Klimm *et al. BMC Genomics* (2020) 21:756
https://doi.org/10.1186/s12864-020-07144-2

BMC Genomics

RESEARCH ARTICLE Open Access

Functional module detection through integration of single-cell RNA sequencing data with protein–protein interaction networks

Florian Klimm[1,2*], Enrique M. Toledo[3], Thomas Monfeuga[3], Fang Zhang[3], Charlotte M. Deane[4] and Gesine Reinert[4]

Abstract

Background: Recent advances in single-cell RNA sequencing have allowed researchers to explore transcriptional function at a cellular level. In particular, single-cell RNA sequencing reveals that there exist clusters of cells with similar gene expression profiles, representing different transcriptional states.

Results: In this study, we present scPPIN, a method for integrating single-cell RNA sequencing data with protein–protein interaction networks that detects active modules in cells of different transcriptional states. We achieve this by clustering RNA-sequencing data, identifying differentially expressed genes, constructing node-weighted protein–protein interaction networks, and finding the maximum-weight connected subgraphs with an exact Steiner-tree approach. As case studies, we investigate two RNA-sequencing data sets from human liver spheroids and human adipose tissue, respectively. With scPPIN we expand the output of differential expressed genes analysis with information from protein interactions. We find that different transcriptional states have different subnetworks of the protein–protein interaction networks significantly enriched which represent biological pathways. In these pathways, scPPIN identifies proteins that are not differentially expressed but have a crucial biological function (e.g., as receptors) and therefore reveals biology beyond a standard differential expressed gene analysis.

Conclusions: The introduced scPPIN method can be used to systematically analyse differentially expressed genes in single-cell RNA sequencing data by integrating it with protein interaction data. The detected modules that characterise each cluster help to identify and hypothesise a biological function associated to those cells. Our analysis suggests the participation of unexpected proteins in these pathways that are undetectable from the single-cell RNA sequencing data alone. The techniques described here are applicable to other organisms and tissues.

*Correspondence: f.klimm@gmail.com
[1]Department of Mathematics, Imperial College London, London SW7 2AZ, UK
[2]Mitochondrial Biology Unit, University of Cambridge, Cambridge CB2 0XY, UK
Full list of author information is available at the end of the article

Background

Liver metabolism is at the centre of many non-communicable diseases, such as diabetes and cardiovascular disease [1]. In healthy organisms, the liver is critical for metabolic and immune functions and gene-expression studies have revealed a diverse population of distinct cell types, which include hepatocytes in diverse functional cell states [2]. As diabetes is a complex and heterogenous disease, the study of liver physiology at single-cell resolution helps us to understand the biology [3]. At a single-cell level, however, large-scale protein interaction data is not yet available [4]. The available data are in the form of single cell gene expression levels. Such single cell data provide processing challenges but it is plausible to enhance their analysis through the integration of protein interactions. In this study, we develop scPPIN a method for the integration of single-cell RNA-sequencing data with complementary PPINs. Our scPPIN analysis of liver single cell data and PPINs reveals biological pathways in cells of different transcriptional states that hint at inflammatory processes in a subset of hepatocytes.

In recent years, much attention has been given to scRNA-seq techniques as they allow researchers to study and characterise tissues at a single-cell resolution [5–7]. Most importantly, scRNA-seq reveals that there exist clusters of cells with similar gene expression profiles, commonly referred to as 'cell states' [8]. Multiple approaches have been created to reveal these cell clusters, driven by the transcriptional profile of each cell [9, 10]. Computational tools can identify biomarkers for such cell clusters [11]. Specifically, the analysis of differentially expressed genes (DEGs) between these cell clusters has been shown to reveal different cell types [12], diseased cells [13], and cells that resist drug treatment [14]. Due to technological advances the quality and availability of scRNA-seq data has increased dramatically in the last decade [15]. This makes the development of computational approaches for interpreting scRNA-seq data an active field of research [16] of which one research direction is the identification of gene regulatory networks in scRNA-seq data (e.g., SCENIC [17], PIDC [18]).

These approaches do not make systematic use of available protein–protein interaction data. One can represented such data as PPINs and use PPINs to, for example, identify essential proteins [19–21] and to predict disease associations [22, 23] or biological functions [24–26]. For this, researchers have used tools from network science and machine learning. Many of these methods build on the well-established evidence that in PPINs, proteins with similar biological functions are closely interacting with each other. These groups of proteins with common biological functions are called *modules* [27, 28].

It is understood that gene-expression is context-specific and thus varies between tissues [29], changes over time [30], and differs between healthy and diseased states [31]. It follows therefore that different parts of a PPIN are active under different conditions [32]. Analysing PPINs in an integrated way, together with bulk gene-expression data, provides such biological context, helps to reveal context-specific active functional modules [33, 34], and can identify proteins associated with disease [35].

Based on the success of methods where PPINs have been integrated with bulk expression data, we have developed scPPIN, a novel method to integrate scRNA-seq data with PPINs. It is designed to detect active modules in cells of different transcriptional states. We achieve this by clustering scRNA-seq data, performing a DEGs analysis, constructing node-weighted PPINs, and identifying maximum-weight connected subgraphs with an exact Steiner-tree approach. Our method is applicable to the broad range of organisms for which PPINs are available [36].

The scPPIN method can be used to analyse mRNA-seq data from any tissue or organ type. As a case study, we investigate scRNA-seq data from human liver spheroids because this tissue is important in many diseases and it is known to have diverse cell types with different cellular metabolic processes. This makes the application of our method particularly relevant, because we expect the identification of very different active modules in different cell clusters — a hypothesis that our investigation partially confirms.

Our method identifies proteins involved in liver metabolism that could not be detected from the scRNA-seq data alone. Some of them have been shown to be involved in the liver of other organisms and for others this study is the first indicator of a specific function in liver. Furthermore, we can associate cells in a given transcriptional state with enriched biological pathways. In particular, we find that cell clusters have different biological functions, for example, translational initiation, defence response, and extracellular structure organisation. To test the scPPIN method on scRNA-seq data from a different tissue, we investigate human adipose tissue in the Supplementary Results 6. We detect other functional modules than in the liver data and also find different biological functions enriched.

