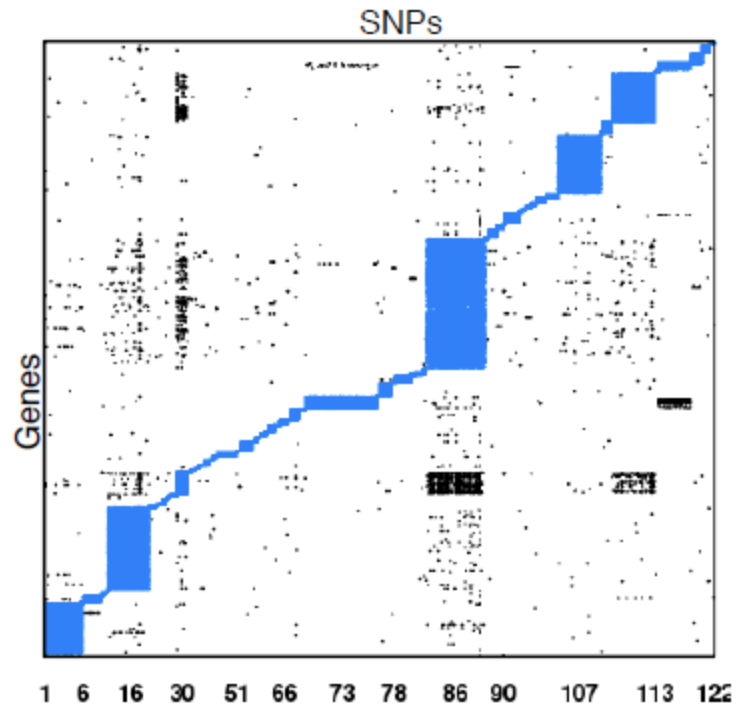


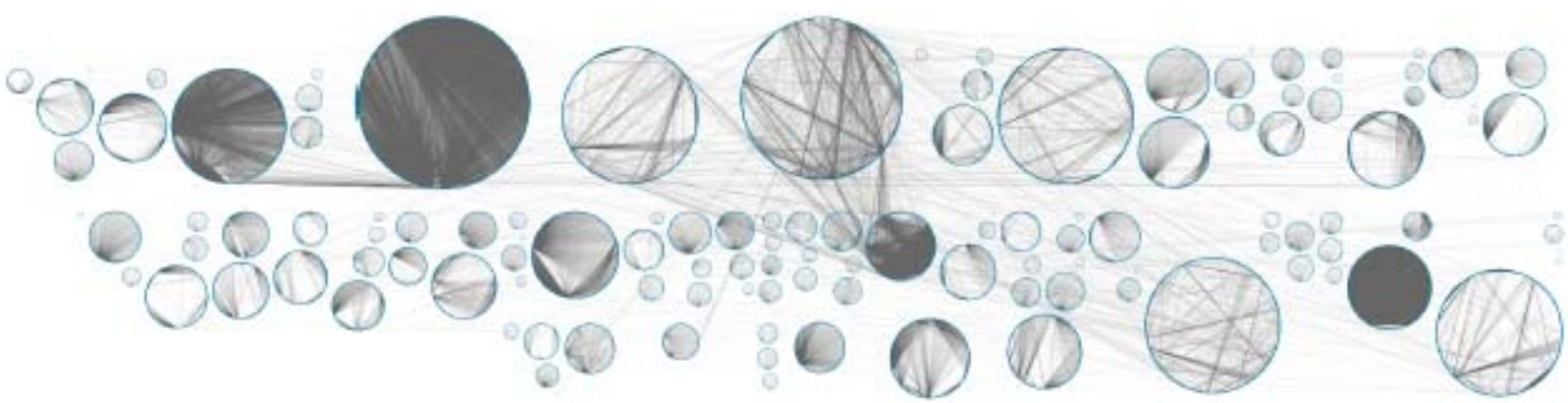
Characterization

Genomic localization of eQTLs
Enrichment in biological functions, pathways, association with diseases

