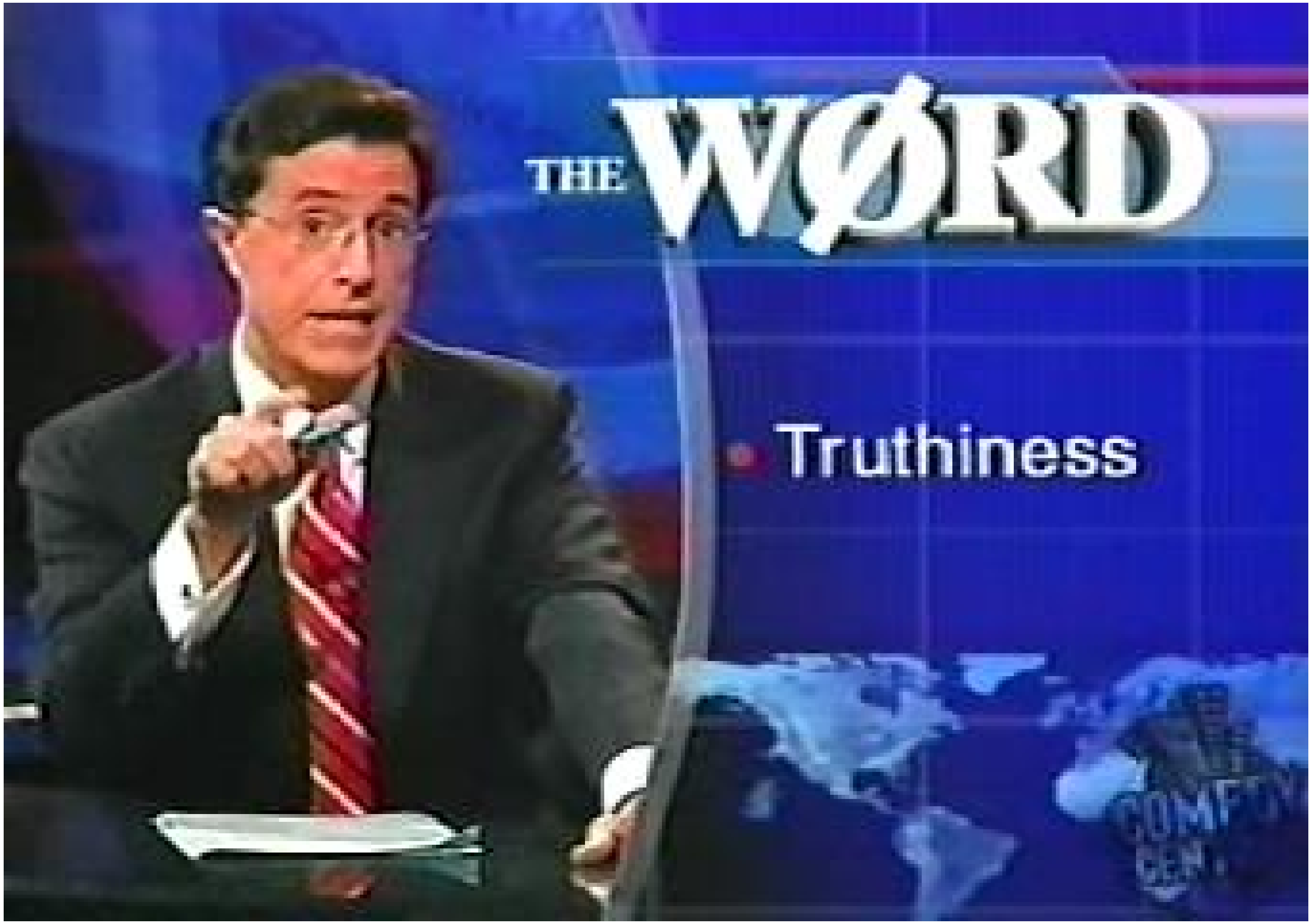
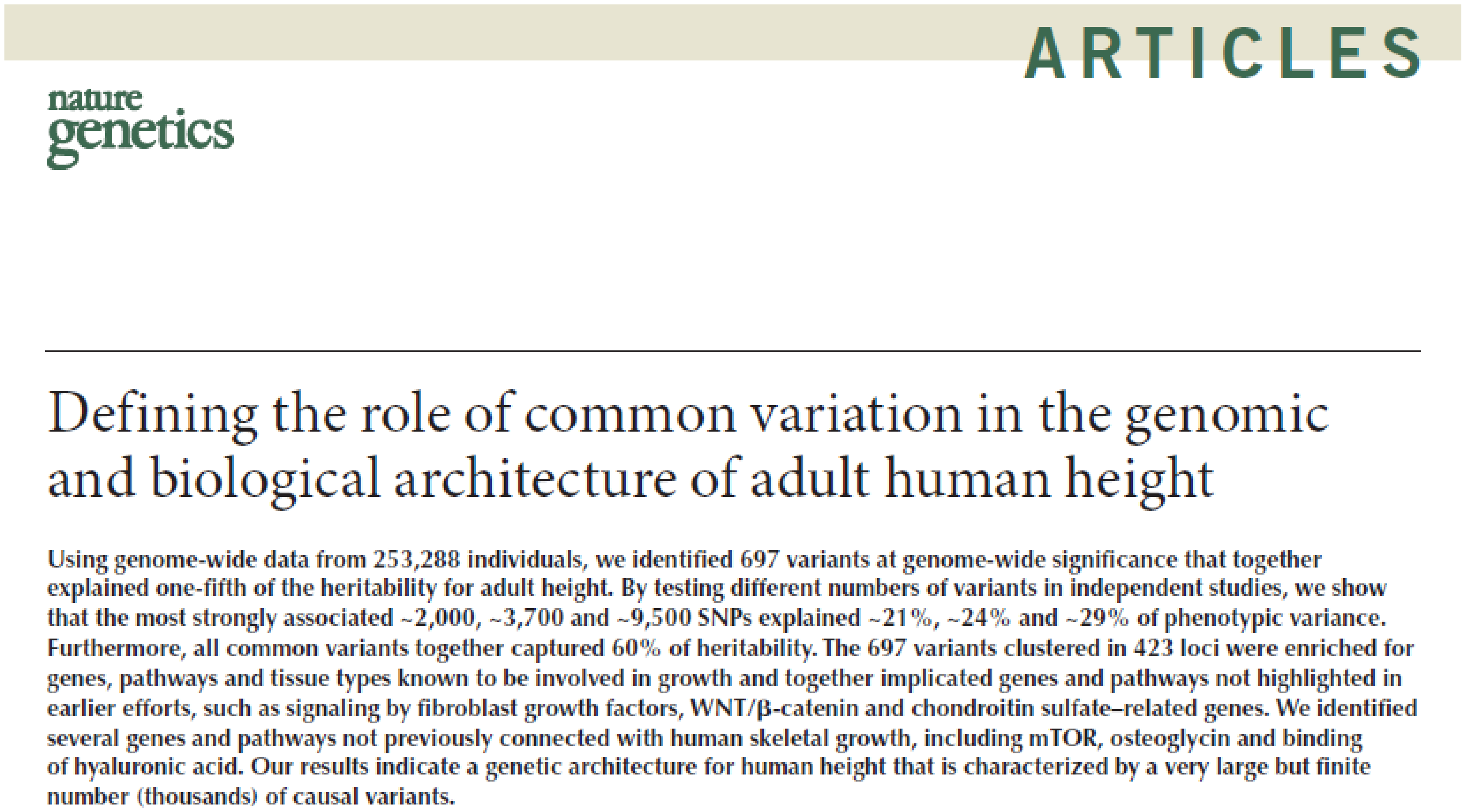
USING NETWORKS TO RE-EXAMINE THE GENOME-PHENOME CONNECTION

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

The WØRD



When you feel it in your gut, you know it must be right.



