Reassessing the modularity of gene co-expression networks using the Stochastic Block Model

Manuscript v1.0 - May 31, 2023

Diogo Melo[1,2,*], Luisa F. Pallares[3], and Julien F. Ayroles[1,2,*]

[1] Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA
[2] Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA
[3] Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany

[*] Correspondence: Diogo Melo <damelo@princeton.edu>, Julien F. Ayroles <jayroles@princeton.edu>

Abstract

Finding communities in gene co-expression networks is a common first step toward extracting biological insight from such complex datasets. Most community detection algorithms expect genes to be organized into assortative modules, that is, groups of genes that are more associated with each other than with genes in other groups. While it is reasonable to expect that these modules exist, using methods that assume they exist a priori is risky, as it guarantees that alternative organizations of gene interactions will be ignored. Here, we ask: can we find meaningful communities without imposing a modular organization on gene co-expression networks, and how modular are these communities? For this, we use a recently developed community detection method, the weighted degree corrected stochastic block model (SBM), that does not assume that assortative modules exist. Instead, the SBM attempts to efficiently use all information contained in the co-expression network to separate the genes into hierarchically organized blocks of genes. Using RNA-seq gene expression data measured in two tissues derived from an outbred population of *Drosophila melanogaster*, we show that (a) the SBM is able to find ten times as many groups as competing methods, that (b) several of those gene groups are not modular, and that (c) the functional enrichment for non-modular groups is as strong as for modular communities. These results show that the transcriptome is structured in more complex ways than traditionally thought and that we should revisit the long-standing assumption that modularity is the main driver of the structuring of gene co-expression networks.

Introduction

Gene co-expression networks inform our understanding of cell and organismal function by encoding associations between genes. Associations between expression levels can indicate common function, and the number of connections can point to central or regulatory genes (Dam et al., 2018). Due to the large dimensionality of gene expression data, often composed of several thousands of gene expression measures, a major tool in the analysis of co-expression is gene clustering: separating the genes into related groups, which can then be explored separately (D'haeseleer, 2005). This drastically reduces the number of genes we need to consider at the same time and allows for the identification of hubs or centrally connected genes that can be used to inform further experimental validation (Imenez Silva et al., 2017; Langfelder & Horvath, 2008).

The question is, given a co-expression network, how should we cluster the genes? The general idea behind several methods is to look for similar genes, as these are expected to be involved in related biological functions. However, several definitions of similarity have been used. The most basic measure of similarity borrows from classical morphological integration theory and attempts to find gene modules based on their correlations. In this context, genes in the same module are expected to be highly correlated and perform similar functions, while genes in different modules are expected to have low correlations (Magwene, 2001; Olson & Miller, 1958; Wagner et al., 2007). Here, we refer to this classic pattern of higher within- than between-group correlation as assortativity, and to the groups as assortative modules. Other methods use the correlations to create other measures of similarity, which are then used as input to clustering algorithms. Weighted gene co-expression

network analysis (WGCNA, Langfelder & Horvath, 2008) uses a power transformation of the correlation between gene expressions (or a topological similarity measure built with these transformed correlations) (Dong & Horvath, 2007; B. Zhang & Horvath, 2005) as a similarity measure that is then separated into assortative modules using hierarchical clustering. One of the main objectives of WGCNA is finding hub genes, which have high connectivity within modules and are clearly identified by hierarchical clustering. Other methods borrow from network analysis and attempt to explicitly maximize the Newman Modularity (Newman, 2006) of the weighted gene network. For example, Modulated Modularity Clustering (MMC, Stone & Ayroles, 2009) uses an adaptive algorithm to find a non-linear distance between genes based on their correlations that maximizes the number of modules uncovered by maximizing modularity. Although these methods differ in their definition of similarity, they all impose an assortative structure on the gene expression network, in which similar genes are expected to be more correlated with each other than with other genes.

Clustering genes in tightly correlated modules aligns with the intuition that groups of genes performing similar functions should be highly correlated. However, imposing assortativity will necessarily ignore alternative organizations, if they exist, and could prevent us from fully understanding of how transcriptional networks are organized. To avoid this problem, we use a more general measure of similarity that allows us to find meaningful gene groups that are not necessarily assortative but still have clear biological interpretation. This measure is implemented in the weighted nested degree corrected stochastic block model (wnDC-SBM, or SBM for brevity, Peixoto, 2017, 2018), which has shown promising results in similar applications (see Baum et al. (2019) and Morelli et al. (2021)). The SBM is different from other clustering methods in that it does not attempt to find assortative modules (i.e., modules with higher within- than between-module correlation). Instead, any information contained in the gene co-expression network can potentially be used to inform the clustering. To be sure, the SBM can capture an assortative modular pattern if it is present, but it is general enough to also capture other network organizations (L. Zhang & Peixoto, 2020). Furthermore, even if, in the context of the SBM, assortativity is not the main driver of gene partitioning, it can still be used to interpret the clusters we obtain. By measuring the modularity of the identified clusters we can compare networks with respect to their modularity without the problem of comparing a measure that was maximized in order to find the clusters in the first place. This opens the possibility of an unbiased comparison of the degree of modularity in different transcriptional networks (e.g., different cell types, tissues, species, etc), which is a question that remaines unexplored so far.

Here, using a multi-tissue RNAseq dataset from *Drosophila melanogaster*, we show first, that the SBM, a model with no free parameters, can find many more gene clusters than competing methods. Second, that such gene clusters are biologically meaningful as revealed by highly specific gene ontology enrichment. Third, that biological meaning is not restricted to assortative modules as traditionally thought but extends to the non-assortative parts of the transcriptome. Our results highlight the importance of using clustering algorithms that don't rely on assortativity metrics to explore the structure of transcriptomes in a comprehensive and unbiased manner.

Methods

Gene expression measures

Elsewhere (Pallares et al., 2023), we quantified whole-genome gene expression in hundreds of outbred *Drosophila melanogaster* female flies using a high-throughput RNAseq library preparation protocol (TM3seq, Pallares et al., 2020). To build the gene co-expression networks, here we use a subset of the full dataset

that includes: samples for two tissues, head (n = 212) and body (n = 252), individuals with the best coverage (average gene counts: head = 4.65M, body = 4.58M), and genes with moderate to high expression (average CPM > 5 and detected in every sample, head n = 5584 genes, body n = 5533 genes). The expression matrices used to generate co-expression networks correspond to VOOM-transformed gene counts (Law et al., 2014) where the effect of known and unknown (Leek & Storey, 2007) covariates was removed using the function removeBatchEffect from the R package limma (Law et al., 2014). Details on the collection of RNAseq data, library preparation, and processing of raw RNAseq counts can be found in Pallares et al. (2023).

Gene co-expression network

Using the gene expression measures for both tissues we generate co-expression network graphs. In theory, we could proceed using a full network in which all pairs of genes are connected but fitting the SBM with this fully connected graph is computationally too expensive. So, to reduce the connectivity of the network, we impose a stringent Benjamini-Hochberg false discovery rate (FDR) cut-off on the edges, removing edges with a large p-value associated with the correlation between the corresponding genes. As edges are removed, some genes with only non-significant correlations become disconnected from the rest of the network and can be removed. By gradually reducing the FDR threshold, we reduce the density of the gene network while attempting to keep as many genes as possible, until we arrive at a viable set of genes and connections with which to fit the SBM. We chose an FDR of 1% for the head and 0.1% for the body datasets which kept most of the genes (94.2% in the head:5261, and 92.6% in the body:5124) while reducing the graph density to a manageable level for use in the SBM. This set of genes is used in the three clustering methods compared in this study: WGCNA, MMC, and SBM.