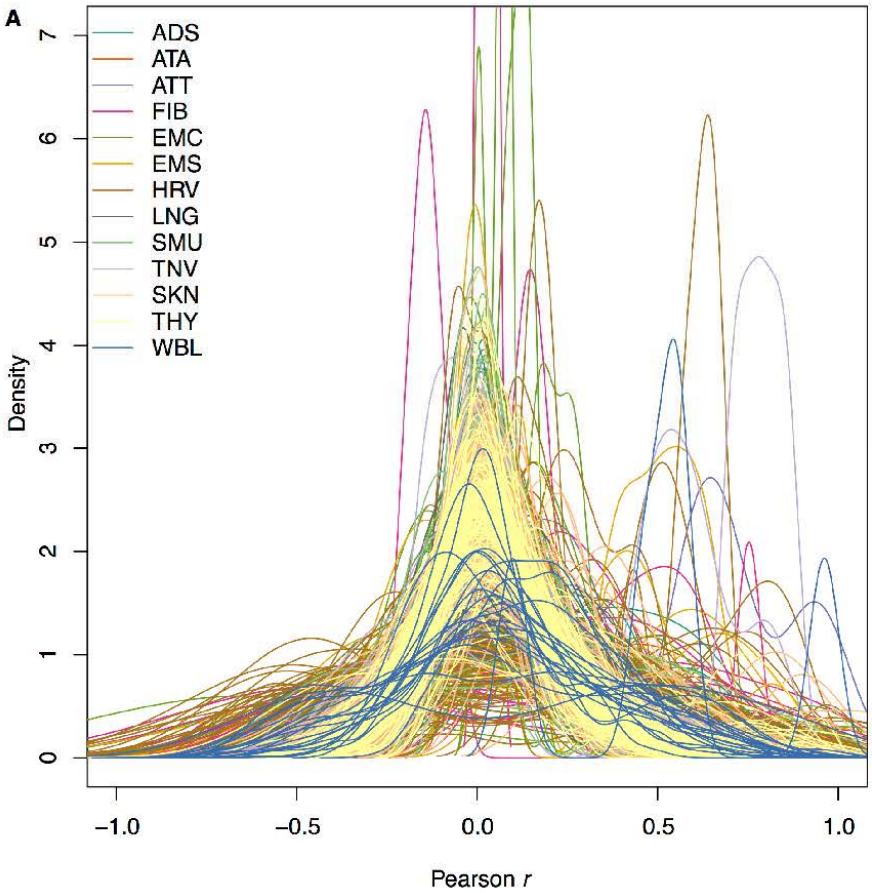
eQTL networks are highly modular

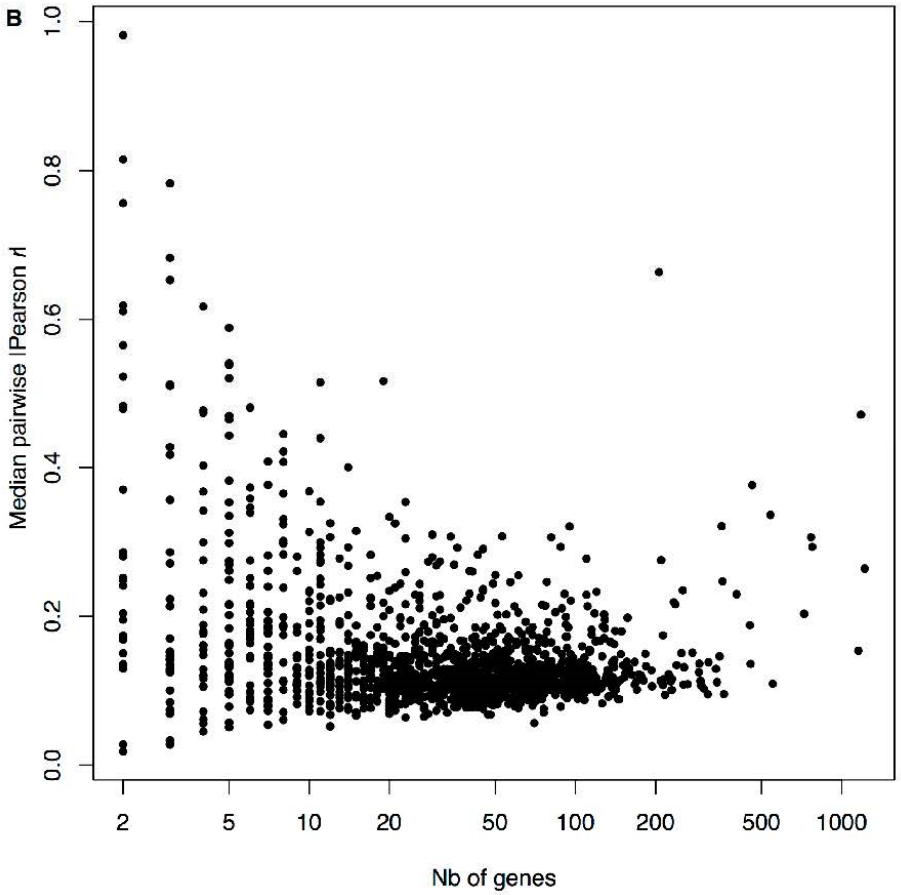




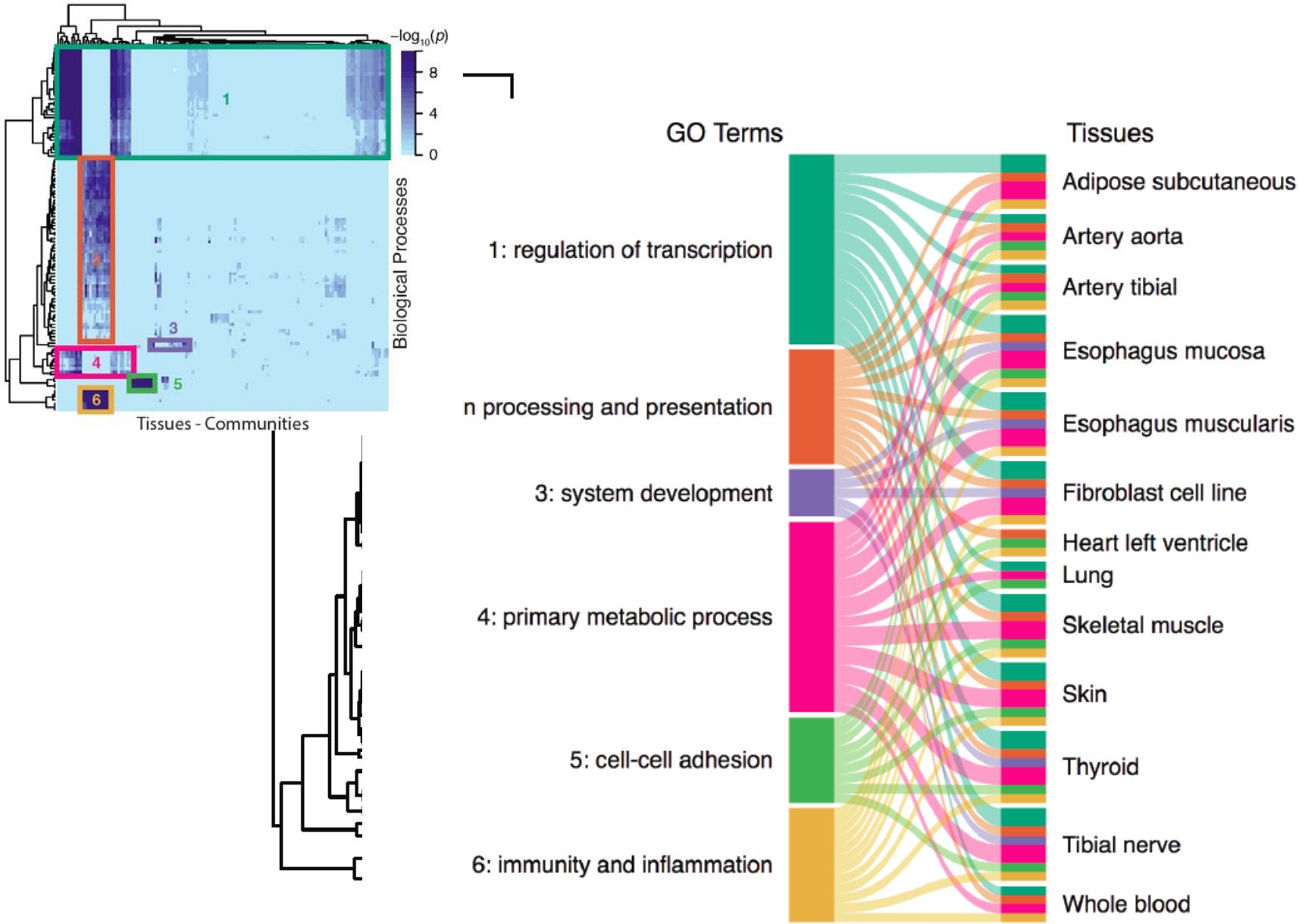


