**In search of a core cellular network in single cell transcriptome data**

**Introduction**

Life on Earth has gone through many transitions in organizational complexity (Smith and Szathmary 1997). Among these, the evolution of multicellularity stands out as a key milestone. This transition has occurred independently multiple times across the tree of life and paved the way for tremendous phenotypic expansion and biological diversification (Parfrey and Lahr 2013). Although this has led to the evolution of cell-type-specific regulatory pathways that define cells with vastly different functions, all cells in multicellular organisms also carry out common functions that are essential for cell survival. Whether these common functions are supported by a common core of genes functioning in all cells, coordinated to ensure survival in the face of diverse functional demands, remains a central question in biology (Lim et al. 2013; Hart and Alon 2013). In particular, do all cells utilize the same set of genes to accomplish common functions, and do these genes function in a fixed and coordinated fashion—a core regulatory network?

To identify the molecular basis of common functions, researchers have used various strategies such as identifying constitutively active genes over temporal or spatial scales, and across environments. These genes are typically referred to as ‘housekeeping genes’ and are thought to perform essential functions. They tend to share aspects of sequence structure, chromatin environment and evolutionary history (Eisenberg and Levanon 2013; Rancati et al. 2018). For example, housekeeping genes are evolutionarily ancient (Zhu et al. 2008), exhibit a high level of evolutionary conservation (Zhang and Li 2004), and are enriched for several functions, including metabolism, RNA binding, protein degradation and cytoskeleton functions (Zhang and Li 2004; Lehner and Fraser 2004). While core functions are often described based on 'housekeeping genes', we recognize that genes do not work in isolation, but work with each other to carry out biological processes. Individual molecular abundances alone cannot adequately capture biological organizations. High-throughput methods that generate high-dimensional ‘omic’ data have greatly increased our understanding of molecular and cellular function and organization, in particular through the analysis of molecular networks (Barabasi and Oltvai 2004; Proulx et al. 2005; Thompson et al. 2015; Promislow 2005). Studying core functions from a network perspective may provide novel insights into the principles of biological organization.

Previous studies have explored different types of networks to find commonality. For example, co-expression network analysis of whole-organism level bulk transcriptome data in human and *Arabidopsis* has found a substantial number of gene pairs whose co-expression spans multiple datasets (Lee et al. 2004; He and Maslov 2016). In both analyses, gene pairs expressed across samples were enriched in translation, DNA replication, and regulation of transcription functions, all generally considered to be core cellular functions. Recent studies of tissue-level transcriptome data have typically focused on tissue-specific networks (Greene et al. 2015; Sonawane et al. 2017). For example, a recent study constructed tissue-specific PPI networks using co-immunoprecipitation within each of seven mouse tissues, and discovered core cellular modules, present in all mouse tissues, composed of evolutionarily ancient proteins (Skinnider et al. 2021).

A major drawback of most previous studies is that the networks were inferred from bulk data, which profiles heterogeneous cell populations of an organism or in a tissue. Bulk samples face two main limitations for network construction. First, differences in cellular compositions between samples may confound covariation analysis (Farahbod and Pavlidis 2020). Second, measurements that are averaged over thousands of cells in bulk samples make it difficult to detect interactions between genes in individual cells, such as the presence of co-expression patterns and the cell-specificity of these interactions. Co-expression in particular is an indicator of functional relationships (Hughes et al. 2000), which from a network perspective can provide valuable insight into cell function (Barabasi and Oltvai 2004).

With the advent of single-cell sequencing technologies, we have an unprecedented opportunity to reveal gene relationships in specific cellular contexts and probe cellular-level networks (Trapnell 2015; Tanay and Regev 2017). Despite the compelling motivation and a wealth of diverse data, identifying core networks through the analysis of gene co-expression networks in single-cell sequencing data remains both a promise and a challenge. Several factors contribute to these challenges. First, single-cell data has a high level of sparsity and noise, adding uncertainties in conventional analysis (Lähnemann et al. 2020). Second, it remains an open question how to choose an appropriate measure to quantify gene co-expression strength (Stone et al. 2021; Ly and Vingron 2021). Third, with a large number of heterogeneous cells profiled in a typical single-cell experiment, at which level to construct gene co-expression networks and how to extract their common co-expression patterns.

Here, we address a number of these challenges by proposing a computational pipeline tailored to single-cell sequencing data. We first systematically evaluated a set of gene co-expression measurements to select the most appropriate one. We then constructed cell cluster-specific gene co-expression networks. To detect common patterns from multiple large-scale networks, we adopted a matrix factorization (MF) informed technique. MF is a popular method in high-dimensional data analysis, which has been successfully applied in cell clustering, gene module identification, and cell trajectory inference (Stein-O’Brien et al. 2018). Finally, we validated these common patterns in each individual network through permutations and defined the most recurrent ones as common modules which potentially compose a core cellular network.

As a proof of principle, we applied our pipeline to a single-cell RNA sequencing (scRNA-seq) dataset derived from whole fly brains (Davie et al. 2018), aiming to answer several fundamental questions: Can we identify shared co-expression patterns across different cell types, how common are these connections occur across different cell types, do these shared co-expressed genes define a core cellular network, and if so, what properties does this core network manifest?

