Network Inference in Biological Systems: Using Data to Discover Complex Relationships

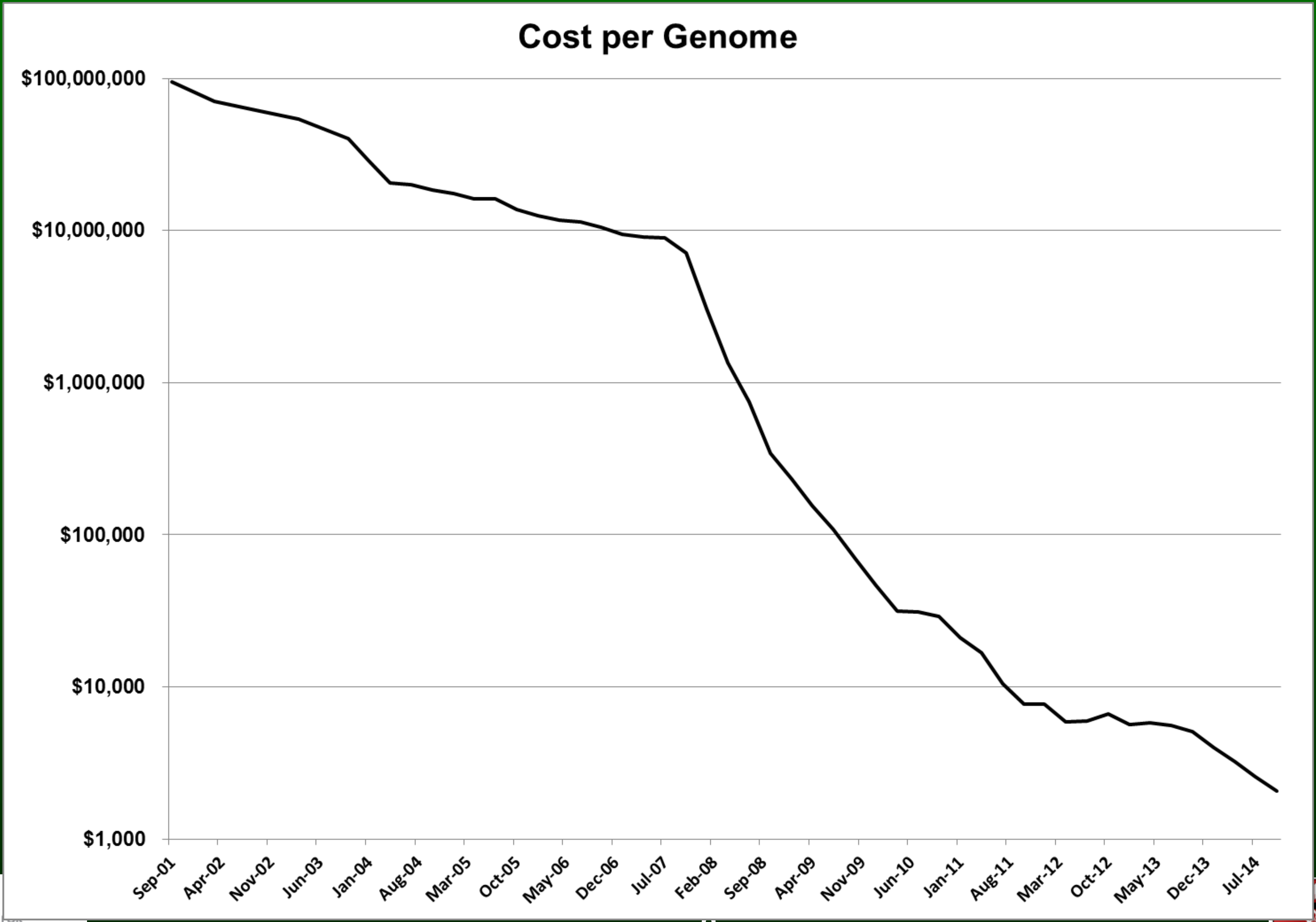
John Quackenbush
July 23, 2014

The Computational Biology and Functional Genomics Laboratory

at the Dana-Farber Cancer Institute and Harvard School of Public Health

VE RI TAS

Costs of Generating Data Have Plummeted



CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

eQTL Analysis

Use genome-wide SNP data 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?

eQTL Networks: A simple idea

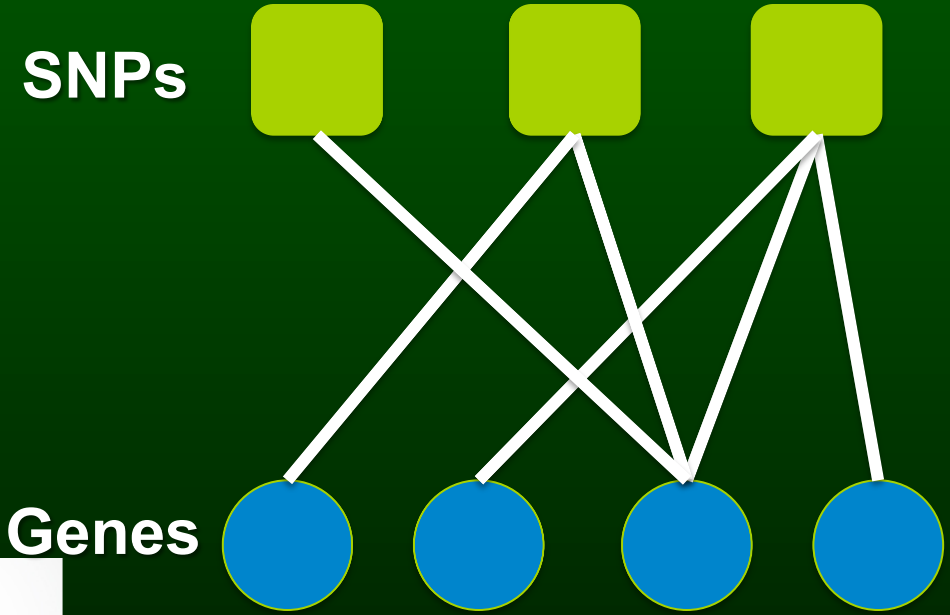
- eQTLs should group together with core SNPs regulating particular cellular functions
- Perform a “standard eQTL” analysis:

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

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

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

eQTL Networks: A simple idea

- Create a bipartite graph where SNPs and genes are nodes and significant eQTL associations are edges.
- Use “leading eigenvector” clustering to find “communities” in the graph