eQTL networks are not co-expression networks

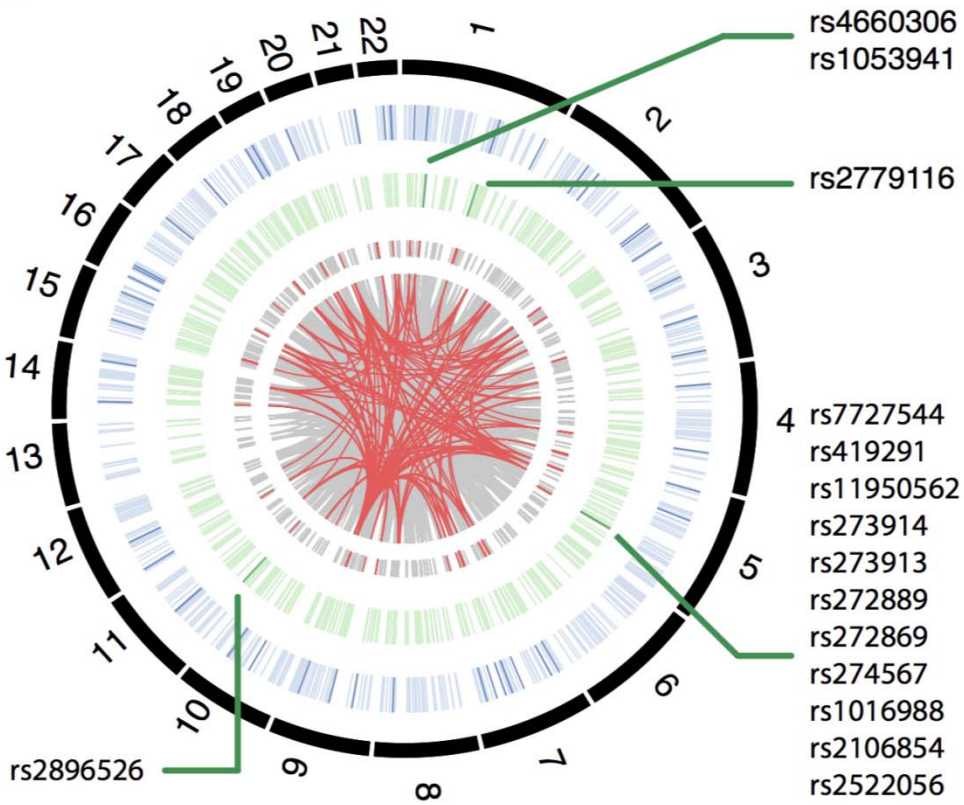


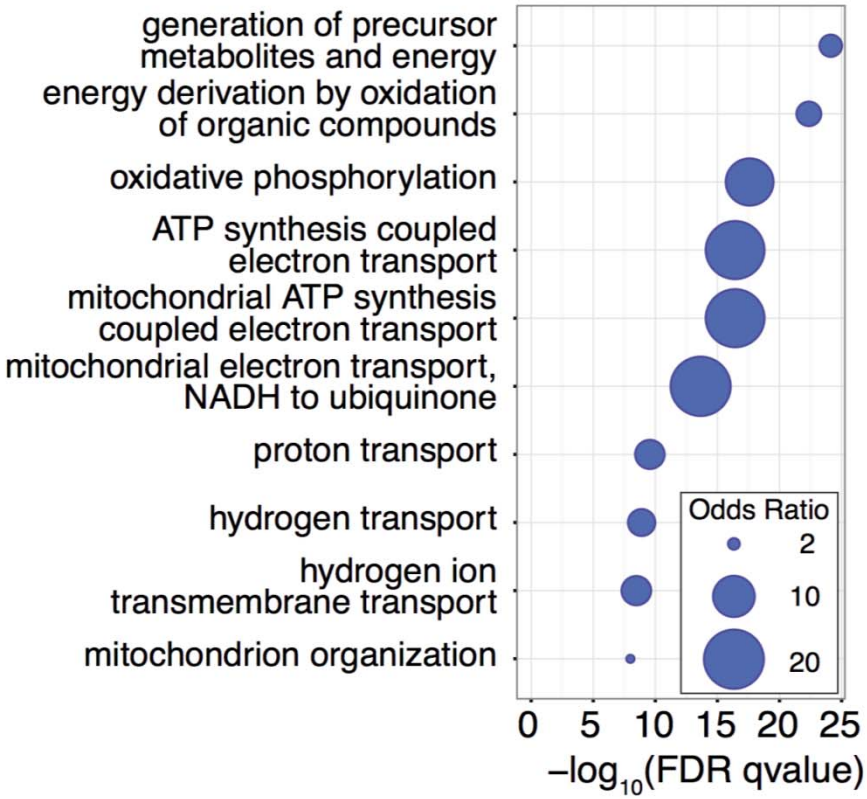