Edge weights

Each method uses different edge weights for the network graph. Both WGCNA and MMC can use the fully connected graph, so we maintain all edges in these methods. We use the toplogical overlap matrix (TOM) similarity in WGCNA, and the Spearman correlation derived distance in MMC. We use the low-density graph described above for the SBM, with the edge weights given by the inverse hyperbolic tangent transformed Spearman correlations between gene expressions. This transformation allows the edge weights to be modeled by normal distributions in the SBM, as we discuss below.

Stochastic Block Model

The Weighted Nested Degree Corrected Stochastic Block Model (Karrer & Newman, 2011; Peixoto, 2017) is a Bayesian generative model that attempts to find the partition with the highest posterior probability given the observed network and edge weights. Broadly speaking, this is achieved by dividing the network into groups of genes, called blocks, and modeling the weight and existence of a link between two genes in a network solely on their belonging to a particular block. So, genes with similar patterns of connections tend to be clustered in the same block. The degree correction refers to a modification of the standard Stochastic Block Model that allows genes with different degrees to be clustered in the same block (see Peixoto, 2017 for details).

If b is a particular partition of the genes in the weighted gene network A, we write a model that generates A with probability given by $P(A|b,\theta)$, where θ stands in for any extra parameter we need besides the group partition b. With this model, we can write the posterior probability of the block partition b given the observed network:

$$P(b|A) = \frac{P(A|\theta, b)P(\theta, b)}{P(A)}$$

where P(A) is a normalization constant. As for the additional parameters θ, the formulation used here, from Peixoto (2018), uses hard constraints such that there is only one choice of θ that is compatible with A and b, which means that the model has no free parameters. We can then search for the partition (b) that maximizes $P(b|A)$ using computational methods, like Markov Chain Monte Carlo (MCMC) methods.

Description length

The posterior probability of the block partition can be written as:

$$P(b|A) \propto \exp(-\Sigma)$$

Where $\Sigma = -\log[P(A|\theta, b)] - \log[P(\theta, b)]$ is called the description length of the gene network A, and has an information-theoretic interpretation, being the amount of information required to encode the network given θ and b. So, finding the partition that maximizes the posterior probability is the same as minimizing the description length, or, in other words, the chosen partition b is the one that allows us to describe the network using the least information.

The two terms in Σ also allow us to understand why this method offers intrinsic protection against overfitting. The first term $\log[P(A|\theta, b)]$ corresponds to the log-likelihood of the observed network. Increasing the number of blocks allows this likelihood to increase as the degrees of freedom of the model increase. But, the second term, $\log[P(\theta, b)]$ functions as a penalty that increases for complex models with many blocks, and the description length cannot decrease for overly complex models that have more blocks than warranted by the data. So, the selected partition with the minimum description length will necessarily be the simplest partition with similar explanatory power, avoiding overfitting and fully using the available statistical evidence. For example, the SBM would not detect modules that appear in random networks due to statistical fluctuations, in contrast to modularity maximization, which finds spurious modules in random networks (Guimerà et al., 2004; L. Zhang & Peixoto, 2020). We can also use the description length as a principled method for comparing models that simultaneously considers fit to data and model complexity.

Weighted SBM

The weights on the edges can be modeled in the SBM using different distributions depending on the edge weights. When edge weights are correlations, which are continuous numbers that vary between -1 and 1, it is natural to use some transformation to map the correlations onto the real numbers. To do this, we use arctanh transformed correlations as the edge weights and model these weights using normal distributions. In the SBM, the weights are modeled in much the same way as the links between networks, in that the mean and the variance of the observed edge weights between two blocks are a function only of the block structure, i.e., genes in the same block have a similar probability of being connected to other genes and the value of the weights in these edges comes from the same distribution.

Nested SBM

The nested SBM uses a series of non-parametric hierarchical priors that greatly increase the resolution of detected blocks. This nested structure allows for the identification of more and smaller blocks that are statistically

supported than other clustering methods (Peixoto, 2017). This is achieved by treating the gene block partition as the nodes in a nested series of networks, which are then clustered using the same method. So, the genes are clustered in blocks, and these blocks are also clustered in higher-level blocks, and so on, as required to minimize the description length of the gene network (see diagram in fig. 1). The model estimates the number of levels in the hierarchy and the number of blocks in each level. Since the model is generative, we can use posterior samples of the partitions to quantify the uncertainty in any quantity estimated by the model, like the number of levels in the hierarchy, or the number of blocks at each level. For details on the implementation of the SBM, see Peixoto (2017) and Peixoto (2018). All SBM were fitted using the graph-tool python library (Peixoto, 2014). The fitting process consisted of three steps. First, an initial partition of genes into blocks at each level of the SBM hierarchy was obtained using the *NestedBlockState* function. Next, the block partition was refined using the *mcmc_anneal* function, which uses Markov Chain Monte Carlo (MCMC, Peixoto (2017)) and simulated annealing (Kirkpatrick et al., 1983) to find a better network partition (i.e., one with larger posterior probability and smaller description length). Annealing is not always necessary, but in our data it proved to be efficient in reducing computational time. Next, the *mcmc_equilibrate* function was employed to find a partition where subsequent proposals did not improve the posterior probability of the current partition for at least 1000 proposals. At this stage, the block partition was considered equilibrated, allowing for posterior sampling using MCMC. Finally, the posterior sampling was conducted for 1000 iterations using the *mcmc_equilibrate* function, and the median partition of this posterior sample was used for subsequent analysis.

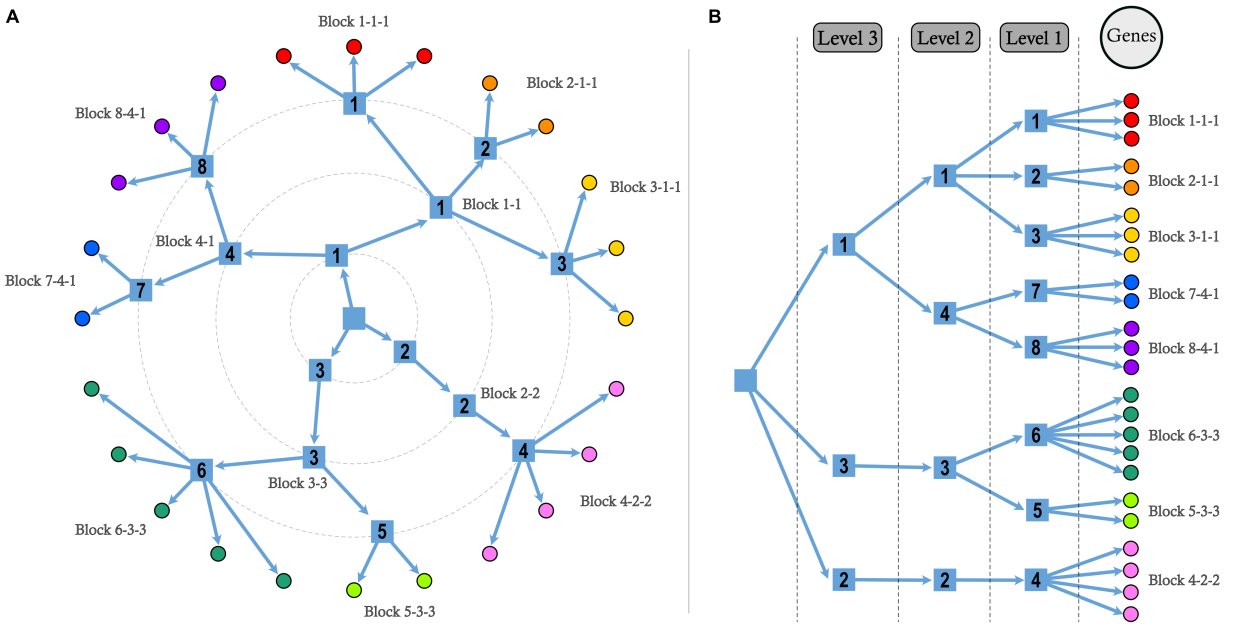

Figure 1: Schematic representation of the clustering in the SBM. Genes are clustered into level-1 blocks, level-1 blocks are clustered into level-2 blocks, and so on. A. Circular representation of the clustering we use in the following figures. Block names are constructed by following the hierarchy, starting at level 1. So in this example, the level-1 block 8 can also be referred to as 8-4-1. B. A tree-like representation that highlights the hierarchy in the nested SBM. Each level-2 block is composed of all the genes in its child level-1 blocks, each level-3 block is composed of all the genes in its child level-2 blocks, and so on.