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

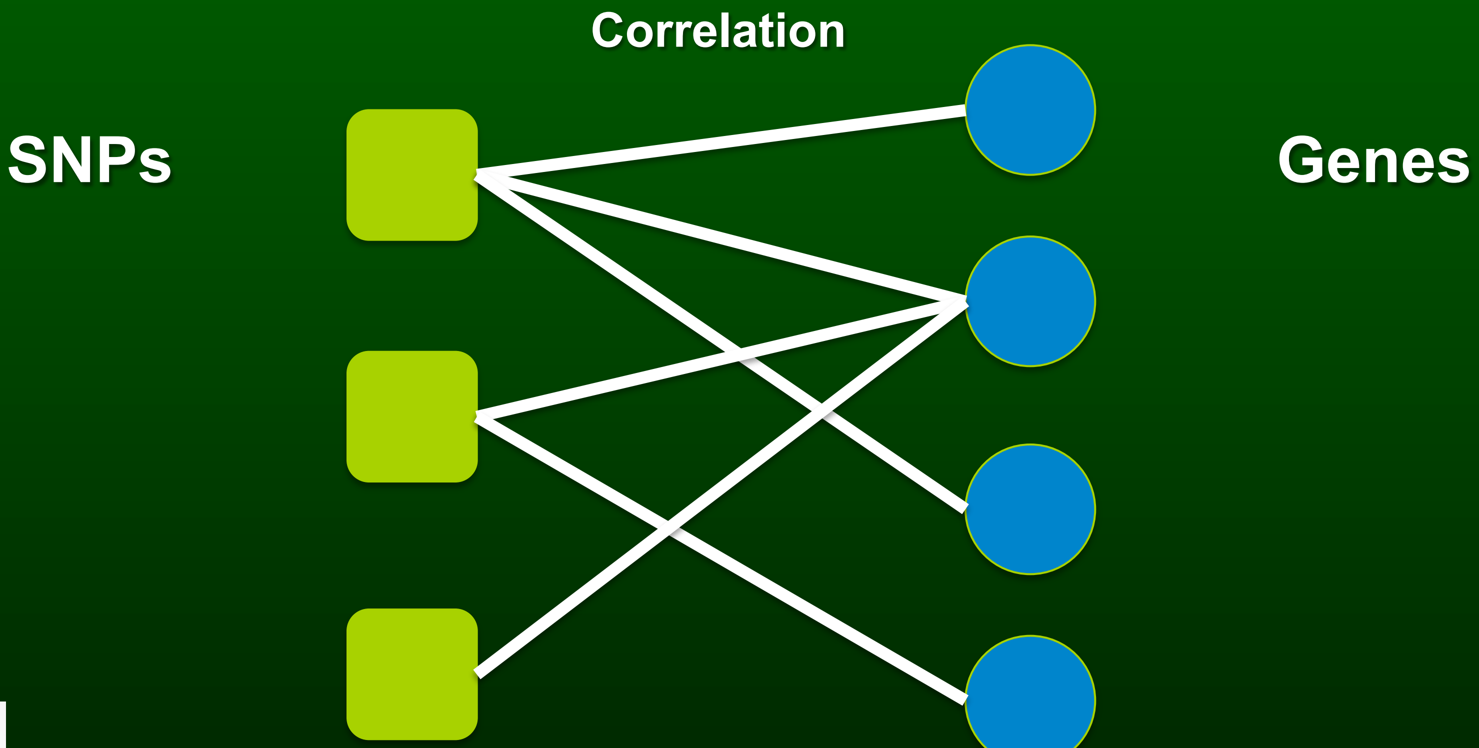
John Platig, Fah Sathirapongsasuti

VE RI TAS

A bipartite network has 2 types of node

Links only connect **different** node types

Node types: SNPs, Genes



CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

John Platig

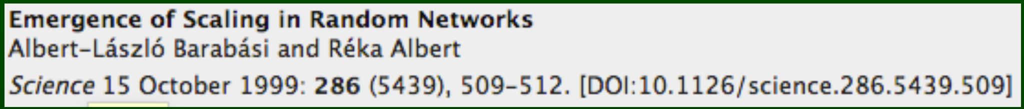
VE RI TAS

Background

- A quantity *x* obeys a power law if it is drawn from a probability distribution:

$$p(x) \propto x^{-a}$$

- Scale-free networks emerge through:
  - (1) expansion through addition of new vertices
  - (2) new vertices attach preferentially to sites that are already well-connected

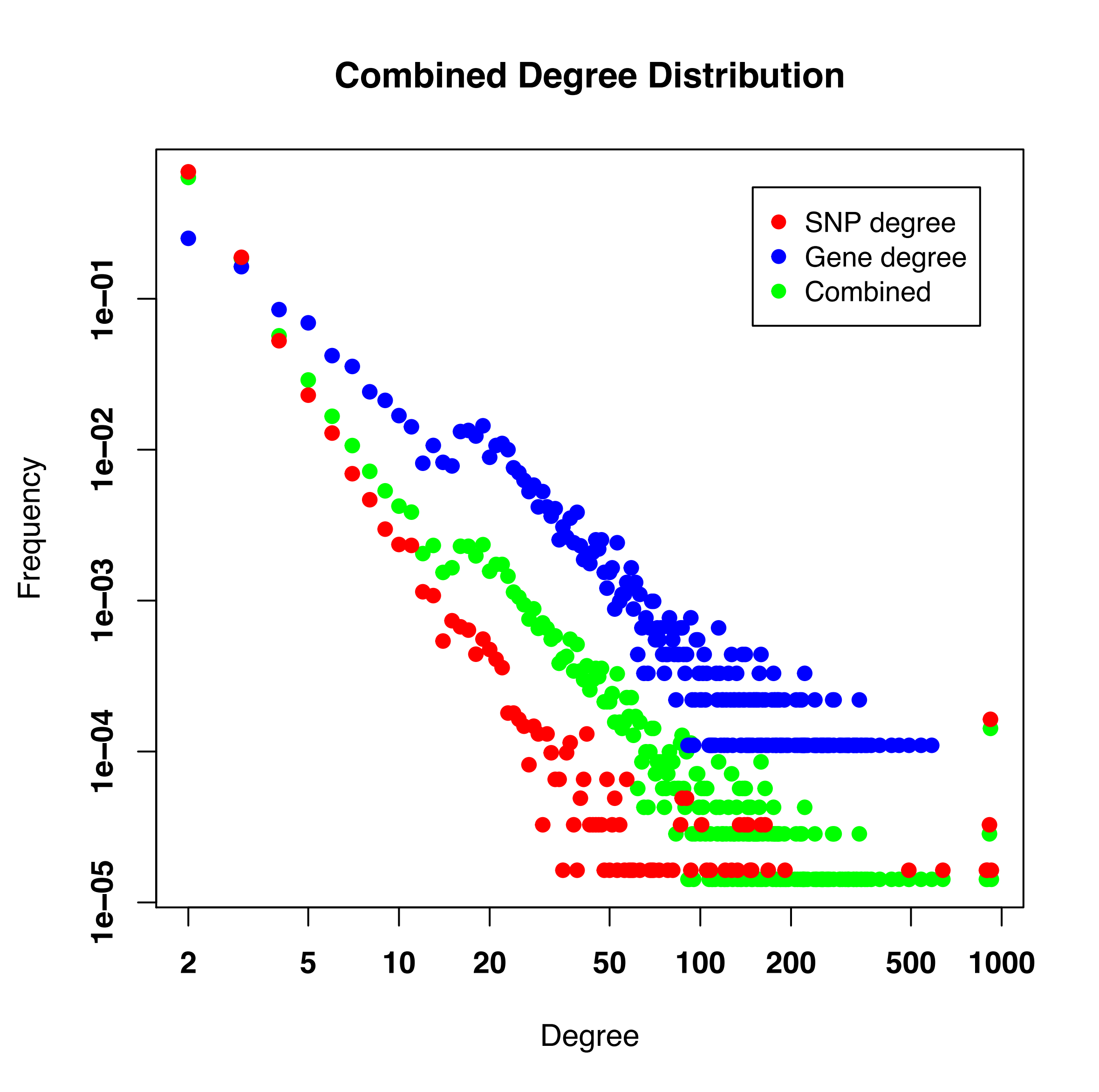


Emergence of Scaling in Random Networks
Albert-László Barabási and Réka Albert
Science 15 October 1999: **286** (5439), 509-512. [DOI:10.1126/science.286.5439.509]