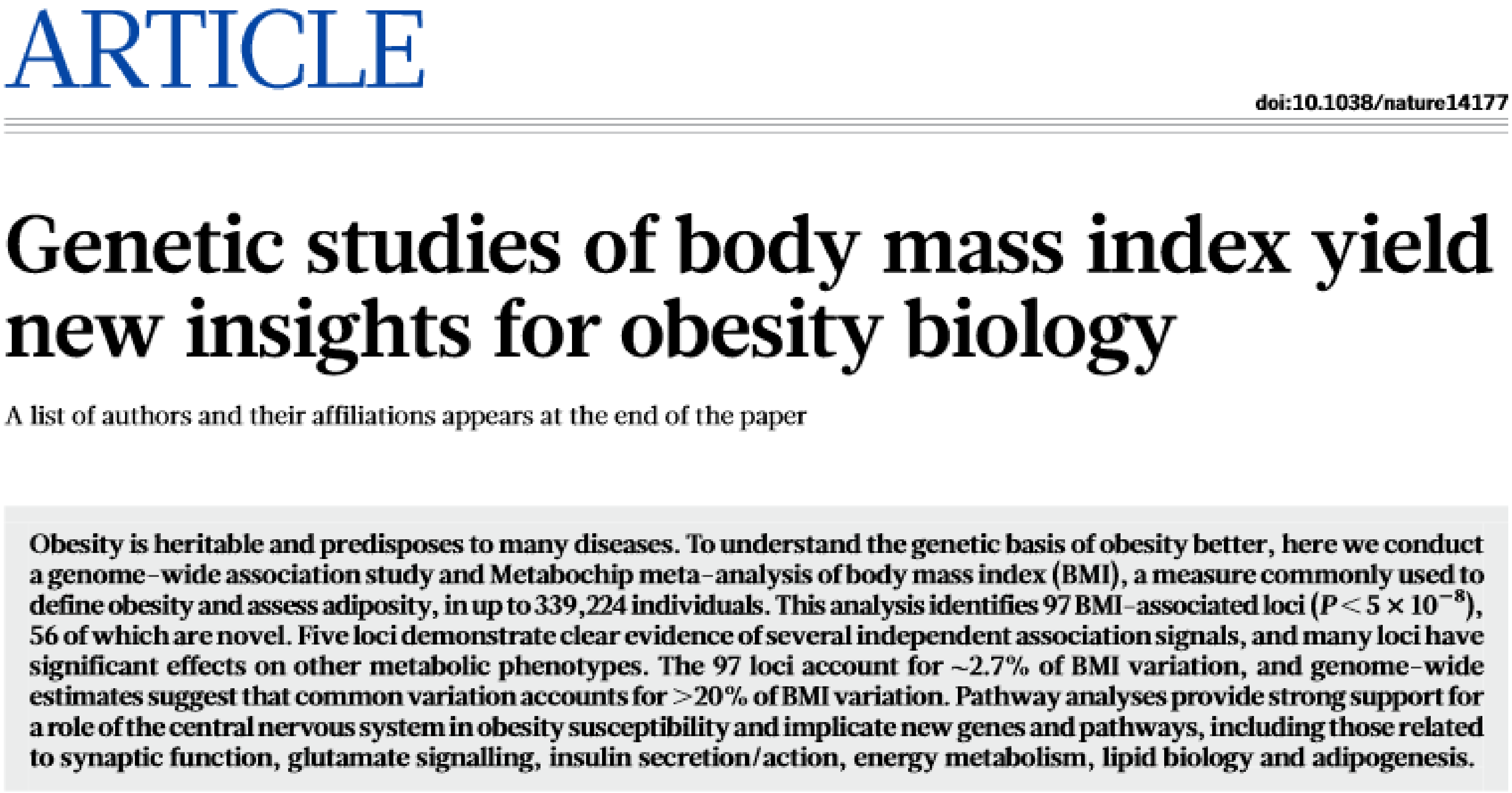
ARTICLES

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?

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

- eQTLs should group into communities with core SNPs regulating particular cellular functions

- Perform a “standard eQTL” analysis using Matrix_EQTL:

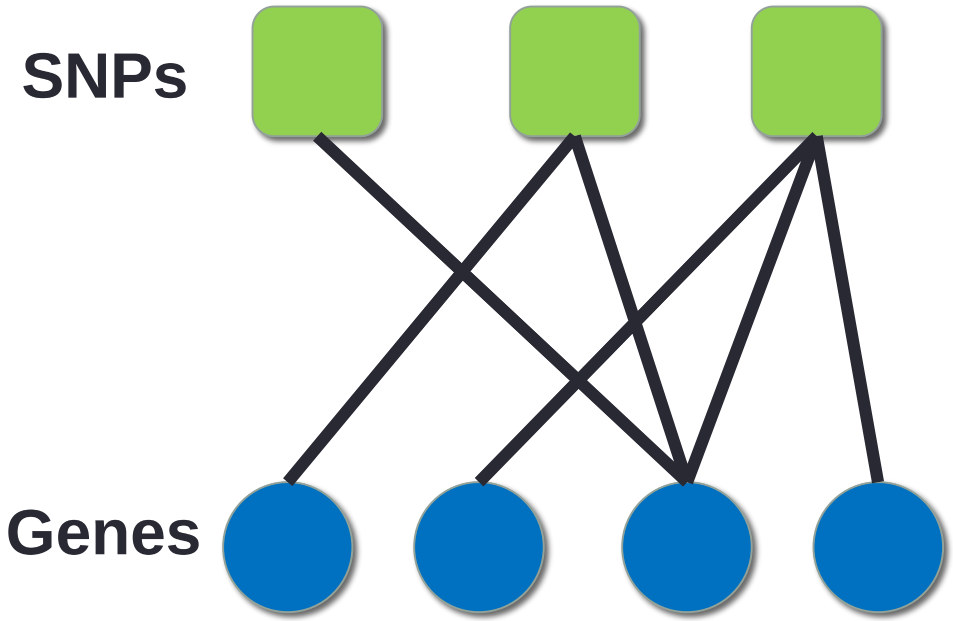
$$Y = \beta_0 + \beta_1 ADD + \varepsilon$$

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

John Platig

Which SNPs affect function?

