Functionalization of a protosynaptic gene expression network

Cecilia Conaco[a,b,1], Danielle S. Bassett[c,d,1], Hongjun Zhou[a,b], Mary Luz Arcila[a,b], Sandie M. Degnan[e], Bernard M. Degnan[e], and Kenneth S. Kosik[a,b,2]

[a]Neuroscience Research Institute, [b]Department of Molecular, Cellular and Developmental Biology, [c]Department of Physics, and [d]Sage Center for the Study of the Mind, University of California, Santa Barbara, CA 93106; and [e]School of Biological Sciences, University of Queensland, Brisbane, Queensland 4072, Australia

Edited by Francisco J. Ayala, University of California, Irvine, CA, and approved April 4, 2012 (received for review February 13, 2012)

Assembly of a functioning neuronal synapse requires the precisely coordinated synthesis of many proteins. To understand the evolution of this complex cellular machine, we tracked the developmental expression patterns of a core set of conserved synaptic genes across a representative sampling of the animal kingdom. Coregulation, as measured by correlation of gene expression over development, showed a marked increase as functional nervous systems emerged. In the earliest branching animal phyla (Porifera), in which a nearly complete set of synaptic genes exists in the absence of morphological synapses, these "protosynaptic" genes displayed a lack of global coregulation although small modules of coexpressed genes are readily detectable by using network analysis techniques. These findings suggest that functional synapses evolved by exapting preexisting cellular machines, likely through some modification of regulatory circuitry. Evolutionarily ancient modules continue to operate seamlessly within the synapses of modern animals. This work shows that the application of network techniques to emerging genomic and expression data can provide insights into the evolution of complex cellular machines such as the synapse.

synapse evolution | community detection | developmental transcriptome | *Amphimedon queenslandica*

In the tree of life, sponges (Porifera), generally recognized as the oldest surviving metazoan phyletic lineage (Fig. 1*B*), occupy a highly informative position for understanding the evolution of features that uniquely characterize animals (1). The synapse, a cellular machine formed through the dynamic assembly of multiple proteins that together perform a specific biological function, is one such metazoan specialization. The synaptic machinery delivers a chemical signal via vesicle fusion at the presynaptic neuronal membrane to postsynaptic receptors, which convert that signal back to an electrical impulse in the postsynaptic neuronal cell. Surprisingly, the genome of the Poriferan demosponge, *Amphimedon queenslandica*, contains an almost complete set of genes homologous to those found in mammalian synapses (Fig. 1*A*), although the organism does not assemble any structure morphologically resembling a synapse (1, 2). Although limited gene innovation and the invention of new protein interaction sites can partially explain how preexisting genes came together to form the synaptic complex (3), the multiple evolutionary steps involved in building a cellular machine through the assembly of an interaction network that can operate as a unit with a discrete biological function remains unknown.

Changes in conserved transcriptional programs arising from modification of instructions encoded in the genome have contributed to our understanding of animal evolution (4–7). Specific patterns of expression can define discrete tissues, cell types, and even functional protein complexes. Genes with similar expression patterns often have similar function (8). Furthermore, when comparing orthologues across divergent species, highly conserved coexpression is a strong predictor of shared function in similar pathways (9–11). These results suggest that functionally related genes might be under similar expression constraints (12). Thus, changes in coexpression relationships for any group of genes may contain information on the assembly and evolution of cellular machines. To understand the evolutionary transition leading to the emergence of a functional synapse, we used network analysis to identify unique patterns of synaptic gene coexpression in representative species from diverse phylogenetic positions. We show that "protosynaptic" genes have an inherent modular structure and that the coregulatory links between these modules characterize species with functional synapses. In contrast, ancient eukaryotic cellular machines, such as the proteasome and nuclear pore, already operate in early metazoans, and their associated genes display highly correlated expression patterns over development. These findings suggest that reorganization of gene expression, most likely through the modification of transcriptional regulation, was a key factor in the evolution of cellular machines such as the synapse.

Results

To study functionalization of the synaptic gene network (Fig. 2*A* and Fig. S1*A*), we obtained the expression profiles of sponge synaptic gene homologues by sequencing the *A. queenslandica* transcriptome at four developmental stages from larva to adult. For comparison, expression data were also obtained for the same set of synaptic genes from five representative animals with varying complexities in tissue organization (Fig. 1*B*). Animal species included in this study were the cnidarian coral, *Acropora millepora*; invertebrate bilaterians, *Caenorhabditis elegans* (nematode) and *Drosophila melanogaster* (arthropod); and vertebrates, *Danio rerio* (zebrafish) and *Xenopus tropicalis* (frog) (13–17). The correlation matrix for synaptic gene homologues from each species was constructed by computing the Pearson correlation coefficient between all pairs of gene expression profiles across development (Fig. 3*A*). The correlation matrix represents a network in which the genes are nodes and the correlations between gene expression patterns are edges. We averaged all elements of the correlation matrix to obtain a measure of connectivity or coregulation, R (Fig. 3 *B*, *D*, *F*, and *H*). By using a community detection algorithm (18–20), the modularity, Q, of each network was computed by determining the

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "In the Light of Evolution VI: Brain and Behavior," held January 19–21, 2012, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and audio files of most presentations are available on the NAS Web site at www.nasonline.org/evolution_vi.

Author contributions: K.S.K. designed research; C.C., H.Z., and M.L.A. performed research; D.S.B., S.M.D., and B.M.D. contributed new reagents/analytic tools; C.C., D.S.B., and K.S.K. analyzed data; and C.C., D.S.B., S.M.D., B.M.D., and K.S.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE29978).

[1]C.C. and D.S.B. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: kosik@lifesci.ucsb.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1201890109/-/DCSupplemental.