- Hubs dominate the topology of scale-free networks

- eQTL hotspots are genomic regions that play an important role in regulating gene expression

Results: COPD



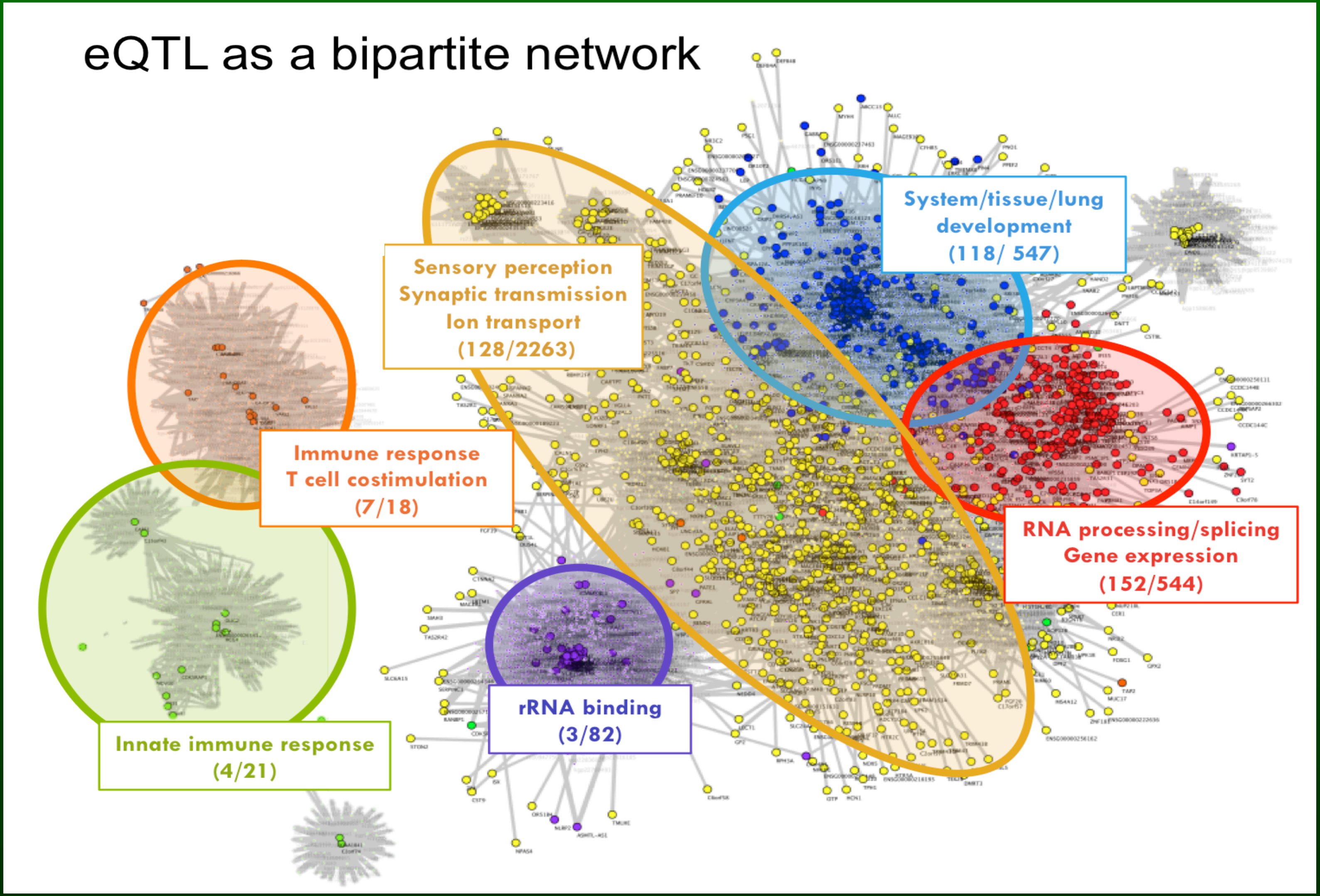
CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

VE RI TAS

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

eQTL Networks: A simple idea



eQTL as a bipartite network

Sensory perception
Synaptic transmission
Ion transport
(128/2263)

System/tissue/lung
development
(118/ 547)

Immune response
T cell costimulation
(7/18)

RNA processing/splicing
Gene expression
(152/544)

Innate immune response
(4/21)

rRNA binding
(3/82)

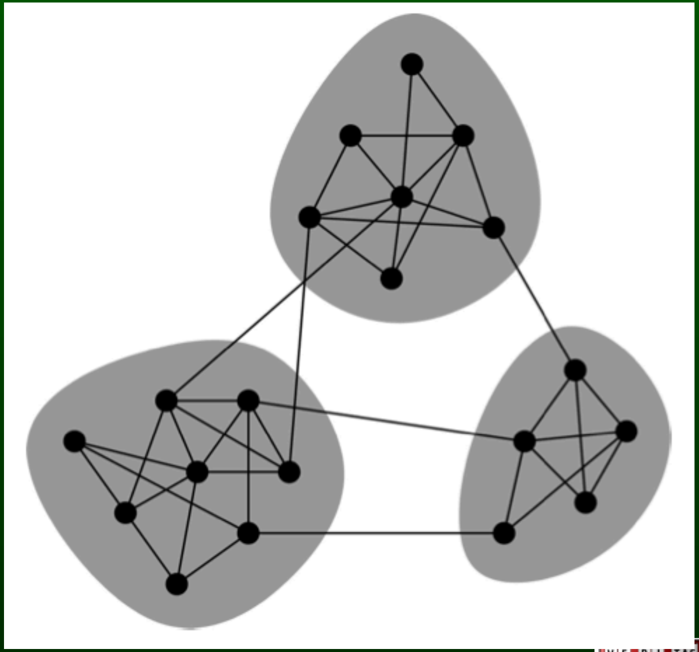
Communities are groups of highly intra-connected nodes

- 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)

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

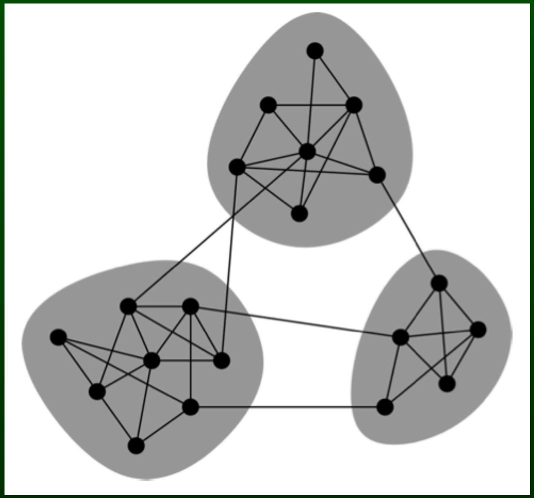
John Platig

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**.