This case study demonstrates that scPPIN provides insights into the context-specific biological function of PPINs. Importantly, these insights would not have been revealed from either data type (PPIN or scRNA-seq) alone. As this technique is, in principle, applicable to a wide range of organism and tissue types, it could reveal functional modules in these, too.

Results

In this paper, we present scPPIN, a method that detects functional modules in different cell clusters. The method

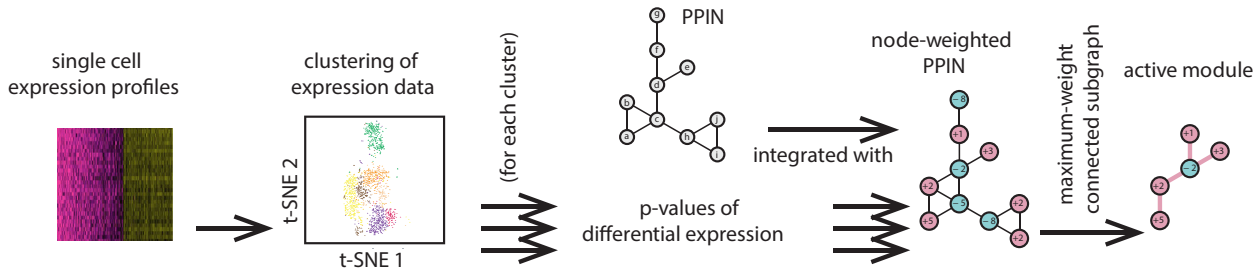

Fig. 1 Our method consists of the following steps. (**1**) clustering of scRNA-seq data (e.g., with SEURAT [9]). For each cluster, we (**2**) compute *p*-values of differential expression and use them to (**3**) estimate node scores by using an approach presented in [33]. (**4**) We combine these node scores with a PPIN to construct node-weighted PPINs for each cluster. (**5**) We compute functional modules as maximum-weight connected subgraphs

involves multiple analysis steps (see Fig. 1 for an overview and the "Methods" section for a detailed discussion). First, we preprocess the scRNA-seq profiles. Second, we use an unsupervised clustering technique from SEURAT to identify sets of cells in similar transcriptional states. Third, for each cluster, we identify DEGs using a Wilcoxon rank sum test. Fourth, for each gene in every cluster, we compute additive scores from these *p*-values (see [33] and Supplementary Note 1). Fifth, for each cluster, we map these gene scores to their corresponding proteins in a PPIN, constructed from publicly available data from BIOGRID [36]. Lastly, we identify functional modules as maximum-weight connected subgraphs in these node-weighted networks. To test whether this integrative analysis of scRNA-seq data with a PPIN has been successful, we compare the detected modules with a 'biological ground truth' in the form of GO-enrichment annotations.

In order to demonstrate SCPPIN, we investigate newly measured scRNA-seq data of liver hepatocytes (see "Methods" section for a description of the experimental setup and preprocessing steps). Using a standard modularity-maximisation algorithm, we obtain ten cell clusters of which seven consist of hepatocytes (see Fig. 2), which make up a majority of the liver tissue. Hepatocytes are known to show a functional diversity and are, for example, involved in the carbohydrate metabolism [2].

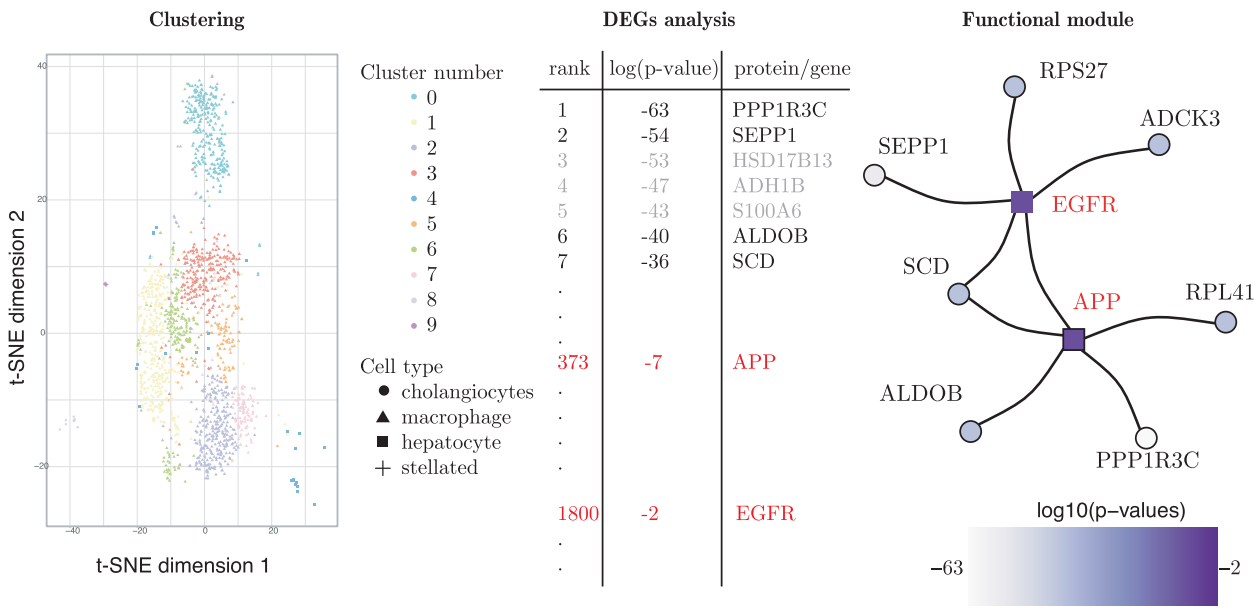

Fig. 2 (Left) Clustering the scRNA-seq data reveals ten clusters of which seven are hepatocyte cells. We visualise the cells in a two-dimensional space as obtained from a *t-distributed stochastic neighbour embedding* (t-SNE) of their original high-dimensional space [37], which is defined by the expression of the $d = 6983$ genes. (Middle) For each cluster, we perform DEGs analysis to identify genes that most differentially expressed in a given cluster. Here, we show DEGs for hepatocyte cluster 6 (H6). (Right) We use SCPPIN to integrate the *p*-values from DEGs analysis with the PPIN and identify a functional module for the H6 cluster. We find genes that are significantly differentially expressed ($p < 0.05$; disks) and proteins that are not strongly differentially expressed ($p \geq 0.05$; squares). Colour indicates *p*-value from low (white) to high (purple)

We focus on hepatocytes because it allows us to study the heterogeneity of cellular function in this single cell type.