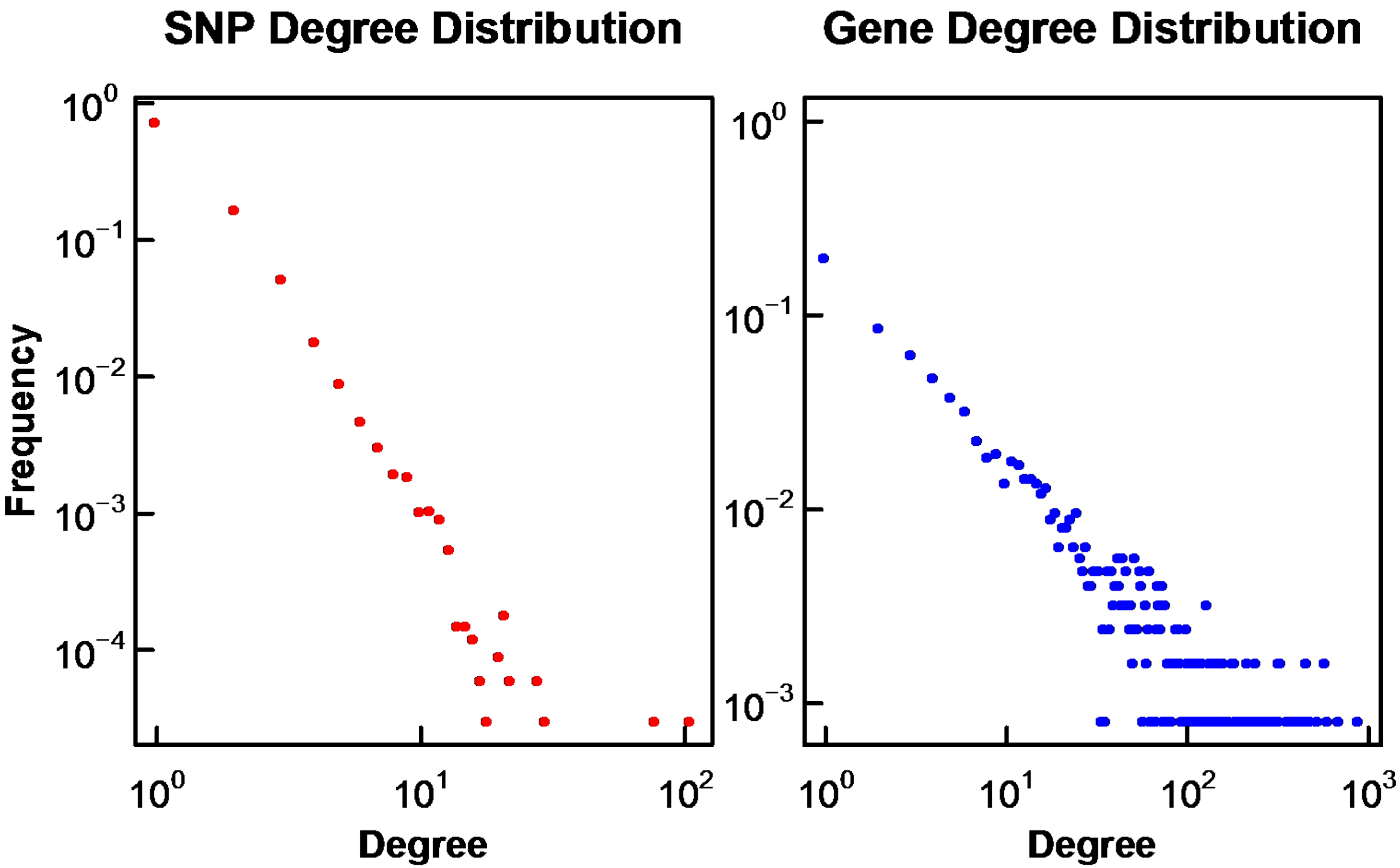
Many strong eQTLs are found near the target gene. But what about multiple SNPs that are correlated with multiple genes?



Can a network of SNP-gene associations inform the functional roles of these SNPs?