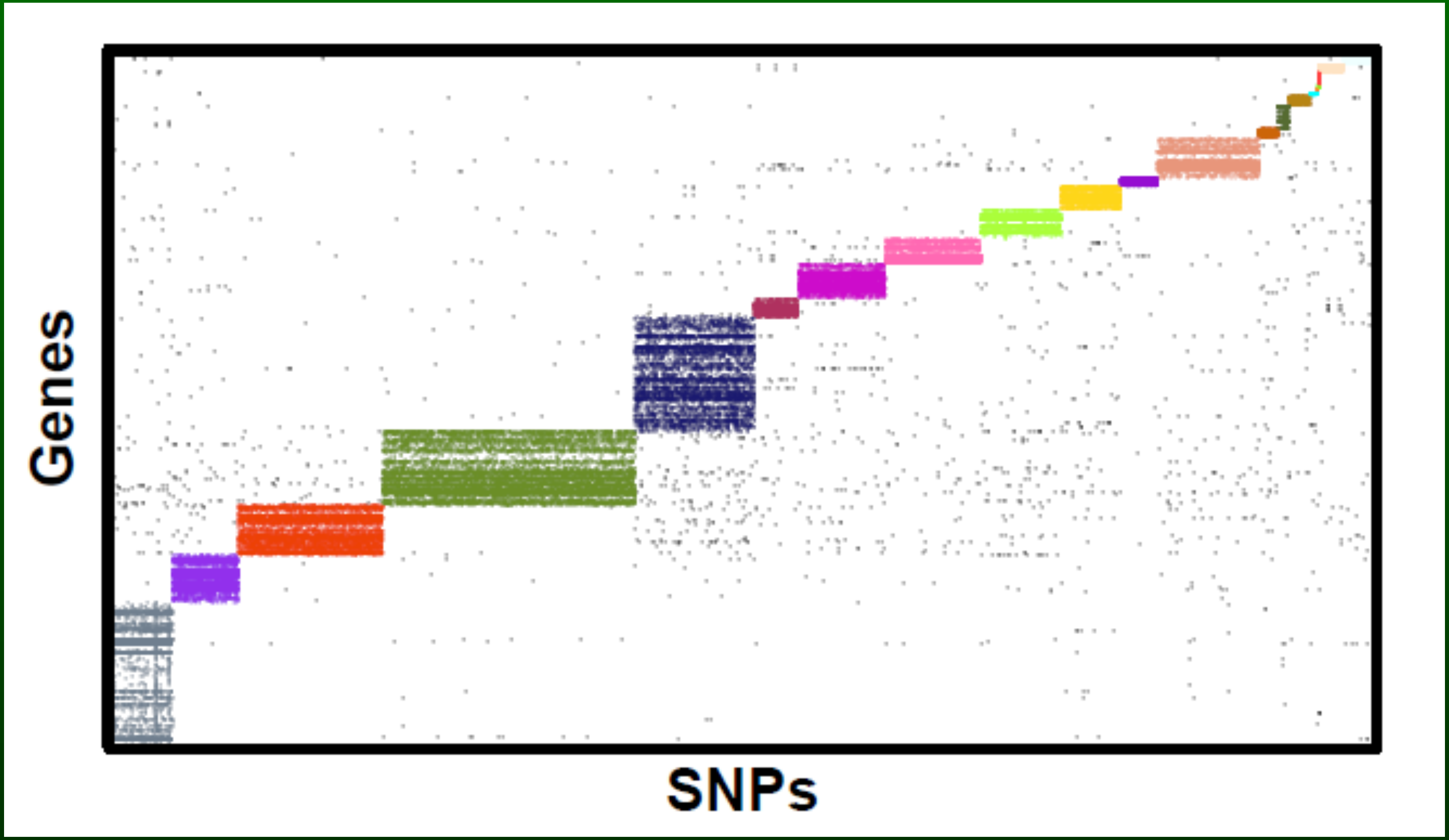
Bipartite networks require a different null model

