NETSCOPE: Information-Theory based Network Discovery and Analysis

Tido Bergmans[1*], Tousif Jamal[1*] and Tansu Celikel[1,2]

[1] Department of Neurophysiology, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands
t.bergmans@neurophysiology.nl, t.jamal@neurophysiology.nl
[2] School of Psychology, Georgia Institute of Technology, Atlanta - GA, USA
celikel@gatech.edu

Abstract. Molecular, cellular, structural, and functional networks in the brain each represent different levels of organization and complexity. These networks are interconnected, and their interactions underlie the brain's abilities to process information, regulate bodily functions, and mediate behavior. While understanding the emergence of higher-level networks (structural and functional) from the interactions of lower-level (molecular and cellular) ones is crucial, a universal analytical method applicable to all brain network types has been lacking. Here we introduce an open-source toolbox (NETSCOPE) that identifies weighted network architectures using mutual information (MI) and variation of information (VI). We demonstrate the accuracy of the resulting networks using synthetic data and by recreating five molecular networks in S. cerevisiae. We finally deploy NETSCOPE to identify cell type specific transcriptional networks and reconstruct the brain-wide neural networks that encode touch in the mouse brain. The code is made available in Python, Google Colab, Jupyter Notebook, MATLAB, and Octave. Beyond its core functionality, this network discovery method holds potential for broader applications, including the development of bioinspired sparse artificial networks.

Keywords: mutual information, molecular networks, cellular networks, functional networks, network sparsification

1 Introduction

The brain's remarkable capabilities in information processing is grounded in the intricate web of molecular, cellular, structural, and functional networks. These networks, each representing a distinct level of organization and complexity, are not isolated entities but are deeply interconnected, with their interactions forming the basis of the brain's multifaceted functions. Despite the critical importance of understanding how higher-level networks (structural and functional) emerge from the interplay of lower-level ones (molecular and cellular), the field has long been challenged by the absence of a universal analytical framework that can

seamlessly navigate across these diverse network types, rather than big data that provides insight into the organization of these networks.

Thanks to the increased accessibility of next-gen sequencing and high throughput-omics approach, for example, it is now possible to study molecular network interactions systematically from genes to metabolites; even though the computational methods to reconstruct molecular networks are still in their infancy, systems theoretical approaches already contribute to our improved understanding of genetic disorders [1], oncogenomics [2], brain plasticity [3–5] and cellular circuits [6–8]. By studying the molecular mechanisms of a cell as a network, the focus of the analysis can shift from differential expression of individual genes to functional correlations between groups of genes, revealing genetic pathways and their interactomes.

Molecular network inference is typically performed by calculating similarity measures for every pair of genes in a sampled dataset [9, 10]. Since similarity measures determine what genes are connected, the choice of similarity measure greatly influences the resulting network. Pearson's correlation, for example, is often used as a similarity measure because it is computationally efficient, but it can only recognize linear dependencies and is sensitive to outliers [13]. Deep learning techniques, such as DeepSEM uses a deep generative model to create a gene regulatory network of single-cell RNA sequencing [11]. Toolboxes such as GeNeCK [12] create a network interface to visualize the gene regulatory network and compare different similarity measure methods.

A promising avenue for addressing this challenge lies in the application of mutual information metrics. Mutual information, a concept borrowed from information theory, offers a powerful means to quantify the information shared between different elements of coupled systems, making it particularly suited for analyzing the intricacies of brain networks from molecular to functional levels. Unlike traditional methods that may only capture linear relationships or require prior assumptions about network structure, mutual information provides a flexible and nonparametric approach that can adapt to the inherent complexities of brain organization and is invariant under reparameterization. Therefore, any prior transformations of the data (e.g. log-transformation or discretization methods) do not influence the resulting (computed) networks. This makes MI a useful tool for cross-analyzing molecular data from multiple datasets.

Previous approaches to MI-based molecular network inference include ARACNE [15], parmigene [16] and PANA [17]. These approaches construct gene co-expression networks from transcriptional data using MI to infer connectivity between any given pairs of genes. While ARACNE uses Data Processing Inequality (DPI) to remove indirect edges from the resulting network, parmigene seeks minimal computational costs using parallel computing. The PANA algorithm uses functional gene clusters as a basis for network inference. Commonly, these methods allow visualization of binary networks where two nodes are connected when their MI is above a certain threshold. This threshold may be established by shuffling the transcriptional data and determining the p-value from the resulting MI values. Such a binary or unweighted network can be used to discover patterns of con-

nectivity and find clusters of nodes that are co-expressed. Biological networks, however, are not binary networks. The weight of the connection between any given two nodes carries information about functional and structural connectivity in the network. Using MI as a measure of similarity, and calculating Variation of Information (VI) rather than distance [15], can provide the necessary weight information and enable quantitative analysis of connectivity in weighted and directed networks. Because network statistics such as average path length and betweenness centrality can only be computed from a weighted network where edge weight corresponds to distance, this expansion of the MI would offer new opportunities to integrate graph-based analytical methods to study biological networks. NETSCOPE, a plug-and-play toolbox for constructing functional networks at the molecular level, provides the aforementioned functionality. The toolbox, natively compatible with Google Colab, Anaconda, MATLAB and Octave, includes a data processing pipeline that uses information theory to infer networks from sampled data. Functions could be directly compiled in Colab or in Python using Oct2Py [16]. The toolbox can be used for analysis and visualization of the resulting network, including pathway analysis, clustering, and data exporting to network visualization software.

NETSCOPE takes advantage of mutual information (MI) and Variation of Information (VI) to uncover correlations in biological data, such as gene expression data, transgenic line data, and construct networks. The unique properties of MI and VI enable a rapid and robust data processing pipeline, making NETSCOPE applicable for studying biological networks. To validate the effectiveness of the toolbox, we conducted rigorous tests using synthetic data with known network topology, ensuring its sensitivity and selectivity. Furthermore, we reconstruct transcriptional networks in different yeast species [20] and compare the results against established genetic networks in yeast. These comparisons demonstrate that NETSCOPE can accurately reconstruct genetic networks from gene expression data, regardless of parameter choices. Furthermore, we showcase the versatility of NETSCOPE in molecular network analysis. By examining the connectivity of cell type-specific networks in the mouse somatosensory cortex, and transgenic lines of virus injections in primary somatosensory cortex, we provide evidence of its utility. Additionally, NETSCOPE enables the creation and visualization of sex-dependent networks for housekeeping genes in the $L2/3$ layer of the mouse somatosensory cortex and the mapping of energy and density of the transgenic lines in the somatosensory cortex. Through these evaluations and applications, NETSCOPE demonstrates its efficacy in unraveling the complex dynamics of molecular networks. The network generated by NETSCOPE has a unique characteristic: it is a weighted network where the weights correspond to the calculated mutual information between pairs of nodes. This weighted network provides valuable insights into the intricate nature of network connections, including multilayered dependencies among nodes. This higher-order analysis of the network unveils a wealth of detailed information, allowing a deeper understanding of its complex structure.

2 Methods

NETSCOPE is a Python, Octave/MATLAB toolbox that can be used to construct and analyze correlation networks (Fig. 1). The toolbox has a modular organization. After data import, normalization, and entropy calculations, mutual information is calculated between any given node (e.g. transcript) in the network. "Network analysis" routines allow sparsification of the network by identifying indirect connections and provide tools for graph-theory-based network analysis. Users can optionally export the data for further analysis, e.g. in Gephi [21], for custom visualizations.

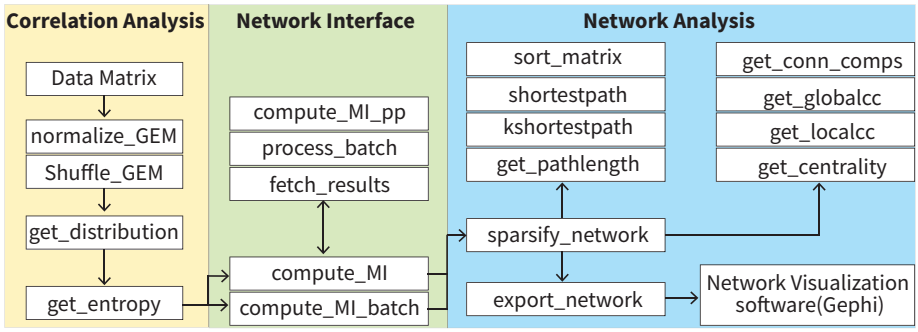


Fig. 1. An overview of NETSCOPE functions and workflow. The data processing pipeline can be separated into three modules: data preparation, network inference, and network analysis. Function lists in each module and data processing hierarchy are shown.

The toolbox comprises a data processing pipeline that imports a high-dimensional dataset organized as a matrix where columns are samples (e.g. cells) and rows are (molecular) targets (e.g. genes), and constructs a network based on correlations; see Fig. 2 for the implementation of the toolbox for network analysis from gene expression data. The toolbox includes functions to evaluate network structure and identify hubs, clusters, and pathways in the network. The following sections describe the different steps in the pipeline in more detail.