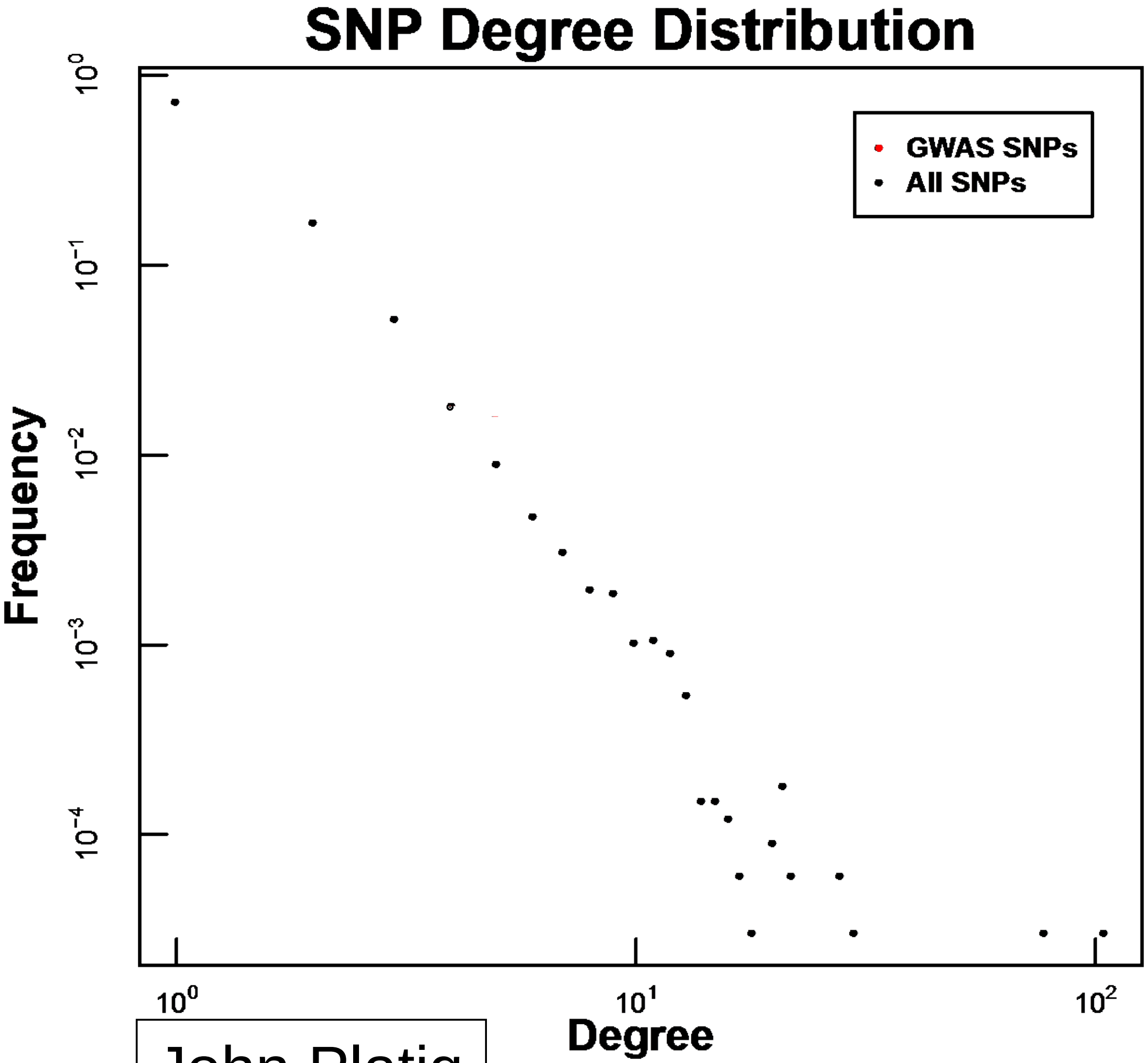
John Platig

Results: COPD



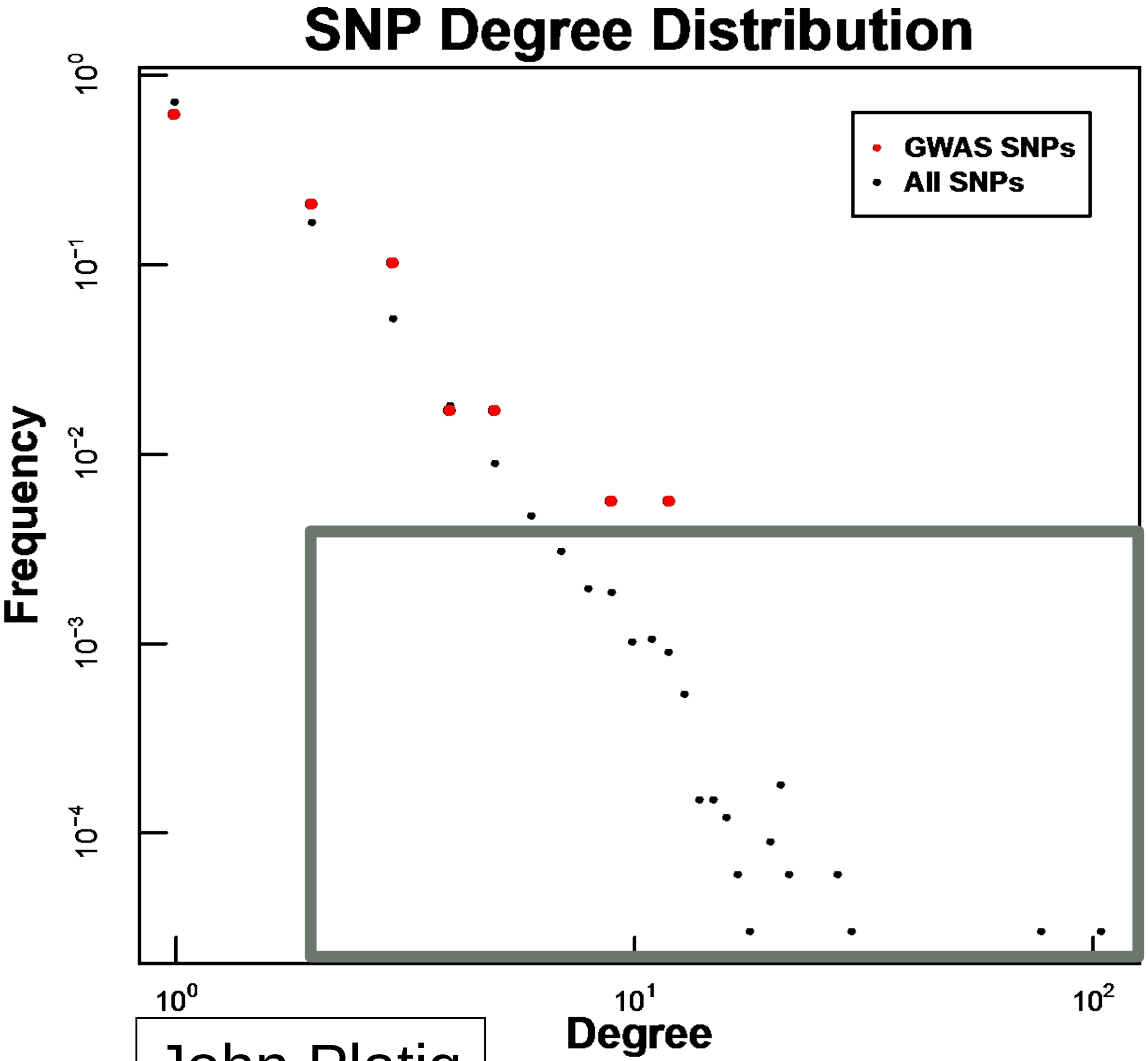
John Platig

What about GWAS SNPs?



John Platig

What about GWAS SNPs?

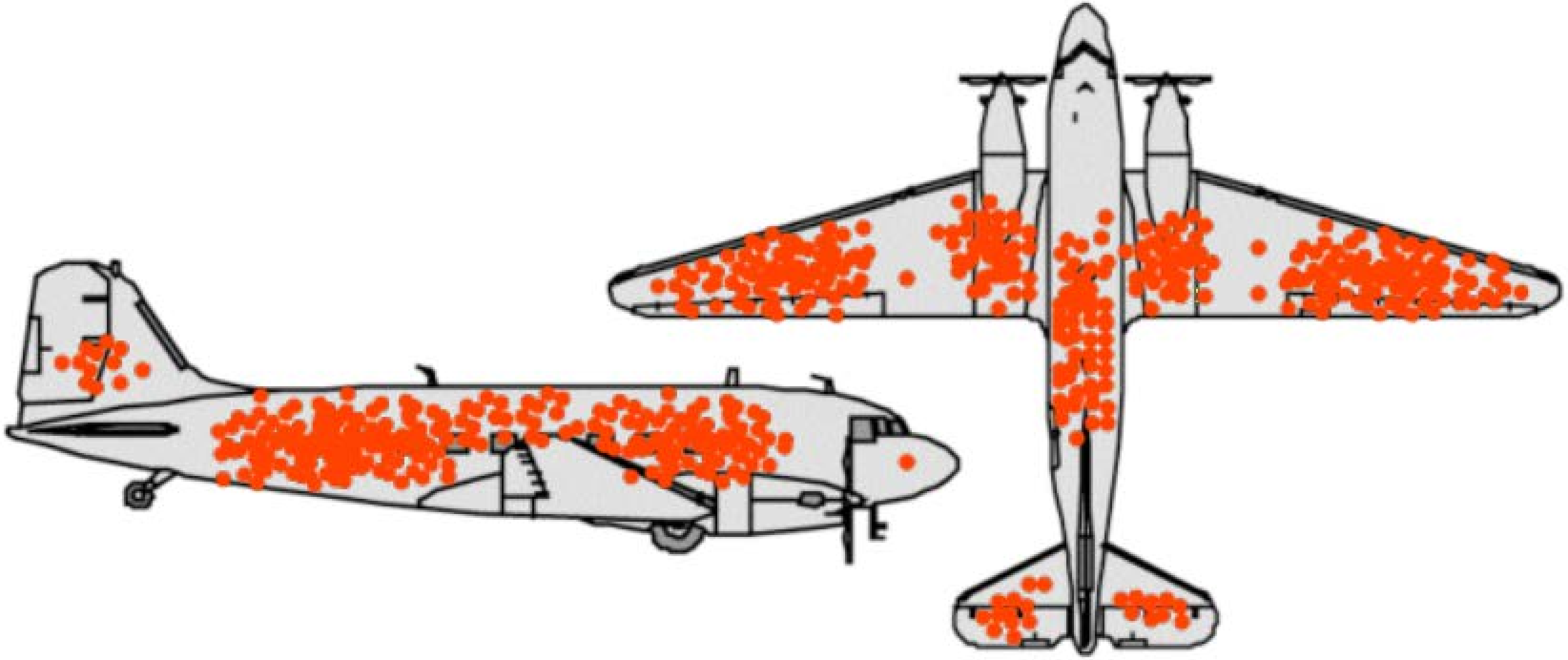


The "hubs" are
a
GWAS desert!

John Platig

What are the critical areas?

Abraham Wald: Put the armor where the bullets aren’t!



http://cameronmoll.com/Good_vs_Great.pdf

Network Structure Matters?

- Are “disease” SNPs skewed towards the top of my SNP list as ranked by the overall out degree?
- No!
  - The collection of highest-degree SNPs is devoid of disease-related SNPs
  - Highly deleterious SNPs that affect many processes are probably removed by strong negative selection.

