BMC Bioinformatics

RESEARCH

Open Access

SUBATOMIC: a SUbgraph BAsed mulTi-OMIcs clustering framework to analyze integrated multi-edge networks

Jens Uwe Loers[1,2,3] and Vanessa Vermeirssen[1,2,3*]

*Correspondence:
vanessa.vermeirssen@ugent.be

[1] Lab for Computational Biology, Integromics and Gene Regulation (CBIGR), Cancer Research Institute Ghent (CRIG), Ghent, Belgium
[2] Department of Biomedical Molecular Biology, Ghent University, Ghent, Belgium
[3] Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

Abstract

Background: Representing the complex interplay between different types of biomolecules across different omics layers in multi-omics networks bears great potential to gain a deep mechanistic understanding of gene regulation and disease. However, multi-omics networks easily grow into giant hairball structures that hamper biological interpretation. Module detection methods can decompose these networks into smaller interpretable modules. However, these methods are not adapted to deal with multi-omics data nor consider topological features. When deriving very large modules or ignoring the broader network context, interpretability remains limited. To address these issues, we developed a SUbgraph BAsed mulTi-OMIcs Clustering framework (SUBATOMIC), which infers small and interpretable modules with a specific topology while keeping track of connections to other modules and regulators.

Results: SUBATOMIC groups specific molecular interactions in composite network subgraphs of two and three nodes and clusters them into topological modules. These are functionally annotated, visualized and overlaid with expression profiles to go from static to dynamic modules. To preserve the larger network context, SUBATOMIC investigates statistically the connections in between modules as well as between modules and regulators such as miRNAs and transcription factors. We applied SUBATOMIC to analyze a composite *Homo sapiens* network containing transcription factor-target gene, miRNA-target gene, protein–protein, homologous and co-functional interactions from different databases. We derived and annotated 5586 modules with diverse topological, functional and regulatory properties. We created novel functional hypotheses for unannotated genes. Furthermore, we integrated modules with condition specific expression data to study the influence of hypoxia in three cancer cell lines. We developed two prioritization strategies to identify the most relevant modules in specific biological contexts: one considering GO term enrichments and one calculating an activity score reflecting the degree of differential expression. Both strategies yielded modules specifically reacting to low oxygen levels.

Conclusions: We developed the SUBATOMIC framework that generates interpretable modules from integrated multi-omics networks and applied it to hypoxia in cancer. SUBATOMIC can infer and contextualize modules, explore condition or disease specific modules, identify regulators and functionally related modules, and derive novel gene

functions for uncharacterized genes. The software is available at https://github.com/CBIGR/SUBATOMIC.

Keywords: Composite subgraphs, Multi-edge networks, Network analysis, Multi-omics, Modules, Gene regulation, Topology, Hypoxia, Gene function prediction, Gene regulatory networks

Introduction

Eukaryotic gene regulation involves a complex interplay between different types of biomolecules to safeguard correct gene expression in space and time. Transcription factors (TFs) bind to specific sequences in the DNA, such as promoter and enhancer regions, to activate or repress gene expression [1–3]. Co-factors bind to TFs and interact with the transcriptional machinery [2, 4]. At the epigenetic level, the accessibility of chromatin is the degree to which molecules such as TFs, RNA-polymerases or chromatin organizing proteins are able to establish a physical contact with the underlying DNA via promotor, enhancer and insulator regions [5]. The accessibility is dynamic and changes in response to external stimuli as well as developmental signals lead to notable differences in expression between various cell types [5, 6]. Several classes of non-coding RNA (ncRNA) also have an impact on gene regulation. MicroRNAs (miRNA) suppress protein translation or induce messenger RNA (mRNA) degradation, mostly by binding to the 3'-UTR of target messenger RNAs [7, 8]. Moreover, they are regulated by DNA methylation, histone modifications, and more than 140 forms of RNA modifications [9]. In turn, miRNAs themselves target epigenetic-associated enzymes such as DNA methyltransferases, ten-eleven translocation genes, and histone deacetylases [9, 10]. Long non-coding RNAs (lncRNAs) act as signal molecules that mediate transcription of downstream genes, as decoy molecules to repress biological processes and pathways by binding TFs and blocking their regulatory activity [11, 12], or compete with mRNAs for miRNA binding [13, 14]. Additionally, several genes, especially regulatory proteins such as TFs and miRNAs, have undergone duplication events during their evolution, leading to gene redundancy and/or the acquisition of novel biological functions over time [15, 16].