eQTL communities are functionally enriched



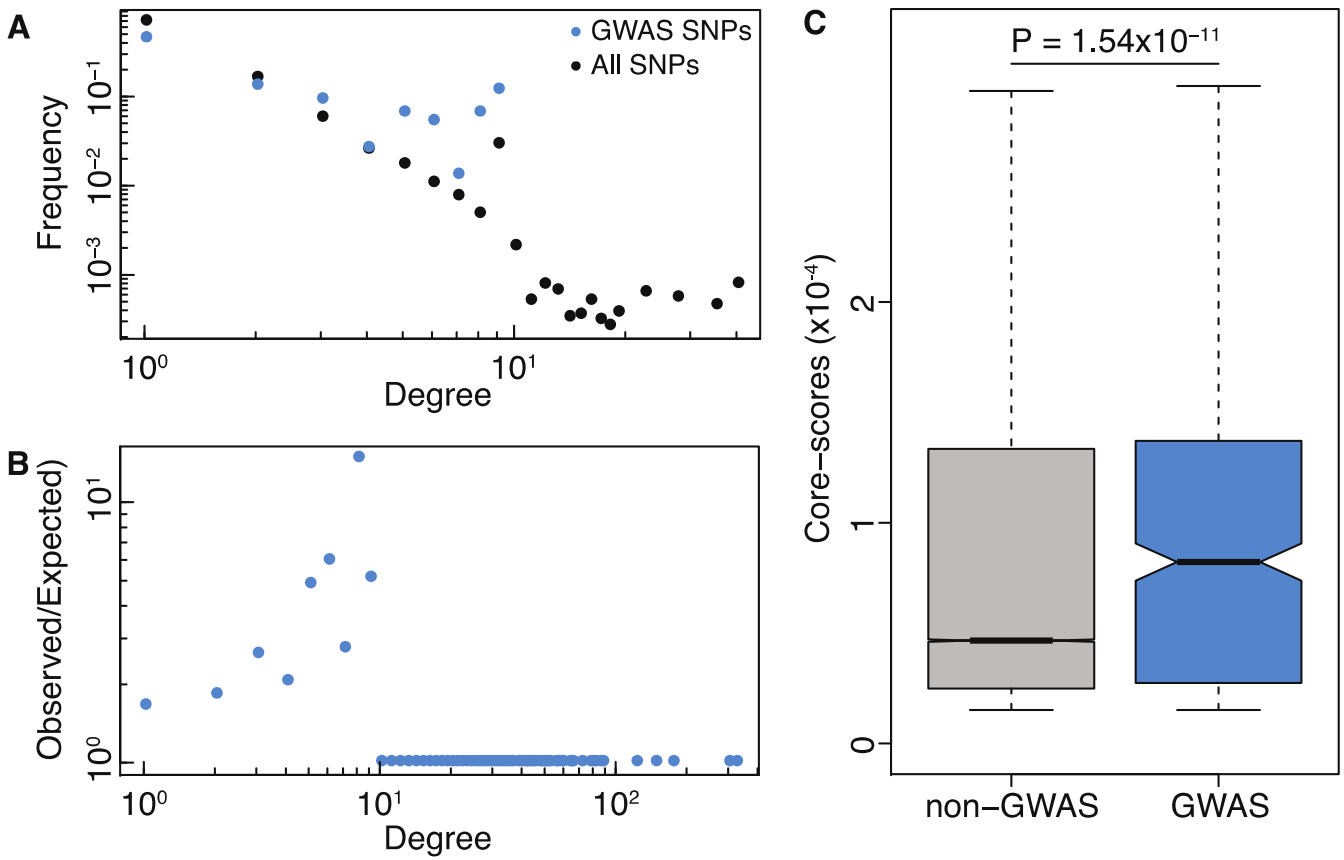
Functionally Enriched Communities Span Many Chromosomes