We identified four common modules which collectively form a core cellular network in the fly brains. We functionally annotated these modules and found that these common modules have different usages in brain cell types and some of them show age-related changes. To validate our findings, we analyzed a second single-nucleus RNA sequencing (snRNA-seq) dataset from fly head samples (Li et al. 2021). We showed that the common modules are recapitulated in this separate dataset and also pointed out those dataset-specific common modules likely resulting from the different experimental techniques. Gaining such information not only allows us to understand the composition and function of the detected core network, but also provides insight into the molecular organization principles of gene modules underlying cell phenotypes. To our knowledge, this is the first study searching for a core cellular network among cell types using single-cell sequencing data in the fly brains.

**Results**

Our results are summarized as follows. First, we describe our pipeline. Second, we show the core network recovered from a fly brain single-cell RNA-seq dataset applying our pipeline. Third, we illustrate the biological properties of the identified common gene modules. Finally, we cross-validate the common modules in a separate fly head single-nucleus RNA-seq dataset.

**Pipeline overview**

Gene co-expression network analysis is a widely used approach in uncovering gene relationships, but how to conduct such analysis properly in single-cell data given its sparse and noisy measurements remains an issue. Moreover, with multiple large-scale gene co-expression networks, how to efficiently and accurately detect common functional biological modules is challenging. We designed a pipeline to circumvent these challenges (**Method)**. **Figure 1** illustrates the working principle of our pipeline. This pipeline starts with constructing cell cluster-specific gene co-expression networks. We systematically evaluated seven metrics quantifying the strength of gene co-expression and found Pearson's Correlation Coefficient (PCC) showed relatively robust and stable performance (**Supplementary Note 1**). In a recent study which benchmarked several gene correlation measures in network inference from single-cell data also found PCC behaved better than other metrics (Stone et al. 2021). Thus, we chose PCC in constructing cell cluster-specific gene co-expression networks.

To identify common gene modules from a set of gene co-expression matrices, we employed a multi-view Nonnegative Matrix Factorization (mvNMF) technique which builds on the flexible matrix factorization framework (**Supplementary Note 2)**. The mvNMF algorithm was proposed to find shared patterns from multiple input matrices by computing a consensus basis matrix (Liu et al. 2013). Directly applying mvNMF to a large number of genome-scale gene networks can be time-consuming and computational expensive. We resorted to a more efficient approach by compressing information from large numbers of networks into two feature matrices. The first feature matrix shows the average gene co-expression strength across networks and the second one records their uniformity patterns. This compression reduces the computational task from joint matrix factorization on >100 matrices to a simpler one while retaining relevant signals for detecting shared network structures. We determined the parameters values used in the algorithm following a data-dependent manner based on random matrix theory (**Supplementary Note 3).**

After processing the two feature matrices by mvNMF, a consensus matrix is obtained, encoding how different genes contribute to each consensus factor. This consensus matrix forms the basis for candidate module extraction. The resultant candidate modules are evaluated in each gene co-expression network through permutations, and those modules that occur in at least 80% of all examined networks are defined as common modules.

We performed simulations to assess our pipeline’s performance. The result showed that our pipeline was able to detect both common modules or the common part of overlapping modules in simulated correlation matrices (**Supplementary Note 4**).

**Analysis of scRNA-seq fly brain atlas data reveal a core network**

To illustrate the application of our pipeline in detecting shared patterns across networks, we analyzed a fly brain scRNA-seq atlas dataset to find common modules in various cell types (Davie et al. 2018). This datasets contain 17,473 gene expressions in 56,902 cells with sex labels of 115 cell clusters. After data pre-processing, we retained 60 cell clusters in network analysis as they contain more than 100 cells in both females and males (**Method, Figure S1**).

To facilitate the search of common gene modules, we identified 3015 commonly expressed genes across cell clusters to be used in network construction. We constructed gene co-expression networks for each cell cluster in each sex separately and fed them into our pipeline. The output consensus basis matrix contained 25 columns, from which four gene modules were extracted. Permutation tests showed that all four modules were detected in every cell cluster in both females and males. Thus, we kept these four modules as common gene modules. Due to the fact that these four modules are present in every examined cell type, collectively, they form a core cellular network that exists in the fly brain.

These four modules differ in size, ranging from 10 to 166 gene members (**Figure 2, Table S1**). Mapping these four modules back to the strength feature matrix, we observed gene block structures supporting tightly gene coordinations. There is minimal gene overlap between modules aside from module 1 and module 4. While these two modules share 41 gene members, the shared percentage in either module alone is less than 30% (**Figure S2**). Considering some genes have multiple functional roles or be involved in several biological pathways, and the two modules were not dominated by the shared gene members, we kept them as separate gene modules.

To illustrate how module signals manifest in the feature matrices, we subset each feature matrix with genes of a given module and computed the sub-matrix average values. Contrasting one feature against the other for each module, the result showed that these two features are highly correlated (**Figure 2,** Pearson’s Correlation Coefficient, r = 0.99), suggesting a dense gene module preserved in more networks tends to have larger gene co-expression strength.