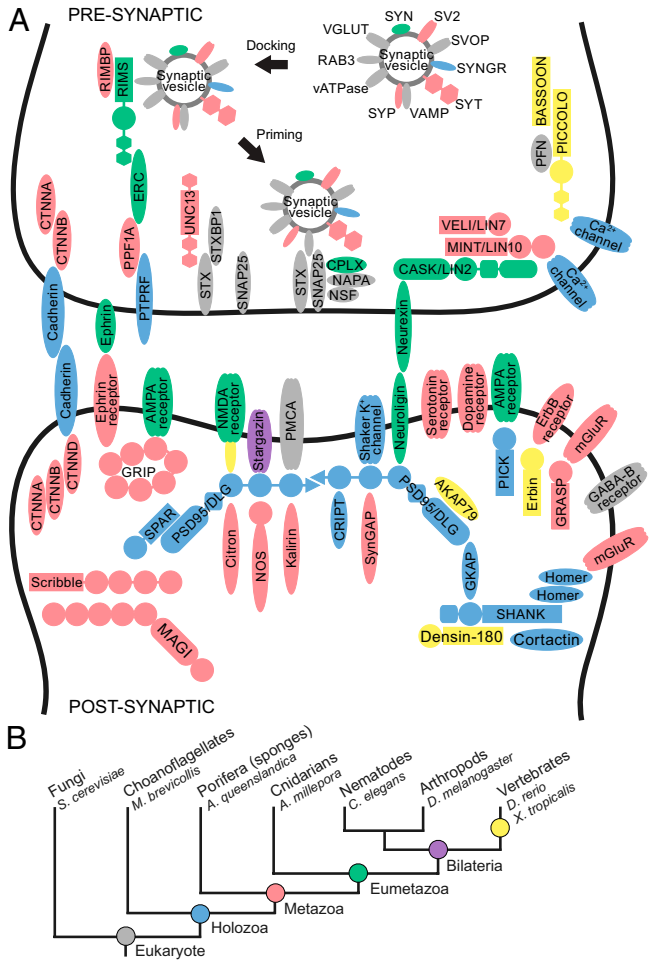

Fig. 1. Origins of synaptic genes. (*A*) Homologues of genes in the human synaptic complex were identified in the genomes of selected organisms representing key phylogenetic steps in animal evolution. Colors indicate the inferred ancestor of origin for each gene, as indicated in *B*. (*B*) Evolutionary relationships among animal phyla. The names of representative species are shown.

optimal partition of the network into communities whose nodes were more connected to other nodes inside of their own community than expected in a random null model (Fig. 3 *C*, *E*, *G*, and *I*). The modularity, Q, can be interpreted as a measure of the cohesiveness of coregulation: higher Q values indicate more segregation between coregulated groups. To determine the statistical significance of our results, we computed the same properties (R and Q) for various random control models.

The synaptic gene expression profiles were more highly correlated in eumetazoan species than in the sponge (Fig. 3*B*). This is apparent in the cnidarian coral, *A. millepora*, which possesses nerve cells organized into a simple diffuse net. The bilaterian synaptic gene networks showed even greater coregulation compared with sponge or coral. Synaptic genes showed significantly increased correlation compared with permuted and random controls in all species (Fig. 3*B* and Table S1). To verify the observed differences in expression coregulation, we performed pairwise comparisons of subsets of synaptic genes common between species. Comparison of genes found in sponge and the other five species showed that the increased correlation in eumetazoans was significant ($P < 1 \times 10^{-5}$, two-tailed t test; Table S2). Pairwise comparison of average coregulation for genes common between coral and each of the other species further revealed significantly greater correlation in bilaterian organisms ($P < 1 \times 10^{-10}$, two-tailed t test).

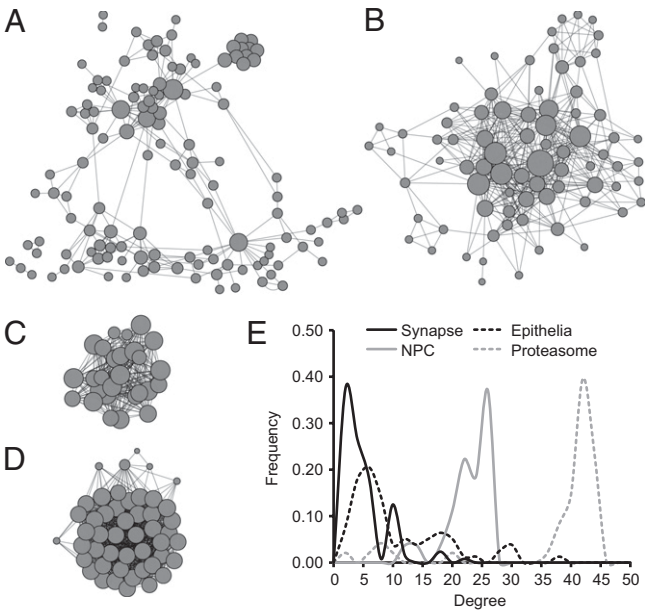

Fig. 2. Structure of protein interactions within the (*A*) synaptic, (*B*) epithelial, (*C*) NPC, and (*D*) 26S proteasome networks. Each node represents a gene. Node size represents the number of interactions formed by a protein and edge length is proportional to the strength of evidence for a functional link between two proteins. Network structures are based on the human interactome annotated in STRING (37) and visualized by using Cytoscape (38). (*E*) Degree distribution patterns of gene networks based on the human interactome. The frequency of nodes that exhibit the indicated number of connections (degree) is shown.