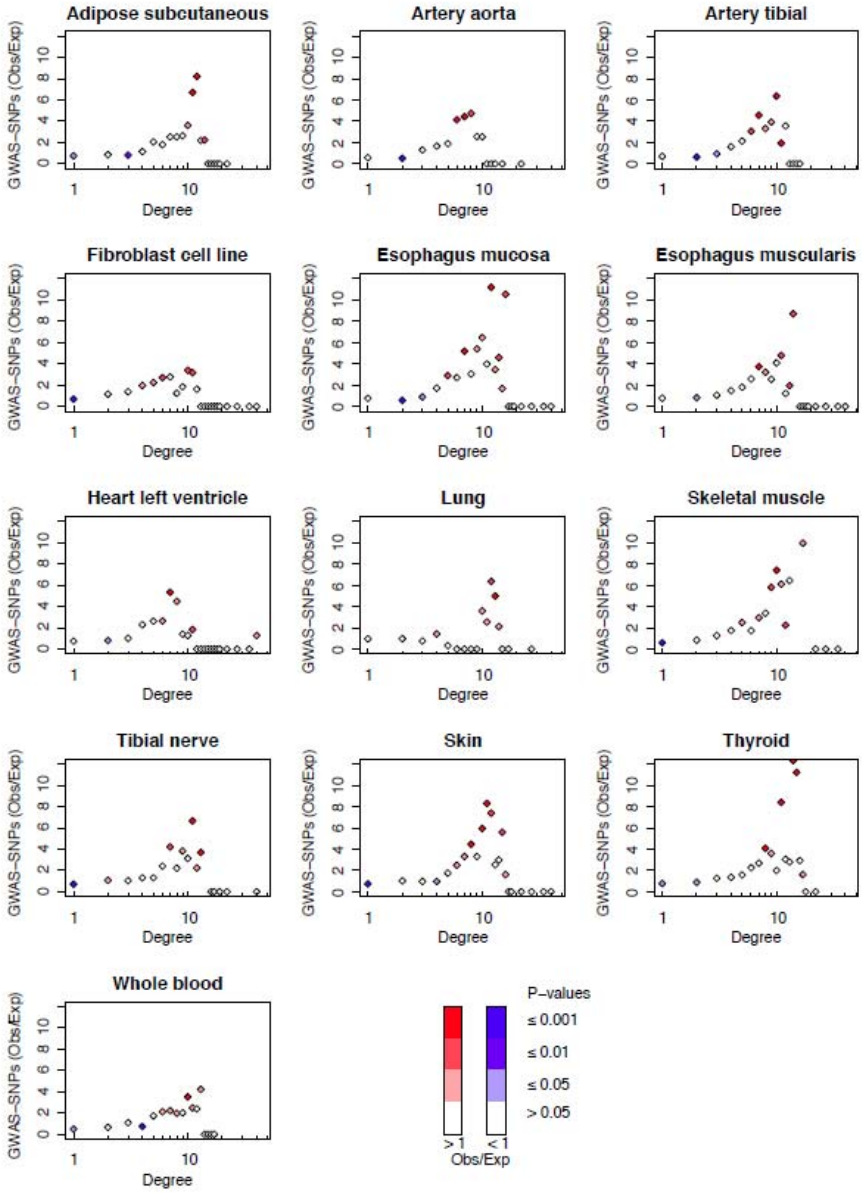


Heart Left Ventricle Community 86

GWAS SNPs are cores, but not hubs

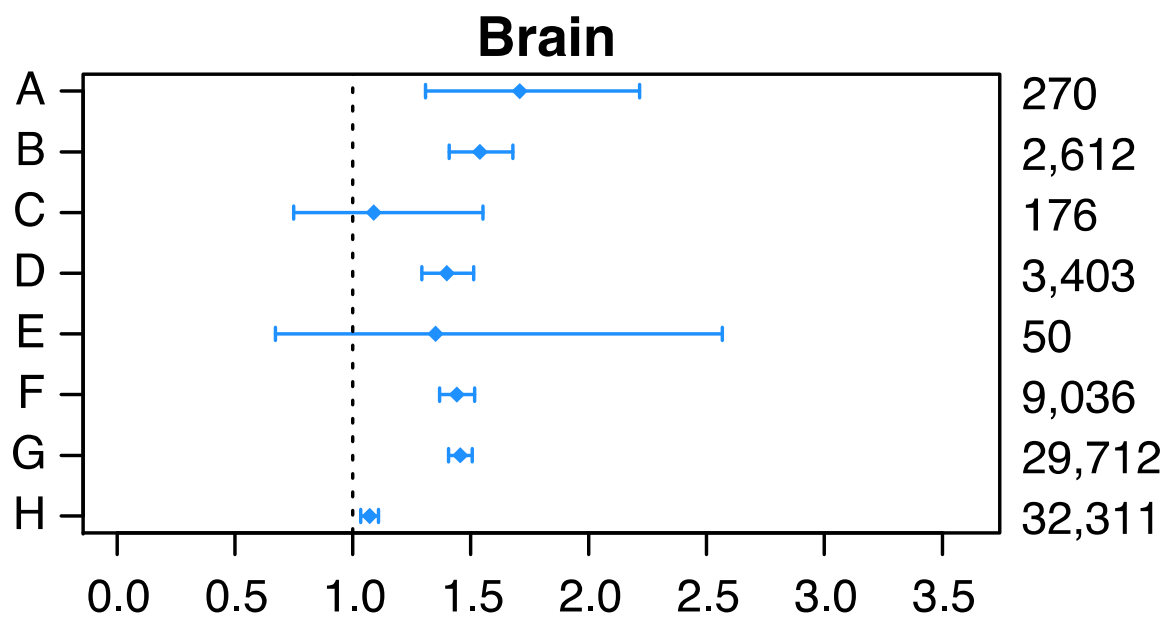


