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

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

[1]Department of Statistics, University of Oxford, Oxford OX1 3LB, United Kingdom
[2]Discovery Technology and Genomics, Novo Nordisk Research Centre Oxford, Oxford OX3 7FZ, United Kingdom

July 10, 2019

Abstract

Recent advances in single-cell RNA sequencing (scRNA-seq) have allowed researchers to explore transcriptional function at a cellular level. In this study, we present scPPIN, a method for integrating single-cell RNA sequencing data with protein–protein interaction networks to detect 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 a case study, we investigate RNA-sequencing data from human liver spheroids but the techniques described here are applicable to other organisms and tissues. scPPIN allows us to expand the output of differential expressed genes analysis with information from protein interactions. We find that different transcriptional states have different subnetworks of the PPIN significantly enriched which represent biological pathways. In these pathways, scPPIN also identifies proteins that are not differentially expressed but of crucial biological function (e.g., as receptors) and therefore reveals biology beyond a standard differentially expressed gene analysis.

*Corresponding author. Email: f.klimm@gmail.com

1 Introduction

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 includes 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 to understand the involved biology [3]. At a single-cell level, however, large-scale protein interaction data is not available [4]. In this study, we develop scPPIN a method for the integration of single-cell RNA-sequencing data with complementary protein–protein interaction networks (PPINs). The scPPIN analysis 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 single-cell RNA sequencing (scRNA-seq) techniques because they allow researchers to study and characterise tissues at a single-cell resolution [5, 6, 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]. The analysis of differentially expressed genes (DEGs) between these cell clusters can reveal different cell types [11], diseased cells [12], and cells that resist drug treatment [13]. Due to technological advances the quality and availability of scRNA-seq data increased dramatically in the last decade [14]. This makes the development of computational approaches for interpreting scRNA-seq data an active field of research [15] of which one research direction is the identification of gene regulatory networks in scRNA-seq data (e.g., SCENIC [16], PIDC [17]).

These approaches do not make systematic use of available data on protein–protein interaction, which one may represented as PPINs [18]. One may use PPINs to, for example, identify essential proteins [19, 20] and to predict disease associations [21, 22] or biological functions [23, 24, 25]. For this, researchers use 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. One calls these groups of proteins with common biological functions *modules* [26, 27].

It is understood that gene-expression is context-specific and thus varies between tissues [28], changes over time [29], and differs between healthy and diseased states [30]. It follows therefore that different parts of a PPIN are active under different conditions [31]. Analysing PPINs in an integrated way, together with bulk gene-expression data, provides such biological context, helps to reveal context-specific active functional modules [32, 33], and can identify proteins associated with disease [34].

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 organisms for which PPINs are available, which is the case for a broad range of organisms [35].

We may use scPPIN 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 has diverse cell types with different cellular metabolic processes. This makes the application of our method particularly intriguing, 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 the liver metabolism that could not be detected from the scRNA-seq data alone. Some of them have been shown to be important in the liver of other organisms and for others this study is the first indicator of important 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.

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. We therefore anticipate that this technique might also reveal novel insights for other organisms and tissue types.