These pairwise correlation values were significantly greater than coregulation within three separate random control models ($P < 0.05$, two-tailed t test; *Materials and Methods*). However, Q values for most of the synaptic gene networks did not show the consistent decrease relative to controls that would be expected in a set of genes that were coherently coregulated. This suggests that the synaptic gene network is composed of subsets of genes with distinguishable differences in their developmental expression patterns, similar to what we would expect from a random collection of genes taken from the transcriptome. These distinct modules may be performing disparate activities that are necessary for the overall function of the synaptic machinery (Fig. 3*C* and Table S1).

The detection of coregulated gene communities is a data-driven process that is not biased by any prior knowledge of function. We sought to determine whether functionally defined subsets of synaptic proteins corresponded to the gene communities found in the coregulation modules. Nodes in the synaptic protein interaction network of each species were colored according to the coregulation module from which they were derived (Fig. 4*A*). Module composition (i.e., node colors) of the three largest functional complexes were tabulated (Fig. 4*B*). Those genes which comprise the postsynaptic density tended to fall within a single module for most eumetazoans. This same tendency was also true for the synaptic vesicle genes in most bilaterians. In contrast, sponge synaptic genes in these functional complexes showed a more heterogenous expression pattern that appeared to follow a different regulatory logic than that of functional synaptic networks, as reflected by the greater diversity in module composition within each biological complex. One striking exception is the vacuolar ATPase complex (vATPase), which is tightly coregulated even in sponge, suggesting a gain of functionality long before animal divergence (21). It should be noted, however, that, although we did not see similar module enrichment patterns for these functional complexes in the frog, we did observe a strong correlation of synaptic gene expression in this species (Fig. 3*A*).

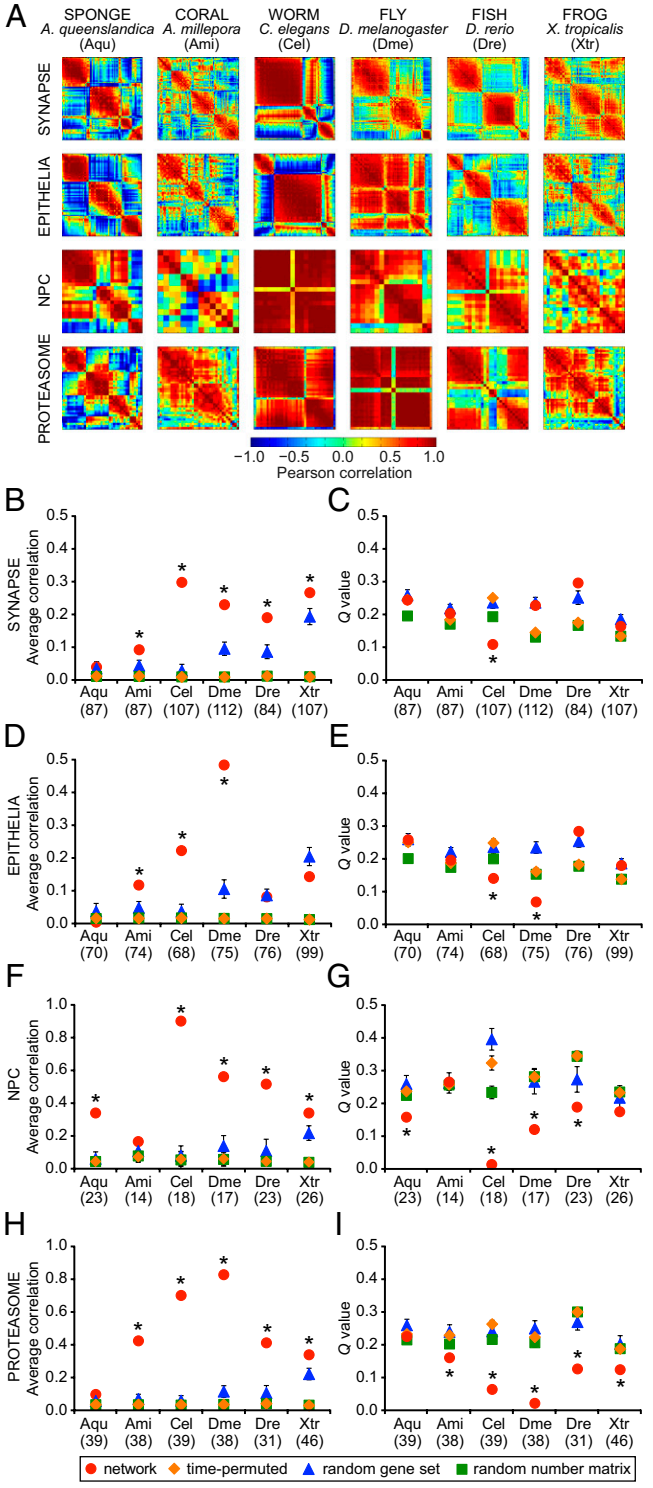

Fig. 3. Correlation and modularity analysis for gene networks in six organisms. (*A*) The strength of genetic coregulation for any two genes in a network was estimated by computing the Pearson correlation coefficient of their expression across developmental stages. Heat maps represent *N*×*N* correlation matrices for genes in each network in each species (red, positive correlation; blue, negative correlation). (*B*, *D*, *F*, and *H*) Average correlation, *R*, was computed from the matrices in *A*. (*C*, *E*, *G*, and *I*) The presence of distinct coregulated modules was estimated by the *Q* value (19). The computations for each true network (red circles) were also performed on control data sets: time-permuted (1,000 randomly scrambled versions of the correlation matrix, orange diamonds), random gene set (100 gene sets of size *N* randomly sampled from the entire transcriptome, blue triangles), and random number matrix (100 matrices generated with the same gene number and developmental stages as the true network, green squares). The number of genes included in the analysis for each network in each species is shown in parentheses. Error bars represent SD of *R* and *Q*; some SDs are smaller than the marker size. Asterisks indicate a significant difference from the random gene set control ($P < 0.05$, two-tailed *t* test).