High-throughput technologies, like RNA-seq, ChIP-seq and mass spectrometry yield an enormous amount of high-quality data in the context of gene regulation. These data are available in databases that continuously grow by adding and integrating novel data types and datasets. One example is the resource 'Discriminant Regulon Expression Analysis' (DORothEA) [17, 18]. It contains signed TF-target interactions based on literature-curated resources, ChIP-seq peaks, gene expression-based inference and TF binding sites information [17, 18]. DORotheEA is embedded in the OmniPath database, which includes many additional interaction types such as miRNA-target interactions, lncRNA-target interactions, ligand-receptor binding, and protein–protein interactions [19, 20]. HumanNet is a human gene network resource that captures co-functional and physically binding interactions: the co-functional network (COF) includes co-essentiality and co-expression interactions, while the protein–protein interaction network contains literature-curated and high-throughput interactions of physically binding proteins [21]. Many more databases exist for diverse types of molecular interactions and their size and number continuously grow.

To understand gene regulation in depth, we need to comprehend how different molecular interactions together coordinate phenotype-specific gene expression. Indeed, several studies have shown that considering complementary molecular interactions increases our understanding of regulatory processes. Co-expressed genes and genes encoding physically interacting proteins are often regulated by the same set of TFs [22, 23]. Genes encoding TFs that control miRNA expression have a higher chance to be post-transcriptionally repressed by the miRNA [24]. Genes co-regulated by miRNAs show weaker functional links compared to TF-regulated genes [25]. These complex, diverse interactions between several biomolecules in gene regulation can be modeled at a systems level in gene regulatory networks (GRNs). GRNs map the molecular interactions between regulators, mainly TFs, and their target genes, based on relevant high-throughput data, with or without using computational inference [26, 27]. Integrated GRNs take into account different types of molecular interactions implicated in gene regulation [28]. Currently, proficient methods for integrating multi-omics data into these GRNs are still lacking, as well as methods for the analysis, and biological interpretation of intricate, integrated networks.

Biological networks are often hard to interpret as a whole. They possess a high number of nodes and edges merged into a giant 'hairball' structure that makes a meaningful visualization and their functional interpretation extremely challenging [29, 30]. Many approaches have been developed to tackle this problem. Their shared principle is to decompose these hairball structures into smaller interpretable subnetworks, often referred to as modules or communities. Methods can be co-expression based, topology-based, pan-sample based and multi-edge based including tools such as WGCNA, SimMod, ModulOmics, and LemonTree [31–41]. WGCNA clusters genes with high expression correlation and summarizes modules using their module eigengene [31, 32]. SimMod uses a mixed integer non-linear programming model to integrate WGCNA-based co-expression networks with physical and genetic interactions into multi-omics communities [33]. ModulOmics identifies de-novo cancer driver pathways and modules by integrating protein–protein interactions, mutual exclusivity of mutations and copy number variations (CNVs), transcriptional co-regulation, and co-expression [40]. Other methods integrated TF-target gene interactions and protein–protein interactions with a 'function-to-structure' based method by deriving modules based on genes with a shared GO annotation [41]. LemonTree infers co-expression modules in multiple runs, merges these in consensus modules, and finally connects these modules to regulators, such as TFs, miRNAs or CNVs, using multi-omics data [36]. While existing approaches strongly increased the interpretability of large multi-omics networks, some challenges remain. One common limitation is that the number of derived modules is usually very small and they contain a lot of genes. Most large modules correlate well with biological properties or phenotypes but lack detailed and causal interpretation. Moreover, modules are interpreted as completely separated entities and do not share any genes. However, when considering topology in multi-edge networks, genes can appear in different topological contexts and thus in different modules. On the other hand, given that many small and interpretable modules exist, keeping track of the inter-module relationships is crucial to not miss out on a broader network interpretation. Thus, a method that considers network topology, edge-causality, and condition-specific data (e.g. expression) while

producing small and interpretable modules in a specific network context can substantially complement the existing inference methods.