Implement "BRIM" algorithm to find communities

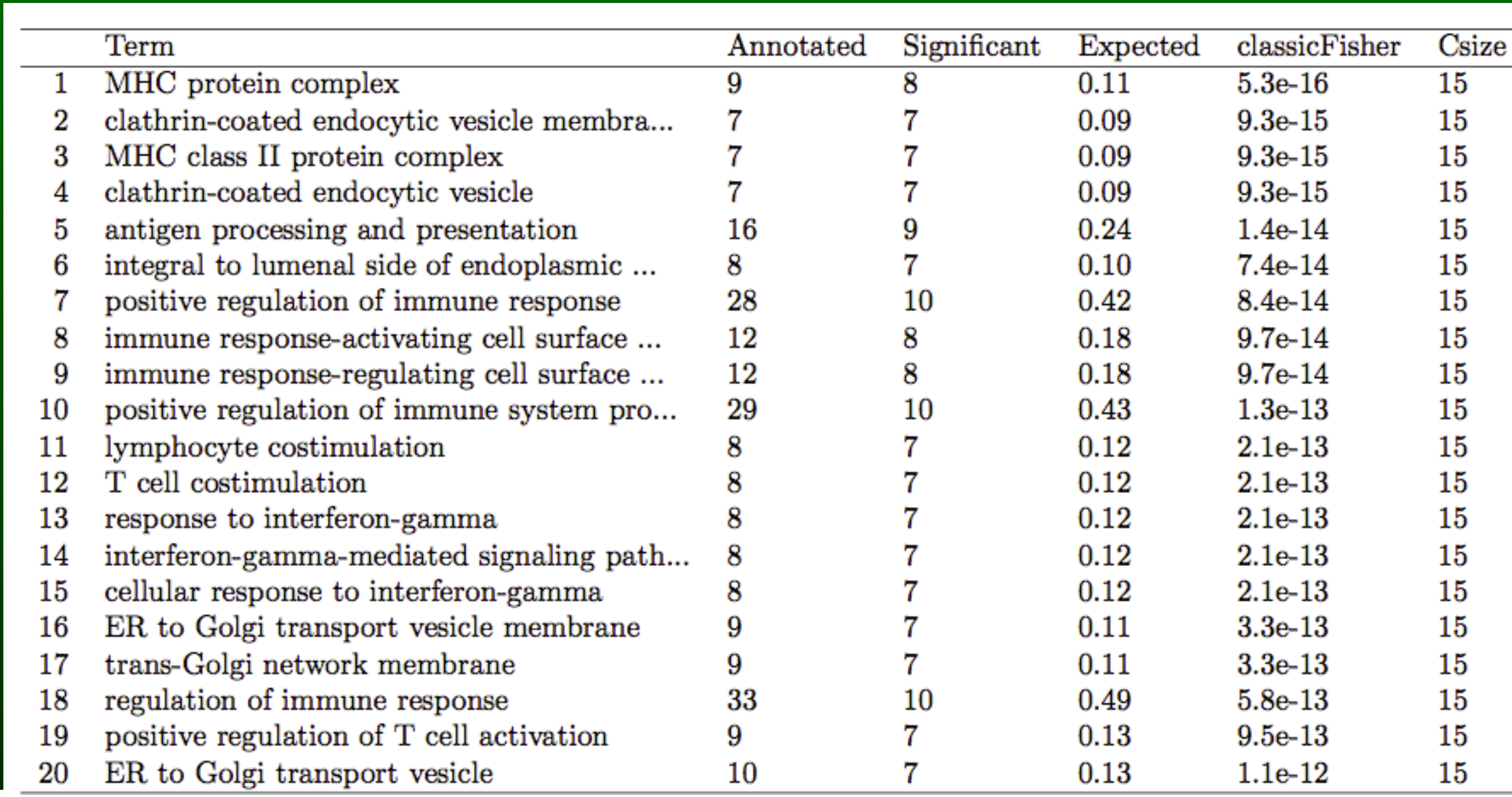


Newman 2006 (PNAS)

John Platig


Genes
SNPs

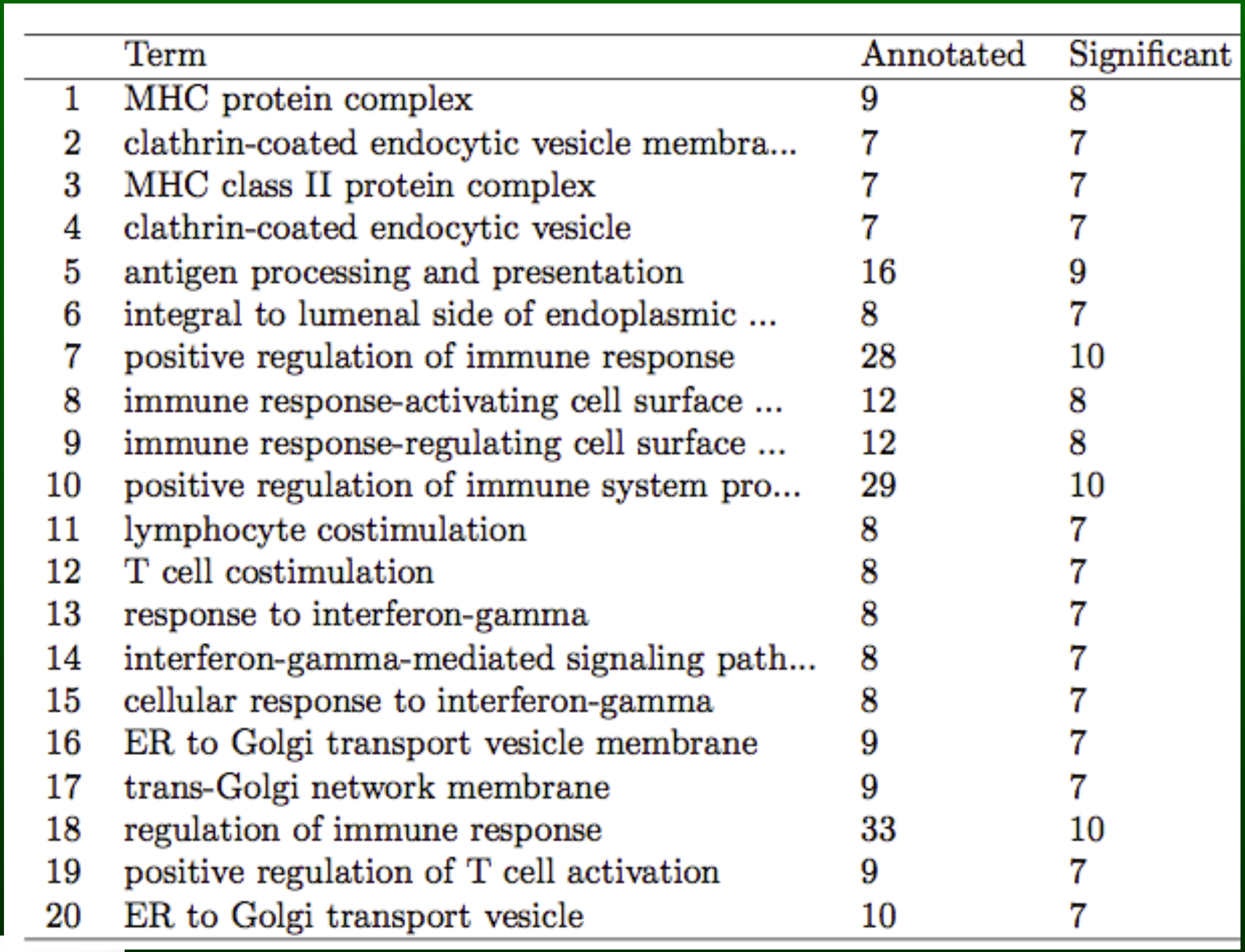
BRIM produces GO enriched Communities



	Term	Annotated	Significant	Expected	classicFisher	Csize
1	MHC protein complex	9	8	0.11	5.3e-16	15
2	clathrin-coated endocytic vesicle membra...	7	7	0.09	9.3e-15	15
3	MHC class II protein complex	7	7	0.09	9.3e-15	15
4	clathrin-coated endocytic vesicle	7	7	0.09	9.3e-15	15
5	antigen processing and presentation	16	9	0.24	1.4e-14	15
6	integral to lumenal side of endoplasmic ...	8	7	0.10	7.4e-14	15
7	positive regulation of immune response	28	10	0.42	8.4e-14	15
8	immune response-activating cell surface ...	12	8	0.18	9.7e-14	15
9	immune response-regulating cell surface ...	12	8	0.18	9.7e-14	15
10	positive regulation of immune system pro...	29	10	0.43	1.3e-13	15
11	lymphocyte costimulation	8	7	0.12	2.1e-13	15
12	T cell costimulation	8	7	0.12	2.1e-13	15
13	response to interferon-gamma	8	7	0.12	2.1e-13	15
14	interferon-gamma-mediated signaling path...	8	7	0.12	2.1e-13	15
15	cellular response to interferon-gamma	8	7	0.12	2.1e-13	15
16	ER to Golgi transport vesicle membrane	9	7	0.11	3.3e-13	15
17	trans-Golgi network membrane	9	7	0.11	3.3e-13	15
18	regulation of immune response	33	10	0.49	5.8e-13	15
19	positive regulation of T cell activation	9	7	0.13	9.5e-13	15
20	ER to Golgi transport vesicle	10	7	0.13	1.1e-12	15

