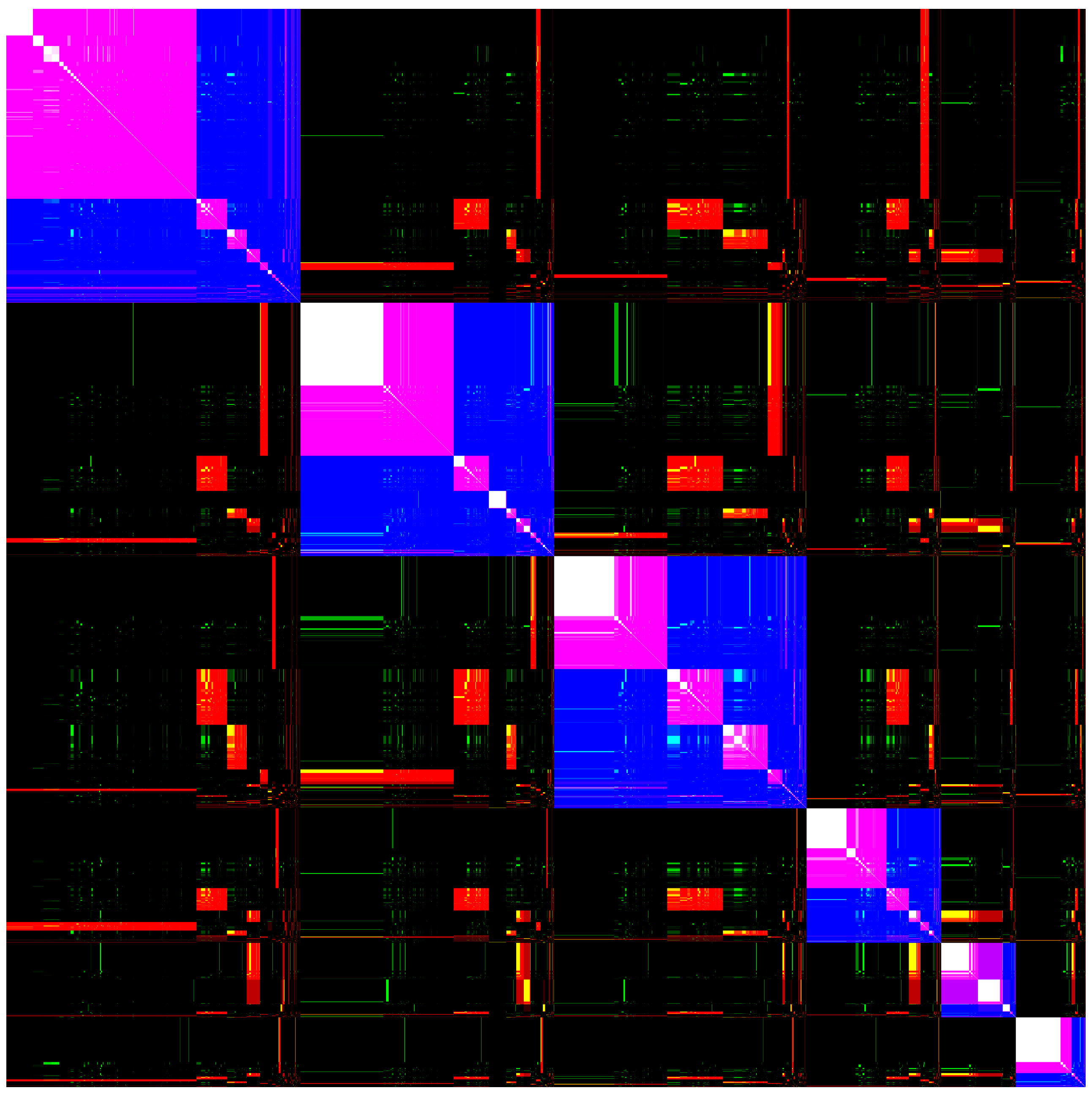
Robust Detection of Hierarchical Communities from *Escherichia coli* Gene Expression Data[1]

Santiago Treviño III†, Yudong Sun†, Tim F. Cooper‡ and Kevin E. Bassler †

†Department of Physics, University of Houston, Houston TX, USA ‡Department of Biology and Biochemistry, University of Houston, Houston TX, USA