Like the synaptic network, the epithelial network also lacks a morphological correlate in the sponge. In epithelial cells, the adherens junction links to apical-basal polarity genes and Wnt/planar polarity genes (Fig. 2*B* and Fig. S1*B*). Although *A. queenslandica* expresses many orthologues of epithelial genes, the sponge exhibits only rudimentary features of a functional epithelia (22, 23). As in the synaptic gene set analysis, we extracted the expression patterns of epithelial genes from six species and calculated the average correlation, *R*, and modularity, *Q*, of the coregulation network (Fig. 3 *D* and *E*). The epithelial network in all species that were tested showed significantly greater *R* when compared pairwise vs sponge ($P < 1 \times 10^{-8}$, two-tailed *t* test; Table S2). As in the synaptic network, the modularity of epithelial networks was not consistently lower compared with random controls for most of the species tested.

Neurons and epithelial cells and their defining cellular machines appear in eumetazoans after sponges diverged from other animals. We asked whether genes drawn from more ancient machines present in all eukaryotes might show a different pattern of expression characteristic of machines that were functionalized before the origin of animals. We performed a similar modularity optimization on transcriptome data for homologues of genes in the nuclear pore complex (NPC) and the 26S proteasome (Fig. 2 *C* and *D* and Fig. S1 *C* and *D*). These networks are highly interconnected and exhibit a negatively skewed degree distribution, which differs from the relatively large hubs and positively skewed degree distribution observed in mammalian synaptic and epithelial networks (Fig. 2*E*).

The nuclear envelope is a defining feature of eukaryotic cells (24). Transport of molecules between the nucleus and cytoplasm is mediated by the NPC, which is made up of approximately 30 nucleoporin genes. Coregulation analysis of nucleoporin homologues represented in the transcriptome set revealed higher average correlation and generally lower modularity compared with the synaptic or epithelial networks in the same species (Fig. 3 *F* and *G*). Most of the NPC networks showed consistently greater *R* and lower *Q* compared with permuted or random size-matched data, suggesting that the components of the NPC act as a single functional unit (Table S1). In contrast, greater modularity of the synaptic and epithelial polarity networks suggests a requirement for some modularity in the operation of these machines, perhaps as a result of the presence of ancient submachines, such as the vATPase community.

The 26S proteasome is a well conserved protein degradation machine composed of products from more than 31 genes (25). Coregulation analysis of homologues of proteasomal genes revealed that, like the NPC, the proteasome has higher average correlation and lower modularity compared with the synaptic or epithelial networks within each species (Fig. 3 *H* and *I*). All eumetazoans showed significantly higher correlation when compared pairwise vs. sponge ($P < 1 \times 10^{-52}$, two-tailed *t* test; Table S2). Coregulation and modularity of proteasomal genes differed significantly from permuted or random data, except in the sponge (Table S1). Nevertheless, in all species tested, including the sponge, the proteasome gene set emerged as a distinct community when analyzed together with NPC genes (Fig. 5) and is therefore likely to represent a functionally significant module.

In a unicellular eukaryote, like the yeast, *Saccharomyces cerevisiae*, the NPC and proteasome gene networks exhibit high correlation and low modularity that is quite similar to the average values observed for the metazoans (Table S1). These findings further support the hypothesis that gene networks that establish