Out of these four modulus, module 2 showed the strongest signal. This module consists of 10 genes (**Table S1)**. Six genes encode heat shock proteins. The two long non-coding RNA genes, *-element* and *hsr-*, are well-known to be actively induced under the heat shock stress (Kang et al. 2018; Prasanth et al. 2000). The other two genes *stv* and *DnaJ-1* act as co-factors that regulate and interact with heat shock proteins. These genes all play critical roles in the protein folding process. Protein-folding is a vital cellular process for normal protein function and misfolded proteins contribute to the pathology of many diseases. In the brain, abnormal protein folding and aggregation are related to a number of neurodegenerative diseases, such as Alzheimer’s disease, Parkinson’s disease, Huntington’s disease (Ross and Poirier 2004). The recovery of this module as a core module in the fly brains highlights its universality and criticality in maintaining normal brain functions.

Module 3 showed the second strongest signal. This module mainly contains ribosomal protein genes and translation elongation factors, involved in cytoplasmic translation and ribosome biogenesis processes. Module 4 is the largest module which is dominated by gene members encoding key players in ATP metabolic processes, such as NADH dehydrogenase, cytochrome c oxidase, and ubiquinol-cytochrome-c reductase, supporting the notion that the pathways of ATP production need to respond rapidly to changes in energy demand at the synapse to sustain neuronal activity.

Module 1 is functionally annotated to brain-related functions, such as proton transmembrane transport and synaptic signaling (**Figure 2**). To identify characteristic genes of this module, we subset the strength feature matrix with its gene members and calculated the sum of all edge weights for each gene to approximate its hubness. The top genes with highest hubness scores are *TM4SF*, *cpx*, and *Sap47*. Gene *TM4SF* is a member of the tetraspanin family signatured by four transmembrane domains. It was found to be expressed at high levels in the fly head and its family members were shown to be involved in synapse formation, function and plasticity (Fradkin et al. 2002). Gene *cpx* encodes a synapse-specific SNARE-binding protein, which is a required component of the core fusion machinery in neurotransmitter release (Huntwork and Littleton 2007). Mutations in this gene lead to abnormal synaptic vesicle fusion and this gene has been shown to be conserved in *Drosophila* and mammals (Cho et al. 2010). It was also reported as a modifier of human-tau toxicity in a fly disease model (Feuillette et al. 2020). Gene *Sap47* encodes a conserved protein associated with synaptic vesicles, which is required for intact synaptic and behavioral plasticity (Blanco-Redondo et al. 2019).

In sum, we found four common modules in the fly brain atlas data whose sizes cover a wide range. Functional annotation showed that these modules are involved in basic cellular functions and brain related biological processes. Collectively, they form a core cellular network that exists in the fly brains.

**Common modules activity contribute to cell identity and show age-related dynamics**

The presence of common modules in every cell cluster supports their essentially maintaining cellular functions, but different cells may tune them differently to achieve a balance between housekeeping functions and cell-specific ones. To investigate the employment of common modules in different cell clusters, we used the AUCell algorithm to quantify common module activities in cells (Aibar et al. 2017). This algorithm benefits as it is independent of data normalization procedures and reports module activities at a single cell level. After quantifying the module activity in each cell, we used cell averages per cluster to represent module activities at a cell cluster level. A heatmap of the module activities across cell clusters showed that different cell clusters have different usage of common modules (**Figure 3A**). Module 3 was in general at high activity levels in all cell clusters, while module 2 was detected as most active in a subset of cell clusters. Module 1 and module 4 showed moderate activity levels across most of the clusters with the exception that module 1 was at low activity in glia cells.

We noticed that the combinations of different common module activity levels are informative about cell phenotypes or identities. Most female and male cells of the same cell cluster are grouped together. Non-neuronal cells, such as astrocyte-like, ensheating glia cells, and plasmatocytes formed a separate branch group away from other neuronal cells. Cell cluster 7 was singled out from both neuronal cells and non-neuronal cells. Given its high activity in module 2 with annotated protein folding functions, this cell cluster may represent stressed cells during single-cell sample preparation or sequencing experiments. Within neuronal cells, cell types of fly mushroom bodies were close to each other, such as Mushroom body output neurons (MBON) and Kenyon cells. Together, these common modules compose tightly coordination systems which work together to contribute to cellular phenotypes or identities.

**Common modules activity show age-related changes**

As the single-cell dataset contains fly brain cells collected at different age groups, there is a chance to investigate whether the activities of common modules show age-related dynamics. We examined 3 cell clusters which contain at least 10 cells in both females and males in every age group. We found that the activity of module 4, functionally annotated to ATP metabolic process, decreased with age, and the activity of module 2, related to the ribosomal biogenesis, remained active irrespective of age (**Figure 3B**). This result echoes previous findings (Davie et al. 2018). A recent study using single-cell transcriptome analysis on mouse aging brains showed that commonly aging-upregulated genes across cell populations were dominated by ribosomal protein genes and the most commonly aging-downregulated genes were mitochondrial respiratory chain complex and glycolysis-related genes (Ximerakis et al. 2019). Taking together, the activity dynamic patterns of common modules in the fly brain may reflect the translational needs and metabolic changes induced by aging.