GWAS SNPs not Hubs—in every tissue

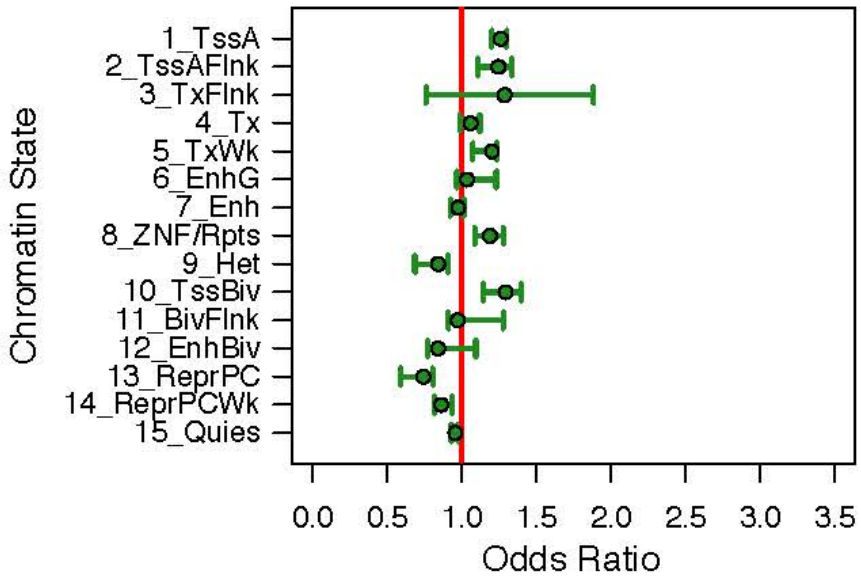


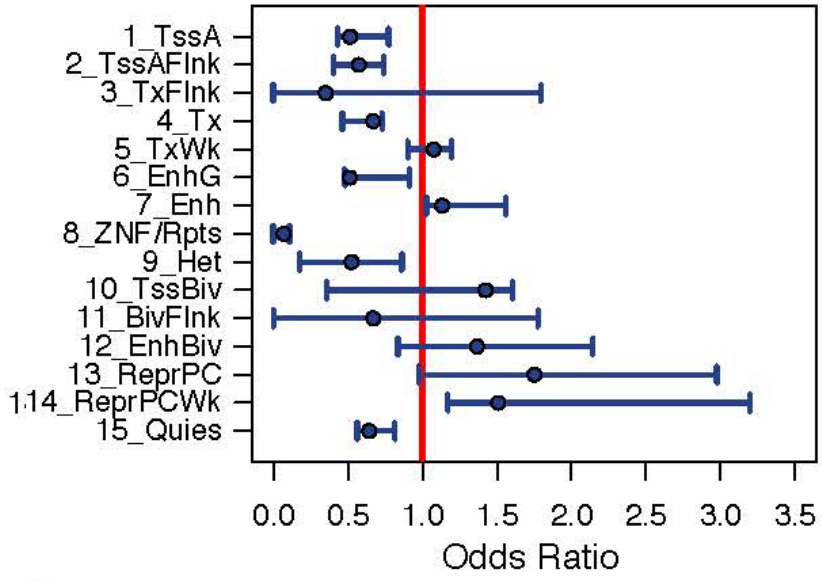
Core SNPs are more likely to be functionally annotated