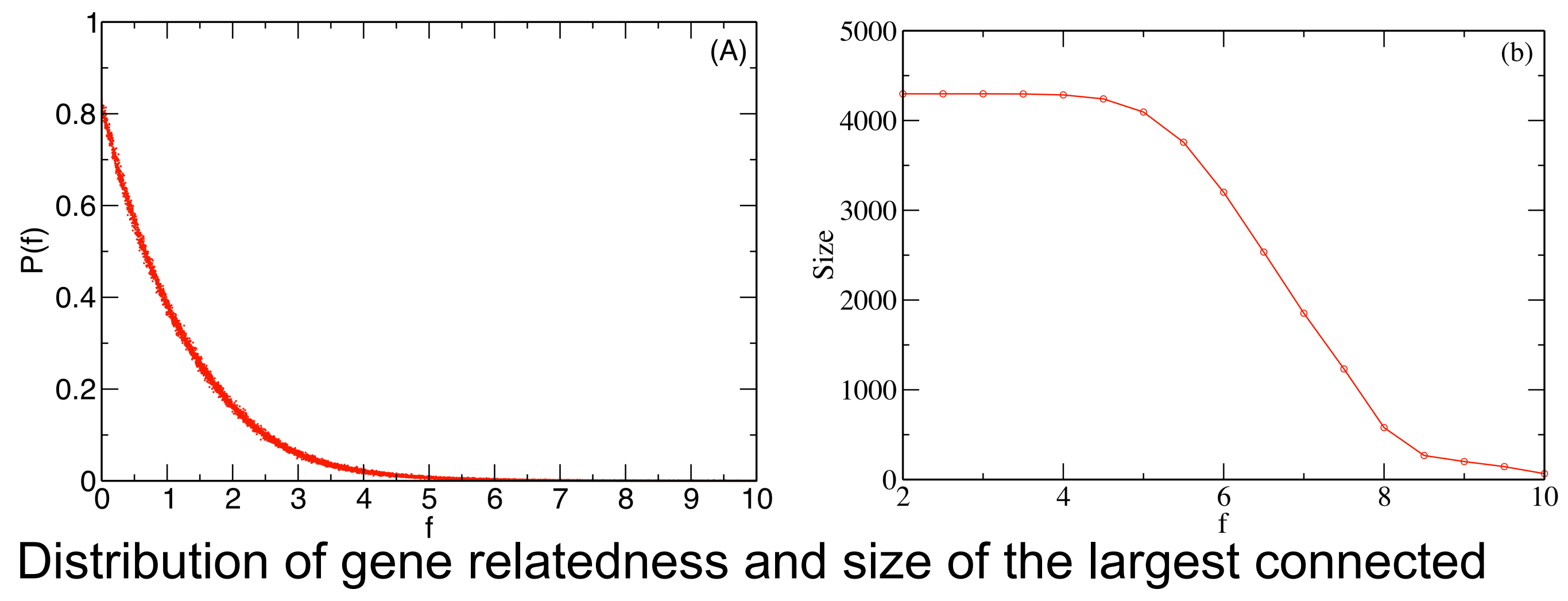
Background

One of the fundamental themes in biology is the hierarchical organization of its constituents. At higher levels of a hierarchy new properties emerge due to the complex interaction of constituents at lower levels. Determining if and how genetic regulatory networks are hierarchically structured would aid in understanding the properties and functional processes of the networks. With the increasing availability of genetic expression data, developing methods to infer and detect functional communities within the network is an important goal of systems biology. Unfortunately, noise in expression data creates variability in the inferred network and the stochastic nature of community detection creates variability in the functional communities detected with existing methods.



Correlation matrix showing community structure found in the *E. coli* network with relatedness threshold values f_{min} = 2,4, and 6. The matrix element in position (X,Y) is colored blue, red, or green if genes X and Y are in the same communities at threshold values 2, 4 or 6, respectively. The density of the color indicates the strength of the correlation in the ensemble. Additionally if two genes are found together at multiple threshold values the element is a combination of the colors assigned at each threshold value.

Inferring gene relatedness networks from expression data



Distribution of gene relatedness and size of the largest connected component in the *E. coli* CLR network