**Analysis of single-nucleus RNA-seq fly head atlas data validate common modules**

To validate the finding in single-cell fly brain dataset, we applied our pipeline to a separate dataset on fly head but using a different technique, a single-nucleus fly head dataset (Li et al. 2021). We find that this single-nucleus dataset is more sparse than the single-cell dataset, with a 95.37% of zero in the gene expression count matrix, while this number in the single cell dataset is 90.55%. After quality control and data preprocessing, 54 cell clusters in both females and males were used in gene co-expression analysis and 1351 genes were identified as commonly expressed genes in this dataset (**Figure S1**).

Running our pipeline led to four candidate modules identified. Permutation test showed that while none of these 4 modules were detected in all 54 female and 54 male cell clusters, they all showed up in more than 80% of the examined cell clusters(**Figure S3**), thus, we kept all of them as common gene modules. The minimal module size is 7 and the largest is 157. Functional annotation of these modules showed that these modules are involved in basic cellular functions and brain-related ones (**Figure 4, Figure S4**). Out of these 4 modulus, module 3 has the strongest gene co-expression signal. This module contains 7 genes and is involved in protein-folding functions, consistent with the result from the single-cell dataset.

To verify our findings from the previous single-cell dataset, we compared the two sets of common modules in terms of module gene compositions. The result showed a high level of agreement between these two datasets (**Figure 5**). For example, modules 3 in single-cell dataset corresponds to module 1 in single-nucleus dataset with 25 genes overlapped, both were annotated to ribosome biogenesis functions. Protein folding related gene modules were recovered in both datasets, with 5 genes being the same. Chemical synaptic transmission as brain-related functions were encoded by module 1 in single-cell dataset and module 4 in single-nucleus dataset, sharing 43 gene members.

We also noticed some modules appear to be specific to one dataset. Module 2 in the single-nucleus dataset contains 7 genes and is related to the phototransduction process, which had two genes intersected in two separate modules of the single-cell dataset. Six out of the 7 gene members in this module have been experimentally demonstrated to be involved in the phototransduction pathway (Wang and Montell 2007; Mazzotta et al. 2013). This module’s activity distribution among cell clusters showed it was most active in cell clusters of the fly visual system, such as photoreceptor cells, pigment cells, and optic chiasma glial cells (**Figure S5**). Another dataset-specific module is module 4 in the single-cell dataset, which was annotated to the ATP metabolic process and contains 166 genes, only 7 of them overlapped with the single-nucleus dataset. This is within our expectation, as energy production is executed in the cytoplasm, which is captured by the single-cell experiment sample preparations but not the single-nucleus one.

In sum, we independently repeated our analysis on another fly head single-nucleus dataset and confirmed those common modules status. Also, we noted the dataset-specific common modules, some of them could be ascribed to different experiment techniques.

**Discussion**

To what extent do all cells in an organism rely on a common core of interacting genes? To investigate this question, we examined cell cluster-specific gene co-expression networks using fly brain scRNA-seq data. We proposed and demonstrated the use of a novel computation pipeline in exploring common gene modules using co-expression relationships of gene pairs across different brain cell clusters. We identified a core gene co-expression network composed of four modules in the fly brains. We showed examples that some gene members of the four modules have been experimentally demonstrated to play critical roles in certain biological processes. The identified modules were further investigated in terms of their relationship with cell identities and the aging process. We found differential usage of common modules across brain cell clusters and aging-related dynamics of module activities.

Our study is distinct in several ways from previous work interrogating core networks. First, instead of relying on expression levels of individual genes to identify genes common across cell types, we examined covariation between genes as the measure of functional commonality, which provides not only a stricter criterion to infer gene function (Hughes et al. 2000), but also likely captures conserved gene regulatory networks (Hart and Alon 2013; Stuart et al. 2003; Segal et al. 2003). Second, we identified covarying gene pairs using scRNA-seq data, which unlike bulk tissue data, can be defined by cell type, even within a single biological sample. In contrast to bulk transcriptomic analysis, where the cellular specificity of each interaction is largely ambiguous, scRNA-seq enabled us to build cell type-specific networks at a resolution that has hitherto not been possible.

**The continuous axes of cell identity**

Many studies in single cells focused on the specificity part, looking for cell type-specific genes or dividing cells in smaller and finer clusters. We note that the sharing part also encodes uniqueness information. For example, commonly expressed genes in single cells also entail information about cell identity, which is realized through forming tightly coordinated modules to be wired into different combinatorial activity patterns. This supports the concept of a continuous axis of cell identity, emphasizing the contribution of continuous expression gradient in determining cell identity aside from switch-like binary marker genes (Crow et al. 2020; Tanay and Regev 2017). Further work on cataloging basal gene expression programs and their activation pattern across diverse cell types would help develop empirical and mechanistic models of cell types and reveal the biological organization logic underlying cell phenotypes.