We identify DEGs in each of the hepatocyte clusters. In Fig. 2, middle panel, we show the p-values of differential expression for some of the genes in hepatocyte cluster H6. Usually, the top-ranked genes in each of the clusters can be seen as 'marker genes', i.e., one may use these genes to associate cells with a certain transcriptional state. For H6, for example, *protein phosphatase 1 regulatory subunit 3C* (PPP1R3C) has the smallest of all p-values, namely 10^{-63}. It therefore could serve as a potential biomarker and is a known regulator of liver glycogen metabolism [38]. While such a DEG analysis reveals important genes in certain cell states it is not straightforward to identify the crucial biological pathways. Next, we demonstrate that integrating p-values from a DEGs analysis with PPIN information can reveal a more comprehensive picture of the biological processes. In the right panel of Fig. 2, we show a functional module identified by SCPPIN. We detect a subnetwork consisting of nine proteins. This module consist of seven proteins with small p-values (among them PPP1R3C) that are connected to each other via the *amyloid precursor protein* (APP) and *epidermal growth factor receptor* (EGFR), which have p-values $\sim 10^{-7}$ and 10^{-2}, respectively. Both proteins are integral membrane proteins and do not show significant differential expression in this cell cluster as they rank 373 and 1800 out of all differentially expressed genes. The EGFR signalling network has been identified as a key player in liver disease [39]. The precise function of APP is unknown but it is involved in Alzheimer's disease and also has been hypothesised to be involved in liver metabolism [40].

These findings demonstrate that SCPPIN can help to automate the further investigation of results from a DEG analysis by identifying parts of the PPIN that correspond to genes that are significantly differentially expressed. Furthermore, it also identifies proteins corresponding to genes that are not significantly differentially expressed in a particular cluster. These genes are candidates of a biological connector function between differentially expressed genes.

Influence of the false discovery rate

We have demonstrated that SCPPIN can reveal functional modules inside a PPIN and associate them with cells of a certain transcriptional state. Now we explore whether there is only one functional module for a given cell state or whether there are functional modules of different sizes. We anticipate in this case that the latter is true, as it has been shown that functional modules may exists at multiple scales [28].

There is only one free parameter in SCPPIN, the false discovery rate (FDR), which is defined as the proportion, out of all genes which are declared significantly differentially expressed, that are false positive and indeed not significantly differently expressed. Given the distribution of p-values of differential expression and a FDR, we can compute node weights that yield the intended FDR (see Supplementary Note 1). Intuitively, increasing the FDR identifies a larger subgraph of the PPIN as an active module. In the following, we explore this systematically, for the hepatocyte cluster H6 that we investigated above.

The size $M \in [1, N]$ of the detected modules is non-decreasing with the FDR. While the size M is non-decreasing, our method is non-monotonous, i.e., proteins identified for a certain FDR are not necessarily detected for all larger FDRs. For small FDRs, we detect a module of size $M = 1$, which is exactly the protein with the smallest p-value[1]. For FDRs close to one, we detect a maximum weight subgraph which is spanning almost the whole network.

In Fig. 3, we show the size M of the optimal subnetworks for cluster H6 as a function of the FDR. As expected, the M(FDR) is non-decreasing. For FDR $< 10^{-26}$, we detect a single node, which represents PPP1R3C, the protein with the smallest p-value ($\sim 10^{-63}$). For larger FDRs, we detect subnetworks of larger size that contain proteins that are associated with larger p-values and could not have been identified with the gene-expression data alone. For FDR $= 10^{-25}$, for example, we detect the subnetwork of size $M = 9$ (shown in Fig. 2).

For FDR $< 10^{-22}$, we detect an even larger functional module, which partially overlaps with the one identified for FDR $< 10^{-23}$, as it also includes EGFR as connector between proteins with small p-values. The second connector is *ELAV-like protein 1* (ELAVL1) with $p \approx 0.06$. The precise function of ELAVL1 is unknown but it is believed to play a role in regulating ferroptosis in liver fibrosis [41]. For even larger FDRs, we identify a module with $M = 42$ nodes out of which 9 are not identified from the gene-expression data alone. We observe all the before-mentioned connectors, as well as, *hepatocellular carcinoma-associated Antigen 88* (ECI2) and *S100 calcium-binding protein* A4 (S100A4). The latter regulates liver fibrogenesis by activating hepatic stellate cells [42]. Overall, the number of proteins we identify additionally with our method is moderately increasing with the FDR. In Supplementary Note 5, we show these M(FDR) curves for the hepatocyte clusters.

Functional modules for different clusters

In the "Influence of the false discovery rate" section, we investigated the influence of the FDR on the detected modules for a single cell cluster. As we obtained seven hepatocyte clusters (see Fig. 2), we can compute DEGs and

[1] If this is non-unique, multiple optimal modules of size $M = 1$ exist and can be detected. In none of our examples was this the case.