John Platig

BRIM produces GO enriched Communities

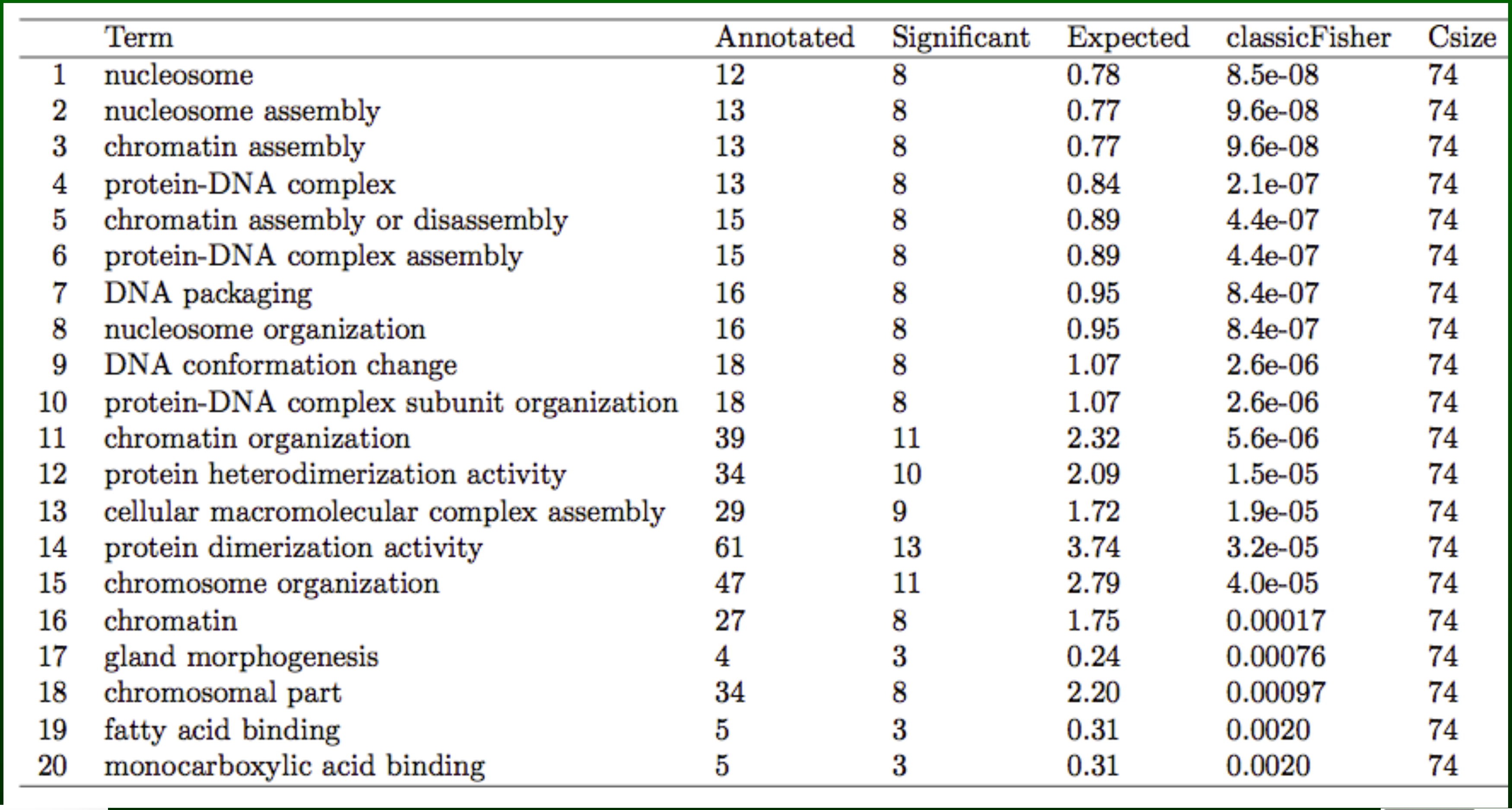


	Term	Annotated	Significant
1	MHC protein complex	9	8
2	clathrin-coated endocytic vesicle membra...	7	7
3	MHC class II protein complex	7	7
4	clathrin-coated endocytic vesicle	7	7
5	antigen processing and presentation	16	9
6	integral to lumenal side of endoplasmic ...	8	7
7	positive regulation of immune response	28	10
8	immune response-activating cell surface ...	12	8
9	immune response-regulating cell surface ...	12	8
10	positive regulation of immune system pro...	29	10
11	lymphocyte costimulation	8	7
12	T cell costimulation	8	7
13	response to interferon-gamma	8	7
14	interferon-gamma-mediated signaling path...	8	7
15	cellular response to interferon-gamma	8	7
16	ER to Golgi transport vesicle membrane	9	7
17	trans-Golgi network membrane	9	7
18	regulation of immune response	33	10
19	positive regulation of T cell activation	9	7
20	ER to Golgi transport vesicle	10	7