**Limitation**

In this study, we sought a core of interacting genes found across cell types in the fly brain. While the work described here benefits from access to high quality single-cell transcriptome data, there are still several caveats worth noting. First, we inferred co-expressed gene pairs from gene expression data statistically. Gene co-expression is not equivalent to gene co-regulation, which may be more indicative of functional relationships. Further experimental work is needed to validate the functional implications of these gene pairs. Second, we currently only include two feature matrices in the matrix factorization procedure. Extending this framework to include more feature matrices may increase the sensitivity of the pipeline in detecting common modules. Third, our current framework focuses on identifying common gene modules across cell clusters, extending our analysis to find condition-specific modules would help reveal more factors in understanding the functional organization of brain cell types.

**Conclusion**

Identifying the common factors in multiple biological networks is a significant task, for the reason that cells of different conditions, such as heterogeneous cell types in the same tissue, cell populations at different development time points, cells of various disease states, and homologous cell types in different species, may have shared patterns or underlying common cellular functional organizations. Understanding their sharing and specificity by pinpointing the underlying module structures help elucidate the organizing principles of cellular networks and provide deeper insight into molecular functions. Our study provides a clear route for identifying common modules in single-cell transcriptome data and highlights the importance of investigating network-level dynamics in biological systems.

**Methods**

**Datasets**

1. Single-cell RNA-seq fly brain atlas dataset

We downloaded the fly brain atlas data from NCBI Gene Expression Omnibus with GEO accession ‘GSE107451’. This dataset was generated using single-cell libraries on the 10X genomics platform, profiling fly brain cells from 8 age groups in both sexes (Davie et al. 2018). The original dataset contains 17,473 gene expression data in 56,902 high-quality female or male brain cells grouped into 115 cell clusters. As a quality control step, we excluded all mitochondrial genes in the dataset and removed 668 cells in a cell cluster named ‘Hsp’ as they represent stressed cells (Jasper Janssens, personal communication). We then removed cells that had either less than 200 expressed genes, less than 500 total unique molecular identifier counts, or a total fraction of mitochondrial gene expression exceeding 30%. These criteria led to the removal of another 42 cells, leaving 56,192 cells. These cells were assigned to 115 cell clusters. We selected 60 cell clusters which had at least 100 cells in both females and males, 29 of which were annotated to known brain cell types by the authors in the original publication. We used the ‘ID Validator’ tool on the flybase website to validate all gene symbols to the ‘FB2021\_05, released October 15, 2021’ version.

To find genes effectively expressed across cell types, we selected gene sets which express in more than 10 cells in each cell cluster in each sex separately, then extracted genes which were in more than half of these gene sets. This procedure led to 3015 genes identified as commonly expressed genes used in the following gene co-expression analysis (**Figure S1**). We used the R package ‘scran’ to perform data normalization before downstream analysis.

1. Single-nuclei RNA-seq fly head atlas dataset

We downloaded the fly head atlas data from the Fly Cell Atlas website (<https://flycellatlas.org/>). This dataset was generated using single-nuclei libraries on the 10X genomics platform (Li et al. 2021). The downloaded dataset contains 13,056 gene expression data in 100,527 high-quality female or male brain cells grouped into 82 cell clusters. We excluded all mitochondrial genes in the dataset and removed cells that had either less than 200 expressed genes, less than 500 total unique molecular identifier counts, or a total fraction of mitochondrial gene expression exceeding 30%. We excluded two cell clusters labelled as ‘unannotated’ or ‘artefact’ and cells whose sex labels were ‘mix’. These procedures led to 52,947 cells of 80 cell clusters retained. We selected 54 cell clusters which had at least 100 cells in both females and males, all of them were annotated to known brain cell types by the authors in the original publication.

We used the ‘ID Validator’ tool on the flybase website to validate all gene symbols. Applying the same criterion as the single-cell fly brain dataset to find commonly expressed genes, 1351 genes were identified (**Figure S1**), 1180 out of which overlapped with the 3015 commonly expressed genes from the single-cell dataset. We used the R package ‘scran’ to perform data normalization before downstream analysis.

**Pipeline**

1. **The construction of cell type-specific gene co-expression networks**

While there are a number of measures quantifying gene co-expression strength, it remains an open question which one is appropriate for single cell data given its high sparsity and noise (Lähnemann et al. 2020). To select an informative metric, we systematically evaluated the consistency and stability of seven gene co-expression related measures (**Supplementary Note 1**).

These seven measures are: Pearson’s Correlation Coefficient (PCC), a measure of proportionality () (Quinn et al. 2017)., a co-dependency index (CDI) (Nakajima et al. 2021), a Bayesian correlation (boca) (Sanchez-Taltavull et al. 2020), a correlation measure after controlling for individual genes’ zero proportion (sclink) (Li and Li 2021), a correlation measure using gene differential co-expression pattern (bigSCale2’) (Iacono et al. 2019), and a correlation measure after pooling similar cells into pseudocells (pseudocells) (Feregrino and Tschopp 2021).

In brief, we made data replicates by subsampling cells of the same cell cluster and accessed the consistency of each metric by comparing their highly ranked correlated genes. Based on their performance in our evaluation framework, we chose PCC to measure gene co-expression strength (**Figure S6**). Gene co-expression networks were constructed for each cell cluster separately, using the intersection of each cell cluster’s effectively expressed genes and the identified common expressed gene set.