2 Results

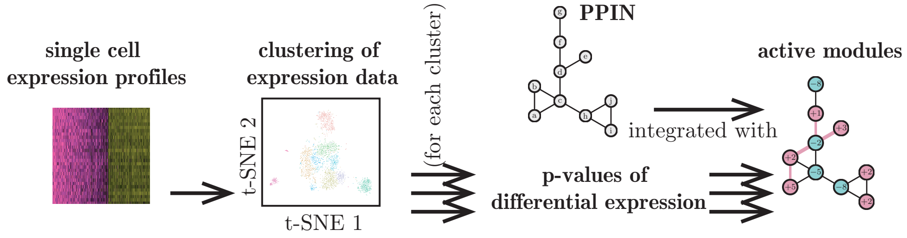
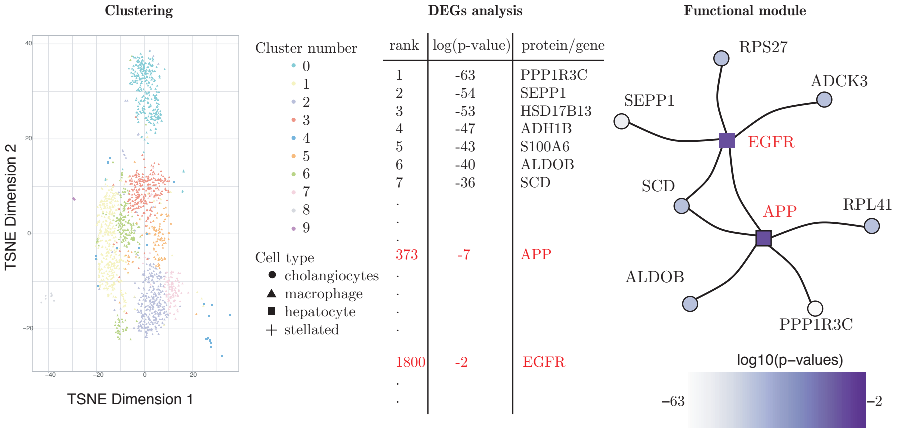


Figure 1: Our method consists of the following steps. (1) clustering of scRNA-seq data (e.g., with Seurat [9]). For each cluster, we (2) compute p-values of differential expression and use them to (3) estimate node scores by using an approach presented in [32]. (4) We combine these node scores with a PPIN to construct node-weighted PPINs for each cluster. (5) We compute functional modules as maximum-weight connected subgraphs.

In this paper, we present scPPIN, a method that allows the detection of functional modules in different cell clusters. The method involves multiple analysis steps (see Fig. 1 for an overview and the Method 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 with a Wilcoxon rank sum test. Fourth, for each gene in every cluster, we compute additive scores from these p-values (see [32] and Supplementary Note 1). Fifth, for each cluster, we map these gene scores to their corresponding proteins in a PPIN, which we constructed from publicly available data from BIOGRID [35]. Lastly, we identify functional modules as maximum-weight connected subgraphs in these node-weighted networks.



rank	log(p-value)	protein/gene
1	-63	PPP1R3C
2	-54	SEPP1
3	-53	HSD17B13
4	-47	ADH1B
5	-43	S100A6
6	-40	ALDOB
7	-36	SCD
.		
.		
.		
373	-7	APP
.		
.		
.		
.		
1800	-2	EGFR
.		
.		

Figure 2: (Left) Clustering the scRNA-seq data reveals ten clusters of which six are hepatocyte cells. (Middle) For each cluster, we may 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 (disks) and proteins that are not strongly differentially expressed (squares). Colour indicates p-value from low (white) to high (purple).

In order to demonstrate SCPPIN, we investigate newly measured scRNA-seq data of liver hepatocytes (see Method Section 8 for a description of the experimental setup and preprocessing steps). Using a standard modularity-maximisation algorithm, we obtain ten cell clusters of which six consist of hepatocytes (see Fig. 2), which make up a majority of the liver tissue. Hepatocytes are known to show a functional diversity, which includes the carbohydrate metabolism [2]. We then focus on hepatocytes because it allows us to study the heterogeneity of cellular function in this single cell type.

Then, we identify DEGs in each of the hepatocyte clusters. In Fig. 2, 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 with 10^{-63} the smallest of all p-values. It therefore could serve as a potential biomarker and is a known regulator of liver glycogen metabolism [36]. While such a DEG analysis reveals important genes in certain cell states it is not straightforward to identify the crucial biological pathways. Here, we demonstrate that integrating p-values from a DEGs analysis with PPIN information can reveal a more comprehensive picture of the biological processes. In 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 [37]. 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 [38].

These findings demonstrate that scPPIN can help to automate the further investigation of results from a DEGs 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.