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

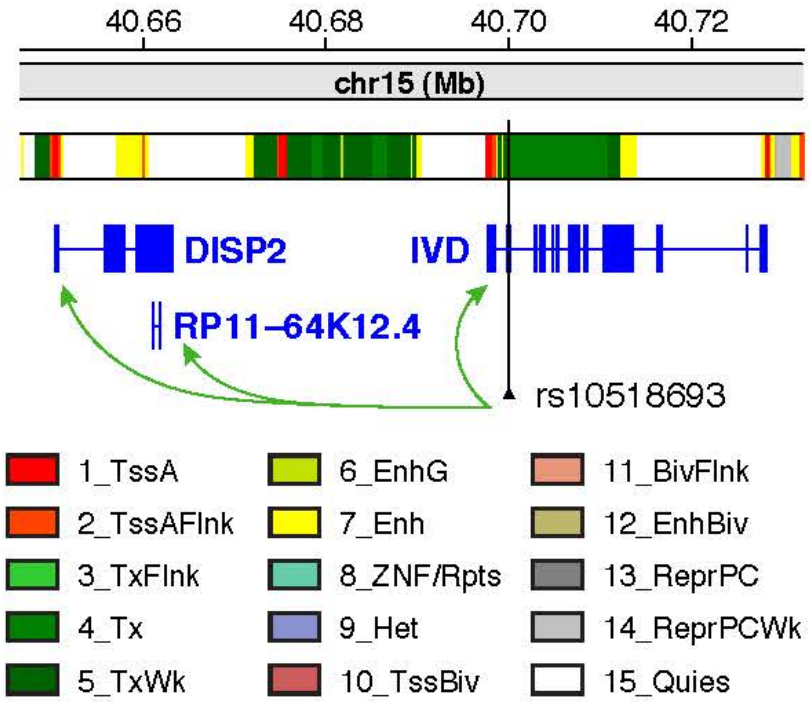
Global Hubs

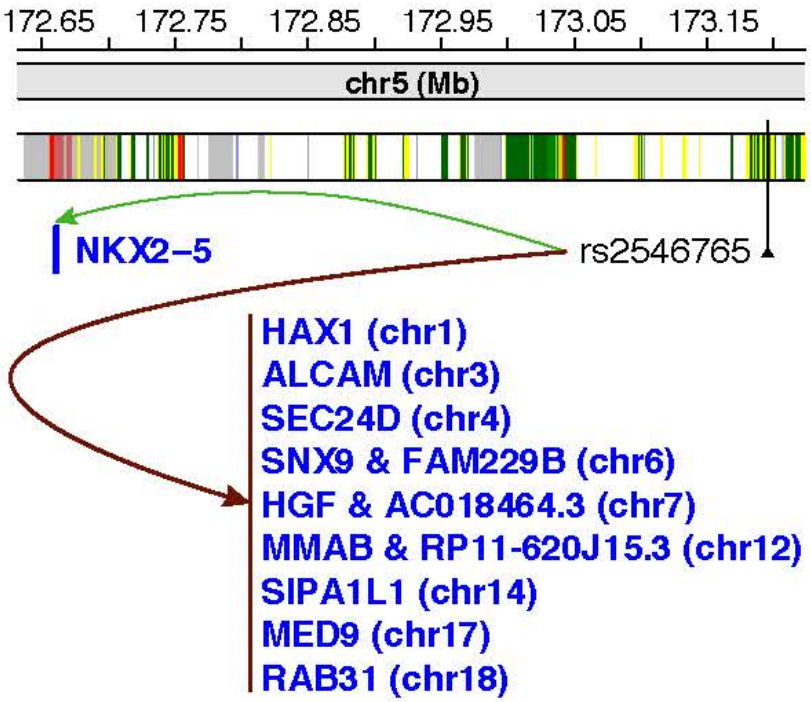
Nongenic Enhancers

Polycomb Repressed Regions

Data from 8 tissues

Core SNPs map to open chromatin





- We find tissue-specific eQTLs map to tissue-specific communities
- And those tissue-specific communities are enriched for SNPs in tissue-specific open chromatin

Interested?

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Bipartite Community Structure of eQTLs

John Platig[1,2]*, Peter J. Castaldi[3,4,5], Dawn DeMeo[3,5,6], John Quackenbush[1,2,3‡]

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

CrossMark
click for updates

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

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

http://arxiv.org/abs/1509.02816; accepted *PLoS Comp Bio*

GTEx eQTLs tell us more

- **There are many *cis*-eQTLs that are shared across tissues.**
- **There are many eQTL communities shared across tissues.**
- **There are also tissue-specific eQTLs and these often to map to tissue-specific functional communities.**
- **These are enriched for SNPs from tissue-specific active chromatin.**
- **Irrespective of tissue, we have modular, functionally enriched communities with Core SNPs most likely to be disease associated.**

Want more?

Cold Spring Harbor Laboratory

bioRχiv beta

THE PREPRINT SERVER FOR BIOLOGY

HOME | CHA

Search

New Results

Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data

Joseph Paulson, Cho-Yi Chen, Camila M Lopes-Ramos, Marieke L Kuijjer, John Platig, Abhijeet R Sonawane, Maud Fagny, Kimberly Glass, John Quackenbush

doi: http://dx.doi.org/10.1101/081802

This article is a preprint and has not been peer-reviewed [what does this mean?].

Cold Spring Harbor Laboratory

bioRχiv beta

THE PREPRINT SERVER FOR BIOLOGY

HOME | CHA

Search

New Results

Sexual dimorphism in gene expression and regulatory networks across human tissues

Cho-Yi Chen, Camila Miranda Lopes-Ramos, Marieke L Kuijjer, Joseph N Paulson, Abhijeet R Sonawane, Maud Fagny, John Platig, Kimberly Glass, John Quackenbush, Dawn L DeMeo

doi: http://dx.doi.org/10.1101/082289

This article is a preprint and has not been peer-reviewed [what does this mean?].

Cold Spring Harbor Laboratory

bioRχiv beta

THE PREPRINT SERVER FOR BIOLOGY

HOME | CHA

Search

New Results

Transcriptional landscape of cell lines and their tissues of origin

Camila Miranda Lopes-Ramos, Joseph N Paulson, Cho-Yi Chen, Marieke L Kuijjer, Maud Fagny, John Platig, Abhijeet R Sonawane, Dawn L DeMeo, John Quackenbush, Kimberly Glass

doi: http://dx.doi.org/10.1101/082065

This article is a preprint and has not been peer-reviewed [what does this mean?].