1. **Feature matrix construction**

To find commonality among a large number of gene co-expression matrices, we compressed them into two feature matrices. This procedure reduces the computational cost and sets the following matrix factorization analysis ignorant of the number of input matrices. The two feature matrices are referred to as the strength matrix and uniformity matrix.

With *N* gene co-expression matrix coming from *N* cell clusters, the co-expression strength of gene *i* and gene *j* in matrix *k* is . The strength matrix, denoted as is defined as:

This feature matrix captures the average gene-gene co-expression strength over a set of networks.

The second feature matrix, referred to as the uniformity matrix and denoted as , is designed to capture if a co-expressed gene pair is uniformly distributed among the networks:

Both values of the strength and the uniformity feature matrices are between 0 and 1, with larger values supporting recurrence of co-expressed gene pairs in multiple cell cluster-specific networks.

In calculation, we took the absolute values of each gene co-expression matrix. To boost the signal of highly co-expressed gene pairs, we re-scaled each gene co-expression strength using:

Then we transformed these values using a logistic function:

, with d = log(9999) and c = -2 log(9999).

After this transformation, for scaled gene co-expression values between 0 and 0.3, is close to 0, and for scaled values between 0.6 and 1, is close to 1. The two feature matrices are calculated on transformed gene co-expression matrices.

1. **Multi-view Non-negative Matrix Factorization**

After summarizing gene co-expression matrices into two representative views as the strength and uniformity matrices, we relied on the matrix factorization principle to find their common basis. Matrix factorization is a commonly used approach to derive interesting biological insights from high-dimensional data and joint matrix factorization enables the identification of a common basis (Stein-O’Brien et al. 2018).

Here, we employed a multi-view non-negative matrix factorization (mvNMF) technique (Liu et al. 2013). We briefly describe the algorithm here, a more detailed description and how it relates to the general matrix factorization problem is shown in **Supplementary Note 2**.

The idea behind mvNMF is to integrate information from different views of the same system to learn common patterns. Different cells provide various views of gene coordinations in a biological system, which being integrated together offer compatible and complementary aspects to probe common gene modules. Each view represented as a matrix, the mvNMF algorithm achieves this goal by deriving a consensus matrix from the multiple input matrices which encodes commonality as consensus factors.

In our context, using the two symmetric feature matrices as inputs, this computation problem is formulated as finding the matrix , , and , which minimize the objective function:

,

subject to constraints, ,

where the denotes Frobenius norm, and represent weights to balance the two views.

In this formulation, the two symmetric data matrices and each is factorized into two components, a view-specific component or and a shared component manifested by the consensus matrix .

The algorithm to solve this optimization problem was proposed in (Liu et al. 2013). It starts with fixing matrix to compute the two view-specific matrices using a multiplicative update rule. After these two feature matrices are obtained, an exact solution exists for calculating the consensus matrix . A matlab code implementing this algorithm was published in (Wang et al. 2018), we adapted it into a custom R code.

Parameters involved in this algorithm include the number of dimensions, or columns for , and the two relative view weights and . We showed how to choose these parameter values in **Supplementary Note 3**. In analyzing the two real datasets, we gave equal importance to the two feature matrix by setting both and to be 0.01. We determined the number of dimensions of the consensus matrix using random matrix theory which was set as 25 for the single-cell dataset and 17 for the single-nucleus dataset.

1. **Module gene member selection**

In the consensus matrix , each column offers a potential gene module shared by the two feature matrices. To extract candidate module gene members, we first selected genes based on their weights in each column separately, and then merged gene sets from different columns with high similarity.

Specifically, to select genes which contribute most to each column vector, we adopted a thresholding approach. We z-score transformed each column of , subtracting each element by its column mean then divided by the column standard deviation. If the resultant z-score of a gene in a column is larger than a given threshold , it is selected as a gene member for that column. is typically set between 2 and 5, smaller leads to large module size which may contain much redundant information and larger makes the modules smaller which may leave some gene signal out. In analyzing the two real datasets, we set to be 3.

After selecting genes for each column, we compared pairwise column gene sets and merged those which share more than 30% their respective gene members. Finally, we removed gene sets which contain less than 5 genes and took the remainings as candidate gene modules.

1. **Candidate module validation and common module identification**

With candidate modules obtained, we assessed their existence in individual gene co-expression networks with a permutation approach. We first quantified a module’s observed signal in a network as (Xiao et al. 2014):

Then we perform permutations to get the signal’s null distribution with random same size modules sampled from the respective network. The P value is calculated as the proportion of random modules that have a larger signal value than the observed one. Raw P values are adjusted using the ‘p.adjust’ function in R with the Benjamin-Hochberg method, and the adjusted P values below 0.01 are regarded as supporting the existence of a module in that respective network. Modules that were detected in more than 80% of all examined networks were extracted as common modules.

1. **Module function annotation**

We performed Gene Ontology (GO) enrichment analysis of genes in each module using the R package ‘clusterProfiler’ (Wu et al. 2021) with a Bonferroni correction and an adjusted P value cutoff of 0.05. Significant GO terms were identified and refined to reduce redundant GO terms via the simplify method from the ‘clusterProfiler’ package.