Gene expression data normalization. The gene expression data contains transcriptional data from a set of genes over a set of cells. These data are stored in a matrix $\mathbf{X}$ where X_{ij} represents the transcript count of gene i in cell (or sample) j. The transcription data is log-transformed and/or normalized with respect to total transcription per cell to account for differences in cell size and metabolic rate. The toolbox provides functions for shuffling or subsampling data for shuffle correction and bootstrapping.

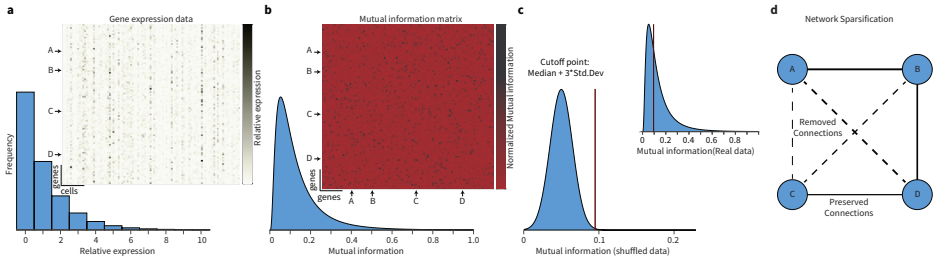


Fig. 2. Basic processing steps for quantification of co-expression network from gene expression data. a The pipeline starts with a data set organized as a matrix, e.g. RNA-sequencing data, where columns are samples (cells, repetitions) and rows are targets (e.g. transcript names). **b** From the expression data, the mutual information (MI) matrix is computed which stores pairwise correlations between genes. This matrix is interpreted as the adjacency matrix of a weighted, undirected network. **c** To establish a significance threshold for MI, a shuffle correction is performed for which MI matrices are computed from randomly shuffled gene expression data (see the section on "Shuffle correction"). A threshold is then based on the median and standard deviation of the shuffled MI distribution. **d** A network sparsification algorithm is applied to remove spurious correlations from the network. This algorithm iterates over the network connections, removing the edges that are the result of an indirect correlation (see the section on "Sparsification by data processing inequality").

Network inference. The inference of gene co-expression networks is based on statistical dependencies between transcription levels of individual genes. Mutual information (MI) is used as a measure of correlation to describe these dependencies and quantifies the amount of information obtained about a stochastic variable through the information describing the other variable [14, 22]; see [23] for a recent expansion. It captures any type of statistical dependency, not just positive or negative linear relations, and is invariant under reparameterization. Mutual information between two genes is computed using the probability distributions of their respective transcription levels, as well as the joint probability distribution. The mutual information is then normalized by the joint entropy of the two genes. The joint entropy is a theoretical maximum for the mutual information so this constrains the normalized mutual information to the interval $[0, 1]$. The probability distributions of the genes are obtained by discretizing the transcription data using Sturges' method [24]. From these distributions, the entropy is calculated according to the equation:

$$H(A) = -\sum_{x} p_A(x) log(p_A(x)) \tag{1}$$

Where $H(A)$ is the entropy of gene A, p_A is the transcription distribution of A, and x is the discretized transcription level. From the same principle, the mutual information $I(A, B)$ between genes A and B is calculated from the distributions p_A, p_B, and joint distribution p_{AB}, and respective transcription levels x, y:

$$I(A,B) = \sum_x \sum_y p_{A,B}(x,y) log\left(\frac{p_{A,B}}{p_A(x), p_B(y)}\right) \tag{2}$$

The pairwise mutual information is stored in a matrix (MI matrix) where every entry (A, B) is the mutual information between genes A and B. This matrix is normalized with respect to the joint entropy of A and B, resulting in a symmetric network matrix with values between zero and one representing weighted connections in the gene co-expression network. The joint entropy can be calculated from the single gene entropies and the mutual information:

$$H(A,B) = H(A) + H(B) - I(A,B) \tag{3}$$

Shuffle Corretion. In order to identify random correlations in the network, a significance threshold is established using shuffle correction. Shuffle correction is performed for each column across rows to preserve overall transcription per cell, but any correlations between genes are removed from the data. From this random gene expression matrix, the MI-based network is inferred according to the steps described in the previous section. This step is repeated N times (default: 10). With the distribution of MI values resembling a normal distribution, the significance threshold can be defined using the median and standard deviation of the distribution, e.g. $\mu + 3 * \sigma$, where μ is the median and σ is S.D. Connections with a mutual information value below this statistically defined cutoff point are considered random correlations and removed from the network.

Sparsification by Data Processing Inequality. The next step is to remove the connections from the network produced by indirect correlations. Indirect correlations appear when the transcription of two genes, e.g. A and C, is coregulated with a third gene B but are otherwise independent. Since there is a correlation in transcriptional behavior between A and B, as well as between B and C, a spurious correlation will also be found between A and C, even though they do not interact, and therefore should not be connected in the network. These indirect connections are removed using the principle of Sparsification by Data Processing Inequality [15]. The Data Processing Inequality states that the mutual information between A and C will be lower than the mutual information between A and B or B and C:

$$\begin{aligned} I(A,C) < I(A,B) \\ I(A,C) < I(B,C) \end{aligned} \tag{4}$$

The sparsification is carried out by iterating over all first-order loops in the network (i.e. all trios of fully connected genes) and removing the weakest connection in the loop.

Network analysis. Various statistics, all of which are included in NETSCOPE, can be computed that help identify genes of interest and understand the overall structure of the network. At the node level, statistics such as degree, local clustering coefficient, and betweenness centrality can be used to identify genes that have an important role in the network, and potentially in transcriptional regulation. On a structural level, Dijkstra's algorithm [25] can be used to compute the (weighted) shortest path from one node to another, or the degree of separation (DOS), i.e. the minimum number of edges needed to get from one node to another. The degree of clustering in the network can be assessed by the global clustering coefficient, the small world coefficient, or PCA-related methods available in the toolbox. Finally, the toolbox contains functionality to export the network in GEXF format, a common data format used by network visualization software, such as Gephi [21].

Since mutual information is a similarity (or "closeness") metric, it cannot be used to describe distances between network nodes. To compute the shortest path from one node to another, the variation of information (VI) was used. The variation of information quantifies the information difference between two distributions [18]; if MI describes how close two nodes are to each other in the network, then VI describes how far they are from each other. The variation of information can be calculated from the joint entropy and mutual information of a gene pair:

$$\begin{aligned} VI(A,B) &= H(A,B) - I(A,B) \\ VI_n(A,B) &= 1 - I_n(A,B) \end{aligned} \tag{5}$$

where $I_n(A,B)$ is the normalized mutual information. The maximum value for VI is 1 which is the case when MI equals zero and two genes are completely uncorrelated. Since two disconnected nodes in the network should have a "distance" of infinity (instead of 1), the distance metric in the network was defined as the ratio of VI to MI:

$$dist(A,B) = \frac{VI(A,B)}{I(A,B)} = \frac{1 - I_n(A,B)}{I_n(A,B)} \tag{6}$$

3 Results

In the following two sections, we assess the quality of networks that were constructed with the toolbox. First, a network was constructed from artificial (synthetic) data with a known ground-truth network organization. Expression profiling was performed on the synthetic data, and the underlying network was inferred using NETSCOPE. Second, a network was constructed from RNA-sequencing data from Saccharomyces cerevisiae, a species of yeast whose genome has been studied extensively [26–28]. The resulting network was compared to five previously found co-expression networks in *S*. cerevisiae and a network pathway analysis was done to identify potential genes of interest.

Ground truth network. A topologically simple, undirected, unweighted ground truth network was created and used to generate synthetic expression data. The ground truth network consists of 100 disjoint graphs of 10 nodes connected in a linear pattern. From a random seed, expression data were generated that preserve this topology. The data was fed into the data processing pipeline to infer a co-expression network. The resulting network was shuffle corrected with a cut-off point of $\mu + 3 * \sigma$, sparsified and then compared to the ground truth network to assess its accuracy and to evaluate the limits of the method in terms of minimum data dimensionality. This was repeated by varying the number of samples in the expression data and the levels of noise added to the expression data. The number of samples ranged from 100 to 8000 and the noise was sampled from a normal distribution with a standard deviation of 10% to 50% of the mean. To assess the quality of the inferred networks as a function of data density and the number of samples, the true/false positive rates were used, where a false positive is a connection found between two nodes that were not connected in the original network.

The results (expectedly) showed that the quality of the network increases with the number of samples and decreases with the amount of noise added to the system (Fig. 3a). Distributions of MI values in the 8000 sample networks as a function of the degree of separation (DOS) in the original network revealed that the directly connected nodes have a median MI > 0.2 for all noise levels, with MI decaying as the distance between nodes increases (Fig 3b). The decay rate increases with the amount of added noise, and disconnected nodes always have **MI** around zero.

S. cerevisiae networks. To test the performance of NETSCOPE with biological systems, networks were constructed from RNA-sequencing data from S. cerevisiae [20]. A co-regulatory network was constructed and filtered after shuffle correction from 7126 genes sequenced in 672 samples. The subsets of this network were compared to five corresponding co-expression networks found in previous studies, which are listed in Table 1. The network data were made available by YeastNet v_3 [29].