2.1 Influence of the False Discovery Rate

Thus far, we demonstrated that scPPIN can reveal functional modules inside a PPIN and associate them with cells of a certain transcriptional state. Now we explore that there is not necessarily one functional module for a given cell state but functional modules of different size that scPPIN may identify.

There is only one free parameter in scPPIN, the false discovery rate (FDR). 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 before.

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 the optimal subnetworks for cluster H6 as a function of the FDR. As expected, the M(FDR) is non-decreasing. For

[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.

$\mathrm{FDR} < 10^{-26}$, 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 $\mathrm{FDR} = 10^{-25}$, for example, we detect the subnetwork of size $M = 9$ (shown in Fig. 2).

For $\mathrm{FDR} < 10^{-22}$, we detect an even larger functional module, which partially overlaps with the one identified for $\mathrm{FDR} < 10^{-23}$, 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 [39]. 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 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 [40]. Overall, the number of proteins we identify additionally with our method is moderately increasing with the FDR. In Supplementary Note 7, we show these $M(\mathrm{FDR})$ curves for all six hepotacyte clusters.

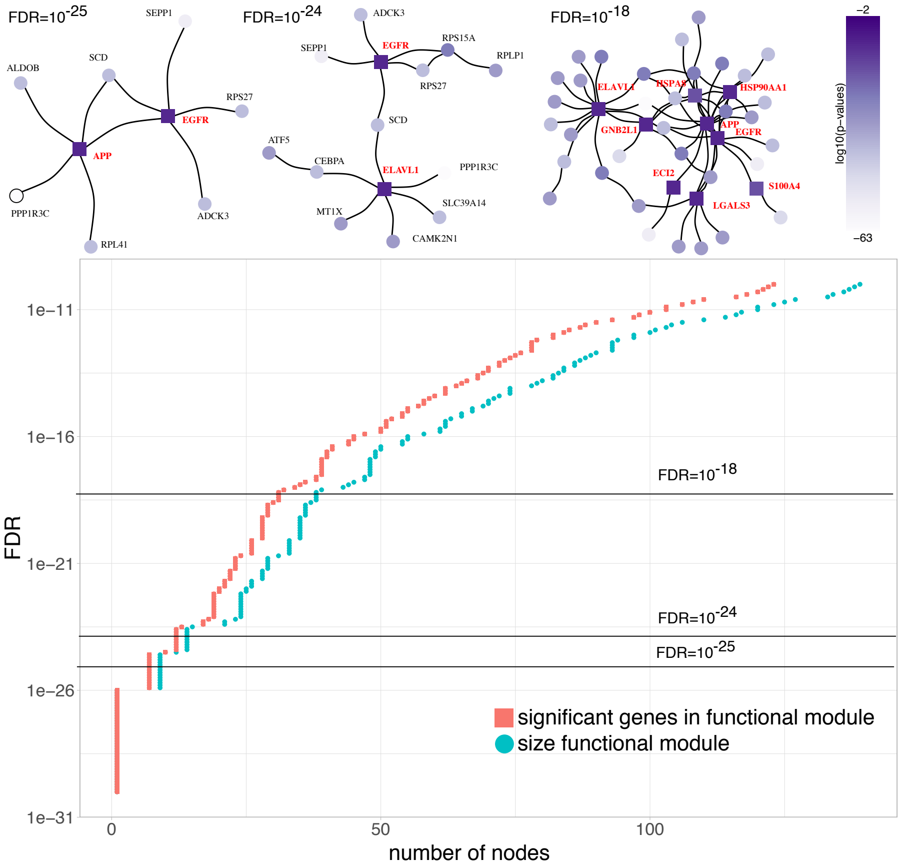


Figure 3: (Upper Panel) Network plots of the detected modules for three choices of the FDR (10^{-25}, 10^{-24}, and 10^{-18}). Nodes' colour indicates p-value from low (white) to high (purple). We indicate proteins that would not have been discovered from the gene-expression data alone as squares and 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.