ATP6V1G2
ATRNL1
HLA-DQA2
HLA-DQB1
HLA-DQB2
HLA-DRA
HLA-DRB1
HLA-DRB4
HLA-DRB5
MAGEA2B
MICB
NCR3
PLEKHG6
PSORS1C1
TAP2

John Platig

BRIM produces GO enriched Communities



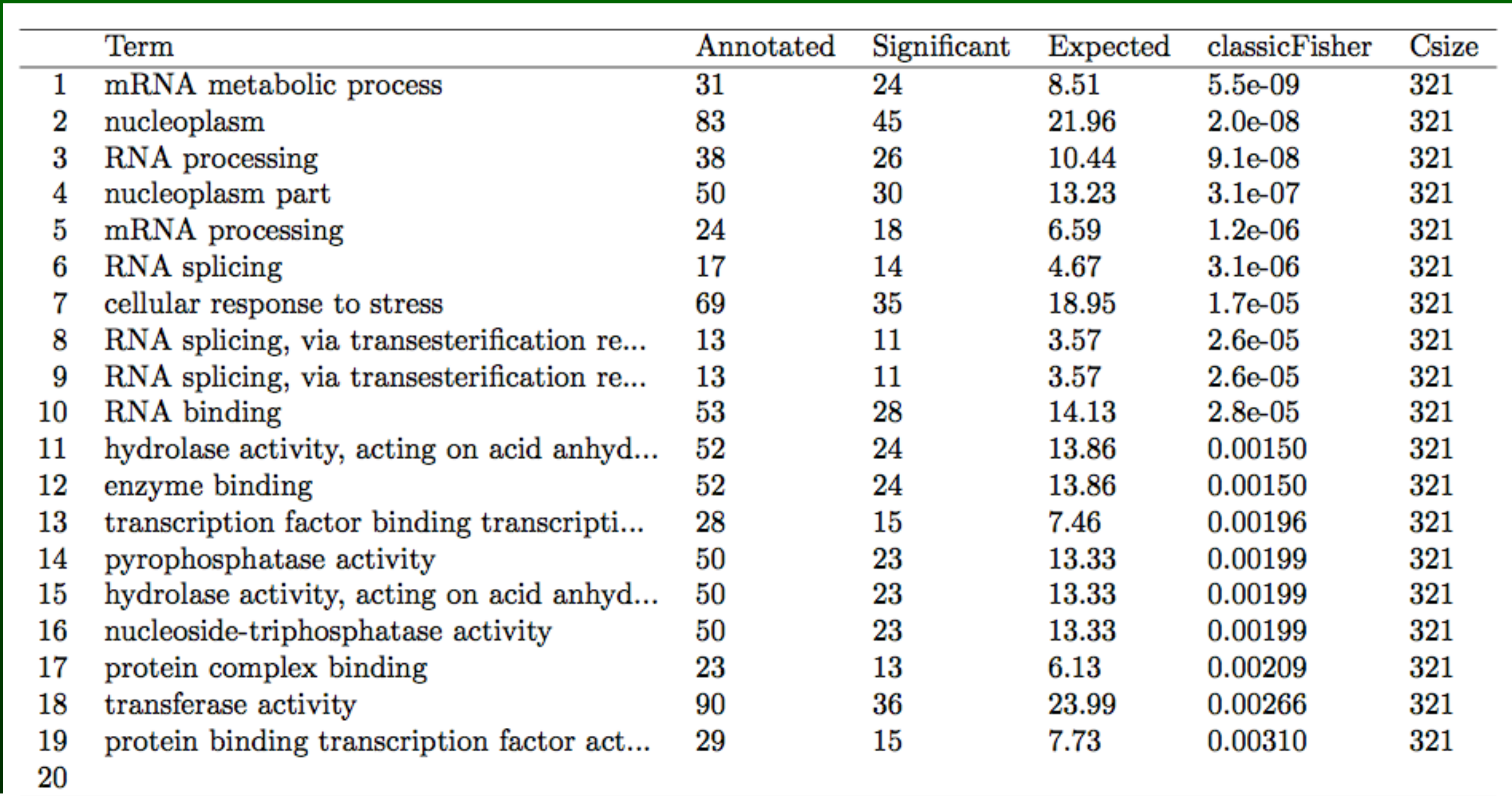
	Term	Annotated	Significant	Expected	classicFisher	Csize
1	nucleosome	12	8	0.78	8.5e-08	74
2	nucleosome assembly	13	8	0.77	9.6e-08	74
3	chromatin assembly	13	8	0.77	9.6e-08	74
4	protein-DNA complex	13	8	0.84	2.1e-07	74
5	chromatin assembly or disassembly	15	8	0.89	4.4e-07	74
6	protein-DNA complex assembly	15	8	0.89	4.4e-07	74
7	DNA packaging	16	8	0.95	8.4e-07	74
8	nucleosome organization	16	8	0.95	8.4e-07	74
9	DNA conformation change	18	8	1.07	2.6e-06	74
10	protein-DNA complex subunit organization	18	8	1.07	2.6e-06	74
11	chromatin organization	39	11	2.32	5.6e-06	74
12	protein heterodimerization activity	34	10	2.09	1.5e-05	74
13	cellular macromolecular complex assembly	29	9	1.72	1.9e-05	74
14	protein dimerization activity	61	13	3.74	3.2e-05	74
15	chromosome organization	47	11	2.79	4.0e-05	74
16	chromatin	27	8	1.75	0.00017	74
17	gland morphogenesis	4	3	0.24	0.00076	74
18	chromosomal part	34	8	2.20	0.00097	74
19	fatty acid binding	5	3	0.31	0.0020	74
20	monocarboxylic acid binding	5	3	0.31	0.0020	74

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

John Platig

VE RI TAS

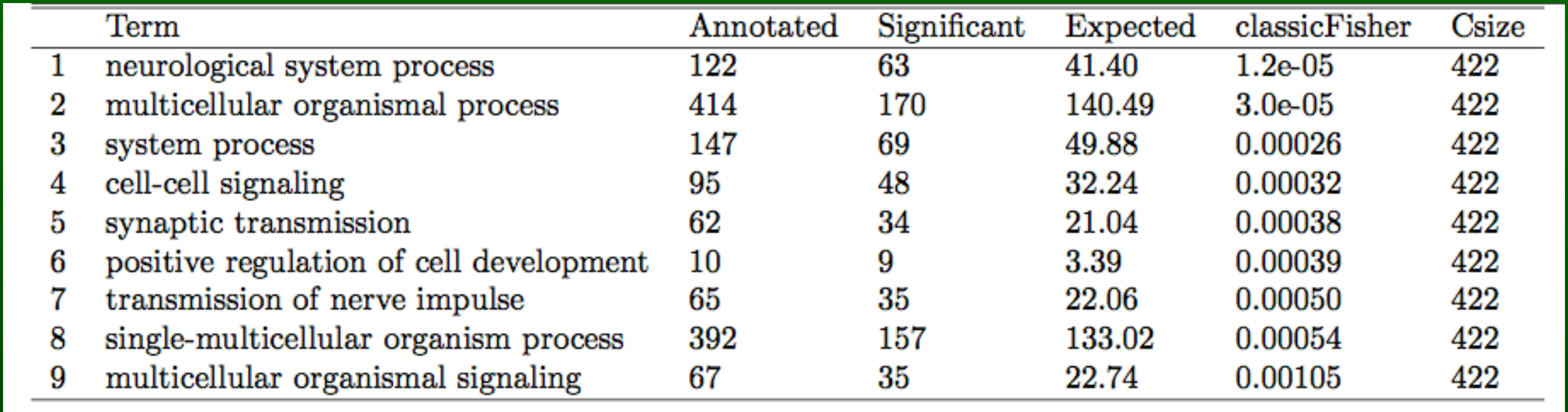
BRIM produces GO enriched Communities



	Term	Annotated	Significant	Expected	classicFisher	Csize
1	mRNA metabolic process	31	24	8.51	5.5e-09	321
2	nucleoplasm	83	45	21.96	2.0e-08	321
3	RNA processing	38	26	10.44	9.1e-08	321
4	nucleoplasm part	50	30	13.23	3.1e-07	321
5	mRNA processing	24	18	6.59	1.2e-06	321
6	RNA splicing	17	14	4.67	3.1e-06	321
7	cellular response to stress	69	35	18.95	1.7e-05	321
8	RNA splicing, via transesterification re...	13	11	3.57	2.6e-05	321
9	RNA splicing, via transesterification re...	13	11	3.57	2.6e-05	321
10	RNA binding	53	28	14.13	2.8e-05	321
11	hydrolase activity, acting on acid anhyd...	52	24	13.86	0.00150	321
12	enzyme binding	52	24	13.86	0.00150	321
13	transcription factor binding transcripti...	28	15	7.46	0.00196	321
14	pyrophosphatase activity	50	23	13.33	0.00199	321
15	hydrolase activity, acting on acid anhyd...	50	23	13.33	0.00199	321
16	nucleoside-triphosphatase activity	50	23	13.33	0.00199	321
17	protein complex binding	23	13	6.13	0.00209	321
18	transferase activity	90	36	23.99	0.00266	321
19	protein binding transcription factor act...	29	15	7.73	0.00310	321
20						