As expected, compared to the synthetic data, the MI values were significantly lower in the yeast networks and the median value of normalized MI in each of the 5 networks was less than 0.05 (Fig. 4a). To numerically compare NETSCOPE's network reconstruction against previously published network structure in yeast, we calculated the receiver operating characteristics (ROC). Because ROC is sensitive to noise level, we calculated three thresholds to binarize the MI matrix (Fig. 4a; three candidate thresholds established by shuffle correction are shown in color) and quantified the successful identification (i.e. true positive) against the false positive ratio (Fig. 4b). The expected results showed that with the threshold inversely proportional to the true and false positive discovery rate; with the increased threshold both discovery rates are reduced. Conservatively $\sim 60\%$ of all previously identified networks could be correctly discovered using NETSCOPE with a false discovery rate of $\sim 20\%$ (Fig. 4) using the conservative

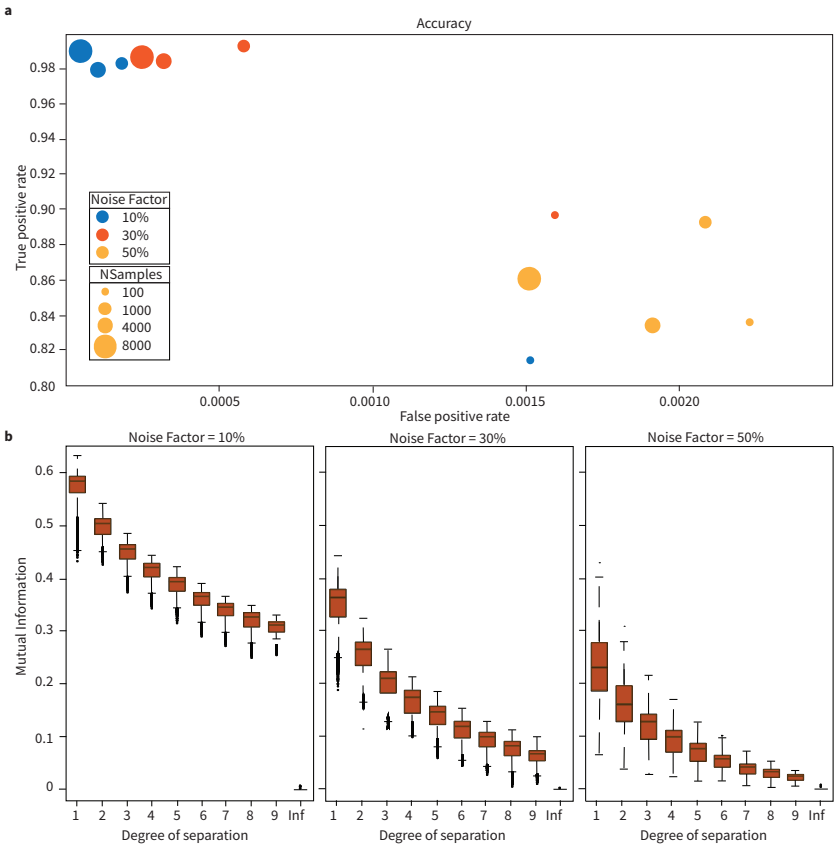


Fig. 3. Basic processing steps for quantification of co-expression network from gene expression data. Networks constructed from synthetic gene expression data. Networks were constructed from synthetic expression data with varying numbers of samples and variable amounts of noise added. **a** Accuracy of constructed networks when compared to their ground truth respective. Marker color denotes noise levels (red=10%, blue=30%, yellow=50%) and marker size denotes the number of samples used $(100, 1000, 4000, 8000)$. **b** MI as a function of separation (DOS) in the networks. DOS is the number of edges between a pair of genes; a value of 1 means they are directly connected; a value of 'inf' means there is no possible path between them.

Table 1. Sources of the co-expression networks in Saccharomyces cerevisiae

No.	Reference	Title
1	Mitchell et al. 2009 [34]	Adaptive prediction of environmental changes by microorganisms.
2	Wade et al. 2009 [35]	The Snf1 kinase and proteasome-associated Rad23 regulate UV-responsive gene expression.
3	Shalem et al. 2011 [36]	Transcriptome kinetics is governed by a genome-wide coupling of mRNA production and degradation: a role for RNA Pol II.
4	Yona et al. 2012 [37]	Chromosomal duplication is a transient evolutionary solution to stress.
5	Orlando et al. 2008 [38]	Global control of cell-cycle transcription by coupled CDK and network oscillators.

normalized MI threshold (i.e. $11 * \sigma$). The discovery rate improves to $> 80\%$ with lower thresholds (e.g. $7 * \sigma$) at the expense of an increased false discovery rate.

NETSCOPE differs from previous approaches to MI-based molecular network inference (ARACNE [12], parmigene [13] and PANA [14]) as such it allows weighted network analysis for pathway discovery (see the variation of information in Methods). The direct utility of identification of weights in the network is that it allows computation of the shortest path among any given two targets within the network, hence expanding the utility of MI analysis to pathway discovery. This approach can be expanded to the analysis of each target as a hub by calculation of the betweenness centrality which is defined as the fraction of all shortest paths that go through a particular node [27]. Fig. 4c shows the distributions of betweenness centrality for the nodes in the S. cerevisiae networks, as well as the top 5 highest scoring nodes for this metric. Genes that have a high centrality in multiple networks are in bold print. These results show that the mutual information analysis described above could be used to identify molecular networks extract correlational transcriptional networks could be used to identify molecular networks.

Cell type-specific networks in the mouse brain. In order to assess the similarity or difference between molecular networks in different cell types, networks were constructed from gene expression data from cells of six different neuronal and non-neuronal cell types in the mouse somatosensory cortex. Gene expression data was previously obtained using single-cell RNA-sequencing, and contain transcript counts of a total of 19972 genes in 1691 cells of the somatosensory cortex [31]. For computational efficiency, a random subset of 1000 genes was selected. Six cell-specific networks were constructed by using only cells from the particular type for network construction: astrocytes/ependymal ($N = 143$), endothelial/mural cells ($N = 202$), interneurons ($N = 164$), microglia ($N = 84$), oligodendrocytes ($N = 699$) and pyramidal cells ($N = 399$). The networks were

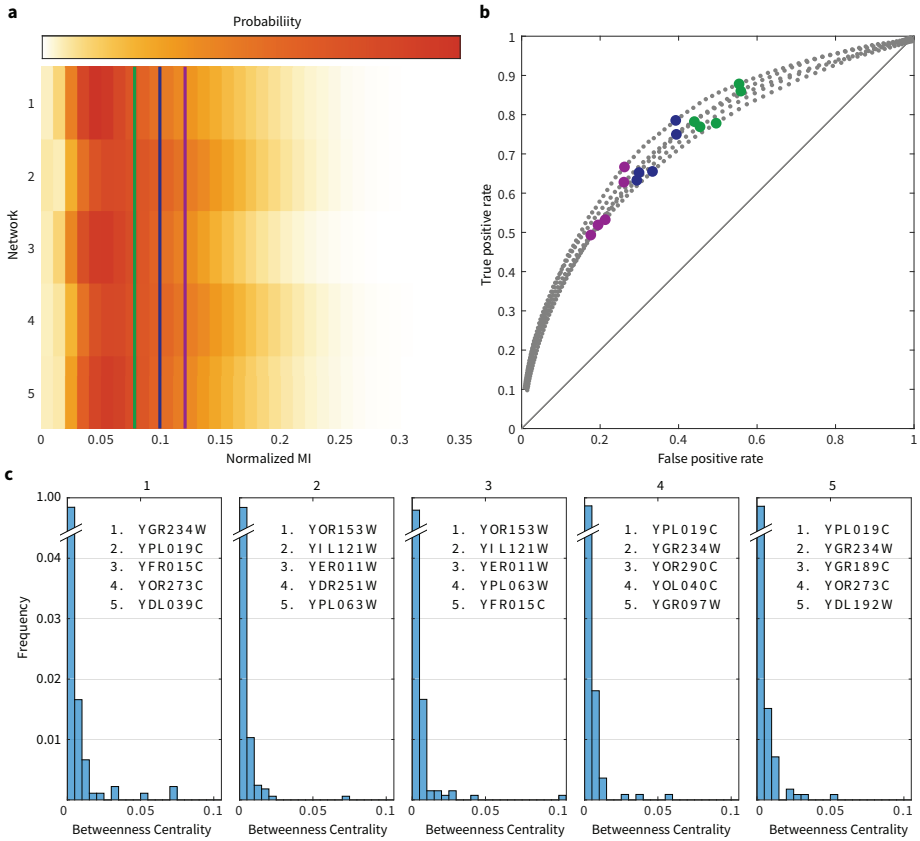


Fig. 4. Five gene co-expression networks in S. cerevisiae constructed from RNA-seq data. a Distributions of MI values of the constructed networks. The three candidates' μ plus 7 (green), 9 (blue) and 11 (purple) times σ of the shuffled MI distribution are shown. **b** Receiver operating characteristic (ROC) curve. True and false positive rates show the similarity between the constructed networks and previously established networks (see Table 1) for different significance thresholds. The colored data points correspond with the thresholds shown in **a**. **c** Distributions of the nodes' betweenness centrality in the five networks. Every plot lists the top five genes with the highest betweenness centrality. Genes that are preserved across different networks are bold-printed.

thresholded ($\mu + 3\sigma$) after shuffling (see above) and sparsified. Fig. 5a shows the MI matrices of the cell type-specific networks after clustering using Principal Component Analysis (PCA). For each network, PCA was applied to the MI matrix and nodes were sorted by their highest-scoring component. This way, nodes with a similar connectivity profile are close to one another, revealing clustering in the cell-type specific molecular networks. The order of the nodes is different for each network, illustrating differences in local connectivity.