2.2 Functional modules for different clusters

Previously, we investigated the influence of the FDR on the detected modules for a single cell cluster. As we obtained six hepatocyte clusters (see Fig. 2), we can compute DEGs and 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 each of the six clusters. The detected cluster differ in size with the largest consisting of 52 nodes (cluster H2) and the smallest consisting only of a single node (Cluster H6). This 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 an important role in different biological contexts.

To systematically access these biological contexts, we perform an 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 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* [41, 2].

The analysis of different cell states in the scRNA-seq with scPPIN indicates that indeed 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.

3 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. The presented method scPPIN builds on advances of DEG analysis, which are standard tools for the interpretation of scRNA-seq data. As

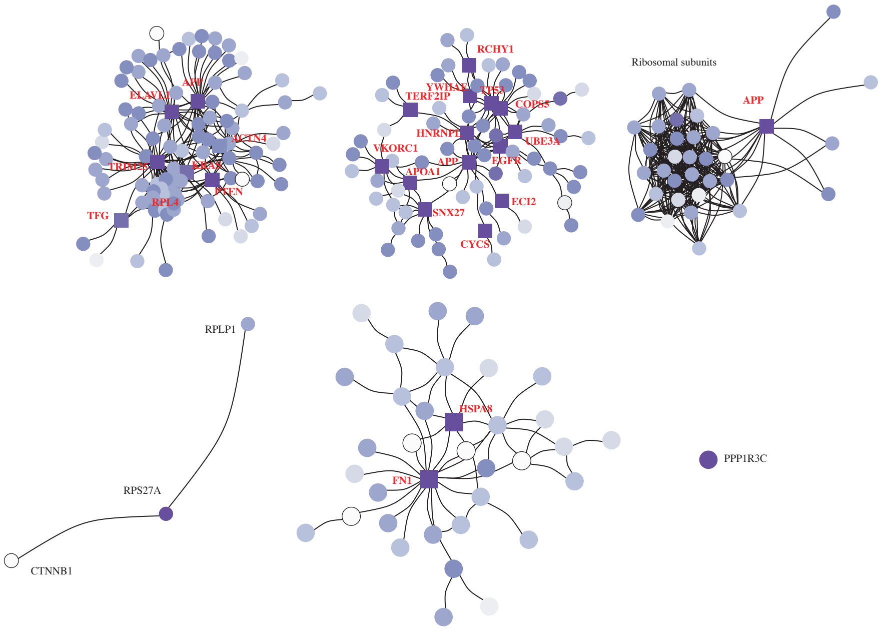


Figure 4: Detected modules for all six hepatocyte clusters for FDR = 10^{27}. We find that the detected modules vary strongly in size with the smallest consisting of a single protein and the largest consisting of 51 proteins. Colour indicates p-value of associated gene from low (white) to high (purple). We show nodes as squares if the could 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 Supplementary Material for illustration with all protein names.

a case study, we investigated data from healthy human livers. We find that the six identified cell clusters have different subnetworks of the PPIN with 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 important 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, important drivers of cell fate such as P53 and KRAS, but also proteins of so-far unknown function such as TERF2IP and TFG.

In a more general setting, scPPIN can be used to systematically analyse

cluster	module size M	top enriched GO terms
H1	51	translational initiation nuclear-transcribed mRNA catabolic process SRP-dependent cotranslational protein targeting to membrane
H2	52	response to stress defense response platelet deganulation
H3	27	SRP-dependent cotranslational protein targeting to membrane nuclear-transcribed mRNA catabolic process translational initiation
H4	3	*none*
H5	30	extracellular structure organization regulated exocytosis alcohol metabolic process
H6	1	*none*

Table 1: For each of the six clusters we give the three most enriched GO terms.