1. **Module activity quantification**

For each identified module from the dataset, we applied the AUCell algorithm (Aibar et al. 2017) to the gene expression count data to quantify its activation level within each individual cell.

Given one cell *x*, genes are first ranked in descending order according to their expression level in *x*. Then, the recovery curve (ROC) for module *M* is derived where the x-axis is the expression-level based ranking of all genes and the y-axis the number of genes recovered from the gene members of module *M*.

The activity score of module *M* in cell *x* is defined as the area under the curve (AUC) for the top ranked genes. The assumption of this measure is that modules with higher activity in a biological process tend to possess high-ranking gene expression levels, therefore harbor a higher activity score. The advantage of the AUCell algorithm is its independence of gene expression units or normalization methods. We calculated module activity score in each individual cell and then took their average for a cell cluster. After this procedure, a module activity score by cell cluster matrix is obtained.

**Simulations**

We tested the performance and accuracy of our pipeline using two simulation schemes. The first scheme simulated multiple symmetric gene covariation networks harboring different numbers of common modules, the second scheme simulated overlapping modules, with each module containing a common part and data-specific part. We simulated weighted networks by adding Gaussian noise with two sets of parameters. We used several metrics to evaluate the performance, including true positive rate (TPR), false positive rate (FPR) and Matthew’s correlation coefficient (MCC) (Xiao et al. 2014). Detailed information about the simulations is shown in **Supplementary Note 4** and **Figure S7.**

**References**

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**: 1083–1086.

Barabasi A-L, Oltvai ZN. 2004. Network biology: understanding the cell’s functional organization. *Nat Rev Genet* **5**: 101–113.

Blanco-Redondo B, Nuwal N, Kneitz S, Nuwal T, Halder P, Liu Y, Ehmann N, Scholz N, Mayer A, Kleber J. 2019. Implications of the Sap47 null mutation for synapsin phosphorylation, longevity, climbing proficiency and behavioural plasticity in adult Drosophila. *J Exp Biol* **222**: jeb203505.

Cho RW, Song Y, Littleton JT. 2010. Comparative analysis of Drosophila and mammalian complexins as fusion clamps and facilitators of neurotransmitter release. *Mol Cell Neurosci* **45**: 389–397.

Crow M, Suresh H, Lee J, Gillis J. 2020. Coexpression reveals conserved mechanisms of transcriptional cell identity. *bioRxiv*.

Davie K, Janssens J, Koldere D, De Waegeneer M, Pech U, Kreft Ł, Aibar S, Makhzami S, Christiaens V, González-Blas CB. 2018. A single-cell transcriptome atlas of the aging Drosophila brain. *Cell* **174**: 982–998.

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet* **29**: 569–574.

Farahbod M, Pavlidis P. 2020. Untangling the effects of cellular composition on coexpression analysis. *Genome Res* **30**: 849–859.

Feregrino C, Tschopp P. 2021. Assessing evolutionary and developmental transcriptome dynamics in homologous cell types. *Dev Dyn*.

Feuillette S, Charbonnier C, Frebourg T, Campion D, Lecourtois M. 2020. A connected network of interacting proteins is involved in human-Tau toxicity in Drosophila. *Front Neurosci* **14**: 68.

Fradkin LG, Kamphorst JT, DiAntonio A, Goodman CS, Noordermeer JN. 2002. Genomewide analysis of the Drosophila tetraspanins reveals a subset with similar function in the formation of the embryonic synapse. *Proc Natl Acad Sci* **99**: 13663–13668.

Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC. 2015. Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet* **47**: 569–576.

Hart Y, Alon U. 2013. The utility of paradoxical components in biological circuits. *Mol Cell* **49**: 213–221.

He F, Maslov S. 2016. Pan-and core-network analysis of co-expression genes in a model plant. *Sci Rep* **6**: 1–11.

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102**: 109–126.

Huntwork S, Littleton JT. 2007. A complexin fusion clamp regulates spontaneous neurotransmitter release and synaptic growth. *Nat Neurosci* **10**: 1235–1237.

Iacono G, Massoni-Badosa R, Heyn H. 2019. Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome Biol* **20**: 1–20.

Kang I, Choi Y, Jung S, Lim JY, Lee D, Gupta S, Moon W, Shin C. 2018. Identification of target genes regulated by the Drosophila histone methyltransferase Eggless reveals a role of Decapentaplegic in apoptotic signaling. *Sci Rep* **8**: 1–10.

Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A. 2020. Eleven grand challenges in single-cell data science. *Genome Biol* **21**: 1–35.

Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. 2004. Coexpression analysis of human genes across many microarray data sets. *Genome Res* **14**: 1085–1094.

Lehner B, Fraser AG. 2004. Protein domains enriched in mammalian tissue-specific or widely expressed genes. *Trends Genet* **20**: 468–472.

Li H, Janssens J, De Waegeneer M, Kolluru SS, Davie K, Gardeux V, Saelens W, David F, Brbić M, Leskovec J. 2021. Fly Cell Atlas: a single-cell transcriptomic atlas of the adult fruit fly. *bioRxiv*.

Li WV, Li Y. 2021. sclink: Inferring sparse gene co-expression networks from single-cell expression data. *Genomics Proteomics Bioinformatics*.