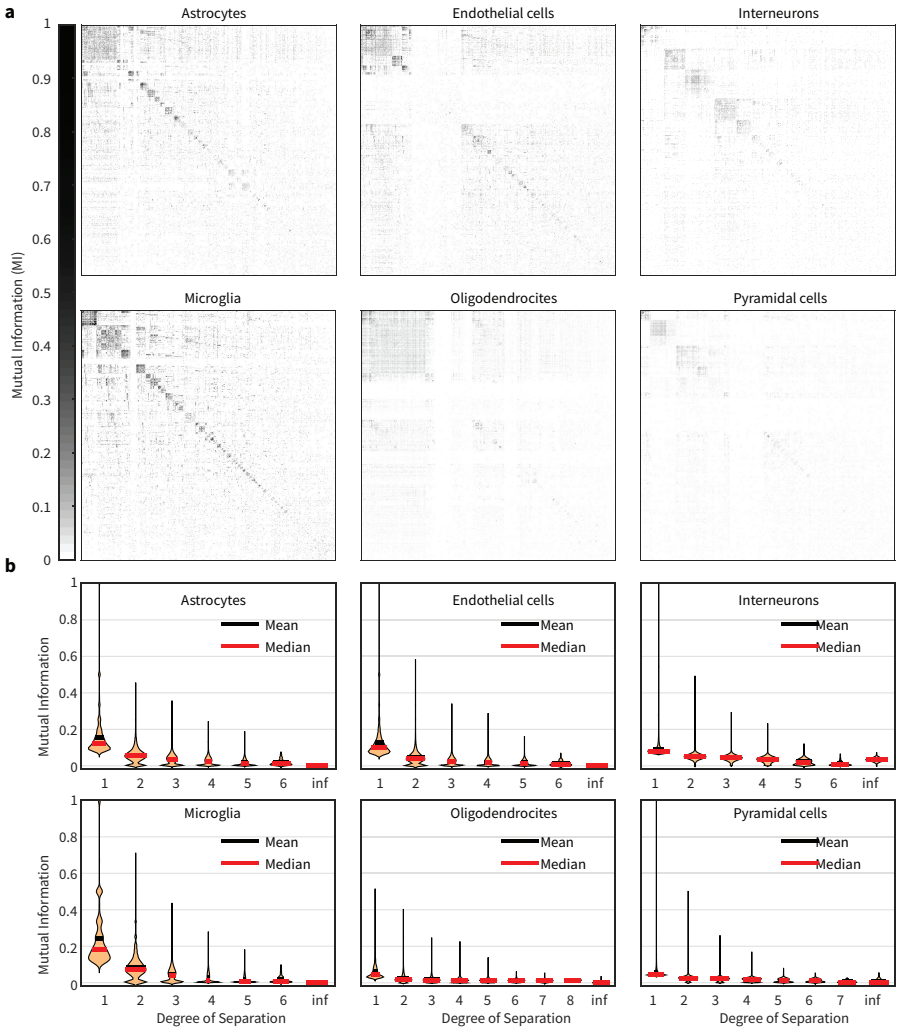


Fig. 5. Cell type-specific gene co-expression networks in the mouse somatosensory cortex. a Mutual Information (MI) matrices showing pairwise correlations of genes in six different cell types in the somatosensory cortex. For every matrix PCA was used to sort the matrices, so that nodes with a similar connectivity profile are next to each other, illustrating differences in clustering between the networks. **b** MI distributions for increasing degree of separation (DOS) for each of the networks. A value of 1 means they are directly connected; a value of 'inf' means there is no possible path between them.

Similarly, to Fig. 3b for synthetic data, Fig. 5b shows the distributions of MI values as a function of DOS in cell type-specific networks. A DOS of 'inf' (infinite) occurs when there is no possible path between two nodes. The figures

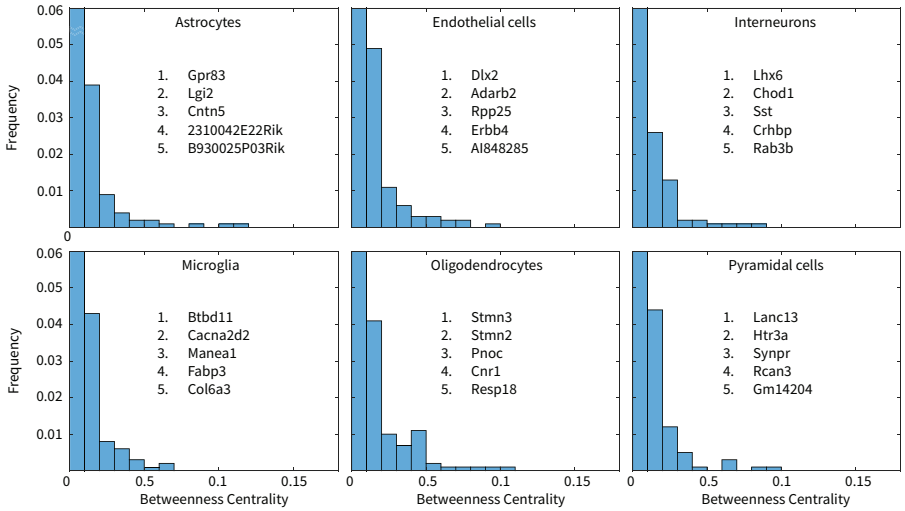


Fig. 6. Cell type-specific gene co-expression networks in the mouse somatosensory cortex. Figures show the distributions of the nodes' betweenness centrality in the six cell type-specific networks. Every plot lists the top five genes with the highest betweenness centrality. Genes that are preserved across different networks are bold printed.

show that MI decreases as DOS increases, indicating that MI is indeed a measure of the 'closeness' of two nodes in the network, which decreases as noise is introduced with every added step along the path. Pathway analysis revealed for each node the betweenness centrality. Fig. 6 shows the distributions of these values for each network. Also displayed in Fig. 6 is, similar to Fig. 4c, the top 5 highest scoring nodes for this metric. Again, genes that are preserved across different networks are bold printed (note that this is not the case for any gene).

Housekeeping Genes. Housekeeping Genes, also referred to as reference genes, are genes that are consistently expressed in all cells and play essential roles in fundamental cellular functions required for maintaining cellular homeostasis. These genes maintain relatively stable expression levels across various tissue and experimental conditions, developmental stages, and environmental factors. In gene expression studies, these genes are commonly utilized as normalization factors to account for sample quality variations and RNA sequencing discrepancies.

To investigate the molecular networks of housekeeping genes in male and female mice, gene expression data from the somatosensory cortex layer 2/3 were utilized to construct networks. The gene expression data had been previously obtained using single-cell RNA sequencing and encompassed transcript counts for a total of 450 genes. The shared genes between males and females were identified, and gene-mutual information matrices were generated based on the gene expression data for each sex. Rows and columns containing solely 0 values were