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

John Platig

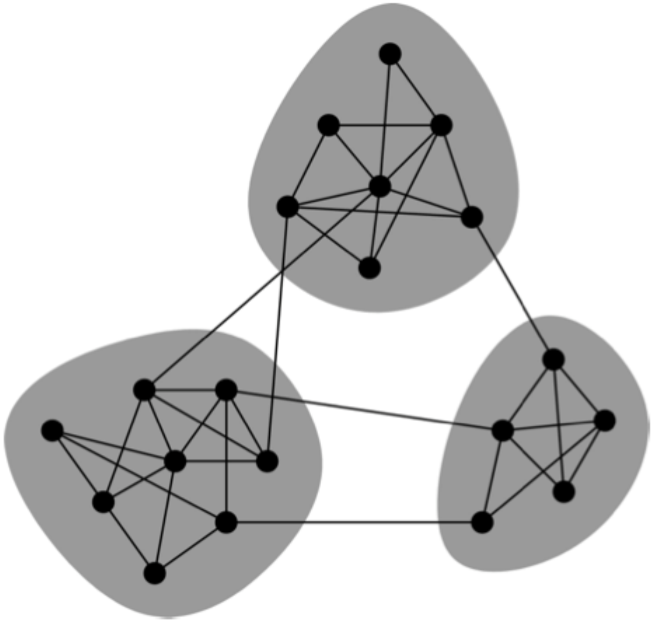
Communities are groups of highly intra-connected nodes

- 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



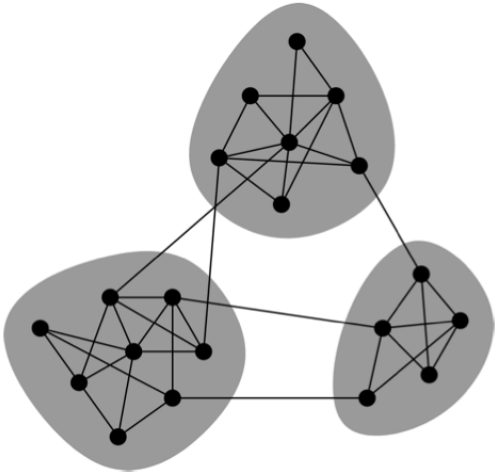
Newman 2006 (PNAS)

John Platig

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**.

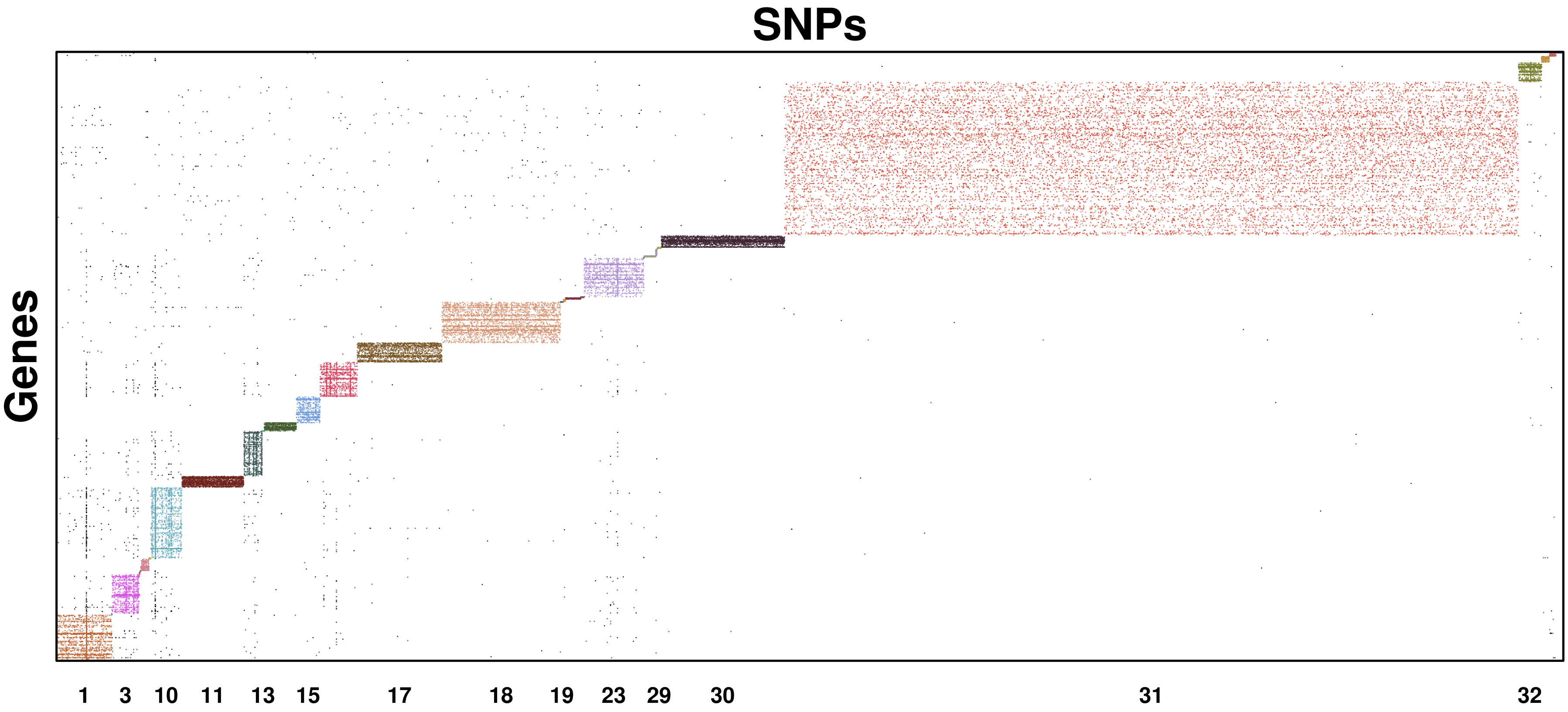
Bipartite networks require a different null model



Newman 2006 (PNAS)