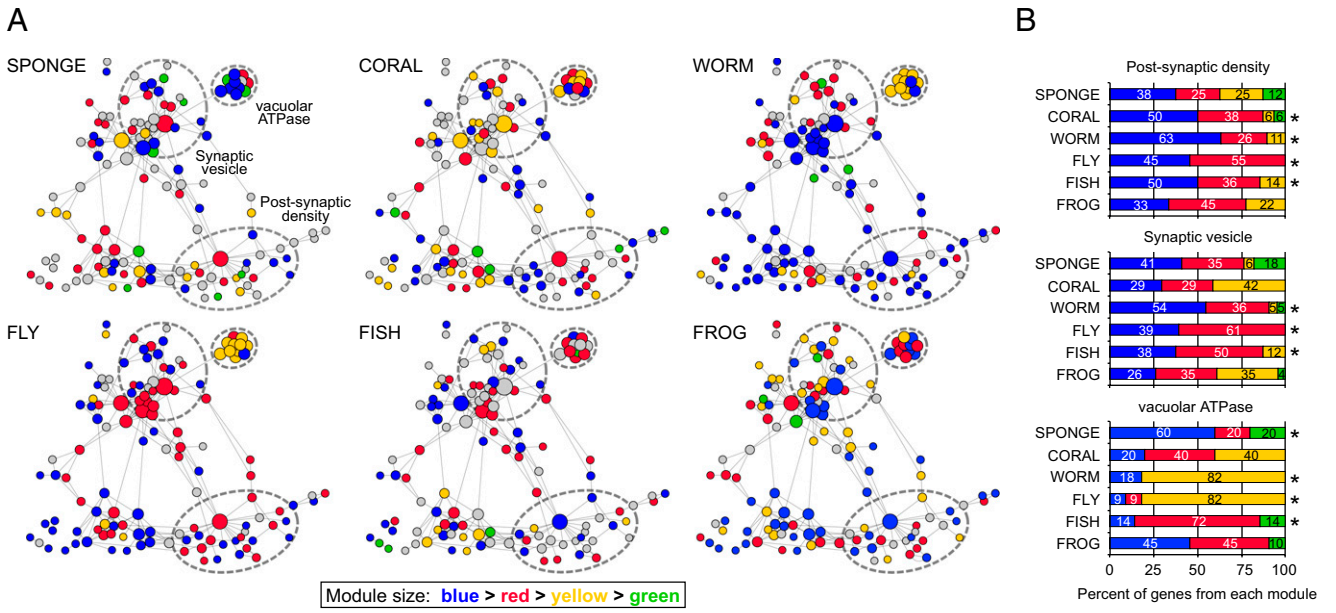

Fig. 4. Functionally defined protein complexes correspond to detected coregulation modules. (*A*) Genes in the synaptic network of each species were assigned to coregulation modules by modularity optimization. Genes were colored according to the module from which they are derived (module size: blue > red > yellow > green). Genes in gray are not represented in the organism or have no available expression data. Dashed circles represent the approximate boundaries of the postsynaptic density, synaptic vesicle, and vATPases. (*B*) Percent of genes in each functional complex that belong to coregulation modules detected by modularity optimization. Colors correspond to the gene modules in *A*. Asterisks indicate complexes for which ≥50% of genes belong to the same coregulation module. Only genes with available expression data in each species were included in the analysis.

their modern function long before the origin of metazoa exhibit significantly higher correlation and lower modularity, consistent with a greater and more homogeneous connectivity between genes.

These results show that data-driven detection of transcriptional expression patterns can reliably reveal a reorganization of gene networks in association with the emergence of their modern collective function from the unknown functions of these same gene sets in the common animal ancestor. This reorganization appears as increased connectivity and a change in the network structure with functional complexes clustering into coregulated modules. In contrast, more ancient machines, such as the proteasome and the NPC, show a cohesiveness of expression as far back as the eukaryotic ancestor.

Discussion

Synaptic proteins must be available in concentrations that drive self-assembly by mass action according to the affinities among their various interaction domains. Among the core features of synapses are scaffolding proteins that position receptors and ion channels in register with synaptic vesicles across the synaptic cleft and link the pre- and postsynaptic elements to intracellular signaling cascades. Coordinated expression of these proteins, as well as the affinity of the interactions, are among the drivers of synapse assembly. Positive selection at specific sites in PDZ scaffolds appear to have roles in determining the binding partners of these highly connected proteins (3), an observation consistent with network growth by link dynamics, i.e., link detachment and attachment (26). Just as mutations in coding sequences can change link dynamics and enable new protein–protein interactions, mutations in *cis* regulatory sequences can lead to the evolution of new transcriptional linkages and coexpression of gene batteries that were not previously associated. In fact, the sponge already possesses homologues of genes that function in bilaterian neurogenesis, although it is yet to be determined if these factors were responsible for a biological unit originating in the sponge ancestor that was selected for an unknown function and later exapted to assemble the synapse (27). These conserved bilaterian developmental and neurogenic genes are associated with spatial patterning of the cnidarian nerve net (28, 29). Further modification of gene regulatory mechanisms in vertebrates placed many synaptic genes under the control of the transcriptional repressor, REST, thus ensuring exclusive and coordinated expression in neurons (30–32).

The hierarchical structure of gene regulatory signaling networks that control the body plan are thought to evolve by changes in *cis* regulatory regions resulting in changes in timing, level, and location of gene expression (33). In contrast, the network edges of cellular machines represent physical interactions rather than a cascade of signaling events (34). Nevertheless, the resolution of a signaling or interaction network depends on the extent of coregulatory data available to inform the graph edges. Our analysis required that we compare the coregulation and modularity of the same set of genes; however, inclusion of genes linked to the synaptic network that are not shared between the comparison groups would likely improve the coregulation signal as gene innovation and duplication can affect network structure through dynamic interactome rewiring (35). Although these limitations increase the likelihood of detecting biologically spurious correlations and may contribute to the apparent modularity observed in some random gene sets, the ability of the community detection algorithm to partition genes into their respective cellular machines indicates a functional correlate of the structural communities derived simply from transcriptional coregulation (Fig. 5). The generation of more transcriptomes at finer temporal and spatial resolution and the sequencing of genomes from other basal metazoans, as well as improved homologue detection, may strengthen or weaken an alternative explanation that the gene expression patterns in *A. queenslandica* represent a loss of more ancient gene regulatory patterns.

Evolutionary growth of gene interaction networks is a key facet of organismal complexity. Several publications have claimed that gene expression networks are scale-free (4), and although no rigorous proof of the claim exists, many gene expression networks do display a tail in their degree distributions, indicating the presence of large hubs. Interestingly, one particular model of scale-free network growth suggests that (*i*) networks expand continuously by