Modularity and Assortativity

Although the nested SBM does not attempt to find the partition of genes that maximizes modularity (see definition below), when using this method we can ask if the inferred partition is modular or not by calculating the Newman modularity at each level of the hierarchy. Newman Modularity is calculated at each nested level

using:

$$M = \frac{1}{2E}\sum_r e_{rr} - \frac{e_r^2}{2E}$$

where e_{rr} is, by convention, twice the sum of edge weights internal to group r, e_{rs} is the sum of edge weights between groups r and s, $e_r = \sum_s e_{rs}$, and E is the sum of all weights. Newman modularity quantifies the intuition that genes in the same module should be more connected than across modules by comparing the within-group connections (e_{rr}) to the expected value of the connections across all the groups ($\frac{e_r^2}{2E}$). The higher the difference between correlations within- and between-groups, the higher the value of M.

We further decompose the contribution of each Level-1 block to the modularity by defining the assortativity of a block as:

$$q_r = \frac{B}{2E}\left(e_{rr} - \frac{e_r^2}{2E}\right)$$

where B is the number of blocks. Using this definition, modularity is just the average assortativity, $M = \frac{1}{B}\sum_r q_r$, and blocks with positive assortativity contribute to increasing modularity, while blocks with negative assortativity decrease it. Assortativity values vary between -1 for a fully non-assortative block (all edges are to other blocks), and 1 for a fully assortative one (all edges are internal to the block).

WGCNA and MMC

We use WGCNA to cluster the genes into modules using the topological overlap measure (TOM) similarity with a soft threshold of 6 in a signed similarity measure. WGCNA produces modules by cutting the hierarchical clustering tree at different heights, and we use the dynamic cutting option to create the modules. We use a signed network (as opposed to ignoring the sign of the correlation between genes) because inspection of the gene network graph reveals large groups of genes linked by negative correlations in our data, suggesting a large-scale structure that would be obscured by using the unsigned method. Signed similarity has been shown to lead to more robust modules (Mason et al., 2009), and in tuning WGCNA we were able to cluster more genes and find more modules using the signed method. MMC has no option to use the sign of the correlation, so we use the absolute value of the Spearman correlations.

Gene Ontology enrichment

We assess the biological relevance of the clustering obtained by each method by comparing their gene ontology (GO) enrichment. We filter enrichment using a Benjamini-Hochberg FDR rate of 5%, with a minimum of 4 genes in the enriched set. All gene ontology analyses were done using the clusterProfiler R package v4.2.2 (Wu et al., 2021) and the Org.Dm.eg.db database package v3.15 (Carlson, 2022).

Table 1: Fraction of blocks at each level of the SBM hierarchy that show significant GO enrichment at the 5% FDR level with a minimum of 4 genes in the enriched set.

Tissue	Level 1	Level 2	Level 3	Level 4	Level 5
Head	65% (53/82)	100% (21/21)	100% (6/6)	100% (3/3)	100% (2/2)
Body	65% (51/78)	95% (20/21)	100% (9/9)	100% (3/3)	100% (2/2)

Results

Gene clustering

To assess the consequences of assuming that communities in transcriptional are networks assortative, we compared the performance of clustering algorithms that rely on modularity maximization (WGCNA and MMC) to the performance of the SBM. For this, we run the three clustering algorithms on the same gene co-expression matrices. MMC failed to cluster most genes, placing almost all genes into the same large module. Given this poor performance on our data, we do not discuss MMC further and instead focus on comparing SBM to WGCNA. To distinguish between gene clusters derived from the SBM and WGCNA, we refer to the former as 'blocks', and to the latter as 'modules'. Gene clustering for all methods is presented in Supporting Information Table S1.

Using the SBM, in both head and body, we were able to cluster all genes, identifying a nested partition with 5 levels (fig. 2). We obtain 2 blocks for both tissues at level 5 (the coarsest); 3 blocks for both tissues at level 4; 6 block for the head and 9 blocks for the body in level 3; 21 blocks for both tissues at level 2; and, finally, 82 blocks for the head and 78 blocks for the body at level 1. In what follows, when discussing specific SBM blocks, we either explicitly define which level of the nested hierarchy we are referring to or give the full path to a given block. For example, level-1 block 12 can also be referred to as 12-7-2-2-1 (see fig. 1 for an illustration on how to interpret these labels).

In contrast with the SBM, WGCNA was able to cluster only 30-40% of the genes. These 2118 genes in the body and 1600 genes in the head were partitioned into 7 modules in both tissues. To assess whether the gene clusters inferred by each algorithm are similar, we compared the results of WGCNA to the SBM blocks at level 3 (fig. 3). We focused on level 3 instead of level 1 (the finest level) because the number of blocks at this level (6 in the head and 9 in the body) are similar to the number of modules in WGCNA (7 modules for both tissues). Overall, the partitions are different, but WGCNA and the SBM do capture some common signals, evidenced by the tendency of Level-3 blocks that share the same Level-4 blocks to be grouped into the same modules in WGCNA. For example, Level-3 blocks 0, 2, 5, and 6 in the body are split between modules 3 and 4, and these blocks are all in the same Level-4 block 0, suggesting some similarity that could explain the WGCNA clustering. Blocks 7 and 9 are both fully assigned to module 2. Also in the body, we find a similar pattern for Level-3 blocks 1, 3, and 4, which are mostly split between modules 1 and 2. In the head, Level-3 block 4 is all assigned to modules 1 and 3. Level-3 blocks 1 and 2 are mostly split between modules 1 and 3, and both are in Level-4 block 2. Importantly, level 3 is an intermediate level in the clustering hierarchy resolved by the SBM, and at finer levels (i.e., level 2 and 1) the gene groups are smaller and functionally more specific.

Modularity and assortativity

Because the SBM does not use modularity maximization to find communities, we were able to use the resulting clustering to measure, in an unbiased manner, the assortativity of individual blocks and the overall degree of modularity of the transcriptional networks in the head and the body. We find that modularity and assortativity are markedly lower in the body (fig. 4). Several blocks in the body have negative assortativity (being more connected across blocks than within), and the maximum value of modularity is 0.035 at level 4 of the nested hierarchy. Even so, several blocks show GO enrichment throughout the distribution of assortativity. In the head, overall modularity is higher, with a peak at 0.14 in level 3. This is still a relatively low value and illustrates how assuming the gene network should be modular can prevent us from finding an informative clustering. All but 5 blocks in the head show positive assortativity, and again GO enrichment is present throughout the assortativity range (fig. 4).