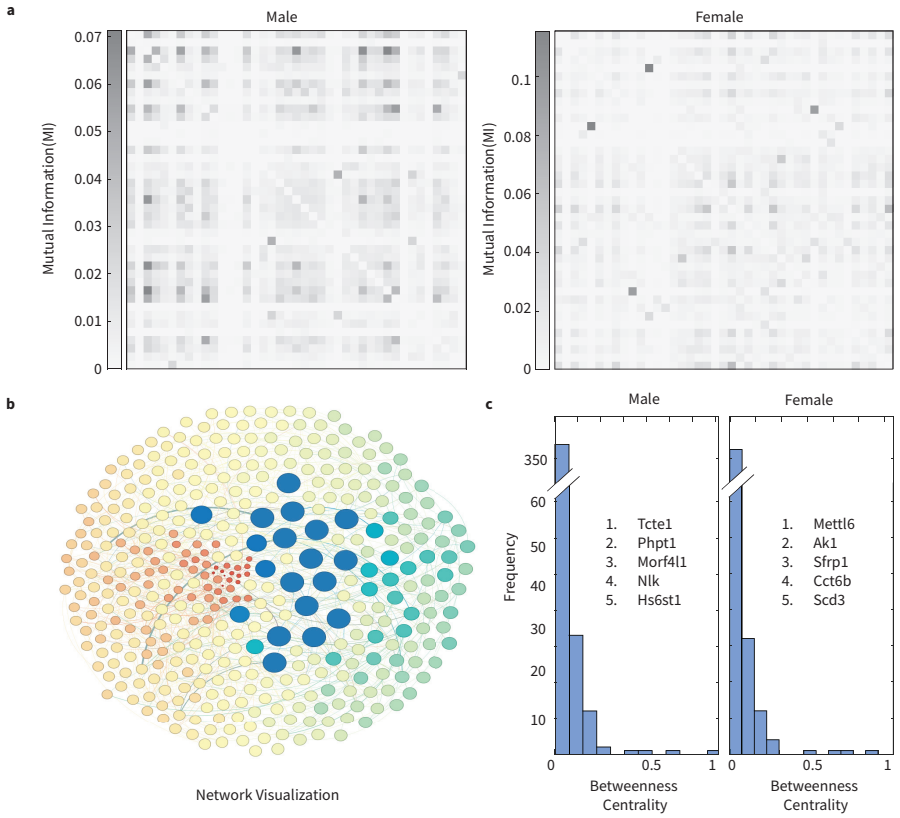


Fig. 7. Housekeeping gene co-expression networks in the mouse somatosensory cortex. a Mutual Information (MI) matrices showing pairwise correlations of genes in Male and Female mice, before network sparsification. **b** Visualization of the Male housekeeping gene correlation network generated from Gephi. The Mi matrix is thresholded using random shuffling, sparsified and then exported to Gephi. In Gephi, the **Force Atlas2** network organization algorithm is applied. The edge between the two nodes represents the MI value between the genes. Independent nodes are removed. The color and the size of each node(4-40) are ranked using the betweenness centrality of the nodes. **nooverlap** algorithm is applied so that no two nodes are overlapping. **c** Histogram of normalized Betweenness Centrality Male and Female housekeeping gene-Networks. The top five genes with the highest betweenness centrality are listed.

eliminated from both matrices. Fig. 8a displays the calculated mutual information matrices representing gene co-expression in the L2/3 somatosensory cortex of male and female mice. No sorting was performed on either of the matrices. The networks derived from these mutual information matrices were created by applying a threshold of $(\mu + \sigma)$ after shuffling (as described above) and subsequently sparsified. Fig. 8b presents the graph visualization obtained from the male mutual information matrix, exported as a network from NETSCOPE to Gephi following thresholding and sparsification. The node size corresponds to the gene's expression count across all male samples. Fig. 8c illustrates the histogram of the betweenness centrality distribution for all genes, with each plot listing the top five genes exhibiting the highest betweenness centrality.

Cellular Networks. To investigate the mapping of transgenic lines in the somatosensory cortex, a comprehensive analysis was conducted using the Allen Brain Atlas. All transgenic line experiments targeting the somatosensory cortex were meticulously sorted, and the corresponding experiment numbers were exported to a CSV file. A total of 65 experiments were identified from the Allen Brain database and all relevant data associated with these experiments were exported to MATLAB. Subsequently, the data were merged into a single file and inputted into NETSCOPE for further processing.

In order to determine the transgenic line map, the data underwent processing in NETSCOPE, where mutual information (MI) was calculated for both Energy and Density measures across all 300 cells. Fig. 8a illustrates the resulting MI matrices, representing the co-expression patterns of cells in the somatosensory cortex of male and female mice. It is important to note that no sorting or any other manipulation was applied to these matrices. The networks were generated from these MI matrices by applying a threshold of $(\mu+\sigma)$ after shuffling, followed by sparsification.

Fig. 8b presents the graph visualization obtained from the Energy MI matrix, which was exported as a network from NETSCOPE to Gephi after undergoing thresholding and sparsification. The size of the nodes in the graph corresponds to the expression count of the gene across all male samples, providing a visual representation of the gene expression levels. Additionally, Fig. 8c displays a histogram showcasing the distribution of betweenness centrality for all genes. Each plot within the histogram highlights the top five genes with the highest betweenness centrality, indicating their significant role in information flow within the network.

4 Discussion

NETSCOPE is a universal Python/Octave/MATLAB toolbox that can be used for the network discovery and reconstruction of networks. NETSCOPE constructs a network with a weighted edge between nodes, rather than only a binary network. The accuracy of the approach was shown using synthetic (artificial)

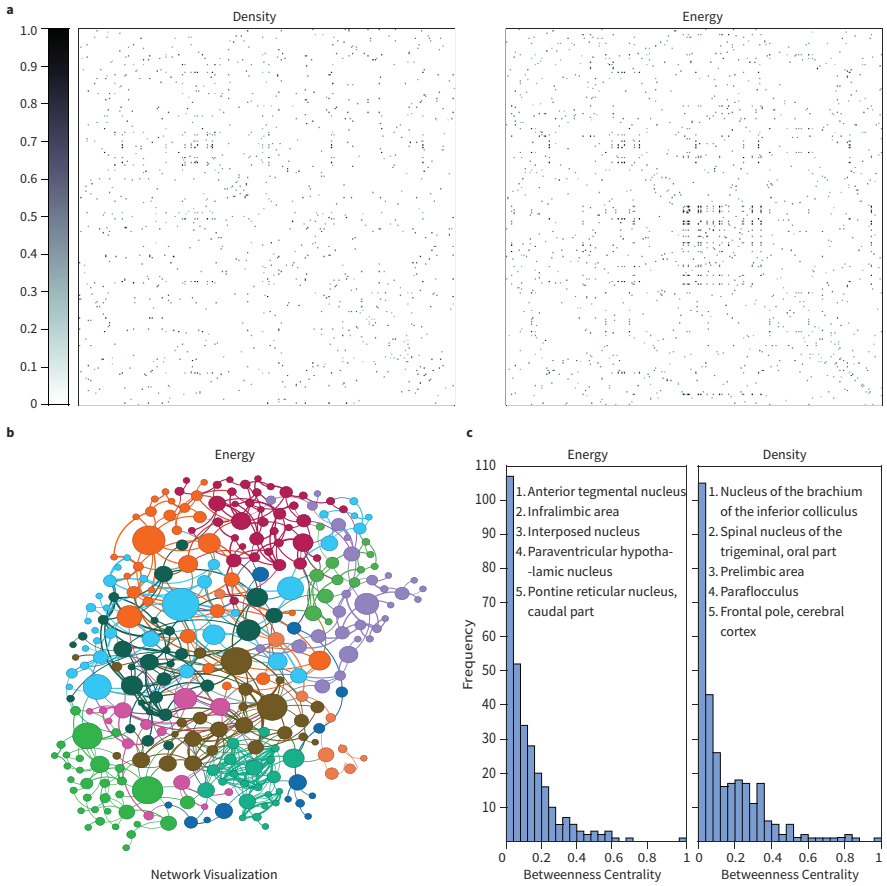


Fig. 8. Cellular Networks in the mouse somatosensory cortex from Allen Brain Institute data. a Mutual Information (MI) matrices showing pairwise correlations of cells in the somatosensory cortex for Energy and Density of the injections, after thresholding ($\mu+\sigma$) and network sparsification. **b** Visualization of the network for the energy generated from Gephi. The Mi matrix is exported to Gephi, **Force Atlas** network organization algorithm is applied. The edge between the two nodes represents the MI value between the genes. Independent nodes are removed. The colour and the size of each node(4-40) are ranked using the betweenness centrality of the nodes. **nooverlap** algorithm is applied so that no two nodes are overlapping. **c** Histogram of normalized Betweenness Centrality for Energy and Density Networks. The top five cells with the highest betweenness centrality are listed.

data and quantified against empirically identified networks in yeast. Expanding co-expression network identification to pathway analysis allows the utility of NETSCOPE for the network discovery (see supplemental fig: 14 and supplemental fig: 13, identification of nodes in networks (Fig. 6) as well as allows analyzing differentially regulated gene networks(see supplemental 9). Differential network analysis helps in network discovery and inference of important genes in the Gene regulatory networks across samples.

NETSCOPE works in a plug-and-play manner without the need for additional toolbox installation or complicated setup; the code is available on GitHub and the expression data can be imported into Colab, Python, Octave or MATLAB from e.g. a comma-separated value (CSV) file. The toolbox offers the flexibility to fine-tune any parameters in the data processing pipeline but also works out of the box using default parameters - the latter of which was the case for the analyses described in this article. Network data can be exported to network software tools for visualization and subsequent analysis (see supplemental Fig.11 and Fig. 12). using some methods in PYTHON/Matlab, NETSCOPE allows calculating network measures, such as **betweenness centrality**; a measure that captures a gene's role in allowing information to pass from one part of the network to the other, **closeness centrality**; a measure that indicates how close a node is to all other nodes in the network, **eigenvector centrality**; a measure that indicates node's importance while giving consideration to the importance of its neighbors, **pageranks** and **hubranks**; algorithmic measures that calculate the importance of each node within the graph, etc (Supplement Fig. 10) to describe the network.