Cold Spring Harbor Laboratory

bioRχiv beta

THE PREPRINT SERVER FOR BIOLOGY

HOME | CHA

Search

New Results

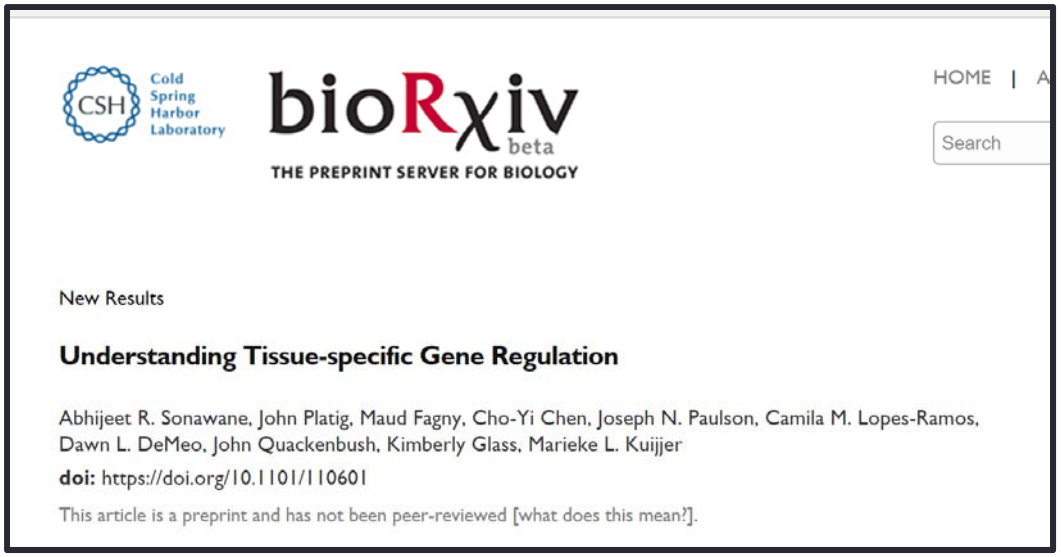
A network-based approach to eQTL interpretation and SNP functional characterization

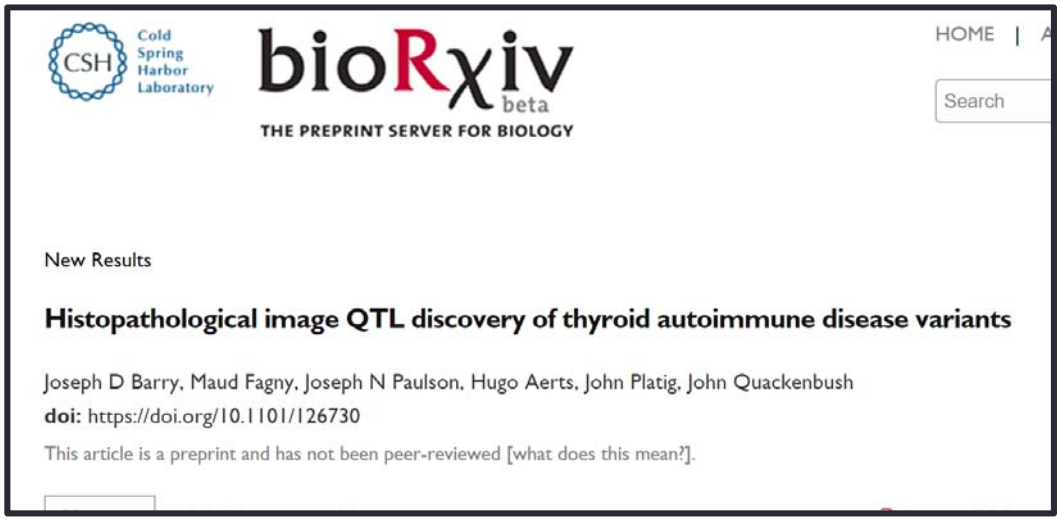
Maud Fagny, Joseph N. Paulson, Marieke L. Kuijjer, Abhijeet R. Sonawane, Cho-Yi Chen, Camila M. Lopes-Ramos, Kimberly Glass, John Quackenbush, John Platig

doi: http://dx.doi.org/10.1101/086587

This article is a preprint and has not been peer-reviewed [what does this mean?].

Want more?





http://tinyurl.com/gtex01

http://tinyurl.com/gtex02

http://tinyurl.com/gtex03

http://tinyurl.com/gtex04

http://tinyurl.com/gtex05

http://tinyurl.com/gtex06

Why Science Matters

The future is here.
It's just not widely distributed yet.

- William Gibson

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)

<johnq@jimmy.harvard.edu>

Acknowledgments

http://compbio.dfci.harvard.edu

DFCI GTEx Team

Joseph Barry
Joey (Cho-Yi) Chen
Shelia Gaynor
Marieke Kuijjer
Camila Lopes-Ramos
Megha Padi
Joseph Paulson
John Platig
John Quackenbush
Daniel Schlauch

CDNM, Brigham and Women's Hospital

Dawn DeMeo
Kimberly Glass
Abhijeet Sonawane

Administrative Support

Nicole Trotman

GEOFFREY BEENE FOUNDATION ALZHEIMER'S INITIATIVE

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

NATIONAL CANCER INSTITUTE

NATIONAL LIBRARY OF MEDICINE

National Heart Lung and Blood Institute

National Institute of Allergy and Infectious Diseases

NVIDIA