John Platig

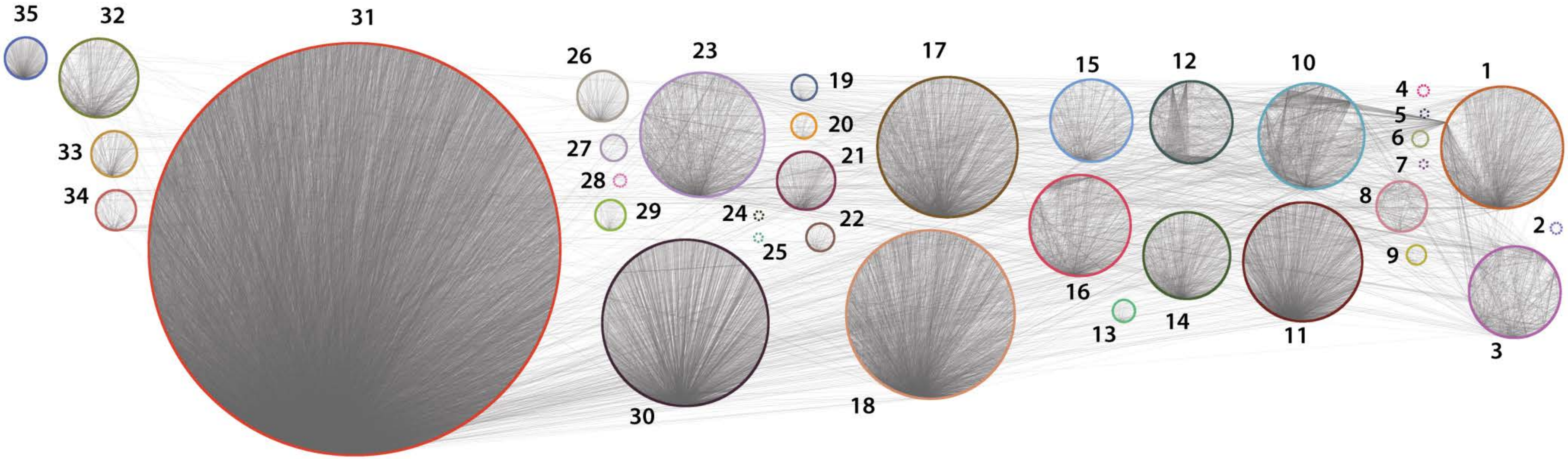
Communities in COPD eQTL networks



John Platig

Communities in COPD eQTL networks

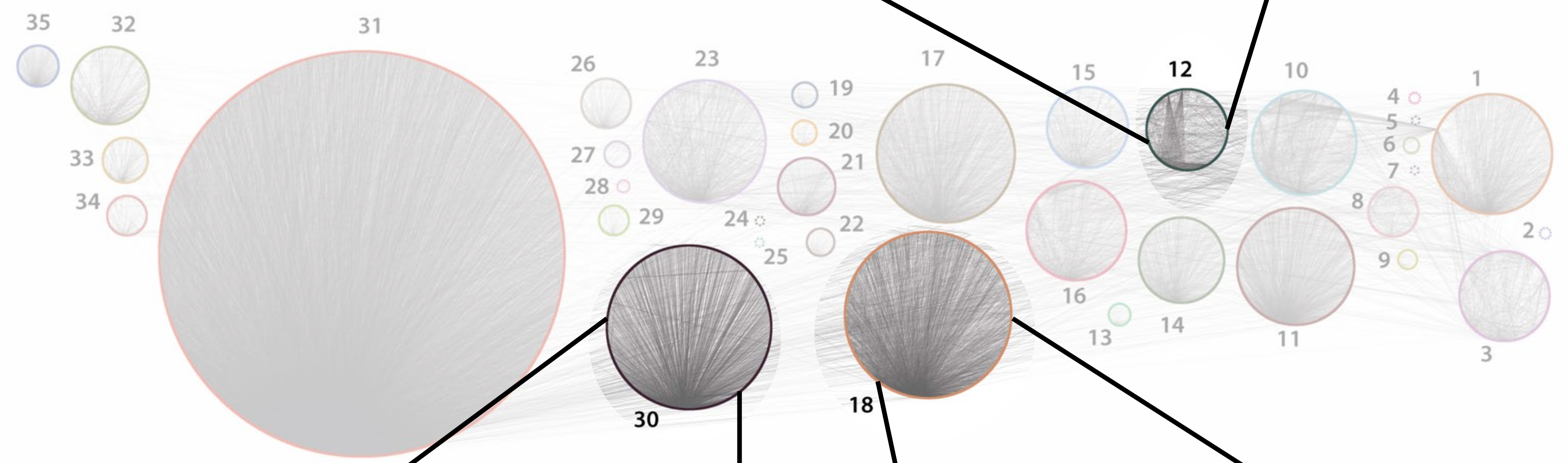
- We identified 35 communities, with Q = 0.77 (out of 1)
- Of 35 communities, 13 are enriched for GO terms ($P<5x10^{-4}$)



John Platig

Communities in COPD eQTL networks

- Microtubule organization
- Cell cycle
- Centrosome



- Immune response
- Stress response
- T cell stimulation

- Chromatin Assembly
- DNA conformation change
- Nucleosome assembly

John Platig

Calculate Local Connectivity

$$Q_i^c = \frac{Q_i}{Q_c}$$

Core Score

$$Q_i = \frac{1}{2m} \sum_{j \in c} \left(A_{ij} - \frac{k_i d_j}{m} \right)$$

Modularity of node *i*

$$Q_c = \frac{1}{2m} \sum_{i,j \in c} \left(A_{ij} - \frac{k_i d_j}{m} \right)$$

Modularity of community *c*

John Platig

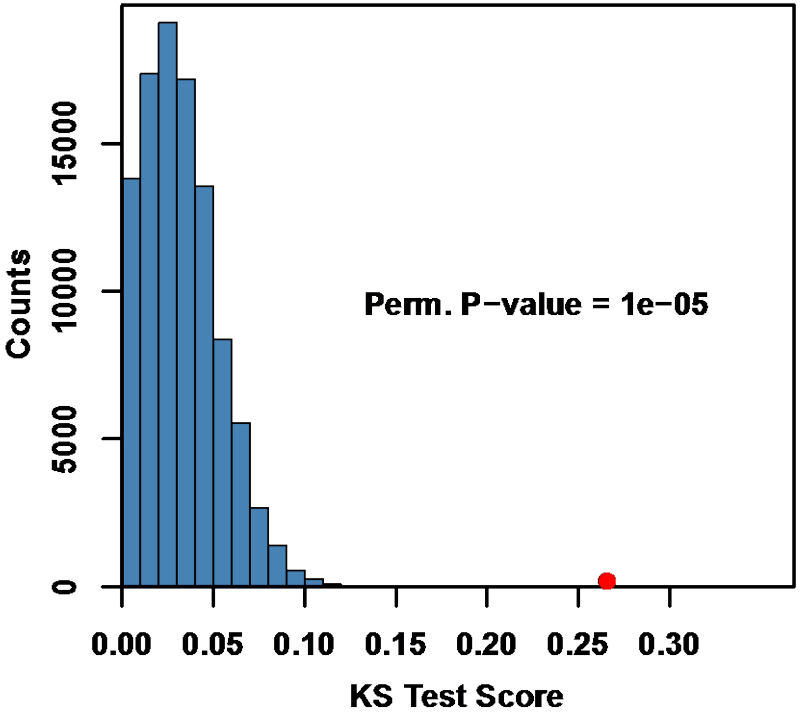
Network Structure Matters!

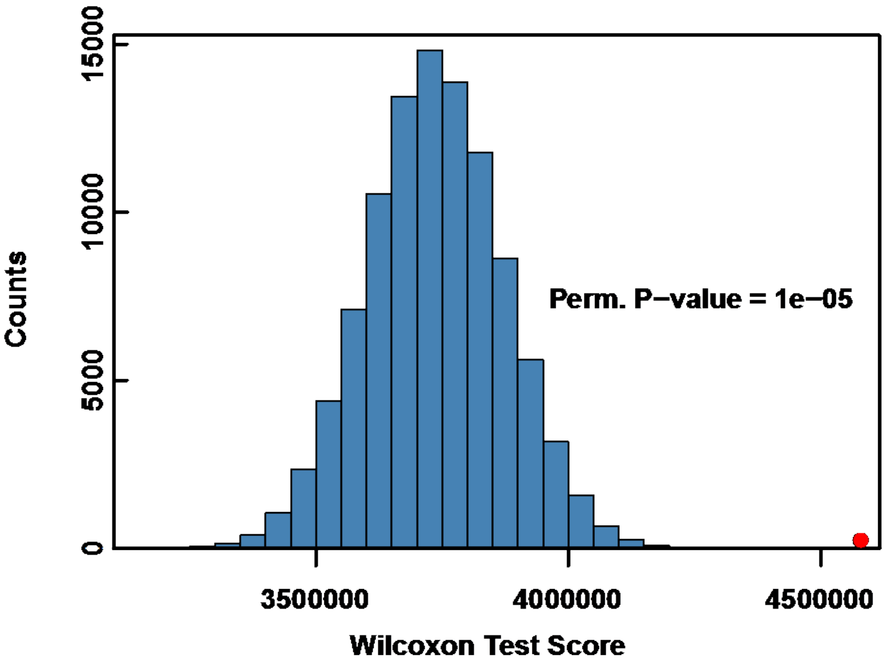
- Are “disease” SNPs skewed towards the top of my SNP list as ranked by the community core score (Q_i^c)?
- Yes!