Drawbacks of the **MI**-based approach include the need for high-dimensional data; Fig. 3 shows that the true positive rate increases with the number of samples and quality of the synthetic data. Especially for higher levels of noise, 1000+ samples are needed for acceptable performance. While **MI** is computationally heavy, parallelization can speed up the process. For the parallel computation of the **MI** matrix, one can use the MATLAB Parallel Processing Toolbox in combination with the parallel processing functions included in NETSCOPE. Due to the universal nature of **MI**, NETSCOPE allows for the cross-comparison of transcription data between cell types, individuals, populations, or species. Rather than comparing individual genes, one can now compare entire transcriptomes by their internal structure, looking for patterns of connectivity that are preserved [32]. These patterns can, for example, shine a light on basic molecular mechanisms at the root of the evolutionary tree, which would be imperceptible through single gene analysis. Going beyond cross-species transcriptomics, NETSCOPE can be used to integrate multi-omics data covering multiple layers of the cellular process - e.g. genomics, epigenomics, proteomics and metabolomics - to reconstruct network representations of the molecular landscape. Understanding the topology of these networks makes it possible to bridge the gap between genotype and phenotype, and predict the downstream effects of genetic defects [33]. It will allow us to link low-level physiological mechanisms to higher-level phenomena such as disease symptoms and behavioral traits, and provide insight

into the mechanisms behind cell fate and how to control it, allowing for cellular reprogramming.

References

1. Lim C-S, Kang X, Mirabella V, Zhang H, Bu Q, Araki Y, et al. BRaf signaling principles unveiled by large-scale human mutation analysis with a rapid lentivirus-based gene replacement method. Genes Dev. 2017; doi: 10.1101/gad.294413.116.
2. Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, et al. Pathway and network analysis of cancer genomes. Nat Methods. 2015; doi: 10.1038/nmeth.3440.
3. Kole K, Komuro Y, Provaznik J, Pistolic J, Benes V, Paul T, et al. Erratum to: Transcriptional mapping of the primary somatosensory cortex upon sensory deprivation. GigaScience. 2018; doi: 10.1093/gigascience/giy040.
4. Kole K, Scheenen W, Tiesinga P, Celikel T. Cellular diversity of the somatosensory cortical map plasticity. Neurosci Biobehav Rev. 2018; doi: 10.1016/j.neubiorev.2017.11.015.
5. Kole K, Lindeboom RGH, Baltissen MPA, Jansen PWTC, Vermeulen M, Tiesinga P, et al. Erratum to: Proteomic landscape of the primary somatosensory cortex upon sensory deprivation. GigaScience. 2018; doi: 10.1093/gigascience/giy041.
6. Huang C, Zeldenrust F, Celikel T. Cortical representation of touch in silico. bioRxiv. 2021; doi: 10.1101/2020.11.06.371252.
7. Zeldenrust F, Calcini N, Yan X, Bijlsma A, Celikel T. Cell type-specific information transfer for sparse coding. bioRxiv. 2020; doi: 10.1101/2020.11.06.371658.
8. Huang C, Englitz B, Reznik A, Zeldenrust F, Celikel T. Information transfer and recovery for the sense of touch. bioRxiv. 2020; doi: 10.1101/2020.12.08.415729.
9. Mc Mahon SS, Sim A, Filippi S, Johnson R, Liepe J, Smith D, et al. Information theory and signal transduction systems: From molecular information processing to network inference. Semin Cell Dev Biol. 2014; doi: 10.1016/j.semcdb.2014.06.011.
10. Christopher A Jackson, Dayanne M Castro, Giuseppe-Antonio Saldi, Richard Bonneau, David Gresham (2020) Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments eLife 9:e51254; https://doi.org/10.7554/eLife.51254
11. Shu, H., Zhou, J., Lian, Q. et al. Modeling gene regulatory networks using neural network architectures. Nat Comput Sci 1, 491–501 (2021). https://doi.org/10.1038/s43588-021-00099-8
12. Zhang, M., Li, Q., Yu, D. et al. GeNeCK: a web server for gene network construction and visualization. BMC Bioinformatics 20, 12 (2019). https://doi.org/10.1186/s12859-018-2560-0
13. Gibson SM, Ficklin SP, Isaacson S, Luo F, Feltus FA, Smith MC. Massive-Scale Gene Co-Expression Network Construction and Robustness Testing Using Random Matrix Theory. PLOS ONE. 2013; doi: 10.1371/journal.pone.0055871.

14. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948; doi: 10.1002/j.1538-7305.1948.tb01338.x.
15. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics. 2006; doi: 10.1186/1471-2105-7-S1-S7.
16. Sales G, Romualdi C. parmigene—a parallel R package for mutual information estimation and gene network reconstruction. Bioinformatics. 2011; doi: 10.1093/bioinformatics/btr274.
17. Ponzoni I, Nueda MJ, Tarazona S, Götz S, Montaner D, Dussaut JS, et al. Pathway network inference from gene expression data. BMC Syst Biol. 2014; doi: 10.1186/1752-0509-8-S2-S7.
18. Meilă M. Comparing clusterings—an information based distance. J Multivar Anal. 2007; doi: 10.1016/j.jmva.2006.11.013.
19. Silvester S. oct2py: Python to GNU Octave bridge –¿ run m-files from python. https://pypi.org/project/oct2py/. Accessed 9 April 2021.
20. Ziemann M, Kaspi A, El-Osta A. Digital expression explorer 2: a repository of uniformly processed RNA sequencing data. GigaScience. 2019; doi: 10.1093/gigascience/giz022.
21. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks. Proc Int AAAI Conf Web Soc Media.
22. Zeldenrust F, de Knecht S, Wadman WJ, Denève S, Gutkin B. Estimating the Information Extracted by a Single Spiking Neuron from a Continuous Input Time Series. Front Comput Neurosci. 2017; doi: 10.3389/fncom.2017.00049.
23. Azarfar A, Calcini N, Huang C, Zeldenrust F, Celikel T. Neural coding: A single neuron's perspective. Neurosci Biobehav Rev. 2018; doi: 10.1016/j.neubiorev.2018.09.007.
24. Herbert A. Sturges. The Choice of a Class Interval. J Am Stat Assoc. 1926;21:65–61926;
25. Dijkstra EW. A note on two problems in connexion with graphs. Numer Math. 1959; doi: 10.1007/BF01386390.
26. Münzner U, Klipp E, Krantz M. A comprehensive, mechanistically detailed, and executable model of the cell division cycle in Saccharomyces cerevisiae. Nat Commun. 2019; doi: 10.1038/s41467-019-08903-w.
27. Dai Z, Huang M, Chen Y, Siewers V, Nielsen J. Global rewiring of cellular metabolism renders Saccharomyces cerevisiae Crabtree negative. Nat Commun. 2018; doi: 10.1038/s41467-018-05409-9.
28. Janjić V, Sharan R, Pržulj N. Modelling the Yeast Interactome. Sci Rep. 2014; doi: 10.1038/srep04273.
29. Kim H, Shin J, Kim E, Kim H, Hwang S, Shim JE, et al. YeastNet v3: a public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae. Nucleic Acids Res. 2014; doi: 10.1093/nar/gkt981.
30. Freeman LC. A Set of Measures of Centrality Based on Betweenness. Sociometry. 1977; doi: 10.2307/3033543.
31. Zeisel A, Muñoz-Manchado AB, Codeluppi S, et al (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347:1138–1142. https://doi.org/10.1126/science.aaa1934

32. Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse. Nat Rev Genet. 2017; doi: 10.1038/nrg.2017.19.
33. Ma A, McDermaid A, Xu J, Chang Y, Ma Q. Integrative Methods and Practical Challenges for Single-Cell Multi-omics. Trends Biotechnol. 2020; doi: 10.1016/j.tibtech.2020.02.013.
34. Mitchell A, Romano GH, Groisman B, Yona A, Dekel E, Kupiec M, et al. Adaptive prediction of environmental changes by microorganisms. Nature. 2009; doi: 10.1038/nature08112.
35. Wade SL, Poorey K, Bekiranov S, Auble DT. The Snf1 kinase and proteasome-associated Rad23 regulate UV-responsive gene expression. EMBO J. 2009; doi: 10.1038/emboj.2009.229.
36. Shalem O, Groisman B, Choder M, Dahan O, Pilpel Y. Transcriptome Kinetics Is Governed by a Genome-Wide Coupling of mRNA Production and Degradation: A Role for RNA Pol II. PLOS Genet. 2011; doi: 10.1371/journal.pgen.1002273.
37. Yona AH, Manor YS, Herbst RH, Romano GH, Mitchell A, Kupiec M, et al. Chromosomal duplication is a transient evolutionary solution to stress. Proc Natl Acad Sci U S A. 2012; doi: 10.1073/pnas.1211150109.
38. Orlando DA, Lin CY, Bernard A, Wang JY, Socolar JES, Iversen ES, et al. Global control of cell-cycle transcription by coupled CDK and network oscillators. Nature. 2008; doi: 10.1038/nature06955.