John Platig

BRIM produces GO enriched Communities



	Term	Annotated	Significant	Expected	classicFisher	Csize
1	neurological system process	122	63	41.40	1.2e-05	422
2	multicellular organismal process	414	170	140.49	3.0e-05	422
3	system process	147	69	49.88	0.00026	422
4	cell-cell signaling	95	48	32.24	0.00032	422
5	synaptic transmission	62	34	21.04	0.00038	422
6	positive regulation of cell development	10	9	3.39	0.00039	422
7	transmission of nerve impulse	65	35	22.06	0.00050	422
8	single-multicellular organism process	392	157	133.02	0.00054	422
9	multicellular organismal signaling	67	35	22.74	0.00105	422

John Platig

Calculate Local Connectivity

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

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

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*

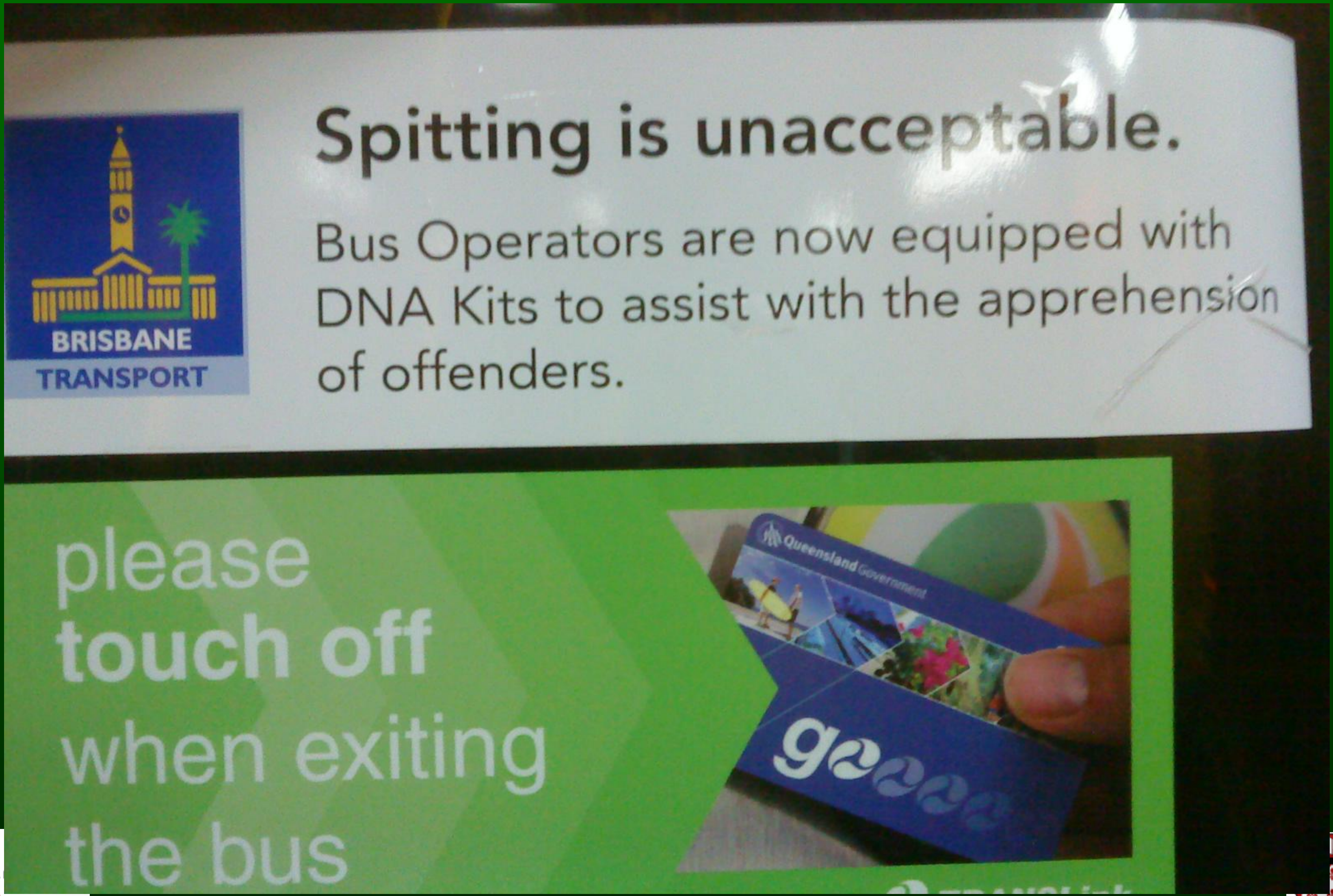
Community Structure Matters

- Are "disease" SNPs skewed towards the top of my SNP list as ranked by the overall out degree?
- No!
  - The highest-degree SNPs are devoid of disease-related SNPs
  - Highly deleterious SNPs that affect many processes are probably removed by evolutionary sweeps.

Community Structure Matters

- Are “disease” SNPs skewed towards the top of my SNP list as ranked by the community core score (Qic)?
- Yes!
  - KS test yields $p < 10^{-16}$,
  - wilcoxon rank-sum yields $p < 10^{-9}$

Genomics is here to stay



Spitting is unacceptable.

Bus Operators are now equipped with DNA Kits to assist with the apprehension of offenders.

please
touch off
when exiting
the bus

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

<johnq@jimmy.harvard.edu>

Array Software Hit Team
Eleanor Howe
John Quackenbush
Dan Schlauch

Gene Expression Team
Fieda Abderazzaq
Stefan Bentink
Aedin Culhane
Benjamin Haibe-Kains
Jessica Mar
Melissa Merritt
Megha Padi
Renee Rubio

University of Queensland
Christine Wells
Lizzy Mason

Center for Cancer Computational Biology
Dustin Holloway
Lan Hui
Lev Kuznetsov
Yaoyu Wang
John Quackenbush
http://cccb.dfci.harvard.edu

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

Students and Postdocs
Martin Aryee
Kimberly Glass
Marieke Kuijjer
Kaveh Maghsoudi
Jess Mar
Megha Padi
John Platig
Alejandro Qiuiroz
J. Fah Sathirapongsasuti

Systems Support
Stas Alekseev, Sys Admin

Administrative Support
Julianna Coraccio

ORACLE®

illumina®

http://compbio.dfci.harvard.edu

NATIONAL CANCER INSTITUTE

NATIONAL LIBRARY OF MEDICINE

National Heart Lung and Blood Institute

NSF

InforSense
The Integrative Analytics Company

VE RI TAS