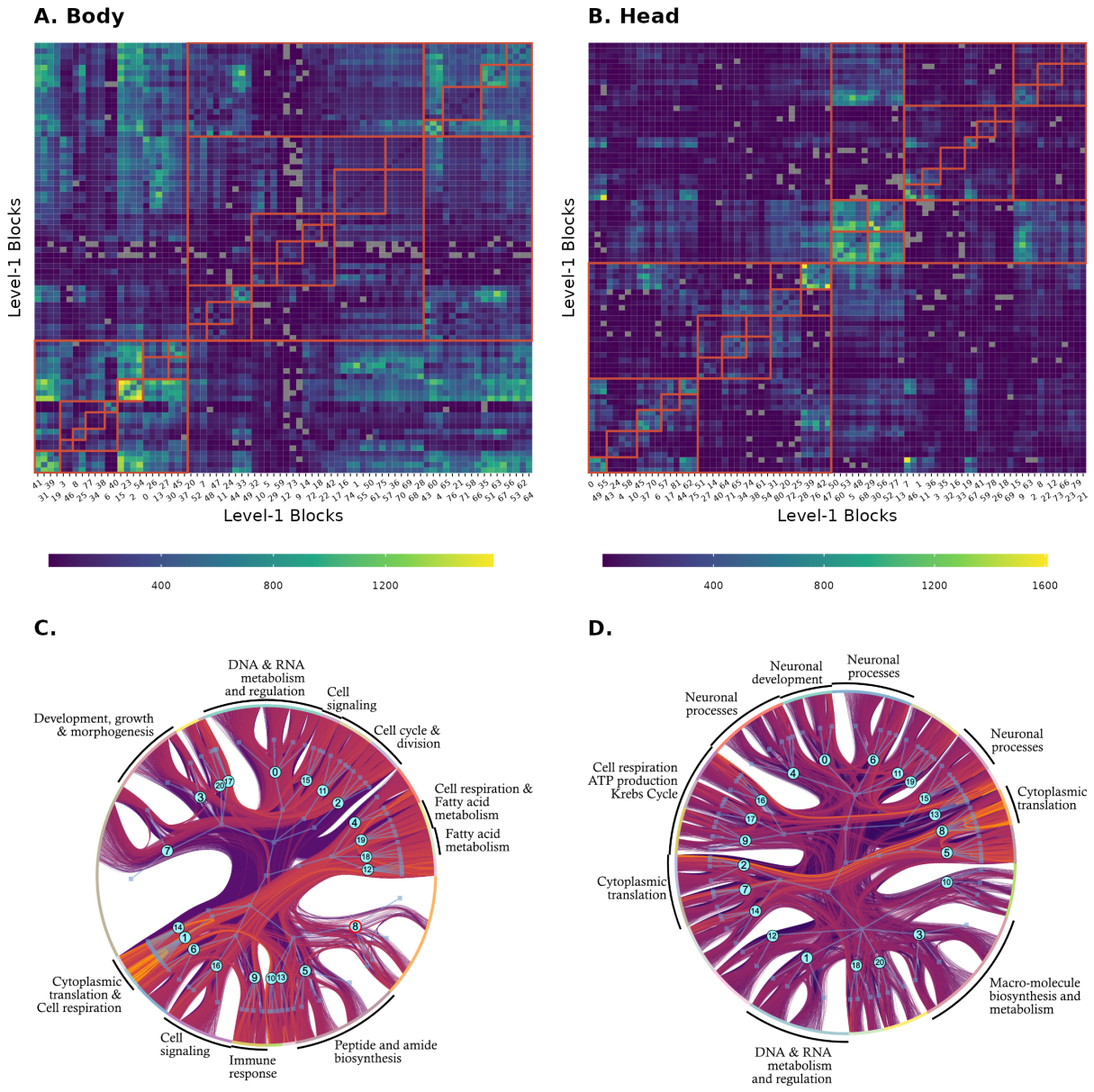

Figure 2: Matrix and graph representations of the SBM clustering. A and B: SBM Level-1 blocks are colored by the number of edges within and between blocks. Gray squares represent pairs of unconnected blocks. The upper levels of the nested hierarchy are shown by the red lines. C and D: A full representation of the fitted block model. Genes are shown at the perimeter, colored by their level 2 blocks. The internal graph shows the hierarchical structure of the fitted SBM. Numbers in blue circles correspond to the level-2 block. Arrows between level-1 blocks and genes are omitted, unlike fig. 1. A subsample of 30.000 edges is shown connecting the genes, and edges are colored according to their transformed weights, with more positive weights plotted on top and more yellow. External labels refer to a non-exhaustive subset of level-2 blocks with clear biological functions inferred from interpreting GO enrichment. Level-2 block 8 in the body, with the blue circle highlighted in red, is the only level-2 block with no GO enrichment.

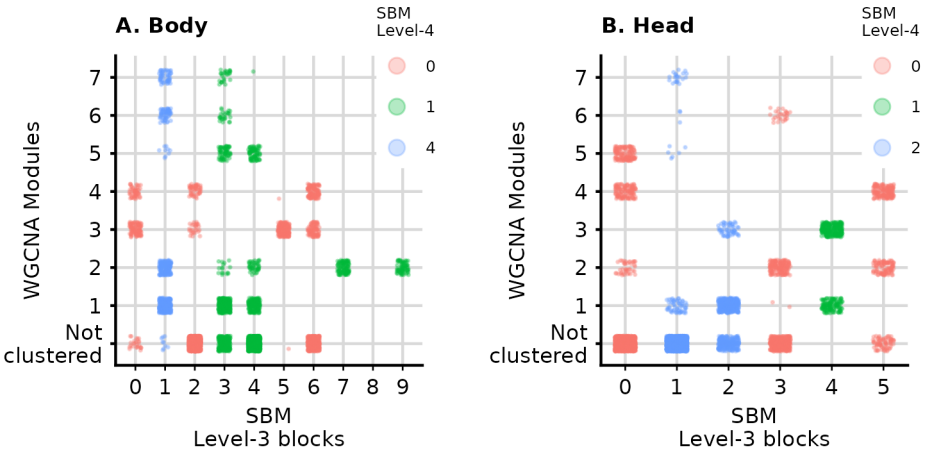

Figure 3: Comparison of the clustering in WGCNA and levels 3 and 4 of the SBM hierarchy for the gene expressions in the body (left) and the head (right). Each point corresponds to a gene. The x-axis corresponds to the Level-3 SBM blocks, and the y-axis the WGCNA modules. Colors correspond to the (coarser) level 4 of the SBM.

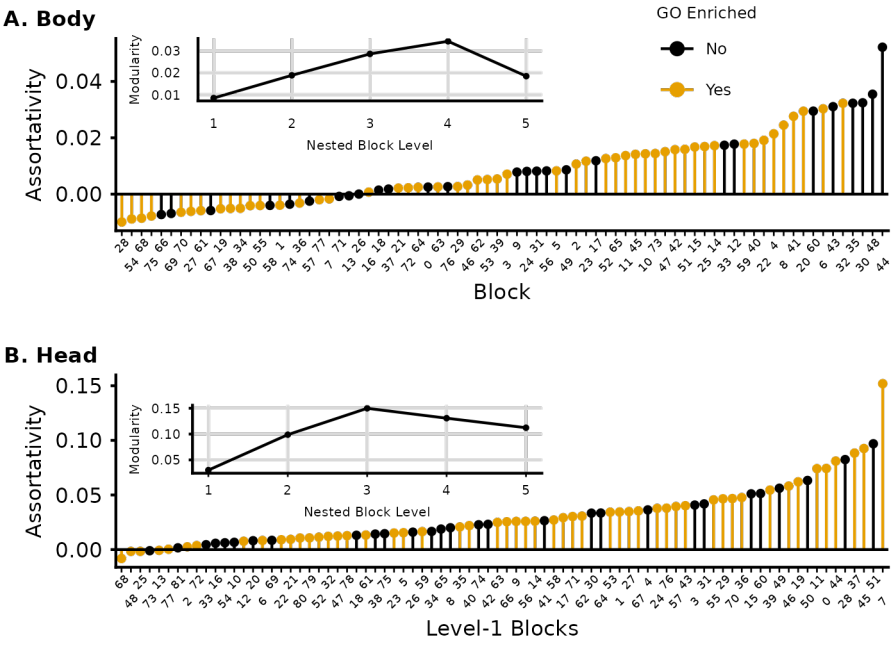

Figure 4: Assortativity measured in the SBM level-1 blocks and Newman Modularity (average assortativity) at each level of the SBM hierarchy (inset). GO enriched blocks are shown in yellow and appear throughout the distribution of assortativity. Modularity is much higher in the head, and it peaks at level 3, dropping in upper levels. Body has a much higher number of non-assortative blocks and lower modularity at all levels. Modularity peaks at level 4 in the body and drops strongly at level 5. Interestingly, the 4 most assortative blocks in the body do not show significant GO enrichment.