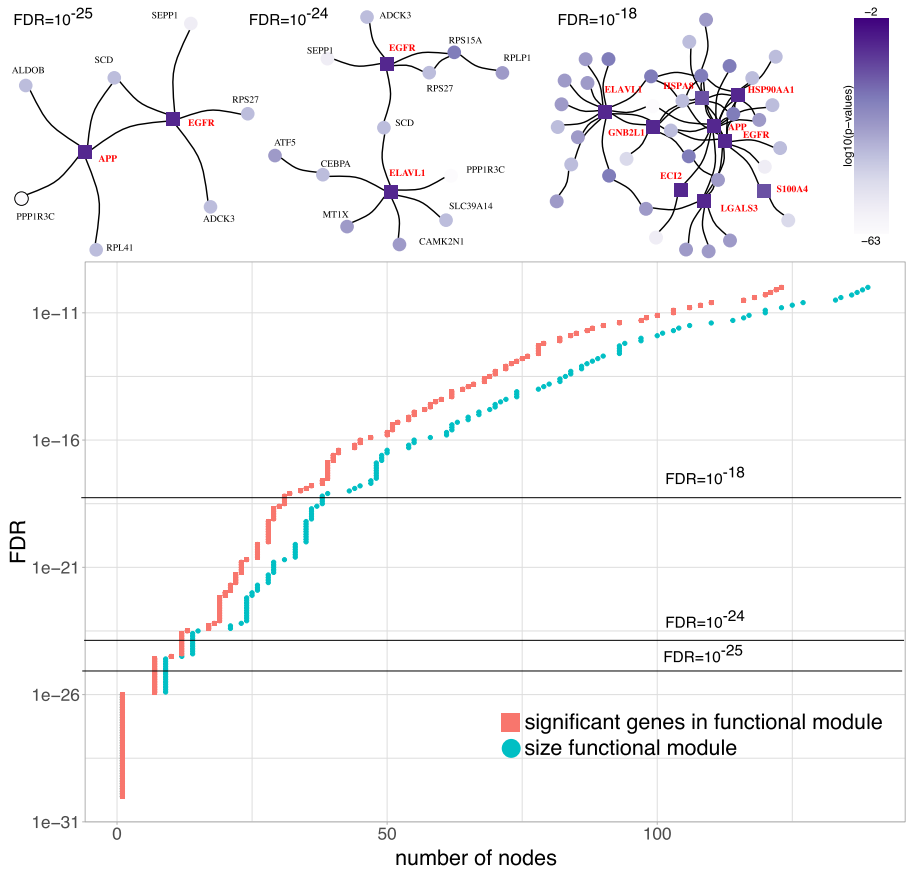

Fig. 3 (Upper Panel) Network plots of the detected modules for three choices of the FDR (10^{-25}, 10^{-24}, and 10^{-18}). The colour of the nodes shows their *p*-value from low (white) to high (purple). We indicate proteins that would not have been discovered from the gene-expression data alone as squares and give their names in bold red font. (Lower Panel) The size of detected modules (blue disks) depends on the FDR. A large fraction of proteins in these modules have significant *p*-values (red squares), however, for all FDR> 10^{-26}, we also identify additional proteins as for a given FDR, the blue disks are to the right of the red disks

thus also active modules for each cluster separately. As each cluster represents a different cell state, different genes are identified by a DEG analysis, which then results in different functional modules. To compare the detected modules, we use SCPPIN with a FDR of 10^{-27} for all of them. In Fig. 4, we show the functional modules for six clusters (we omit the largest cluster due to illustration limitation). The detected clusters differ in size, with the largest consisting of 52 nodes (cluster H2) and the smallest consisting only of a single node (Cluster H6). This heterogeneity occurs because the *p*-values of differential expression are differently distributed for each cluster. Cluster Two has the smallest *p*-values as its gene expression is most different from those in all other clusters, which indicates a special function of these cells in comparison to the rest. As shown for cluster H6 in Fig. 3, increasing the FDR increases also the size of the detected functional module.

In four out of the six modules, we find proteins that we could not have identified with a DEG analysis alone. For cluster H1, these are APP, ELAVL1, TRIM25, ACTN4, PTEN, KRAS, TFG, and RPL4. For cluster H2, these are VKORC1, APOA1, SNX27, CYCS, ECI2, APP, EGFR, UBE3A, HNRNPL, COPS5, TP53, YWHAE, RCHY1, and TERF2IP. For cluster H3, this is APP. For H5, these are HSPA8 and FN1. We find that APP is identified as part of the active module in three of these clusters, which indicates that this membrane-bound protein may play a role in different biological contexts.

To systematically access these biological contexts, we perform a GO-term enrichment test to assess the hypothesis that the detected modules represent biologically relevant pathways (see Methods). We find that all but the two smallest modules have GO terms enriched (see Table 1). The GO terms hint at distinct biological functions for the different cell clusters. Clusters H1 and H3 are involved in