John Platig

Core Scores and GWAS hits?

Are the Core Scores for GWAS disease stochastically larger than a randomly sub-sampled non-GWAS distribution?

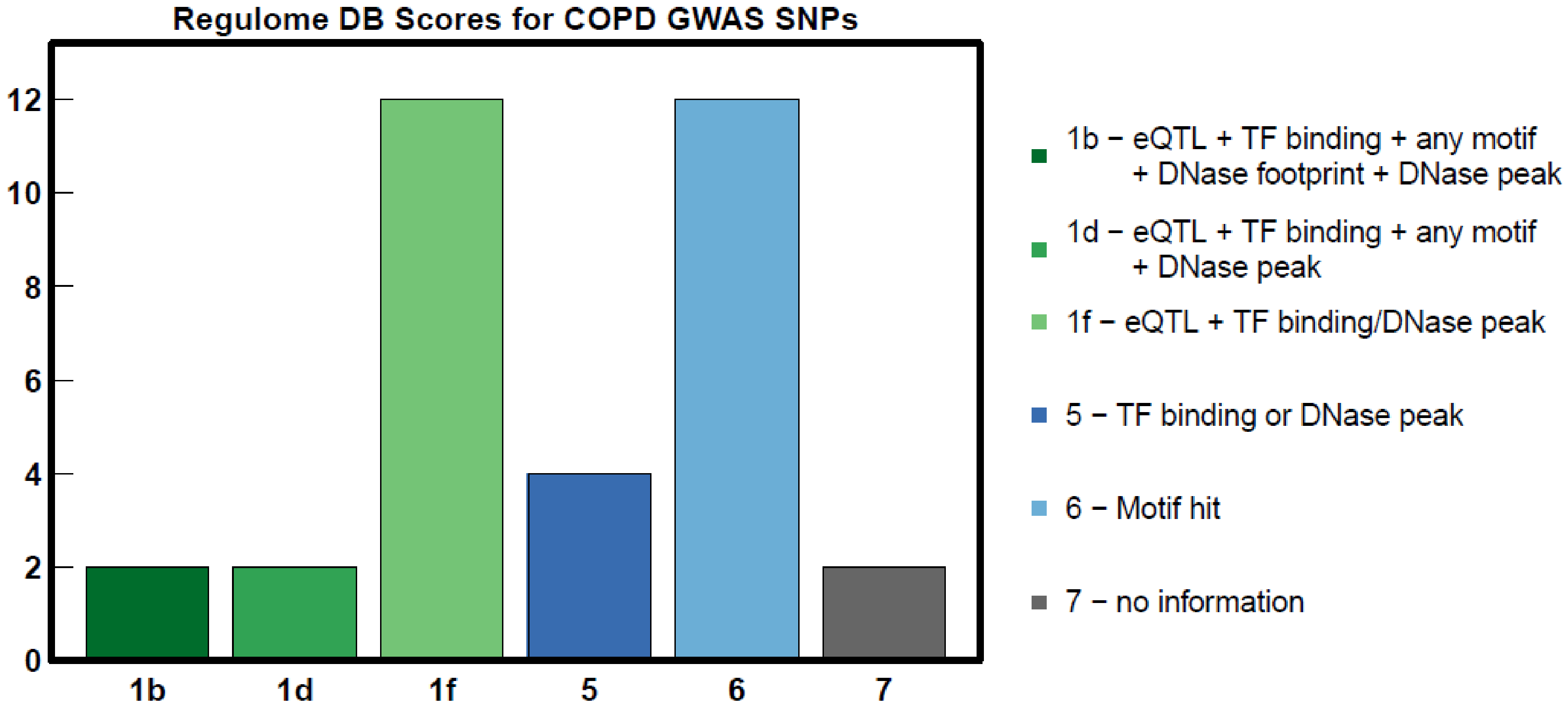




The median core score for GWAS SNPs is 1.7 times higher than the median for the non-GWAS SNPs

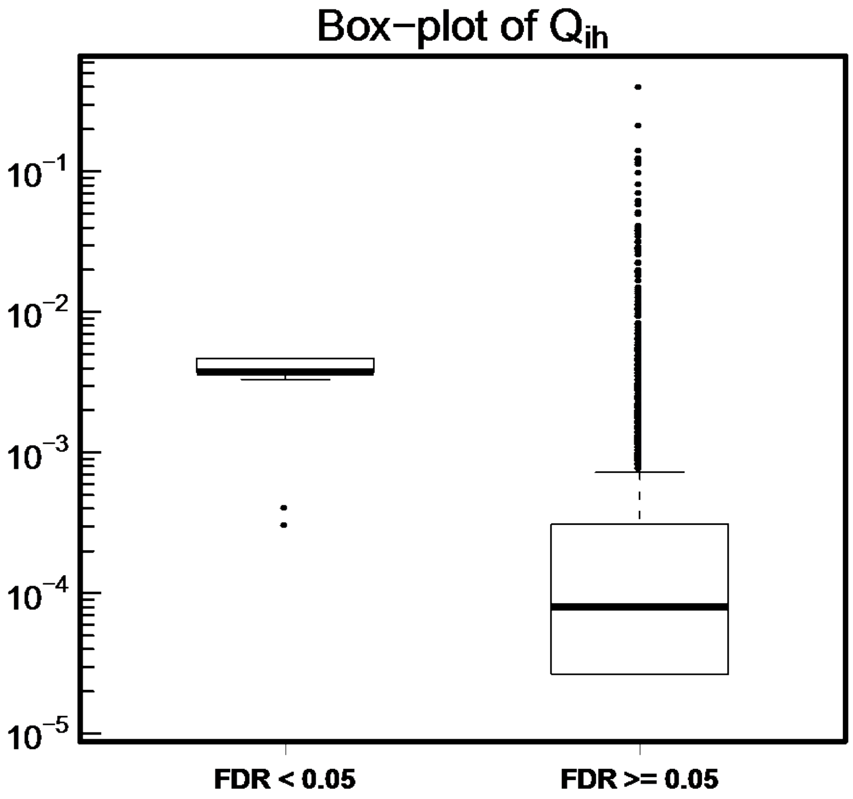
Are Disease SNPs in the eQTL Network functional?