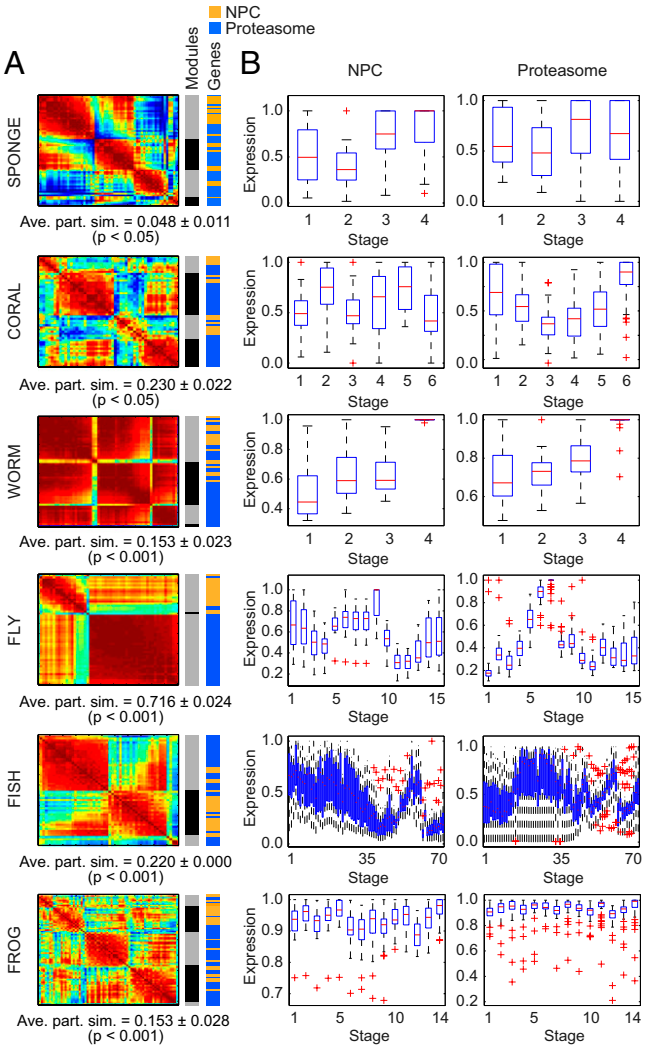

Fig. 5. Modularity optimization detects biologically relevant gene communities. (*A*) Heat maps represent the *N*×*N* Pearson correlation matrices for union networks of NPC and proteasome genes (red, positive correlation; blue, negative correlation). Average partition similarities (ave. part. sim.) computed from permutation testing with 1,000 iterations showed that, compared with the randomly scrambled gene set, genes in the union network clustered into communities that more closely recapitulated the true partition between networks ($P < 0.05$). Color bars to the right of the heat maps indicate the boundaries of detected coregulation modules (Modules) and the relative location of NPC (orange) and proteasome (blue) genes within the detected communities (Genes). (*B*) Box plots show the developmental expression patterns of genes within the NPC (*Left*) and proteasome network (*Right*) for each of the six representative species.

the addition of new nodes, and (*ii*) new nodes attach preferentially to sites that are already well-connected (36). Gene number has not increased by much over the course of metazoan evolution. Thus, the expansion of gene interaction networks, which is required to functionalize metazoan cellular machines, places an exceptionally high premium on enhancing coregulatory patterns between existing genes.

Conclusions

By using genome-wide transcriptome data, we tracked the expression of a common set of synaptic genes in a representative sampling of the animal kingdom. In bilaterians, the expression of synaptic genes is strikingly well coordinated, with smaller coregulation modules detectable within the expression matrix. A particularly prominent module is the vATPase complex found within the presynaptic gene set. Interestingly, synaptic genes in the earliest branching metazoan phyla (Porifera) exhibit a lack of global coregulation compared with eumetazoans with functional nervous systems. Protosynaptic gene expression modules from the sponge, *A. queenslandica*, which lacks synapses and a nervous system, but possesses a nearly complete complement of synaptic genes, are organized into independent communities. These findings suggest that functional synapses evolved through the exaptation of preexisting genes and smaller cellular machines, presumably by modification of regulatory circuitries resulting in coordinated neuronal expression. This work demonstrates that the modularity approach based on network theory provides a very simple and data-driven method for the identification of gene communities, linking this study to a larger array of network diagnostics that could be used in subsequent investigations of the topological organization of gene coexpression networks across species.

Materials and Methods

Expression Data. Genes in the synaptic, epithelial, NPC, and 26S proteasome networks were compiled from the literature (1, 23–25). Protein interaction networks were based on the human interactome annotated in STRING (37) and visualized by using Cytoscape (38). Homologues for genes in these networks were determined by reciprocal best-hit BLAST alignments of human gene sequences to the genome of each species of interest. Expression data for gene homologues was extracted from transcriptomes obtained by RNA sequencing of four developmental stages in sponge, *A. queenslandica* (*SI Materials and Methods*); six experimental treatments of coral larvae, *A. millepora* (16); four developmental stages in worm, *C. elegans* (15); and 15 developmental stages in fly, *D. melanogaster* (14). Microarray data for 70 developmental stages in zebrafish, *D. rerio* (13); and 14 developmental stages in frog, *X. tropicalis* (17), were also included. Microarray expression data for the yeast, *S. cerevisiae*, was obtained from cultures grown to stationary phase (39). To compare expression patterns in transcriptomes obtained by using different methods, the expression for every gene within each dataset was normalized to its maximum value across development (Dataset S1).

Coregulation and Modularity Analysis. For each organism, the strength of genetic coregulation of any two genes throughout development was estimated by computing the Pearson correlation coefficient of expression for those two genes over development. By estimating the coregulation strength for all possible pairs of genes, we constructed organism-specific *N*×*N* coregulation networks in which genes were represented by nodes and connections between genes were weighted by the correlation between their expression levels over development. These coregulation networks were characterized by two diagnostic variables: the average correlation, *R*, and the modularity, *Q*, as defined in the following paragraphs.

The first diagnostic, the average correlation *R*, provides a measure of within-network connectivity which can be interpreted as a measure of coregulation. Significant differences in network coregulation between species were identified using pairwise two-tailed *t* tests of the correlation matrix elements. For these tests, correlation matrices were computed only for the sets of genes that were common between the two species being compared. These union gene sets for pairwise comparisons were constructed without duplicates by using only genes with the best BLAST score to the human protein sequence.