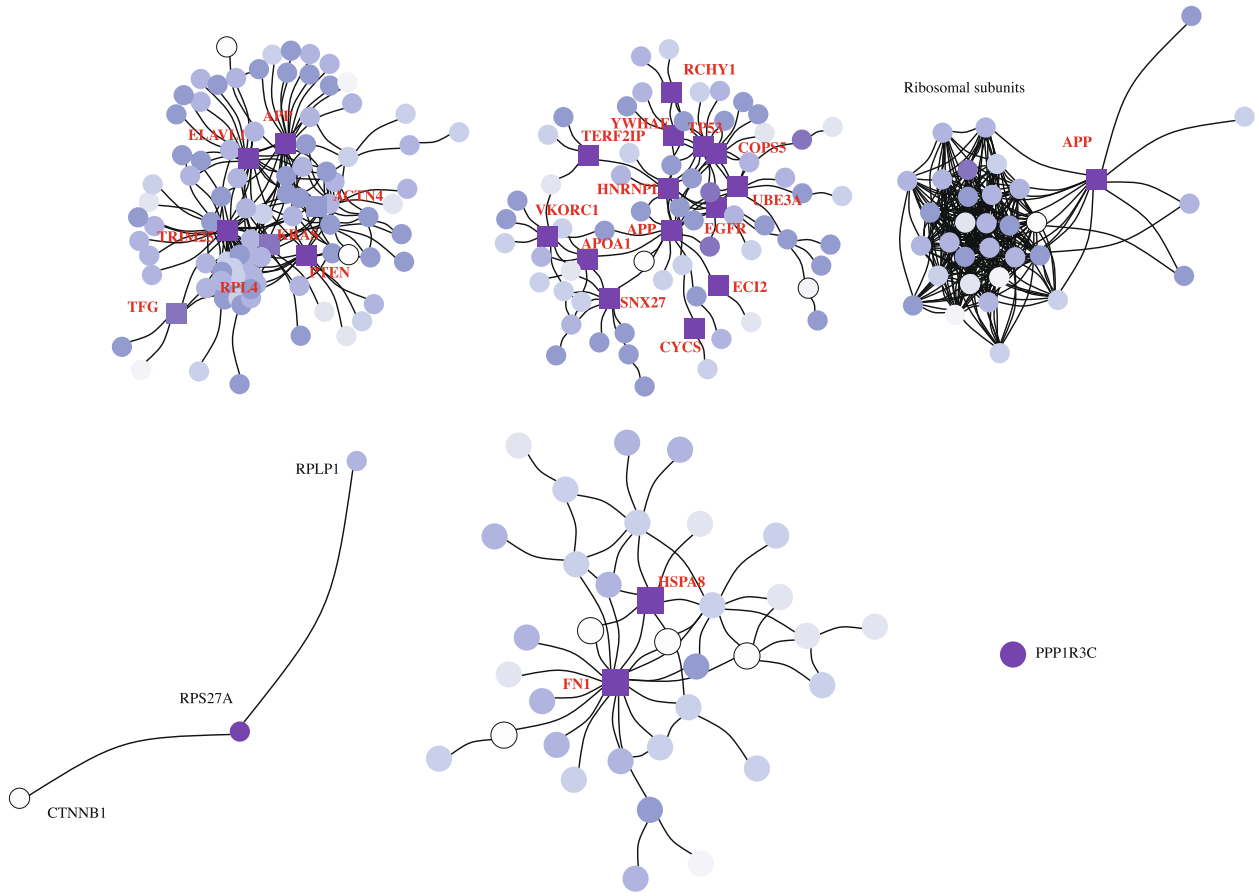

Fig. 4 Detected modules for six hepatocyte clusters for FDR $= 10^{-27}$.We find that the detected modules vary in size, with the smallest consisting of a single protein and the largest consisting of 253 proteins (we omit the largest cluster for visualisation purposes). Colour indicates *p*-value of associated gene from low (white) to high (purple). We show nodes as squares if they would not have been detected without PPIN information. For the larger modules, we only give the names of these proteins that we would not have detected by a DEG analysis. See SI for an illustration which includes protein names

translational initiation, H2 in response to stress, and H5 in the extracellular structure organisation. All of these identified cellular processes represent different hepatocytes functions that have been found in vivo [2, 43].

The analysis of different cell states in the scRNA-seq with scPPIN indicates that genes associated with different parts of the PPIN are active in different transcriptional states. Different biological functions of the cell clusters are reflected by different enriched GO terms. Overall, the integration of scRNA-seq with a PPIN suggests that the cells utilise the underlying PPIN differently to fulfil their diverse biological functions.

Discussion

In this study, we integrated scRNA-seq data with PPINs to construct node-weighted networks. For each cell cluster, detecting a maximum-weight connected subgraph identifies an *active module*, i.e., proteins that interact with each other and taken together the corresponding genes are significantly differently expressed. Our method scPPIN builds on advances in DEG analysis, which are standard tools for the interpretation of scRNA-seq data. As a case study, we investigated data from healthy human livers. We find that the seven identified cell clusters have different subnetworks of the PPIN as functional modules in which the corresponding genes exhibiting most significantly changed expression levels. A GO-term enrichment analysis indicates that these are also associated with different biological functions. Furthermore, these subnetworks identify proteins for which the corresponding genes are not differently expressed in a given cluster but do interact with proteins for which the corresponding genes are strongly differentially expressed. These proteins are candidates for regulatory functions in these cells. It is only through our combination of single-cell data with PPIN data that these candidate proteins can be identified. Often, they are integral membrane proteins such as FN1, EGFR, and APP, drivers of cell fate such as P53 and KRAS, or

Table 1 For each of the seven clusters we give the three most enriched GO terms and the multiple-testing-corrected *p*-values that we received from Fisher's exact tests. For these GO terms, we also give the fold-enrichment as the fraction of genes in the module with this annotation over the total number of genes with this annotation in the whole data set

Cluster	Module size *M*	Top enriched GO terms (fold-enrichment)	$\log_{10}(p\text{-value})$
H1	51	translational initiation (15/158)	−6
		nuclear-transcribed mRNA catabolic process (14/101)	−6
		SRP-dependent cotranslational protein targeting to membrane (14/90)	−5
H2	52	response to stress (32/491)	−2
		defense response (19/200)	−2
		cell maturation (6/18)	−2
H3	27	SRP-dependent cotranslational protein targeting to membrane (24/78)	−26
		nuclear-transcribed mRNA catabolic process (24/86)	−26
		translational initiation (24/105)	−24
H4	3	*none*	
H5	30	extracellular structure organization (9/25)	−3
		exocytosis (12/203)	−2
		alcohol metabolic process (9/69)	−3
H6	1	*none*	
H7	253	SRP-dependent cotranslational protein targeting to membrane (78/88)	−26
		nuclear-transcribed mRNA catabolic process (78/90)	−26
		cotranslational protein targeting to membrane (78/91)	−26

proteins of so-far unknown function such as TERF2IP and TFG.

In a more general setting, SCPPIN can be used to systematically analyse DEGs in scRNA-seq data. The identified networks that characterise each cluster help to identify and hypothesise a biological function associated to those cells. For example, we identified the gene S100A4 in the hepatocyte cluster H6. S100A4 has been identified as a key component in the activation of stellated cells in order to promote liver fibrosis [42]. Although previously identified in a population of macrophages [42, 44], observing the expression of S100A4 in this cluster of hepatocytes may indicate that a subpopulation of hepatocytes promotes fibrogenesis in paracrine. We also identified the amyloid precursor protein (APP) and interaction partners active in multiple hepatocyte clusters. Although little is known about liver-specific functions of APP, in the central nervous system it is a key driver of Alzheimer's disease, as source of the amyloid-β-peptide (Aβ) [45]. Due to the major role of liver in the clearance of plasma Aβ, it would be interesting to study the contribution of Aβ produced in the liver and the impact in the central nervous system. This systemic view of Alzheimer's disease [46, 47] may reveal alternative treatments. A more holistic approach is the comparison of the detected functional modules with disease associations to identify disease modules. We undertake and discuss this approach in Section 6 in the SI.

Despite their success, scRNA-seq techniques have methodological limitations (e.g., zero-inflation [48]). The presented technique might be further improved by considering such specific challenges, e.g., by constructing a different mixture model (see Supplementary Note 1) or implementing an imputation/noise reduction methodology. Furthermore, while we demonstrated exclusively a modularity-based cell clustering, other clustering algorithms might be appropriate, depending on the biological question and the experimental platform [49]. As they may reveal different cell states, SCPPIN may also reveal different functional modules in these.