We previously proposed a data integration framework for the worm *Caenorhabditis elegans* and the plant *Arabidopsis thaliana* that groups specific molecular interactions in composite network subgraphs, clusters these next into biologically relevant, topological modules, highlights connections between modules and regulators, and finally overlays these modules with gene expression profiles to go from static to dynamic modules [28]. We learned that different molecular interactions interrelate in distinct topological modules with specific biological functions to generate a coordinated response in gene regulation. Here, we extended this data integration framework to SUbgraph BAsed mulTi-OMIcs Clustering (SUBATOMIC). SUBATOMIC infers composite subgraph-based modules from diverse interaction databases and gene expression profiles and analyzes them in an updated, generalized, and automated analysis framework. Upon dissecting the composite network into small interpretable modules, we keep track of interactions that connect regulators with modules as well as modules with each other in a superview to preserve their larger network context and facilitate biological interpretation. To make the static network modules dynamic and evaluate their role in specific conditions, we implemented a module activity score. The score can be used to rank modules with regard to their degree of dysregulation upon condition change. The method is applicable to any user-defined set of networks with overlapping nodes for any species of interest.

With its unique approach, SUBATOMIC addresses several gaps in network modularity and multi-omics data analysis. Hyper-edge clustering enables to select for specific topological features in the network, and at the same time to generate small and easily interpretable subnetworks. Moreover, it is possible to investigate all topological features in the ALL modules setting. While analyzing interactions in-between modules and between modules and regulators, users can keep track of other modules and/or regulators that might be involved in similar biological processes, hence embedding the modules in their global network context. The automated pipeline also aids in the functional annotation of modules and their visual exploration in Cytoscape. These properties of SUBATOMIC largely facilitate biological interpretation of multi-omics data. While many tools focus on specific interactions such as protein–protein interaction networks, our method can incorporate any type of directed and undirected interactions independent of the species. Moreover, SUBATOMIC can also explore dynamic networks upon integration of condition-specific data such as transcriptomics, and we provide several metrics to quantify the condition-specific activity of modules. Hence, it is a comprehensive and versatile network analysis tool to investigate specific biological questions using multi-omics data.

We applied SUBATOMIC to integrate six networks from *H. sapiens*, respectively, based on TF-target interactions, miRNA-mRNA interactions, protein–protein interactions, functional interactions, and homologous connections for proteins and miRNAs. The inferred modules allowed us to propose functional hypotheses for insufficiently annotated proteins. As proof of concept, we further contextualized the modules with expression data for cancer cell lines under hypoxic conditions. Hypoxia occurs when a cell or tissue is not sufficiently supplied with oxygen to maintain their homeostatic state

(Gaspar and Velloso, 2018; Hiraga, 2018). This state frequently appears in the tumor microenvironment leading to cellular responses that increase the risk of metastasis and reduce the success of treatment [42, 43]. We identified modules sensitive to hypoxia conditions using both activity and GO term based features. We showed that these responsive modules were highly connected in our superview analysis compared to a random control. We highlighted several examples and guidelines on how to use SUBATOMIC to gain biological insights. The SUBATOMIC pipeline is available on GitHub (https://github.com/CBIGR/SUBATOMIC).

Results

SUBATOMIC: a subgraph based multi-omics clustering framework

We developed SUBATOMIC, a SUbgraph BAsed mulTi-Omics Clustering framework to construct and analyze multi-edge networks (Fig. 1). SUBATOMIC takes networks composed of different interaction types as input. Interactions can be directed such as TF-target interactions and miRNA-target interactions or undirected such as protein–protein interactions. The networks need to have a partial overlapping node set to allow for integration over the different interaction types. Given the multi-edge networks, SUBATOMIC first uses the subgraph enumeration algorithm ISMAGS to decompose it