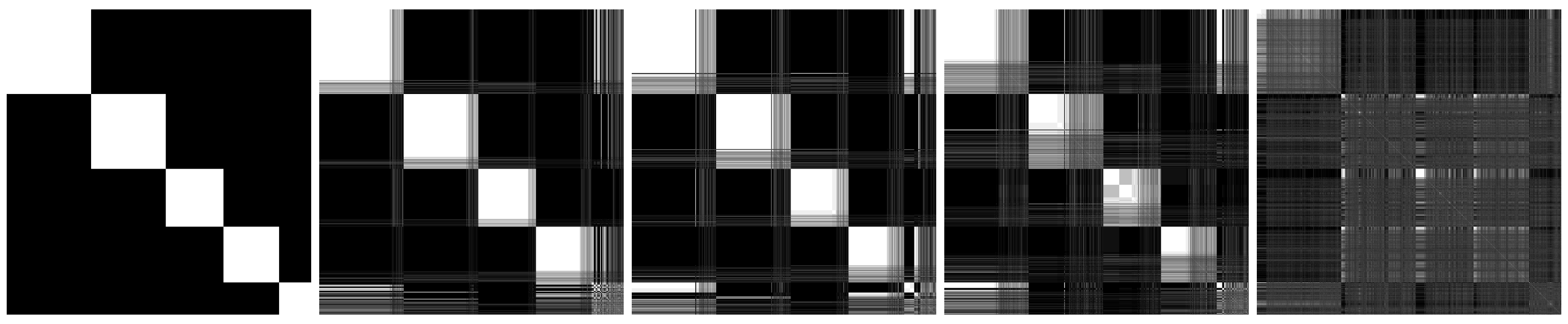
We used the CLR algorithm to infer direct and indirect regulatory interactions between genes based on the similarity of their expression response in 466 experiments in the M^{3D}. To create a network a link was placed between two genes if their corresponding relatedness value was greater than a chosen threshold value, f_{min}. We considered networks inferred from threshold values of f_{min} = 2, 4, and 6. These values correspond to points below, at, and above the *critical threshold value* at which the network is no longer one fully connected component.

Identifying communities and their hierarchical organization with an ensemble approach

To identify communities in each network we used an extension of the leading eigenvalue method that aims to identify a partitioning of nodes into a disjoint set that maximizes Modularity Q. Our extension of the LEM [2] uses a novel variant of the Kernighan – Lin algorithm. At each f_{min} value an ensemble of multiple network partitions was analyzed and a correlation matrix was created to visualize the overall hierarchical organization of the network. This ensemble allows overlapping communities to be identified. We define sets of genes that are always found in the same community as a *core community.*

Community structure is robust to experimental noise

To test the effect of noise on community structure we created several noisy datasets. Each experiment in the noisy dataset contained an expression level for gene X chosen randomly from a normal distribution with mean m(X) and standard error c∙σ(X). A correlation matrix for an ensemble of 10 community partitions detected at each noise level c and threshold f_{min}= 2 was created. We found that noise acts conservatively, decreasing the size of each core community rather than causing association of genes into new communities.



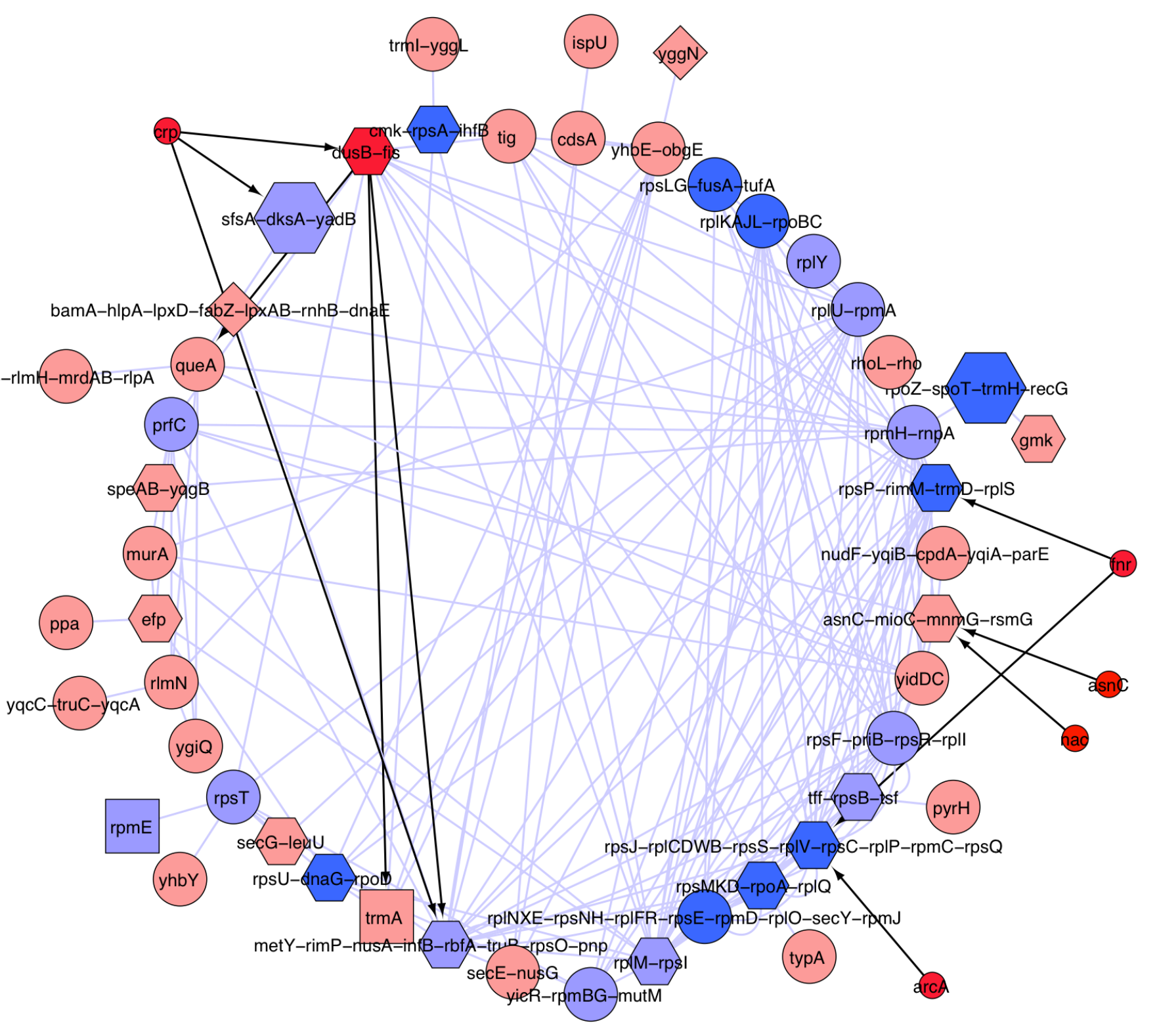
Change in core community structure as noise is increased from c=0 to c=4