In this study, we used DEGs as the foundation to identify the active modules in different cell types. We choose this approach because DEG analysis is a common tool for the identification of biomarkers in scRNA-seq experiments [50]. Our scPPIN method, however, could also be used on other statistics derived from scPPIN data, such as, scores indicating the abundance of gene expression.

There is a rich literature of alternative ways to identify active modules in PPINs [51–53]. Here, we decided to use a maximum-weight connected subgraph approach because it allows an exact solution and is widely-used for bulk RNA-seq analysis [33]. It is an open question whether other approaches, such as, *the maximum clique method* [54] or methods integrating gene-coexpression data [55] are also fruitful in the single-cell setting. Active-module detection methods in general could also be explored to identify potential novel protein interactions because there is an association between functional similarity of protein pairs and whether or not they interact [28]. In this study,

we used GO terms to computationally test the hypothesis that the detected modules facilitate a joint biological function. Despite its shortcomings, such as annotation bias [56], this is a widely-used approach. In future work experimental knockout studies could be used to test the function of these proteins in vivo in cellular context.

In conclusion, we demonstrate that integrating scRNA-seq data with PPINs detects distinct enriched biological pathways and demonstrates a functional heterogeneity of cell clusters in the liver. It suggest the participation of unexpected proteins in these pathways that are undetectable from a gene-expression analysis alone. We provide an R package SCPPIN, so that our method can easily be integrated to current analytical workflows for single cell RNA-seq analysis.

Methods

Protein–protein interaction network

We construct a PPIN from the publicly available BIOGRID database [36], version 3.5.166. The obtained network for *Homo sapiens* has $n = 17,309$ nodes and $m = 296,637$ undirected, unweighted edges. While the PPIN might be directed and edge-weighted [57] (e.g., considering confidence in an interaction [58]), we consider here exclusively undirected networks without edge weights.

Liver spheroid and bioinformatics

Human primary hepatocytes from a mixture of 10 donors grown in a 3D spheroid, were purchased from InSphero AG (Switzerland) and maintained in the culture medial provided by the company. Single cell libraries were prepared with a 10X Genomics 3' kit and sequenced in an Illumina NextSeq 500. Sequencing data demultiplexing and alignment was carried out with CELLRANGER with default parameters [59]. As a quality control, we only kept cells with between 500 and 6000 genes detected. A total of 2597 cells passed this quality control, of which 2123 are hepatocytes. We identified the cell types by using gene markers as identified in [2, 60, 61]. The data is publicly available and we make the process available under https://github.com/floklimm/scPPIN.

Preprocessing

We analyse the scRNA-seq data with the SEURAT R package v2.3.4 [62]. As a preprocessing step, we align the data with a canonical correlation analysis [62] with usage of the first nine dimensions. We identify clusters with the default resolution of one with the function *FindClusters*. To identify cell types, we use gene markers expression and in-house reference datasets.

To compute a p-value of differential expression for each obtained cluster, we use the function *FindAllMarkers* with the argument RETURN.THRESH equal to 1 and LOGFC.THRESHOLD set to 0.0 because we would like to obtain p-values for all genes (significant and non-significant ones). For the same reason, we do not employ a threshold for fold-change in gene expression. We exclude genes that are expressed in less than 10 % of a cluster to avoid comparing sparsely expressed genes.

Node-weighted network construction

The SCPPIN pipeline builds on a method for the identification of functional modules as introduced by Dittrich et al. for analysing bulk gene-expression data [33]. Dittrich et al. compute maximum-weight connected subgraphs to find subnetworks that change their expression significantly in a certain disease. Here, we use a similar approach to identify subnetworks that change significantly in different clusters of cells.

Given a network $G = \{V, E\}$ with node set V and edge set $E \subset V \times V$, we construct a node-weighted network $G_{nw} = \{V, E, W\}$ by assigning each node $i \in V$ a real-valued node weight w_i, which we represent as a function $W : V \to \mathbb{R}$ (see Eq. 1). We construct these node-weighted networks from a PPIN and gene-expression information. The former is in the form of a network and the latter are p-values of differential expression. We assume a bijection between genes and proteins, i.e., each protein is expressed by exactly one gene, which is a simplification of the biological processes. We find this bijection by mapping GeneIDs [36].

We delete all nodes from the PPIN for which no gene-expression data is available. In the Supplementary Note 4, we present an alternative approach that can incorporate proteins with missing expression data.

We assign each node a score

$$W(x) = (\alpha - 1)\left(\log(x) - \log(\tau)\right), \tag{1}$$

which is a function of the p-value x and we vary the *significance threshold* τ to tune the *false discovery rate* (FDR). We estimate α by fitting a *beta-uniform mixture model* to the observed p-values (see Supplementary Note 1). This score $S(x)$ is negative for proteins below the significance threshold τ and positive otherwise.

Mathematical optimisation algorithm

Mathematically, the problem of identifying a subnetwork with maximal change of expression is a *maximum-weight connected subgraph problem*. Algorithmically, it is easier to solve an equivalent *prize-collecting Steiner tree* (PCST) problem [33]. Steiner trees are generalisations of spanning trees [63] and 'prize-collecting' indicates that the nodes have weights. To find a PCST, we use the dual ascent-based branch-and-bound framework DAPCSTP [64, 65]. For all calculations in this paper the algorithm identified an optimal solution in less than 10 s. For details see Supplementary Note 2.

Gene ontology enrichment

We use TOPGO in version 3.8 for the gene ontology enrichment (GO-enrichment) analysis. [66]. We use Fisher's exact test to identify enriched GO terms [67]. All reported GO terms are significant with p-value 0.01 and we use a Benjamini–Hochberg procedure to counteract the multiple-comparison problem.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-020-07144-2.

Additional file 1: Supplementary Information.

Abbreviations

APP: Amyloid precursor protein; DEGs: Differentially expressed genes; ECI2: Hepatocellular carcinoma-associated antigen 88; EGFR: Epidermal growth factor receptor; ELAVL1: ELAV-like protein 1; FDR: False discovery rate; GO-enrichment: Gene ontology enrichment; PCST: Prize-collecting Steiner tree; PPINs: Protein–protein interaction networks; PPP1R3C: Protein phosphatase 1 regulatory subunit 3C; S100A4: S100 calcium-binding protein A4; scRNA-seq: Single-cell RNA sequencing

Acknowledgements

The authors would like to thank Dr. Quin F. Wills for his role that sparked this collaboration. The authors thank Dr. Lyuba V. Bozhilova for helpful discussions. We thank the anonymous reviewers for their fruitful comments.