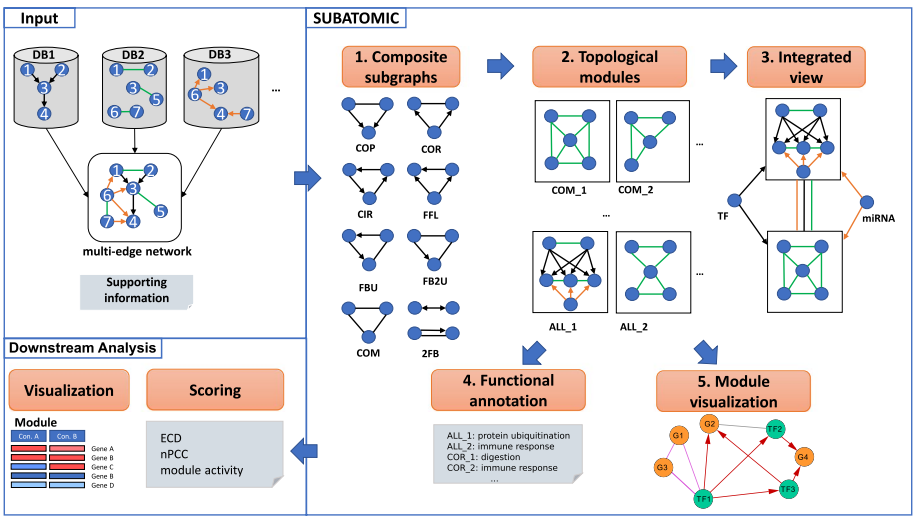

Fig. 1 Overview of the SUBATOMIC workflow. Input: SUBATOMIC takes as input a multi-edge network consisting of directed and/or undirected interactions of different interaction types. Supporting information might contain additional input files such as GO terms, gene annotations, and a list of subgraph definitions that can be used by ISMAGS to screen specifically for these subgraphs (Methods). SUBATOMIC: (1) The multi-edge network is then decomposed into composite subgraphs using ISMAGS for co-pointing (COP), co-regulated (COR), circular (CIR), feed forward loop (FFL), feedback undirected loop (FBU), feedback 2 undirected loop (FB2U), complex (COM) and 2 feedback (2FB) subgraphs. (2) Based on the subgraphs, SCHype generates topological modules for each subgraph type as well as for all subgraph types together (ALL). (3) Modules are connected to each other as well as to regulators to produce an integrated view (further called superview). (4) Modules are functionally annotated with GOATOOLS. (5) We generated files that can be imported into Cytoscape for module visualization. Networks analyzed in this way are considered static if they do not incorporate any condition specific information. Downstream Analysis: Here modules are integrated with condition-specific expression data. Several scores reflect the condition-specific activity of modules: the expression dynamicity score (ECD), the average Pearson correlation of expression values in a module (nPCC) and the module activity score

into a set of 3-node composite subgraphs [44]. Additionally, we incorporated an own script to find specific 2-node subgraphs. Subgraphs are classified according to the type and direction of edges they contain. By integrating directed and undirected edges in 2- and 3-node composite subgraphs, we discriminate eight different subgraph types [28]. In a co-pointing subgraph (COP), an undirected edge connects two regulators and together they regulate a target. The co-regulated subgraph (COR) contains one regulator controlling two interacting target genes. The feed forward loop (FFL) has a regulator that directly regulates a target gene and another regulator, which also controls the target gene. In the circular feedback subgraph (CIR), regulators act upon each other through feedback loops. The feedback-undirected subgraph (FBU) consists of two directed interactions in a cascade that are connected by an undirected interaction. The feedback 2 undirected subgraph (FB2U) combines two undirected interactions and one directed interaction. The complex subgraph (COM) contains only undirected edges. Finally, the two-node feedback subgraph (2FB) couples a directed edge with an undirected edge.