Communities enrich for functionally related genes

P value	GO term num	Com size	GO size	In common	Description
8.41e-42	9288	72	24	24	bacterial-type flagellum
9.57e-39	6826	53	37	25	iron ion transport
8.22e-38	1539	72	28	24	ciliary or flagellar motility
3.67e-35	6412	826	101	79	translation
6.51e-34	3735	826	56	54	structural constituent of ribosome
3.08e-31	3723	826	105	77	RNA binding
1.73e-29	6935	72	22	19	chemotaxis
4.30e-29	3774	72	17	17	motor activity
5.38e-29	9425	72	17	17	bacterial-type flagellum basal body
2.06e-25	19861	72	15	15	flagellum
5.61e-25	5506	53	210	31	iron ion binding
3.72e-24	19843	826	42	40	rRNA binding
6.99e-22	30529	826	36	35	ribonucleoprotein complex
1.72e-21	5840	826	38	36	ribosome
6.62e-21	8652	247	62	32	cellular amino acid biosynthetic process
4.11e-17	5506	139	210	39	iron ion binding
6.66e-16	9055	139	116	29	electron carrier activity
7.30e-15	51539	139	98	26	4 iron, 4 sulfur cluster binding
8.22e-15	15453	300	15	15	oxidoreduction-driven active transmembrane transporter activity light-driven active transmembrane transporter activity
1.85e-13	6865	247	70	27	amino acid transport
6.13e-13	45272	300	13	13	plasma membrane respiratory chain complex I
9.19e-13	30964	300	13	13	NADH dehydrogenase complex
1.97e-12	9060	300	21	16	aerobic respiration
2.15e-12	5515	826	875	251	protein binding calmodulin binding

Top 25 statistically significant matches for f_{min} = 4

To test the biological relevance of the core communities found we compared the overlap of each core community to terms in the gene ontology using a hypergeometric test with Benjamini-Hochberg correction. We found 147, 239, and 288 statistically significant matches between core communities and Gene Ontology (GO) terms for communities identified at f_{min} values of 2, 4, and 6, respectively. Additionally core communities can be analyzed to identify candidate regularity interactions. For example, in the f_{min}=6 core community at right we found a high proportion of ppGpp sensitive promoters suggesting this molecule as a good candidate for regulating the remaining interactions.



An f_{min} = 6 core community

Publications and Support

[1] S. Trevino III , Y. Sun, T.F. Cooper, and K.E. Bassler, “Robust detection of hierarchical communities from Escherichia coli gene expression data,” PLoS Comput. Biol. 8, e1002391 (2012).

[2] Y. Sun, B. Danila, K. Josic, and K.E. Bassler, “Improved community structure detection using a modified fine-turning strategy” EPL 86, 28004 (2009).

The work of ST, YS, and KEB was supported by US NSF grant DMR-0908286, and the work of TC was supported by US NSF grant IOS-1022373 and DARPA grant HR001-09-1-0055.