Gene Ontology enrichment

Most blocks in SBM show significant GO enrichment (Table 1). Enriched level-1 blocks show between 1 and 202 enriched terms, with a mean of 24 terms and median of 13 terms. Level-2 blocks show between 2 and 297 enriched terms, with a mean of 58 terms and median of 38 terms. Furthermore, several blocks show remarkable consistency in their enrichment. For example, Level-3 block 0 in the head is related to neural signaling, sensory perception, and signal transduction. Examining lower levels of the hierarchy, we see that often the daughter blocks at Level-2 are also enriched with generally similar terms, as expected, but these tend to become more specific as we go down the hierarchy. For example, Level-2 blocks 4 and 6: (4-0-0-0) G protein-coupled receptor signaling pathway, detection of light stimulus, phototransduction; (6-0-0-0) synapse organization, axon development, cell-cell signaling, behavior. Many of these enrichments are exclusive to one of the level-1 blocks. Most other Level-2 and Level-1 blocks are readily identifiable as related to development, DNA transcription, cell respiration, cell cycle regulation, immune response, sugar metabolism, among others (fig. 2). All WGCNA modules show GO enrichment (but modules 5, 6, and 7 in the body show only one or two enriched terms, and could be false positives. The more convincing specific enrichments show several related enriched terms). The remaining modules show between 20 and 462 enriched terms, with a mean of 116 terms and median of 58 terms. In general, these enrichments tend to be less specific than the SBM blocks, spanning several biological processes. Supporting Information Table S2 shows GO enrichment for all SBM blocks and WGCNA modules.

Notable individual clusters

Level 2 block 0-0-0 in the head is one of the easiest to interpret, being entirely related to nervous tissue function. Fig. 5 shows the top 8 GO categories for each of the level-1 blocks in block 0-0-0, and the most neuronal enriched WGCNA grouping, module 4. The SBM blocks separate vesicle exocytosis, neuronal differentiation, phototransduction, synaptic signaling, and, interestingly, there is a block related to mRNA processing, which is notable given that alternative splicing is thought to be more common in brain tissues (Su et al., 2018). WGCNA module 4 recovers some of this enrichment but in a less granular way. The cell adhesion and developmental part of the enrichment in block 0-0-0 is separated between WGCNA modules 4 and 5. Some of the level-1 blocks shown in fig. 5 are among the most assortative (above 0.03, see fig. 4 B), and so are prime candidates for detection in WGCNA. The alternative splicing module has a much lower assortativity, so it is not surprising that WGCNA could not detect it.

Some of the most specific enrichments in the SBM are the translation-related blocks. In both body and head, ribosomal proteins are clustered in small and highly enriched level-1 blocks: 6 level-1 blocks in the head and 11 in the body are composed of virtually only ribosome-related protein genes. All are very small, being composed of between 10-30 genes, have low assortativity (fig. 6), and are enriched for very few terms, almost all related to translation. In the body, all of these translation blocks are grouped in level-4 block 1; in the head, they are split between level-4 blocks 1 and 2. Both groups are visible in fig. 2 . There is no equivalent module in WGCNA, but all translation-related genes are in the same much larger modules (module 2 in the head, 295 genes; and module 2 in the body, 345 genes), both of which show enrichment for translation but also several other categories. In the body WGCNA module 2, we see 68 enriched terms related to translation, cell respiration, and several small molecules' metabolic processes; in module 2 of the head tissue, we see 35 enriched terms related to translation, cell respiration, and muscle development. The level-2 clustering of level-1 blocks in the SBM is also informative. In the head, all the translation level-1 blocks are in their own level-2 blocks (8-4-1-1, 7-2-2-1, and 2-2-2-1). In contrast, in the body, the level-1 translation blocks sometimes share level-2 blocks with cell respiration blocks: 1-7-1-1 is composed exclusively of level-1 blocks related to translation, but block 14-9-1-1 is split into translation

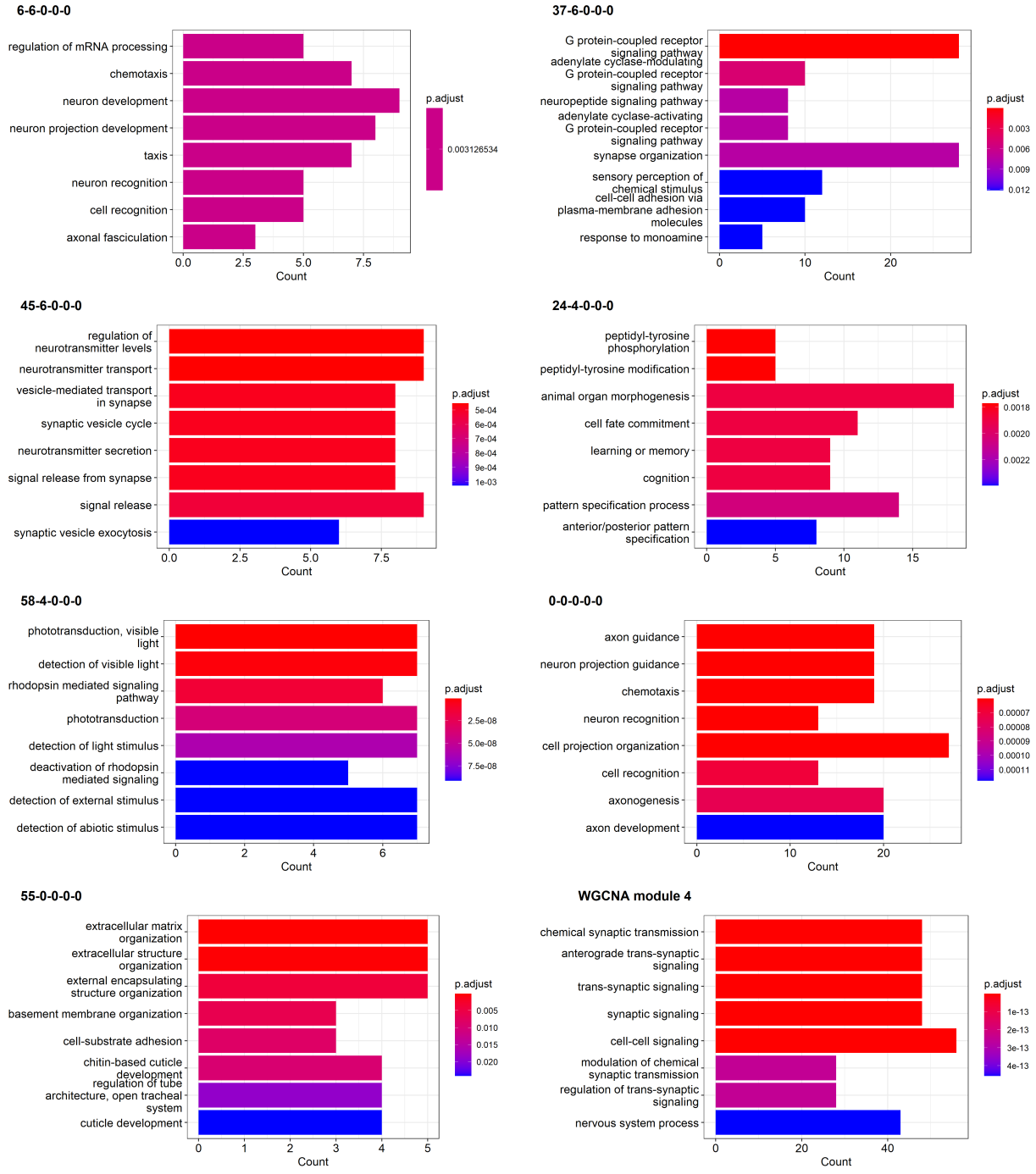

Figure 5: Enriched GO categories in a level-3 block in the head (0-0-0), related to neural signaling. Panels show the corresponding level-1 blocks. Bars correspond to the top 8 GO categories, the x-axis shows the number of genes associated with each term. The last panel shows the most similar WGCNA module, which also contains signaling-related genes, but at a lower resolution and fails to cluster the phototransduction genes, which are in WGCNA module 5 (not shown, but see SI table 2.1).