The second diagnostic, the modularity *Q*, provides a measure of community structure in the coregulation matrix. Importantly, the correlation matrix we used to examine the amount of coregulation (*R*) can equivalently be viewed as a complex network in which gene–gene edges are signed (i.e., positive or negative correlations) and weighted (correlations range from −1 to 1). In each organism's coregulation network, we tested for the presence of uniquely coregulated groups of genes by using the community detection approach (20) of optimizing modularity (18) by using the Louvain method (19) [note that a second heuristic, spectral optimization (40), gave nearly identical results: $r = 0.9960$, $P < 0.01$; Table S3]. We define the correlation matrix *A* and then define w_{ij}^{+} to be an *N*×*N* matrix containing the positive elements of A_{ij} and w_{ij}^{-} to be an *N*×*N* matrix containing only the negative elements of A_{ij}. The quality function to be maximized is then given by the following equation:

$$Q_{\pm} = \frac{1}{2w^{+} + 2w^{-}} \sum_{i} \sum_{j} \left[A_{ij} - \left(\gamma^{+} \frac{w_i^{+} w_j^{+}}{2w^{+}} - \gamma^{-} \frac{w_i^{-} w_j^{-}}{2w^{-}} \right) \right] \delta(g_i, g_j) \quad [1]$$

where g_i is the community to which node i is assigned, g_j is the community to which node j is assigned, γ^{+} and γ^{-} are resolution parameters, and the following equation applies (41):

$$w_i^{+} = \Sigma_j w_{ij}^{+}, w_i^{-} = \Sigma_j w_{ij}^{-}. \quad [2]$$

As evident from Eq. **1**, two free parameters in the optimization of modularity for such a signed, weighted network exist (42): the resolution parameters γ^{+} and γ^{-} (43). For simplicity in the present analysis, we chose the traditional value of γ^{+} of 1.0 and set γ^{-} as 0.1 to dampen the effect of negative correlations. Particular emphasis was placed on the positive correlations in the coregulation matrix for two reasons. First, we noted that most gene sets had significantly more positive correlations, and in fact some gene sets had no negative correlations at all (e.g., worm NPC). To ensure that our analysis was consistent across both organisms and machines, we dulled the influence of negative correlations by setting γ^{-} to be an order of magnitude smaller than γ^{+}. Secondly, we noted that the positive correlations showed considerably more topological organization than the negative correlations (Fig. S2). Further details are provided in *SI Materials and Methods*.

We further examined the dependence of our results on the choice of γ^{+}. We varied γ^{+} from 0 to 2 in intervals of 0.1. We find that, for values of γ^{+} higher than 1, the network disintegrates into a large number of communities (Fig. S3). Our results therefore focus on the smallest yet still coherent modular structures present in these systems.

Robustness and Statistical Validity. To examine the robustness and statistical validity of our findings, we assessed the reliability of the group partitions and tested our results against three separate postmodularity-optimization null models as described in the following paragraphs.

The problem of optimizing the modularity quality function is nondeterministic polynomial-time–hard. It is therefore important to demonstrate that the heuristics that we used produce robust results, i.e., that the partitions found by iterative optimizations are highly similar. For each organism and each machine, we calculated the partition similarity (44) (which is bounded in [0,1]) between 100 separate optimizations. We found that the average partition similarity was >0.8 for most organisms and machines, with the mean over organisms and networks being even higher (Table S1).

In addition to quantifying the reliability of our findings, we examined the statistical validity of our results by comparing the diagnostic variables (R and Q) derived from the true network to those derived from networks constructed from three separate random null models: true random (random number matrix), time-permuted, and random gene set. The true random null-model network is constructed by generating uniformly distributed random numbers for the same number of genes and developmental stages found in the true data set (100 instantiations). A coregulation matrix is then constructed and R' and Q' are calculated. The time-permuted null-model network is constructed by randomly scrambling the order of expression for each gene within the network (1,000 instantiations), recomputing the coregulation matrix, and calculating R' and Q'. The random gene set null-model network was constructed by extracting the expression data for an identically sized randomly chosen set of genes from the whole transcriptome (100 instantiations). Further details are provided in *SI Materials and Methods*. The statistical significance of the true R and Q values was examined by using a one-sample t test in comparison with the R' and Q' values, respectively, for each random null model (Table S1). We noted that the level of background correlation and modularity observed within sets of N genes randomly selected from each of the transcriptomes is variable (Fig. S4). One possible explanation for these differences is that the transcriptome data sets were obtained by using different methods.

Biological Relevance of Detected Modules. We asked whether the modules detected from the coregulation matrix could represent functional entities. We began by calculating the correlation matrix *R2* between the combined gene set of the proteasome and NPC for each species. We optimized the modularity quality function to partition this combined matrix into groups in a data-driven manner. We next asked whether this data-driven partition was statistically similar to the true partition of the genes into the two groups of proteasome genes and NPC genes. To answer this question, we computed the partition similarity between the data-driven partition and the true partition and used permutation testing to determine whether this similarity was statistically significant. The permutation test was implemented by randomly reassigning genes to the two groups of "proteasome" and "NPC," recomputing the correlation matrix $R2'$, partitioning the genes in the correlation matrix into modules, and computing the similarity between this partition and the true partition. This process was repeated 1,000 times to construct a distribution of similarity values expected under the null hypothesis that the coregulation patterns between proteasome and NPC genes do not differ. For each species, the P value to reject this null hypothesis was computed as follows: the number of similarity values derived from the permuted data that were greater than the real similarity value, divided by the number of permutations.

Supporting Information. See *SI Text* for supporting figures, tables, methods, discussion, and data.