DEG 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 [40]. Although previously identified in a population of macrophages [40, 42], we see expression of S100A4 in this clusters of hepatocytes. This indicates, that a subpopulation of hepatocytes promotes fibrogenesis in paracrine. We also identified the amyloid precursor protein (APP) and interaction partners active in multiple hepatocytes 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) [43]. 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 [44, 45] should be taken into consideration in order to try try to find a successful treatment of it.

Despite their success, scRNA-seq techniques have methodological limitations (e.g., zero-inflation [46]). 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.

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 our method can easily be integrated to current analytical workflows for single cell RNA-seq analysis.

4 Acknowledgements

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 reference EP/R513295/1). The authors would like to thank Dr. Quin F. Wills for his role that sparked this collaboration. The authors thank Lyuba V. Bozhilova for fruitful discussions.

5 Author 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.

6 Competing interests

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

7 Code and data availability

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 NCBI's Gene Expression Omnibus [47] and are accessible through GEO Series accession number GSE133948 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE133948`).

8 Methods

8.1 Protein–Protein Interaction Network

We construct a PPIN from the publicly available BioGrid database [35], 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 [48] (e.g., considering confidence in an interaction [49]), we consider here exclusively undirected networks without edge weights.

8.2 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 prepare 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 [50].

8.3 Preprocessing

We analyse the scRNA-seq data with Seurat R package v2.3.4 [51]. As a preprocessing step, we align the data with a canonical correlation analysis [51] 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 references 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 2 % of a cluster to avoid comparing sparsely expressed genes.

8.4 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 [32]. 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 se 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 $W : V \to \mathbb{R}$. We construct these node-weighted networks from a PPIN and gene-expression information. The former is in 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 [35].

We delete all nodes from the PPIN for which no gene-expression data is available. We present an alternative approach that can incorporate proteins with missing expression data in the Supplementary Note 4. We assign each node a score

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

which is a function of the p-value x and we vary the *significance threshold* τ to tune the *false discovery rate* (FDR). We estimating 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.

8.5 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 [32]. Steiner trees are generalisations of spanning trees [52] and 'prize-collecting' indicates that the nodes have weights. To find a PCST, we use the dual ascent-based branch-and-bound framework DAPCSTP [53, 54]. For all calculations in this paper the algorithm identified an optimal solution in less than 10 s. For details see Supplementary Note 2.

8.6 Gene Ontology Enrichment

We use TOPGO in version 3.8 for the gene ontology enrichment (GO-enrichment) analysis. [55]. We use Fisher's exact test to identify enriched GO terms [56]. All reported GO terms are significant with p-value 0.01 and we use a Benjamini–Hochberg procedure to counteract the multiple-comparison problem.

References

[1] Charles D Parry, Jayadeep Patra, and Jürgen Rehm. Alcohol consumption and non-communicable diseases: epidemiology and policy implications. *Addiction*, 106(10):1718–1724, 2011.

[2] Sonya A MacParland, Jeff C Liu, Xue-Zhong Ma, Brendan T Innes, Agata M Bartczak, Blair K Gage, Justin Manuel, Nicholas Khuu, Juan Echeverri, Ivan Linares, et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nature Communications*, 9(1):4383, 2018.

[3] Janaka Karalliedde and Luigi Gnudi. Diabetes mellitus, a complex and heterogeneous disease, and the role of insulin resistance as a determinant of diabetic kidney disease. *Nephrology Dialysis Transplantation*, 31(2):206–213, 2014.

[4] Alexander Dünkler, Reinhild Rösler, Hans A Kestler, Daniel Moreno-Andrés, and Nils Johnsson. SPLIFF: a single-cell method to map protein-protein interactions in time and space. In *Single Cell Protein Analysis*, pages 151–168. Springer, 2015.

[5] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, 50(8):96, 2018.

[6] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research*, 42(14):8845–8860, 2014.

[7] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 2019.

[8] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome Research*, 25(10):1491–1498, 2015.

[9] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411, 2018.