and mitochondrial respiration level-1 blocks. WGCNA also places cell respiration-related genes in the body on the same module 2.

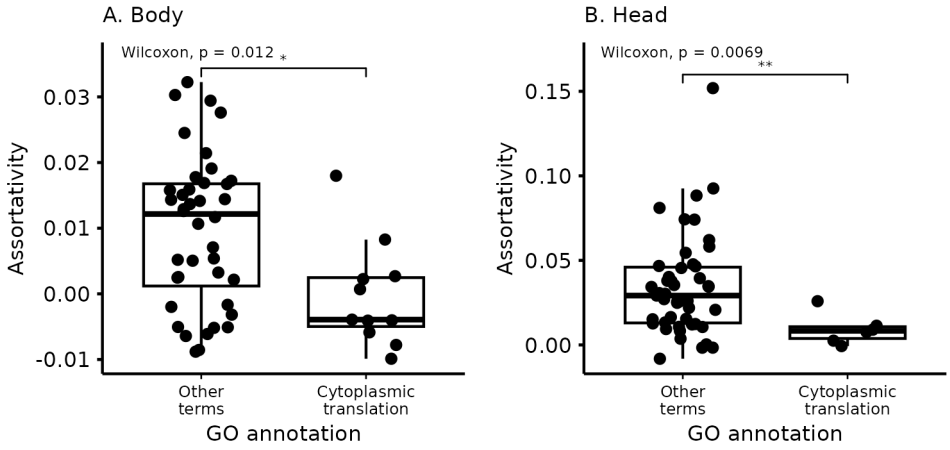

Figure 6: Comparison of assortativity values between level-1 blocks enriched for cytoplasmic translation and all other blocks. Blocks enriched for cytoplasmic translation tend to be less assortative.

Discussion

Here we have used the Stochastic Block Model to explore the organization of gene co-expression networks in female *Drosophila melanogaster*. The SBM, in contrast with other methods explored here, clusters genes by finding the groups that capture as much information on the network of interactions as possible, and was able to (i) cluster all genes into blocks; (ii) identify blocks with both, high resolution (few genes per block) and high functional content (significant GO associations); and (iii) identify blocks that are assortative (higher within- than between-block correlation) as well as non-assortative. This last point exemplifies the novelty of the SBM approach. Using the SBM implies a shift on how we explore co-expression networks: instead of assuming the network is modular and clustering genes based on this assumption, we uncover clusters based on their information content and ask if the resulting groups are modular. Surprisingly, the answer is not always.

Community detection methods

The other clustering approaches we use, that explicitly search for assortative modules, carry important downsides. Methods that use modularity maximization, like MMC, are subject to know statistical problems, surprisingly being prone to both overfitting (finding modular community structure where there is none, Guimerà et al. (2004)) and under-fitting (failing to find modular structure), due to a problem known as the resolution limit, which causes small modules to be incorrectly clustered together in large networks (Fortunato & Barthélemy, 2007). Using WGCNA involves manually tuning several parameters: the choice of using a hard or soft threshold, the exponent in the threshold, the method of separating the genes included in the hierarchical clustering into the modules. These are free parameters that can drastically change the number of genes that are clustered and the number and size of modules. For tuning these parameters, the WGCNA workflow leans heavily on the expectation that gene co-expression networks should be approximately scale-free (Bergmann et al., 2004; Dong & Horvath, 2007; Jeong et al., 2000), but, despite its popularity, this expectation might be unwarranted (Broido & Clauset, 2019; Keller, 2005; Khanin & Wit, 2006; Stumpf et al., 2005). Even with optimal parameters, WGCNA often fails to assign a substantial proportion of genes to any module. While WGCNA is efficient

in finding hub genes, if some functional gene group does not have a hub or has low average similarities, this group will never be identified. Both limitations potentially leave biological insight on the table by ignoring network structures that are different from what the method expects. In contrast, all the freedom in the SBM is restricted to the creation of the network, which we discuss below, and no assumption is made on the structure of the communities in the network. The clustering procedure is completely parameter free, and choices regarding how to model the weights between edges can be made by selecting the model with the shortest description length (Peixoto, 2017). This is a significant advantage for applying the SBM in domains where we lack relevant domain expertise and can't easily tune the parameters for the other clustering methods. Furthermore, the SBM finds a much larger number of communities that are guaranteed to be statistically supported, greatly improving the resolution of the clustering and allowing for more precise biological interpretation of the resulting blocks.

One aspect we did not explore here is the estimation of the gene co-expression network itself, before any attempt at finding communities. Both types of methods used weighted networks: fully connected ones for the WGCNA and MMC pipelines (as per these methods' suggested workflow) and a sparser network for the SBM model fitting, due to computational constraints. Estimating these weights (gene expression correlations) is an error-prone process, as we are estimating many more weights than we have measured individuals, leading to potentially poor estimates (Schäfer & Strimmer, 2005). While the procedures we used here are commonplace, there are more principled ways of building co-expression networks (Peel et al., 2022), and this is an aspect of the usual transcriptomics workflow that could potentially see massive improvements in the near future. Methods like the graphical lasso have been used in this context (Lingjærde et al., 2021; Lyu et al., 2018; Seal et al., 2023), and the expectation is that, when compared to fully connected or thresholded networks, these inferred networks should provide much better estimates of gene-gene connections and weights. Additionally, it is possible to combine community detection via the SBM with network inference, simultaneously using information about community structure to inform the network inference and vice-versa (Peixoto, 2019).

Modularity in gene co-expression networks

Beyond the methodological and practical advantages discussed above, the fact that the SBM does not find gene clusters by attempting to maximize their modularity has major implications for our understanding of biological networks, as it allows us to *measure* the modularity of a given network. In doing so, we find that *D. melanogaster* transcriptomes are organized into assortative as well as non-assortative gene clusters. The latter, however, could not have been identified by methods that assume assortative modules. The possibility of quantifying, in a continuous scale, the degree of modularity (assortativity) of each gene block allowed us to compare the gene co-expression networks derived from head and body tissue, and uncover marked differences in their overall degree of modularity. This opens the possibility of expanding this comparison to different cell types, organs, and even species to get a comprehensive understanding of how modular these biological networks really are.