Authors' contributions

E.M.T performed experiments. E.M.T, T.M., F.K. performed numerical calculations. F.K., C.M.D, and G.R. developed the statistical methods. All authors designed the study and wrote the manuscript. The authors read and approved the final manuscript.

Funding

This the work was supported by a Novo Nordisk–University of Oxford pump priming award under the Strategic Alliance between the two partners. F.K. was funded through this award and is now funded by the EPSRC (funding references EP/R513295/1 and EP/N014529/1). G.R. was supported in part by the EPSRC grant New Approaches to Data Science: Application Driven Topological Data Analysis EP/R018472/1.

Availability of data and materials

The SCPPIN method is available as an R library under https://github.com/floklimm/scPPIN and as an online tool under https://floklimm.shinyapps.io/scPPIN-online/.
The data discussed in this publication have been deposited in the NCBI Gene Expression Omnibus [68] and are accessible through GEO Series accession number GSE133948 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE133948).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

E.M.T., T.M., and F.Z. are/were employees of *Novo Nordisk Ltd*.

Author details

[1]Department of Mathematics, Imperial College London, London SW7 2AZ, UK. [2]Mitochondrial Biology Unit, University of Cambridge, Cambridge CB2 0XY, UK. [3]Discovery Technology and Genomics, Novo Nordisk Research Centre Oxford, Oxford OX3 7FZ, UK. [4]Department of Statistics, University of Oxford, Oxford OX1 3LB, UK.

Received: 14 November 2019 Accepted: 12 October 2020
Published online: 02 November 2020

References

1. Parry CD, Patra J, Rehm J. Alcohol consumption and non-communicable diseases: epidemiology and policy implications. Addiction. 2011;106(10):1718–24.
2. MacParland SA, Liu JC, Ma X-Z, Innes BT, Bartczak AM, Gage BK, Manuel J, Khuu N, Echeverri J, Linares I, Gupta R. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat Commun. 2018;9(1):1–21.
3. Karalliedde J, Gnudi L. Diabetes mellitus, a complex and heterogeneous disease, and the role of insulin resistance as a determinant of diabetic kidney disease. Nephrol Dial Transplant. 2014;31(2):206–13.
4. Dünkler A, Rösler R, Kestler HA, Moreno-Andrés D, Johnsson N. SPLIFF: a single-cell method to map protein-protein interactions in time and space. In: Single Cell Protein Analysis. Springer; 2015. p. 151–68.
5. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50(8):1–14.
6. Saliba A-E, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. 2014;42(14):8845–60.
7. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6):8746.
8. Trapnell C. Defining cell types and states with single-cell genomics. Genome Res. 2015;25(10):1491–8.
9. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20.
10. Andrews TS, Hemberg M. Identifying cell populations with scRNASeq. Mol Asp Med. 2018;59:114–22.
11. Delaney C, Schnell A, Cammarata LV, Yao-Smith A, Regev A, Kuchroo VK, Singer M. Combinatorial prediction of marker panels from single-cell transcriptomic data. Mol Syst Biol. 2019;15(10):9005. https://doi.org/10.15252/msb.20199005.
12. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, Trapnell C. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566(7745):496–502.
13. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, Trombetta JJ. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013;498(7453):236–40.
14. Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, Beqiri M, Sproesser K, Brafford PA, Xiao M, Eggan E. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature. 2017;546(7658):431–5.
15. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13(4):599–604.
16. Rostom R, Svensson V, Teichmann SA, Kar G. Computational approaches for interpreting scRNA-seq data. FEBS Letters. 2017;591(15):2213–25.
17. Aibar S, González-Blas CB, Moerman T, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J, van den Oord J, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14(11):1083–6.
18. Chan TE, Stumpf MP, Babtie AC. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 2017;5(3):251–67.
19. Jeong H, Mason SP, Barabási A-L, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001;411(6833):41–2.
20. Estrada E. Virtual identification of essential proteins within the protein interaction network of yeast. Proteomics. 2006;6(1):35–40.
21. Ali W, Deane CM, Reinert G. Protein interaction networks and their statistical analysis. In: Stumpf MPH, Balding DJ, Girolami M, editors. Handbook of Statistical Systems Biology. Ltd Chichester, UK: John Wiley & Sons; 2011. p. 200–34.
22. Sevimoglu T, Arga KY. The role of protein interaction networks in systems biomedicine. Comput Struct Biotechnol J. 2014;11(18):22–7.
23. Guney E, Menche J, Vidal M, Barábasi A-L. Network-based in silico drug efficacy screening. Nat Commun. 2016;7(1):1-13.
24. Zitnik M, Leskovec J. Predicting multicellular function through multi-layer tissue networks. Bioinformatics. 2017;33(14):190–8.