Lim WA, Lee CM, Tang C. 2013. Design principles of regulatory networks: searching for the molecular algorithms of the cell. *Mol Cell* **49**: 202–212.

Liu J, Wang C, Gao J, Han J. 2013. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM international conference on data mining*, pp. 252–260, SIAM.

Ly L-H, Vingron M. 2021. Effect of imputation on gene network reconstruction from single-cell RNA-seq data. *bioRxiv*.

Mazzotta G, Rossi A, Leonardi E, Mason M, Bertolucci C, Caccin L, Spolaore B, Martin AJM, Schlichting M, Grebler R. 2013. Fly cryptochrome and the visual system. *Proc Natl Acad Sci* **110**: 6163–6168.

Nakajima N, Hayashi T, Fujiki K, Shirahige K, Akiyama T, Akutsu T, Nakato R. 2021. Codependency and mutual exclusivity for gene community detection from sparse single-cell transcriptome data. *Nucleic Acids Res* **49**: e104.

Parfrey LW, Lahr DJG. 2013. Multicellularity arose several times in the evolution of eukaryotes (Response to DOI 10.1002/bies. 201100187). *Bioessays* **35**: 339–347.

Prasanth K V, Rajendra TK, Lal AK, Lakhotia SC. 2000. Omega speckles-a novel class of nuclear speckles containing hnRNPs associated with noncoding hsr-omega RNA in Drosophila. *J Cell Sci* **113**: 3485–3497.

Promislow D. 2005. A regulatory network analysis of phenotypic plasticity in yeast. *Am Nat* **165**: 515–523.

Proulx SR, Promislow DEL, Phillips PC. 2005. Network thinking in ecology and evolution. *Trends Ecol Evol* **20**: 345–353.

Quinn TP, Richardson MF, Lovell D, Crowley TM. 2017. propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci Rep* **7**: 1–9.

Rancati G, Moffat J, Typas A, Pavelka N. 2018. Emerging and evolving concepts in gene essentiality. *Nat Rev Genet* **19**: 34–49.

Ross CA, Poirier MA. 2004. Protein aggregation and neurodegenerative disease. *Nat Med* **10**: S10–S17.

Sanchez-Taltavull D, Perkins TJ, Dommann N, Melin N, Keogh A, Candinas D, Stroka D, Beldi G. 2020. Bayesian correlation is a robust gene similarity measure for single-cell RNA-seq data. *NAR genomics Bioinforma* **2**: lqaa002.

Segal E, Shapira M, Regev A, Pe’er D, Botstein D, Koller D, Friedman N. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**: 166–176.

Skinnider MA, Scott NE, Prudova A, Kerr CH, Stoynov N, Stacey RG, Chan QWT, Rattray D, Gsponer J, Foster LJ. 2021. An atlas of protein-protein interactions across mouse tissues. *Cell* **184**: 4073–4089.

Smith JM, Szathmary E. 1997. *The major transitions in evolution*. Oxford University Press.

Sonawane AR, Platig J, Fagny M, Chen C-Y, Paulson JN, Lopes-Ramos CM, DeMeo DL, Quackenbush J, Glass K, Kuijjer ML. 2017. Understanding tissue-specific gene regulation. *Cell Rep* **21**: 1077–1088.

Stein-O’Brien GL, Arora R, Culhane AC, Favorov A V., Garmire LX, Greene CS, Goff LA, Li Y, Ngom A, Ochs MF, et al. 2018. Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends Genet* **34**: 790–805. https://doi.org/10.1016/j.tig.2018.07.003.

Stone M, McCalla SG, Siahpirani AF, Periyasamy V, Shin J, Roy S. 2021. Identifying strengths and weaknesses of methods for computational network inference from single cell RNA-seq data. *bioRxiv*.

Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science (80- )* **302**: 249–255.

Tanay A, Regev A. 2017. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**: 331–338.

Thompson D, Regev A, Roy S. 2015. Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annu Rev Cell Dev Biol* **31**: 399–428.

Trapnell C. 2015. Defining cell types and states with single-cell genomics. *Genome Res* **25**: 1491–1498.

Wang P, Gao L, Hu Y, Li F. 2018. Feature related multi-view nonnegative matrix factorization for identifying conserved functional modules in multiple biological networks. *BMC Bioinformatics* **19**: 1–13.

Wang T, Montell C. 2007. Phototransduction and retinal degeneration in Drosophila. *Pflügers Arch J Physiol* **454**: 821–847.

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L. 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov* **2**: 100141.

Xiao X, Moreno-Moral A, Rotival M, Bottolo L, Petretto E. 2014. Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules. *PLoS Genet* **10**: e1004006.

Ximerakis M, Lipnick SL, Innes BT, Simmons SK, Adiconis X, Dionne D, Mayweather BA, Nguyen L, Niziolek Z, Ozek C. 2019. Single-cell transcriptomic profiling of the aging mouse brain. *Nat Neurosci* **22**: 1696–1708.

Zhang L, Li W-H. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol* **21**: 236–239.

Zhu J, He F, Hu S, Yu J. 2008. On the nature of human housekeeping genes. *Trends Genet* **24**: 481–484.