ACKNOWLEDGMENTS. We thank Boris Shraiman, Mason Porter, Adel Dayarian, and Marija Vucelja for invaluable suggestions and comments; Scott Grafton and Jean Carlson for facilitating the collaboration; and Marc Kirschner and Leonid Peshkin for sharing *Xenopus tropicalis* microarray data. This work was supported by gifts from Harvey Karp and Gus Gurley (K.S.K.); David and Lucile Packard Foundation (D.S.B.); Public Health Service Grant NS44393 (to D.S.B.); Institute for Collaborative Biotechnologies Contract W911NF-09-D-0001 from the US Army Research Office (to D.S.B.); and the Australian Research Council (S.M.D. and B.M.D.).

1. Srivastava M, et al. (2010) The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* 466:720–726.
2. Sakarya O, et al. (2007) A post-synaptic scaffold at the origin of the animal kingdom. *PLoS ONE* 2:e506.
3. Sakarya O, et al. (2010) Evolutionary expansion and specialization of the PDZ domains. *Mol Biol Evol* 27:1058–1069.
4. Barabási A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5:101–113.
5. Brawand D, et al. (2011) The evolution of gene expression levels in mammalian organs. *Nature* 478:343–348.
6. Oldham MC, Horvath S, Geschwind DH (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci USA* 103: 17973–17978.
7. Oldham MC, et al. (2008) Functional organization of the transcriptome in human brain. *Nat Neurosci* 11:1271–1282.
8. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95:14863–14868.
9. Quackenbush J (2003) Genomics. Microarrays—guilt by association. *Science* 302: 240–241.
10. Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302:249–255.
11. van Noort V, Snel B, Huynen MA (2003) Predicting gene function by conserved co-expression. *Trends Genet* 19:238–242.
12. Carlson MR, et al. (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics* 7:40.
13. Domazet-Lošo T, Tautz D (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468:815–818.
14. Graveley BR, et al. (2011) The developmental transcriptome of Drosophila melanogaster. *Nature* 471:473–479.
15. Hillier LW, et al. (2009) Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. *Genome Res* 19:657–666.
16. Meyer E, Aglyamova GV, Matz MV (2011) Profiling gene expression responses of coral larvae (Acropora millepora) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. *Mol Ecol* 20:3599–3616.
17. Yanai I, Peshkin L, Jorgensen P, Kirschner MW (2011) Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. *Dev Cell* 20: 483–496.
18. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99:7821–7826.
19. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech* 10:P10008.
20. Porter MA, Onnela J-P, Mucha PJ (2009) Communities in networks. *Notices Am Math Soc* 56:1082–1097.
21. Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW (2012) Evolution of increased complexity in a molecular machine. *Nature* 481:360–364.
22. Adams EDM, Goss GG, Leys SP, Launikonis BS (2010) Freshwater sponges have functional, sealing epithelia with high transepithelial resistance and negative transepithelial potential. *PLoS ONE* 5:e15040.
23. Fahey B, Degnan BM (2010) Origin of animal epithelia: Insights from the sponge genome. *Evol Dev* 12:601–617.
24. Wente SR, Rout MP (2010) The nuclear pore complex and nuclear transport. *Cold Spring Harb Perspect Biol* 2:a000562.
25. Voges D, Zwickl P, Baumeister W (1999) The 26S proteasome: A molecular machine designed for controlled proteolysis. *Annu Rev Biochem* 68:1015–1068.

26. Berg J, Lässig M, Wagner A (2004) Structure and evolution of protein interaction networks: A statistical model for link dynamics and gene duplications. *BMC Evol Biol* 4:51.
27. Richards GS, et al. (2008) Sponge genes provide new insight into the evolutionary origin of the neurogenic circuit. *Curr Biol* 18:1156–1161.
28. Marlow HQ, Srivastava M, Matus DQ, Rokhsar D, Martindale MQ (2009) Anatomy and development of the nervous system of Nematostella vectensis, an anthozoan cnidarian. *Dev Neurobiol* 69:235–254.
29. Layden MJ, Boekhout M, Martindale MQ (2012) Nematostella vectensis achaete-scute homolog NvashA regulates embryonic ectodermal neurogenesis and represents an ancient component of the metazoan neural specification pathway. *Development* 139: 1013–1022.
30. Schoenherr CJ, Anderson DJ (1995) The neuron-restrictive silencer factor (NRSF): A coordinate repressor of multiple neuron-specific genes. *Science* 267:1360–1363.
31. Otto SJ, et al. (2007) A new binding motif for the transcriptional repressor REST uncovers large gene networks devoted to neuronal functions. *J Neurosci* 27:6729–6739.
32. Bruce AW, et al. (2004) Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. *Proc Natl Acad Sci USA* 101:10458–10463.
33. Peter IS, Davidson EH (2011) Evolution of gene regulatory networks controlling body plan development. *Cell* 144:970–985.
34. Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* 100:12123–12128.
35. Dreze M, et al.; Arabidopsis Interactome Mapping Consortium (2011) Evidence for network evolution in an Arabidopsis interactome map. *Science* 333:601–607.
36. Udny Yule G (1925) A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *R Soc Lond Philos Trans B* 213:21–87.
37. Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: A Web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* 28: 3442–3444.
38. Shannon P, et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504.
39. Gasch AP, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 11:4241–4257.
40. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E Stat Nonlin Soft Matter Phys* 74:036104.
41. Traag VA, Bruggeman J (2009) Community detection in networks with positive and negative links. *Phys Rev E Stat Nonlin Soft Matter Phys* 80:036115.
42. Gómez S, Jensen P, Arenas A (2009) Analysis of community structure in networks of correlated data. *Phys Rev E Stat Nonlin Soft Matter Phys* 80:016114.
43. Fortunato S, Barthélemy M (2007) Resolution limit in community detection. *Proc Natl Acad Sci USA* 104:36–41.
44. Danon L, Diaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. *J Stat Mech* 2005:P09008.