These results warrant a discussion about the origin of the assumption that gene expression networks are modular. Modularity, understood as the relative independence between groups of complex traits, is often invoked to explain the evolvability of complex phenotypes and has functioned as a unifying concept at several levels of organization with great success (Melo et al., 2016; Wagner et al., 2007; Zelditch & Goswami, 2021). Traits in an organism need to have some level of integration, of interdependence, to form a functioning individual. This necessary interaction between parts poses a problem for understanding the evolution of complex traits, as interdependencies are expected to lead to important evolutionary restrictions (Orr, 2000). Modularity provides a simple solution to this problem as it allows organisms to maintain their function unchanged by coordinating

simultaneous evolutionary changes in all related traits while keeping unrelated traits undisturbed (Ancel & Fontana, 2000; Cheverud, 1996; Wagner & Altenberg, 1996; Wagner & Zhang, 2011). The conceptual usefulness of modularity has informed much of our thinking on how complex traits should be structured, producing a large literature dedicated to finding modules and testing for their existence (Esteve-Altava, 2017). A large part of the literature on modularity developed in the context of morphological traits, and morphological traits being organized into modules can be interpreted as a consequence of the very concrete structural and developmental constraints that lead to the formation and allow proper functioning of these individual body elements (Marcucio et al., 2011; Shirai & Marroig, 2010). These constraints are easy to visualize, as morphological traits like bones and muscle have to fit together in order to function, and individuals in which perturbations are large enough to disrupt these couplings are not viable. The result is a modularity pattern that is kept stable by these structural and functional constraints (Cheverud, 1984, 2004; Porto et al., 2009). However, no such clear structural and physical constraints exist on gene expression, and the interaction between groups of genes can happen through more dynamic and varied mechanisms. While we might expect related genes to be co-expressed and therefore highly correlated, non-linear phenomena can lead to a complete decoupling of the expression levels of co-expressed genes. For example, the effect of gene A on gene B could have a saturation point after which increasing expression of gene A no longer leads to higher levels of gene B, and no correlation is detected in this regime, even if the genes are co-expressed. The marked difference in the level of modularity across the two tissues in our samples illustrates just how variable modularity can be, even within the same species, sex, and population. Furthermore, modularity is not a necessary feature of biological organization (even in the case of evolvability, see Hansen, 2003; Pavlicev & Hansen, 2011; Roseman et al., 2009), and only searching for modularity can blind us to alternative organizations, as we have shown. Indeed, the profound interconnectedness of gene regulation networks has led to a small revolution in our understanding of disease and complex traits (Boyle et al., 2017). The very high dimensionality of gene co-expression networks also allows for genes to be similar in ways that do not lead to high correlations. For example, two genes might be connected to the same genes in different communities, but not among themselves. This similarity would likely be missed by modularity maximization or hierarchical clustering because these genes would not form a classic assortative unit. Meanwhile, the SBM would correctly identify these genes connecting two modules as being similar due to their shared connectivity pattern. Having access to these types of blocks, which are real but non-assortative, could bring new insight into the organization of gene co-expression networks.

Conclusion

Here we find that non-modular blocks are widespread in gene co-expression networks, and that the evidence for their functional relevance is as strong as for modular blocks. This highlights the need to incorporate other sources of information, beyond assortativity, when exploring biological networks. More studies using methods that don't rely on modularity maximization will be needed to determine whether there are general patterns of non-modular organization. For example, here we find that, despite the differences in gene clusters between body and head, the non-modular blocks tend to be associated with cytoplasmic translation. Will this emerge as a general feature of transcriptomes?

Supporting information

Supporting information can be found at https://github.com/diogro/SBM_manuscript. Code for using graph-tools to cluster expression data using the SBM can be found at https://github.com/ayroles-lab/SBM-tools

Author Contributions

Conceptualization: D.M., L.P., and J.A. **Data Curation**: L.P. **Formal Analysis**: D.M. **Funding Acquisition**: D.M., L.P., and J.A. **Investigation**: D.M., L.P., and J.A. **Methodology**: D.M., and L.P. **Project Administration**: J.A. **Resources**: J.A. **Software**: D.M., and L.P. **Supervision**: J.A. **Validation**: D.M. **Visualization**: D.M. **Writing – Original Draft Preparation**: D.M. **Writing – Review & Editing**: D.M., L.P., and J.A.

Acknowledgments

We thank all members of the Ayroles lab for their support. We thank Monique Simon and Cara Weisman for their thoughtful comments on an earlier version of the manuscript. We also thank Tiago Peixoto for help in using graph-tool. D.M. is funded by a fellowship from the Princeton Presidential Postdoctoral Research Fellows Program. L.P. was funded by a Long-Term Postdoctoral Fellowship from the Human Frontiers Science Program and is funded by the Max Planck Society. J.A. is funded by grants from the NIH: National Institute of Environmental Health Sciences (R01-ES029929) and National Institute of General Medical Sciences (NIGMS) (R35GM124881). This study was supported in part by the Lewis-Sigler Institute for Integrative Genomics at Princeton University. We also acknowledge that the work reported in this paper was substantially performed using the Princeton Research Computing resources at Princeton University which is a consortium of groups led by the Princeton Institute for Computational Science and Engineering (PICSciE) and Office of Information Technology's Research Computing.

References

Ancel, L. W., & Fontana, W. (2000). Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool.*, *288*(3), 242–283.

Baum, K., Rajapakse, J. C., & Azuaje, F. (2019). Analysis of correlation-based biomolecular networks from different omics data by fitting stochastic block models. *F1000Res.*, *8*, 465. https://doi.org/10.12688/f1000research.18705.2

Bergmann, S., Ihmels, J., & Barkai, N. (2004). Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.*, *2*(1), E9. https://doi.org/10.1371/journal.pbio.0020009

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, *169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nat. Commun.*, *10*(1), 1017. https://doi.org/10.1038/s41467-019-08746-5

Carlson, M. (2022). *Org.dm.eg.db: Genome wide annotation for fly*.

Cheverud, J. M. (1984). Quantitative genetics and developmental constraints on evolution by selection. *J. Theor. Biol.*, *110*(2), 155–171. https://doi.org/10.1016/S0022-5193(84)80050-8

Cheverud, J. M. (1996). Developmental Integration and the Evolution of Pleiotropy. *Integr. Comp. Biol.*, *36*(1), 44–50. https://doi.org/10.1093/icb/36.1.44

Cheverud, J. M. (2004). Modular pleiotropic effects of quantitative trait loci on morphological traits. In G. Schlosser G & Wagner (Ed.), *Modularity in development and evolution* (pp. 132–153). University of Chicago Press.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nat. Biotechnol.*, *23*(12), 1499–1501. https://doi.org/10.1038/nbt1205-1499

Dam, S. van, Võsa, U., Graaf, A. van der, Franke, L., & Magalhães, J. P. de. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.*, *19*(4), 575–592. https://doi.org/

10.1093/bib/bbw139

Dong, J., & Horvath, S. (2007). Understanding network concepts in modules. *BMC Syst. Biol.*, *1*, 24. https://doi.org/10.1186/1752-0509-1-24

Esteve-Altava, B. (2017). In search of morphological modules: A systematic review. *Biol. Rev. Camb. Philos. Soc.*, *92*(3), 1332–1347. https://doi.org/10.1111/brv.12284

Fortunato, S., & Barthélemy, M. (2007). Resolution limit in community detection. *Proc. Natl. Acad. Sci. U. S. A.*, *104*(1), 36–41. https://doi.org/10.1073/pnas.0605965104

Guimerà, R., Sales-Pardo, M., & Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, *70*(2 Pt 2), 025101. https://doi.org/10.1103/PhysRevE.70.025101

Hansen, T. F. (2003). Is modularity necessary for evolvability? Remarks on the relationship between pleiotropy and evolvability. *Biosystems.*, *69*(2-3), 83–94.