[10] Tallulah S Andrews and Martin Hemberg. Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, 59:114–122, 2018.

[11] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496, 2019.

[12] Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236, 2013.

[13] Sydney M Shaffer, Margaret C Dunagin, Stefan R Torborg, Eduardo A Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, Katrin Sproesser, Patricia A Brafford, Min Xiao, et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, 546(7658):431, 2017.

[14] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature Protocols*, 13(4):599, 2018.

[15] Raghd Rostom, Valentine Svensson, Sarah A Teichmann, and Gozde Kar. Computational approaches for interpreting scRNA-seq data. *FEBS Letters*, 591(15):2213–2225, 2017.

[16] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, et al. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083, 2017.

[17] Thalia E Chan, Michael PH Stumpf, and Ann C Babtie. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems*, 5(3):251–267, 2017.

[18] Waqar Ali, Charlotte M. Deane, and Gesine Reinert. Protein interaction networks and their statistical analysis. In Michael P. H. Stumpf, David J. Balding, and Mark Girolami, editors, *Handbook of Statistical Systems Biology*, pages 200–234. John Wiley & Sons, Ltd Chichester, UK, 2011.

[19] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41, 2001.

[20] Ernesto Estrada. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, 6(1):35–40, 2006.

[21] Tuba Sevimoglu and Kazim Yalcin Arga. The role of protein interaction networks in systems biomedicine. *Computational and Structural Biotechnology Journal*, 11(18):22–27, 2014.

[22] Emre Guney, Jörg Menche, Marc Vidal, and Albert-László Barábasi. Network-based in silico drug efficacy screening. *Nature Communications*, 7:10331, 2016.

[23] Marinka Zitnik and Jure Leskovec. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 33(14):i190–i198, 2017.

[24] Darren Davis, Ömer Nebil Yaveroğlu, Noël Malod-Dognin, Aleksandar Stojmirovic, and Nataša Pržulj. Topology-function conservation in protein–protein interaction networks. *Bioinformatics*, 31(10):1632–1639, 2015.

[25] Tijana Milenković and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, 6:CIN–S680, 2008.

[26] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601, 2015.

[27] Anna CF Lewis, Nick S Jones, Mason A Porter, and Charlotte M Deane. The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology*, 4(1):100, 2010.

[28] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[29] Ioannis P. Androulakis, Eric Yang, and Richard R. Almon. Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annual Review of Biomedical Engineering*, 9:205–228, 2007.

[30] Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, Agnar Helgason, G Bragi Walters, Steinunn Gunnarsdottir, et al. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423, 2008.

[31] Matthew A Reyna, Mark DM Leiserson, and Benjamin J Raphael. Hierarchical hotnet: identifying hierarchies of altered subnetworks. *Bioinformatics*, 34(17):i972–i980, 2018.

[32] Marcus T Dittrich, Gunnar W Klau, Andreas Rosenwald, Thomas Dandekar, and Tobias Müller. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–i231, 2008.

[33] Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719, 2013.

[34] Chao Wu, Jun Zhu, and Xuegong Zhang. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. *BMC Bioinformatics*, 13(1):182, 2012.

[35] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: A general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl_1):D535–D539, 2006.

[36] Minal B Mehta, Swapnil V Shewale, Raymond N Sequeira, John S Millar, Nicholas J Hand, and Daniel J Rader. Hepatic protein phosphatase 1 regulatory subunit 3B (Ppp1r3b) promotes hepatic glycogen synthesis and thereby regulates fasting energy homeostasis. *Journal of Biological Chemistry*, 292(25):10444–10454, 2017.

[37] Karin Komposch and Maria Sibilia. EGFR signaling in liver diseases. *International Journal of Molecular Sciences*, 17(1):30, 2016.

[38] Hong Zheng, Aimin Cai, Qi Shu, Yan Niu, Pengtao Xu, Chen Li, Li Lin, and Hongchang Gao. 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. *Journal of Proteome Research*, 18(3):1218–1227, 2018.