- Map 34 COPD SNPs with GWAS p-values to the eQTL network
- These fell into communities that link to the etiology of COPD
- Of these, 32 had evidence of function based on RegulomeDB



Core Scores for COPD GWAS SNPs

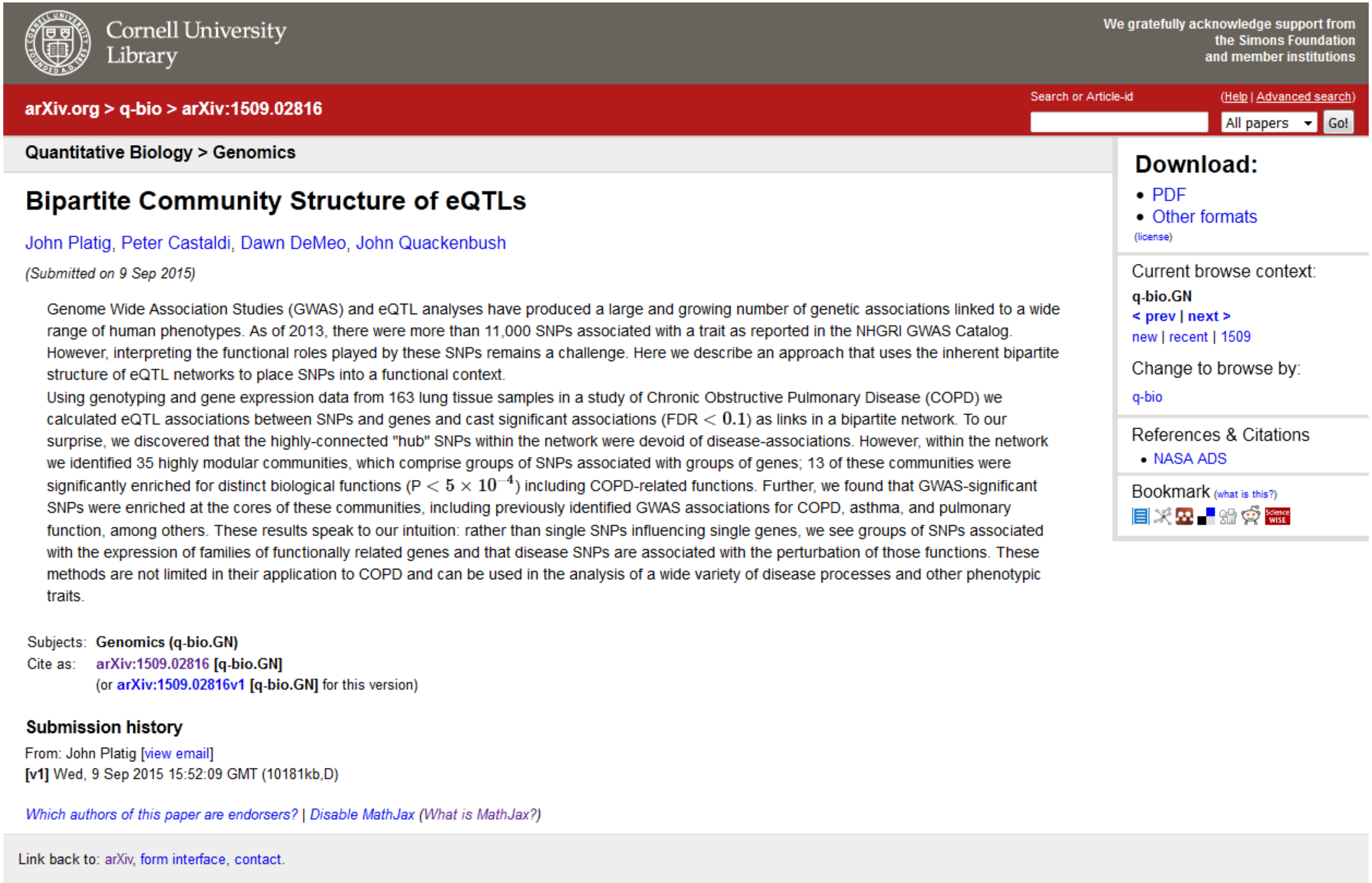
The median core score for the 34 FDR-significant GWAS SNPs is 47 times higher than the median for non-significant SNPs



Truthiness?

- The hubs are devoid of GWAS hits, consistent with strong selection against highly deleterious SNPs/survival bias
- Communities tell us a family of SNPs are associated with regulation of a process consistent with complex traits
- Many communities are apparently preserved across disease states, reflecting processes common to many cell types
- The Core SNPs are highly enriched for disease associations

Interested?



Cornell University Library

We gratefully acknowledge support from
the Simons Foundation
and member institutions

arXiv.org > q-bio > arXiv:1509.02816

Search or Article-id (Help | Advanced search)

All papers ▾ Go!

Quantitative Biology > Genomics

Bipartite Community Structure of eQTLs

John Platig, Peter Castaldi, Dawn DeMeo, John Quackenbush

(Submitted on 9 Sep 2015)

Genome Wide Association Studies (GWAS) and eQTL analyses have produced a large and growing number of genetic associations linked to a wide range of human phenotypes. As of 2013, there were more than 11,000 SNPs associated with a trait as reported in the NHGRI GWAS Catalog. However, interpreting the functional roles played by these SNPs remains a challenge. Here we describe an approach that uses the inherent bipartite structure of eQTL networks to place SNPs into a functional context.

Using genotyping and gene expression data from 163 lung tissue samples in a study of Chronic Obstructive Pulmonary Disease (COPD) we calculated eQTL associations between SNPs and genes and cast significant associations (FDR < 0.1) as links in a bipartite network. To our surprise, we discovered that the highly-connected "hub" SNPs within the network were devoid of disease-associations. However, within the network we identified 35 highly modular communities, which comprise groups of SNPs associated with groups of genes; 13 of these communities were significantly enriched for distinct biological functions (P $< 5 \times 10^{-4}$) including COPD-related functions. Further, we found that GWAS-significant SNPs were enriched at the cores of these communities, including previously identified GWAS associations for COPD, asthma, and pulmonary function, among others. These results speak to our intuition: rather than single SNPs influencing single genes, we see groups of SNPs associated with the expression of families of functionally related genes and that disease SNPs are associated with the perturbation of those functions. These methods are not limited in their application to COPD and can be used in the analysis of a wide variety of disease processes and other phenotypic traits.

Subjects: **Genomics (q-bio.GN)**

Cite as: arXiv:1509.02816 **[q-bio.GN]**

(or **arXiv:1509.02816v1 [q-bio.GN]** for this version)

Submission history

From: John Platig [view email]

[v1] Wed, 9 Sep 2015 15:52:09 GMT (10181kb,D)

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

Link back to: arXiv, form interface, contact.

Download:

- PDF
- Other formats

(license)

Current browse context:

q-bio.GN

< prev | next >

new | recent | 1509

Change to browse by:

q-bio

References & Citations

- NASA ADS

Bookmark (what is this?)

http://arxiv.org/abs/1509.02816; submitted to *Nature Genetics*

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

Systems Support

Stas Alekseev, Sys Admin

University of Queensland

Christine Wells
Lizzy Mason

Center for Cancer Computational Biology

Fieda Abderazzaq
Stas Alekseev
Jalil Farid
Nicole Flanagan
Ed Harms
Alex Holman
Lev Kuznetsov
Brian Lawney
Kshithija Nagulapalli
Antony Partensky
John Quackenbush
Renee Rubio
Yaoyu E. Wang

http://cccb.dfci.harvard.edu

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

Students and Postdocs

Martin Aryee
Stefan Bentink
Kimberly Glass
Benjamin Haibe-Kains
Marieke Kuijjer
Kaveh Maghsoudi
Jess Mar
Melissa Merritt
Megha Padi
John Platig
Alejandro Qiuiroz
J. Fah Sathirapongsasuti

Administrative Support

Julianna Coraccio

NATIONAL CANCER INSTITUTE

NATIONAL LIBRARY OF MEDICINE

National Heart Lung and Blood Institute

NSF