Imenez Silva, P. H., Melo, D., & Mendonça, P. O. R. de. (2017). Insights from systems biology in physiological studies: Learning from context. *Cell. Physiol. Biochem.*, *42*(3), 939–951. https://doi.org/10.1159/000478648

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, *407*(6804), 651–654. https://doi.org/10.1038/35036627

Karrer, B., & Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, *83*(1 Pt 2), 016107. https://doi.org/10.1103/PhysRevE.83.016107

Keller, E. F. (2005). Revisiting "scale-free" networks. *Bioessays*, *27*(10), 1060–1068. https://doi.org/10.1002/bies.20294

Khanin, R., & Wit, E. (2006). How scale-free are biological networks. *J. Comput. Biol.*, *13*(3), 810–818. https://doi.org/10.1089/cmb.2006.13.810

Kirkpatrick, S., Gelatt, C. D., Jr, & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559. https://doi.org/10.1186/1471-2105-9-559

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, *15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, *3*(9), 1724–1735. https://doi.org/10.1371/journal.pgen.0030161

Lingjærde, C., Lien, T. G., Borgan, Ø., Bergholtz, H., & Glad, I. K. (2021). Tailored graphical lasso for data integration in gene network reconstruction. *BMC Bioinformatics*, *22*(1), 498. https://doi.org/10.1186/s12859-021-04413-z

Lyu, Y., Xue, L., Zhang, F., Koch, H., Saba, L., Kechris, K., & Li, Q. (2018). Condition-adaptive fused graphical lasso (CFGL): An adaptive procedure for inferring condition-specific gene co-expression network. *PLoS Comput. Biol.*, *14*(9), e1006436. https://doi.org/10.1371/journal.pcbi.1006436

Magwene, P. M. (2001). New tools for studying integration and modularity. *Evolution*, *55*(9), 1734–1745.

Marcucio, R. S., Young, N. M., Hu, D., & Hallgrimsson, B. (2011). Mechanisms that underlie co-variation of the brain and face. *Genesis*, *49*(4), 177–189. https://doi.org/10.1002/dvg.20710

Mason, M. J., Fan, G., Plath, K., Zhou, Q., & Horvath, S. (2009). Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics*, *10*, 327. https://doi.org/10.1186/1471-2164-10-327

Melo, D., Porto, A., Cheverud, J. M., & Marroig, G. (2016). Modularity: Genes, development and evolution. *Annu. Rev. Ecol. Evol. Syst.*, *47*(1), 463–486. https://doi.org/10.1146/annurev-ecolsys-121415-032409

Morelli, L., Giansanti, V., & Cittaro, D. (2021). Nested stochastic block models applied to the analysis of single

cell data. *BMC Bioinformatics*, *22*(1), 576. https://doi.org/10.1186/s12859-021-04489-7

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Olson, E. C., & Miller, R. L. (1958). *Morphological integration*. University of Chicago Press.

Orr, H. A. (2000). Adaptation and the cost of complexity. *Evolution*, *54*(1), 13–20. https://doi.org/10.1111/j.0014-3820.2000.tb00002.x

Pallares, L. F., Melo, D., Wolf, S., Cofer, E. M., Varada, V., Peng, J., & Ayroles, J. F. (2023). Saturating the eQTL map in drosophila melanogaster: Genome-wide patterns of cis and trans regulation of transcriptional variation in outbred populations. In *bioRxiv* (p. 2023.05.20.541576). https://doi.org/10.1101/2023.05.20.541576

Pallares, L. F., Picard, S., & Ayroles, J. F. (2020). TM3'seq: A Tagmentation-Mediated 3' sequencing approach for improving scalability of RNAseq experiments. *G3 Genes|Genomes|Genetics*, *10*(1), 143–150. https://doi.org/10.1534/g3.119.400821

Pavlicev, M., & Hansen, T. F. (2011). Genotype-Phenotype Maps Maximizing Evolvability: Modularity Revisited. *Evol. Biol.*, *38*(4), 371–389. https://doi.org/10.1007/s11692-011-9136-5

Peel, L., Peixoto, T. P., & De Domenico, M. (2022). Statistical inference links data and theory in network science. *Nat. Commun.*, *13*(1), 1–15. https://doi.org/10.1038/s41467-022-34267-9

Peixoto, T. P. (2014). The graph-tool python library. *Figshare*. https://doi.org/10.6084/m9.figshare.1164194

Peixoto, T. P. (2017). Nonparametric bayesian inference of the microcanonical stochastic block model. *Phys Rev E*, *95*(1-1), 012317. https://doi.org/10.1103/PhysRevE.95.012317

Peixoto, T. P. (2018). Nonparametric weighted stochastic block models. *Phys Rev E*, *97*(1-1), 012306. https://doi.org/10.1103/PhysRevE.97.012306

Peixoto, T. P. (2019). Network reconstruction and community detection from dynamics. *Phys. Rev. Lett.*, *123*(12), 128301. https://doi.org/10.1103/PhysRevLett.123.128301

Porto, A., Oliveira, F. B. de, Shirai, L. T., De Conto, V., & Marroig, G. (2009). The Evolution of Modularity in the Mammalian Skull I: Morphological Integration Patterns and Magnitudes. *Evol. Biol.*, *36*(1), 118–135. https://doi.org/10.1007/s11692-008-9038-3

Roseman, C. C., Kenney-Hunt, J. P., & Cheverud, J. M. (2009). Phenotypic integration without modularity: Testing hypotheses about the distribution of pleiotropic quantitative trait loci in a continuous space. *Evol. Biol.*, *36*(3), 282–291. https://doi.org/10.1007/s11692-009-9067-6

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, *4*, Article32. https://doi.org/10.2202/1544-6115.1175

Seal, S., Li, Q., Basner, E. B., Saba, L. M., & Kechris, K. (2023). RCFGL: Rapid condition adaptive fused graphical lasso and application to modeling brain region co-expression networks. *PLoS Comput. Biol.*, *19*(1), e1010758. https://doi.org/10.1371/journal.pcbi.1010758

Shirai, L. T., & Marroig, G. (2010). Skull modularity in neotropical marsupials and monkeys: Size variation and evolutionary constraint and flexibility. *J. Exp. Zool. B Mol. Dev. Evol.*, *314*(8), 663–683. https://doi.org/10.1002/jez.b.21367

Stone, E. A., & Ayroles, J. F. (2009). Modulated modularity clustering as an exploratory tool for functional genomic inference. *PLoS Genet.*, *5*(5), e1000479. https://doi.org/10.1371/journal.pgen.1000479

Stumpf, M. P. H., Ingram, P. J., Nouvel, I., & Wiuf, C. (2005). Statistical model selection methods applied to biological networks. *Transactions on Computational Systems Biology III*, 65–77. https://doi.org/10.1007/11599128/_5

Su, C.-H., D, D., & Tarn, W.-Y. (2018). Alternative splicing in neurogenesis and brain development. *Front Mol Biosci*, *5*, 12. https://doi.org/10.3389/fmolb.2018.00012

Wagner, G. P., & Altenberg, L. (1996). Perspective: Complex adaptations and the evolution of evolvability. *Evolution*, *50*(3), 967–976. https://doi.org/10.2307/2410639

Wagner, G. P., Pavlicev, M., & Cheverud, J. M. (2007). The road to modularity. *Nat. Rev. Genet.*, *8*(12), 921–931. https://doi.org/10.1038/nrg2267

Wagner, G. P., & Zhang, J. (2011). The pleiotropic structure of the genotype-phenotype map: The evolvability of complex organisms. *Nat. Rev. Genet.*, *12*(3), 204–213. https://doi.org/10.1038/nrg2949

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, xiaochong, Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Zelditch, M. L., & Goswami, A. (2021). What does modularity mean? *Evol. Dev.*, e12390. https://doi.org/10.1111/ede.12390

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, *4*(1), Article17. https://doi.org/10.2202/1544-6115.1128

Zhang, L., & Peixoto, T. P. (2020). Statistical inference of assortative community structures. *Phys. Rev. Research*, *2*(4), 043271. https://doi.org/10.1103/PhysRevResearch.2.043271