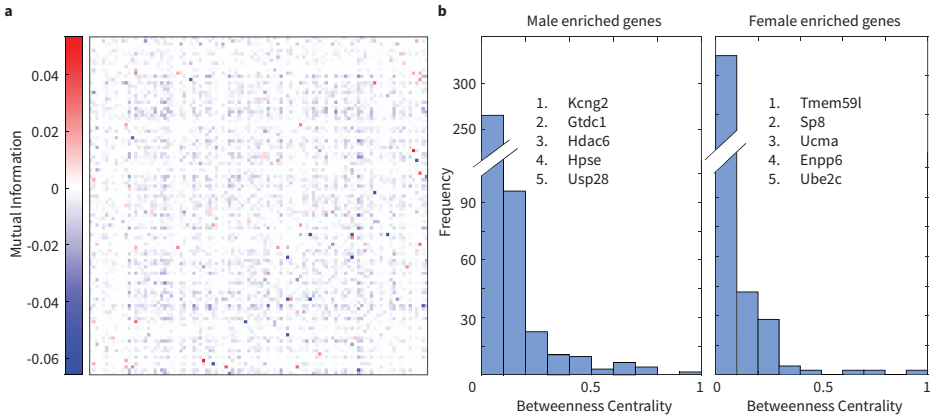


Fig. 9. Differential network for Male and Female. a Mutual Information (MI) matrix showing the difference of the housekeeping gene MI matrices for Male and Female. **b** Histogram of normalized Betweenness Centrality of the network for male enriched genes(i.e. positive values in differential MI matrix) and female enriched genes(i.e. negative values in the differential MI matrix). The top five genes with the highest betweenness centrality are listed.

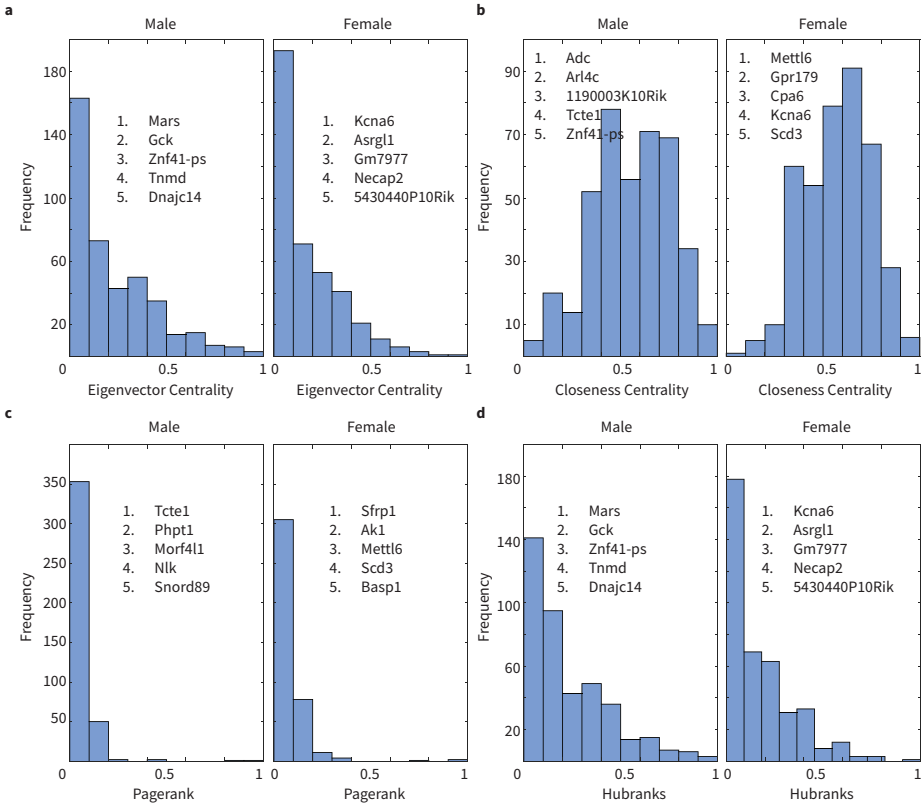
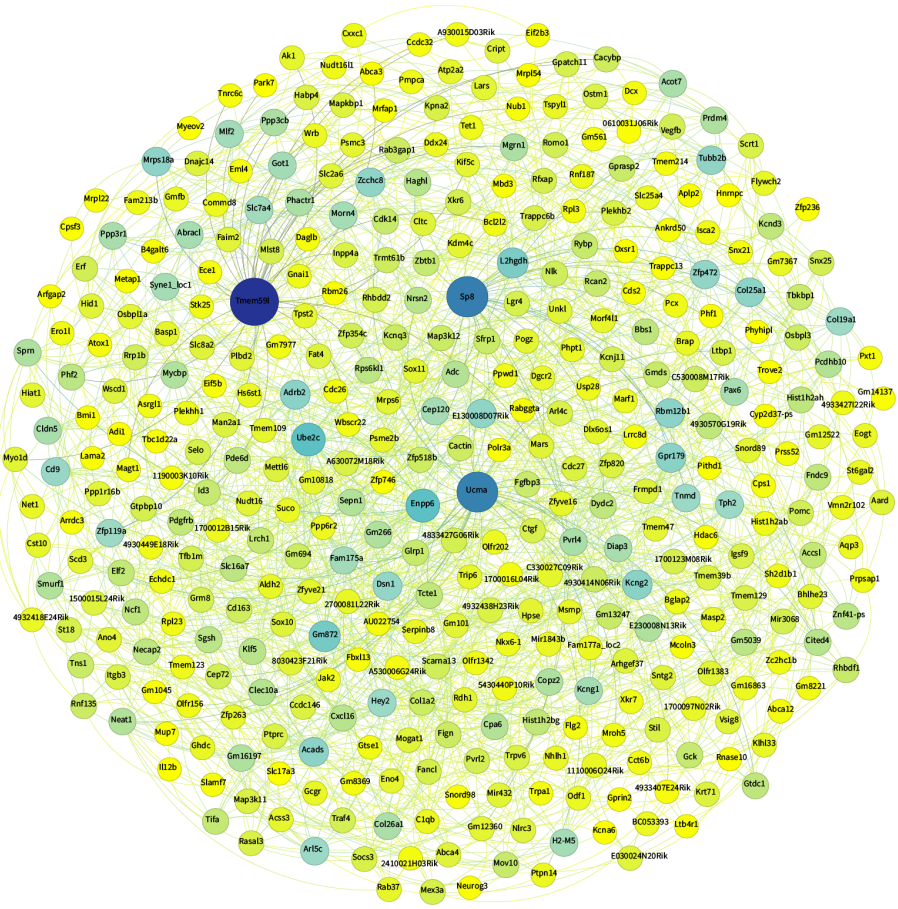


Fig. 10. Network measures of networks generated by NETSCOPE after thresholding and Sparsifying the MI matrics of Housekeeping genes in the mouse somatosensory cortex. a Histogram of the eigenvector centrality for Male and Female housekeeping gene MI networks.**b** Histogram of the closeness centrality for Male and Female housekeeping gene MI networks. **c** Histogram of the normalized pagerank for Male and Female housekeeping gene MI networks. **c** Histogram of the normalized hubranks for Male and Female housekeeping gene MI networks. The top 5 genes with the highest values are listed in each figure.


Tmem59l
Sp8
Ucma
Ube2c
Enpp6
Gm872
Acads
Arl5c
Hey2
Col1a2
Kcng2
Tmem47
Gm13247
Gm5039
Gm16863
Gm8221
Gm12522
Gm14137
Gm7367
Gm7977
Gm266
Gm694
Gm10818
Gm101
Gm1045
Gm16197
Gm8369
Gm12360
Gm2221
Gm561
Gm1045
Gm8221
Gm14137
Gm12522
Gm7367
Gm7977
Gm266
Gm694
Gm10818
Gm101
Gm872
Gm5039
Gm16863
Gm13247