[39] Zili Zhang, Zhen Yao, Ling Wang, Hai Ding, Jiangjuan Shao, Anping Chen, Feng Zhang, and Shizhong Zheng. Activation of ferritinophagy is required for the RNA-binding protein ELAVL1/HuR to regulate ferroptosis in hepatic stellate cells. *Autophagy*, 14(12):2083–2103, 2018.

[40] Lin Chen, Jie Li, Jinhua Zhang, Chengliang Dai, Xiaoman Liu, Jun Wang, Zhitao Gao, Hongyan Guo, Rui Wang, Shichun Lu, et al. S100a4 promotes liver fibrosis via activation of hepatic stellate cells. *Journal of Hepatology*, 62(1):156–164, 2015.

[41] Miri Adler, Yael Korem Kohanim, Avichai Tendler, Avi Mayo, and Uri Alon. Continuum of Gene-Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type. *Cell Systems*, 8(1):43–52.e5, 2019.

[42] Christoph H Österreicher, Melitta Penz-Österreicher, Sergei I Grivennikov, Monica Guma, Ekaterina K Koltsova, Christian Datz, Roman Sasik, Gary Hardiman, Michael Karin, and David A Brenner. Fibroblast-specific protein 1 identifies an inflammatory subpopulation of macrophages in the liver. *Proceedings of the National Academy of Sciences of the United States of America*, 108(1):308–313, 2011.

[43] Christian Haass and Dennis J. Selkoe. Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid beta-peptide. *Nature Reviews. Molecular Cell Biology*, 8(2):101–12, feb 2007.

[44] Jun Wang, Ben J Gu, Colin L Masters, and Yan-Jiang Wang. A systemic view of alzheimer disease—insights from amyloid-β metabolism beyond the brain. *Nature Reviews Neurology*, 13(10):612, 2017.

[45] Neha Sehgal, Alok Gupta, Rupanagudi Khader Valli, Shanker Datt Joshi, Jessica T Mills, Edith Hamel, Pankaj Khanna, Subhash Chand Jain, Suman S Thakur, and Vijayalakshmi Ravindranath. Withania somnifera reverses alzheimer's disease pathology by enhancing low-density lipoprotein receptor-related protein in liver. *Proceedings of the National Academy of Sciences of the United States of America*, 109(9):3510–3515, 2012.

[46] Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):241, 2015.

[47] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[48] Leonid Zosin and Samir Khuller. On directed steiner trees. In *Proceedings of the thirteenth annual ACM-SIAM Symposium on Discrete Algorithms*, pages 59–63. Society for Industrial and Applied Mathematics, 2002.

[49] Lyuba V Bozhilova, Alan V Whitmore, Jonny Wray, Gesine Reinert, and Charlotte M Deane. Measuring rank robustness in scored protein interaction networks. *BioRxiv*, page 502302, 2018.

[50] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017.

[51] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015.

[52] Mariano Beguerisse-Díaz, Borislav Vangelov, and Mauricio Barahona. 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*, pages 937–940. IEEE, 2013.

[53] Matteo Fischetti, Markus Leitner, Ivana Ljubić, Martin Luipersbeck, Michele Monaci, Max Resch, Domenico Salvagnin, and Markus Sinnl. Thinning out steiner trees: a node-based model for uniform edge costs. *Mathematical Programming Computation*, 9(2):203–229, 2017.

[54] Markus Leitner, Ivana Ljubić, Martin Luipersbeck, and Markus Sinnl. A dual ascent-based branch-and-bound framework for the prize-collecting steiner tree and related problems. *INFORMS Journal on Computing*, 30(2):402–420, 2018.

[55] Adrian Alexa and Jorg Rahnenfuhrer. topGO: enrichment analysis for gene ontology. *R package version*, 2(0):2010, 2010.

[56] Ronald Aylmer Fisher. Statistical methods for research workers. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics*, pages 66–70. Springer, 1992.