25. Davis D, Yaveroğlu ÖN, Malod-Dognin N, Stojmirovic A, Pržulj N. Topology-function conservation in protein–protein interaction networks. Bioinformatics. 2015;31(10):1632–9.
26. Milenković T, Pržulj N. Uncovering biological network function via graphlet degree signatures. Cancer Inform. 2008;6:680.
27. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási A-L. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015;347(6224):1257601.
28. Lewis AC, Jones NS, Porter MA, Deane CM. The function of communities in protein interaction networks at multiple scales. BMC Syst Biol. 2010;4(1):100.
29. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.
30. Androulakis IP, Yang E, Almon RR. Analysis of time-series gene expression data: methods, challenges, and opportunities. Annu Rev Biomed Eng. 2007;9:205–28.
31. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, et al. Genetics of gene expression and its effect on disease. Nature. 2008;452(7186):423–8.
32. Reyna MA, Leiserson MD, Raphael BJ. Hierarchical hotnet: identifying hierarchies of altered subnetworks. Bioinformatics. 2018;34(17):972–80.
33. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. Bioinformatics. 2008;24(13):223–31.
34. Mitra K, Carvunis A-R, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet. 2013;14(10):719–32.
35. Wu C, Zhu J, Zhang X. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. BMC Bioinformatics. 2012;13(1):182.
36. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BIOGRID: A general repository for interaction datasets. Nucleic Acids Res. 2006;34(suppl_1):535–9.
37. Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(Nov):2579–605.
38. Mehta MB, Shewale SV, Sequeira RN, Millar JS, Hand NJ, Rader DJ. Hepatic protein phosphatase 1 regulatory subunit 3B (Ppp1r3b) promotes hepatic glycogen synthesis and thereby regulates fasting energy homeostasis. J Biol Chem. 2017;292(25):10444–54.
39. Komposch K, Sibilia M. EGFR signaling in liver diseases. Int J Mol Sci. 2016;17(1):30.
40. Zheng H, Cai A, Shu Q, Niu Y, Xu P, Li C, Lin L, Gao H. Tissue-specific metabolomics analysis identifies the liver as a major organ of metabolic disorders in amyloid precursor protein/presenilin 1 mice of alzheimer's disease. J Proteome Res. 2018;18(3):1218–27.
41. Zhang Z, Yao Z, Wang L, Ding H, Shao J, Chen A, Zhang F, Zheng S. Activation of ferritinophagy is required for the RNA-binding protein ELAVL1/HuR to regulate ferroptosis in hepatic stellate cells. Autophagy. 2018;14(12):2083–103.
42. Chen L, Li J, Zhang J, Dai C, Liu X, Wang J, Gao Z, Guo H, Wang R, Lu S, et al. S100A4 promotes liver fibrosis via activation of hepatic stellate cells. J Hepatol. 2015;62(1):156–64.
43. Adler M, Korem Kohanim Y, Tendler A, Mayo A, Alon U. Continuum of gene-expression profiles provides spatial division of labor within a differentiated cell type. Cell Syst. 2019;8(1):43–525. https://doi.org/10.1016/j.cels.2018.12.008.
44. Österreicher CH, Penz-Österreicher M, Grivennikov SI, Guma M, Koltsova EK, Datz C, Sasik R, Hardiman G, Karin M, Brenner DA. Fibroblast-specific protein 1 identifies an inflammatory subpopulation of macrophages in the liver. Proc Natl Acad Sci U S A. 2011;108(1):308–13.
45. Haass C, Selkoe DJ. Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid beta-peptide. Nat Rev Mol Cell Biol. 2007;8(2):101–12. https://doi.org/10.1038/nrm2101.
46. Wang J, Gu BJ, Masters CL, Wang Y-J. A systemic view of alzheimer disease—insights from amyloid-β metabolism beyond the brain. Nat Rev Neurol. 2017;13(10):612.
47. Sehgal N, Gupta A, Valli RK, Joshi SD, Mills JT, Hamel E, Khanna P, Jain SC, Thakur SS, Ravindranath V. Withania somnifera reverses alzheimer's disease pathology by enhancing low-density lipoprotein receptor-related protein in liver. Proc Natl Acad Sci U S A. 2012;109(9):3510–5.
48. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16(1):241.
49. Menon V. Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data. Brief Funct Genom. 2018;17(4):240–5.
50. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics. 2019;20(1):40.
51. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002;18(suppl_1):233–40.
52. Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. Nat Commun. 2018;9(1):1090.
53. Jalili M, Gebhardt T, Wolkenhauer O, Salehzadeh-Yazdi A. Unveiling network-based functional features through integration of gene expression into protein networks. Biochim Biophys Acta Mol basis Dis. 2018;1864(6):2349–59.
54. Amgalan B, Lee H. WMAXC: a weighted maximum clique method for identifying condition-specific sub-network. PloS ONE. 2014;9(8):104993.
55. Santoni D, Swiercz A, Żmieńko A, Kasprzak M, Blazewicz M, Bertolazzi P, Felici G. An integrated approach (cluster analysis integration method) to combine expression data and protein–protein interaction networks in agrigenomics: application on arabidopsis thaliana. Omics: J Integr Biol. 2014;18(2):155–65.
56. Luecken MD, Page MJ, Crosby AJ, Mason S, Reinert G, Deane CM. CommWalker: correctly evaluating modules in molecular networks in light of annotation bias. Bioinformatics. 2018;34(6):994–1000.
57. Zosin L, Khuller S. On directed steiner trees. In: Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms; 2002. p. 59–63, Society for Industrial and Applied Mathematics.
58. Bozhilova LV, Whitmore AV, Wray J, Reinert G, Deane CM. Measuring rank robustness in scored protein interaction networks. BMC Bioinformatics. 2019;20(1):446. https://doi.org/10.1186/s12859-019-3036-6.
59. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:1–12.
60. Ramachandran P, Dobie R, Wilson-Kanamori J, Dora E, Henderson B, Luu N, Portman J, Matchett K, Brice M, Marwick J, et al. Resolving the fibrotic niche of human liver cirrhosis at single-cell level. Nature. 2019;575(7783):512–8.
61. Aizarani N, Saviano A, Mailly L, Durand S, Herman JS, Pessaux P, Baumert TF, Grün D, et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. Nature. 2019;572(7768):199–204.
62. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502. https://doi.org/10.1038/nbt.3192.
63. Beguerisse-Díaz M, Vangelov B, Barahona M. Finding role communities in directed networks using role-based similarity, Markov stability and the relaxed minimum spanning tree. In: 2013 IEEE Global Conference on Signal and Information Processing; 2013. p. 937–40, IEEE.
64. Fischetti M, Leitner M, Ljubić I, Luipersbeck M, Monaci M, Resch M, Salvagnin D, Sinnl M. Thinning out Steiner trees: a node-based model for uniform edge costs. Math Program Comput. 2017;9(2):203–29.
65. Leitner M, Ljubić I, Luipersbeck M, Sinnl M. A dual ascent-based branch-and-bound framework for the prize-collecting Steiner tree and related problems. INFORMS J Comput. 2018;30(2):402–20.
66. Alexa A, Rahnenfuhrer J. topGO: enrichment analysis for gene ontology. R package version. 2010;2(0):2010.
67. Fisher RA. Statistical methods for research workers. In: Kotz S, Johnson NL, editors. Breakthroughs in Statistics. Springer; 1992. p. 66–70.
68. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.

Publisher's Note