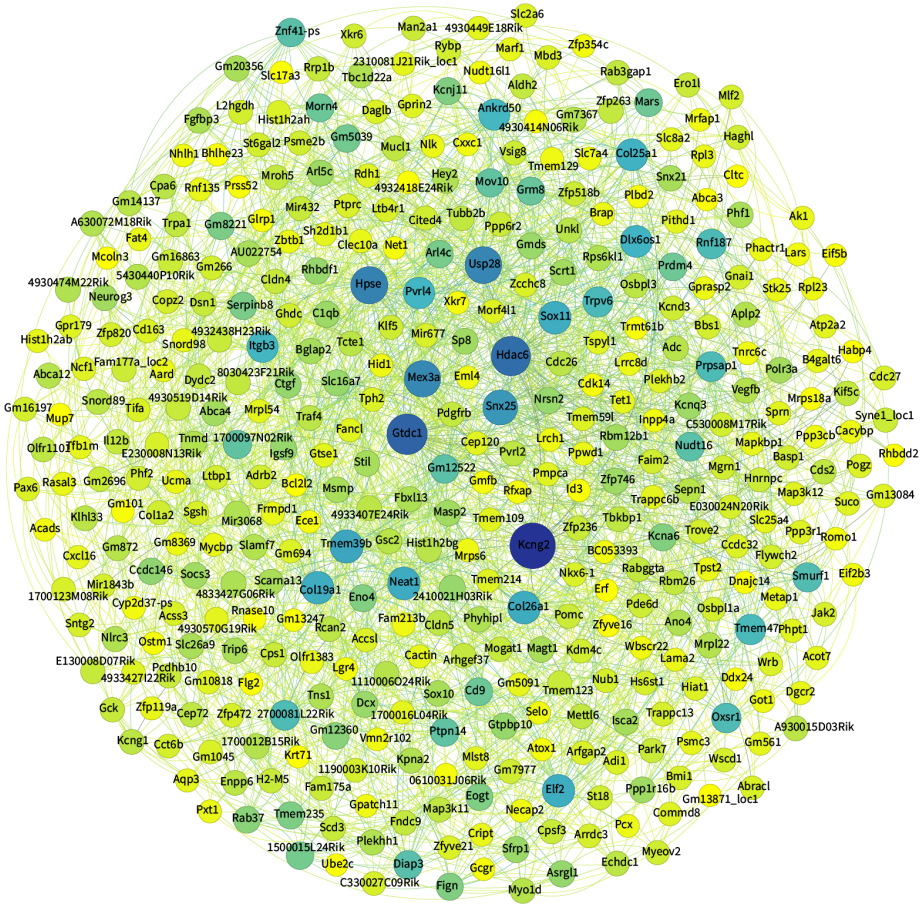


Fig. 12. Housekeeping gene co-expression networks in the mouse somatosensory cortex. a Visualization of the Male housekeeping gene correlation network with all the gene names generated from Gephi. All positive values from the differential Mi matrix are sorted to get the gene co-expression matrix enriched in the male samples. The matrix is thresholded using random shuffling and sparsified before exporting to Gephi. In Gephi, the **Force Atlas2** network organization algorithm is applied. The edge between the two nodes represents the MI value between the genes. Independent nodes are removed. The colour and the size of each node(4-40) are ranked using the betweenness centrality of the nodes. **nooverlap** algorithm is applied so that no two nodes are overlapping

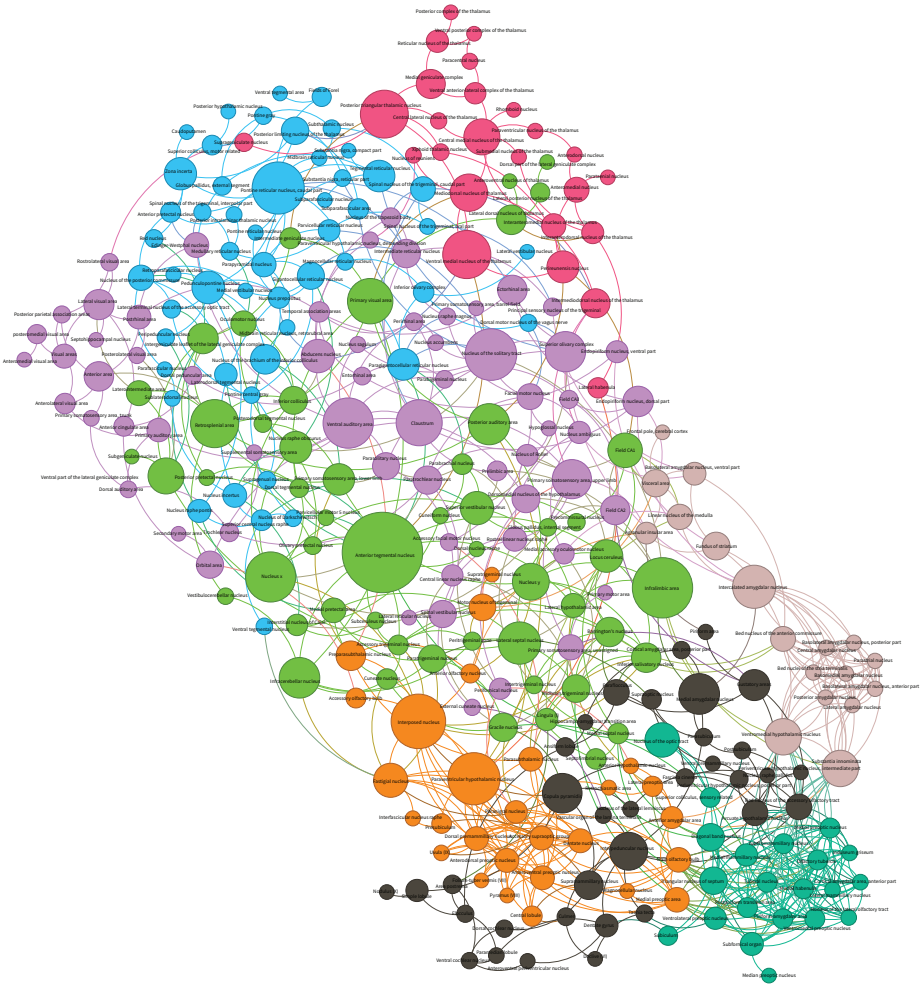


Fig. 13. Visualization of Energy-network of the cellular network created from Allen Brain Atlas data. a Visualization of the Male housekeeping gene correlation network with all the gene names generated from Gephi. All positive values from the differential Mi matrix are sorted to get the gene co-expression matrix enriched in the male samples. The matrix is thresholded using random shuffling and sparsified before exporting to Gephi. In Gephi, the **Force Atlas2** network organization algorithm is applied. The edge between the two nodes represents the MI value between the genes. Independent nodes are removed. The colour and the size of each node(4-40) are ranked using the betweenness centrality of the nodes. **nooverlap** algorithm is applied so that no two nodes are overlapping

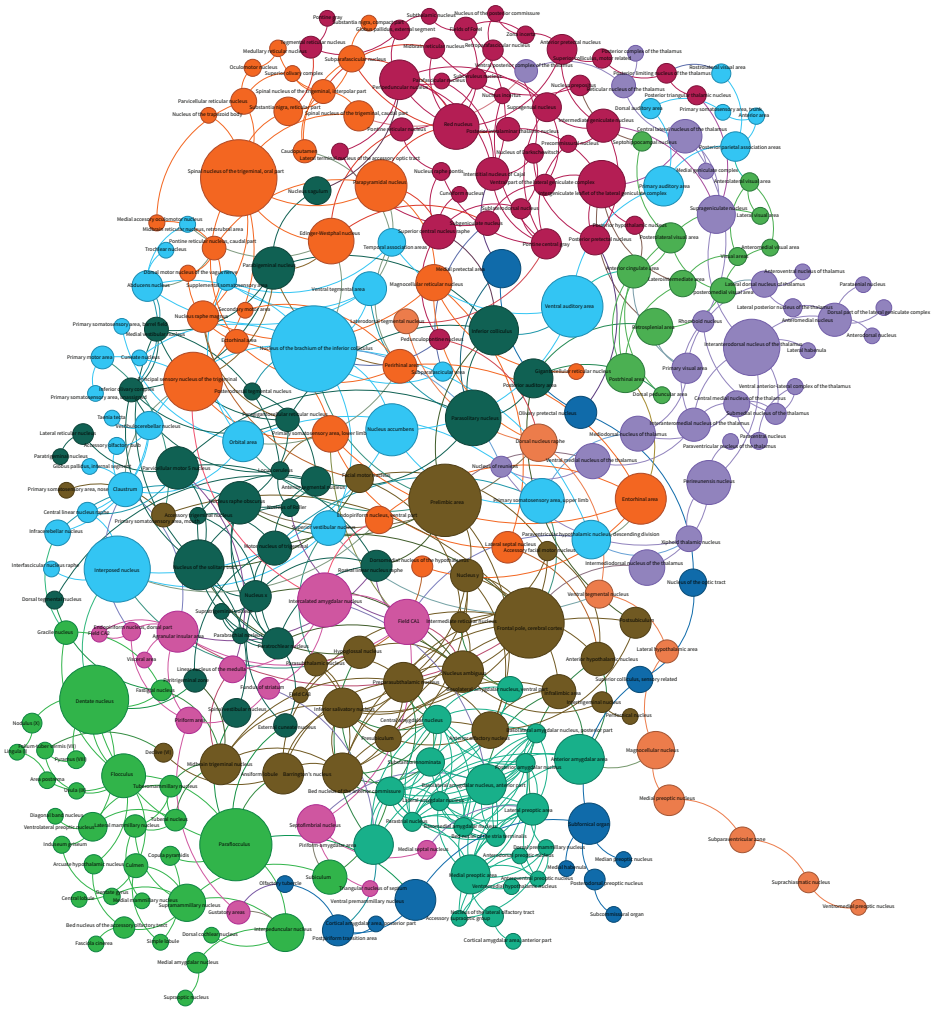


Fig. 14. Visualization of Density-network of the cellular network created from Allen Brain Atlas data. a Visualization of the Male housekeeping gene correlation network with all the gene names generated from Gephi. All positive values from the differential Mi matrix are sorted to get the gene co-expression matrix enriched in the male samples. The matrix is thresholded using random shuffling and sparsified before exporting to Gephi. In Gephi, the **Force Atlas2** network organization algorithm is applied. The edge between the two nodes represents the MI value between the genes. Independent nodes are removed. The colour and the size of each node(4-40) are ranked using the betweenness centrality of the nodes. **nooverlap** algorithm is applied so that no two nodes are overlapping