We followed the ISMAGS nomenclature in representing subgraphs by their specific interaction types and edge signs [44]. Each input network of a specific interaction type is assigned a specific letter: R for TF-gene, M for miRNA-mRNA, P for protein–protein, C for co-functional, and H for homologous interactions. Then each 3-node subgraph obtains a three-letter representation according to the specific interaction type of its edges. For example, a PPP subgraph contains three undirected edges from the protein–protein interaction network and is hence of the COM type. RRP contains two edges from a regulatory TF-gene network and one from the protein–protein interaction network and is hence of the COR type. Subsequently, all subgraphs are clustered by the hypergraph-based spectral clustering algorithm 'Spectral Clustering in Hypergraphs' (SCHype) [45]. SCHype optimizes the edge-to-node ratio on hyperedges that represent the 3-node and 2-node subgraphs during clustering. The resulting modules share common topological features and possess specific biological functions [28, 45]. SUBATOMIC uses SCHype to first generate clusters within each type of subgraph (COM, CIR, FFL, …). Additionally, all subgraphs together are clustered in a module type called 'ALL'. We further filter for subgraphs that contain between 5 and 50 genes similar to our previous approach [28]. Next, SUBATOMIC applies GOATOOLS to functionally annotate the modules based on Gene Ontology [46]. At this point, small and biologically interpretable modules have been obtained, but their network context is not yet considered. To address this, we arrange all modules in superview that connects modules with each other, and finds regulators connected to each module. SUBATOMIC also calculates an output that can be imported in Cytoscape for module network visualization [47]. As a postprocessing step, the topological modules are integrated with expression data to study dynamics of gene regulation over different experimental conditions. In this step several metrics can be calculated to further characterize and prioritize modules in specific conditions. More details on the pipeline can be found in the methods section.

Integrated human regulatory networks

Multi-omics data integration aids in the understanding of dysregulation in complex diseases. Our data integration framework SUBATOMIC not only makes use of multi-omics networks but also connects topological and functional information to leverage their

interpretation. In this study we aimed to construct and analyze multi-omics networks for *H. sapiens* with SUBATOMIC. Therefore, we integrated TF-target gene, miRNA-mRNA, homologous, protein–protein, and co-functional interactions from public resources and finally added expression data from cancer cell lines under hypoxic conditions. We provided a layout of how SUBATOMIC can be used to investigate perturbed gene regulation in a human disease context.

We included five different networks in our analysis that cover distinct interaction types that all influence gene regulation (see Table 1, Methods and Additional file 1). Two networks are directed and model regulatory relationships: the TF-target gene network (R) and the miRNA-mRNA network (M). Three networks are undirected: the homolog network (H), the protein–protein interaction network (P) and the co-functional network (C). The R network includes 53,232 TF-target gene interactions from OmniPath from three different OmniPath sub-databases: DoRothEA (levels A-C), TF-target (curation score > 1) and TF-miRNA interactions [19, 20]. The M network includes 11,085 miRNA-mRNA interactions from OmniPath. To include gene homology, we included 10,847 paralogous interactions between genes from the Ensemble archive and homologous miRNAs with identical seed sequences from miRbase [48, 49]. To include a layer of functional information, we chose two mutually exclusive networks from HumanNet v2 [21, 50]. The 6637 co-functional edges contain co-essentiality, co-expression, and protein domain profile association edges. The 24,773 physical protein–protein interaction network contains edges from high throughput assays such as yeast-two-hybrid and affinity purifications and from literature-curated protein–protein interactions.

More than half of the detected subgraphs are interaction type specific

ISMAGS detected a total of 787,347 3-node subgraphs (Fig. 2). Complex subgraphs were most abundant, covering 82.56% of all composite subgraphs, containing mostly CCC (413,938 – 48.77%) and RRR (85,435 – 10.07%), followed by PCC (80,035 – 9.43%). While the overall fraction of non-COM type subgraphs might be small, rare subtypes can reveal interesting mechanistic insights. Subgraphs shared between different interaction types are less often observed than subgraphs detected within one type of interaction. A total of 534,665 subgraphs contain at least two co-functional interactions, while 150,582 subgraphs contain at least one co-functional and one protein–protein interaction. Subgraphs with edges from the co-functional

Table 1 Overview of the different interaction types included in the multi-edge input network for *H. sapiens*

Interaction type	Letter	Directed	#Nodes	#Edges	#Regulators	#Targets	Database
Regulatory TF-gene interactions	R	Yes	15,014	53,232	526	14,488	OmniPath
Regulatory miRNA-mRNA interactions	M	Yes	4060	11,085	850	3210	OmniPath
Homologous genes	H	No	4862	10,847	–	–	Ensembl
Protein–protein interactions	P	No	10,950	24,773	–	–	HumanNet
Co-functional interactions	C	No	10,682	66,373	–	–	HumanNet

Overview of different interaction types included in the multi-edge input network for *H. sapiens*. The 'letter' column indicates the chosen letter to represent the interaction networks. Input networks were derived from three different databases and contained either directed or undirected edges. For directed edges, the number of regulators and target genes are shown

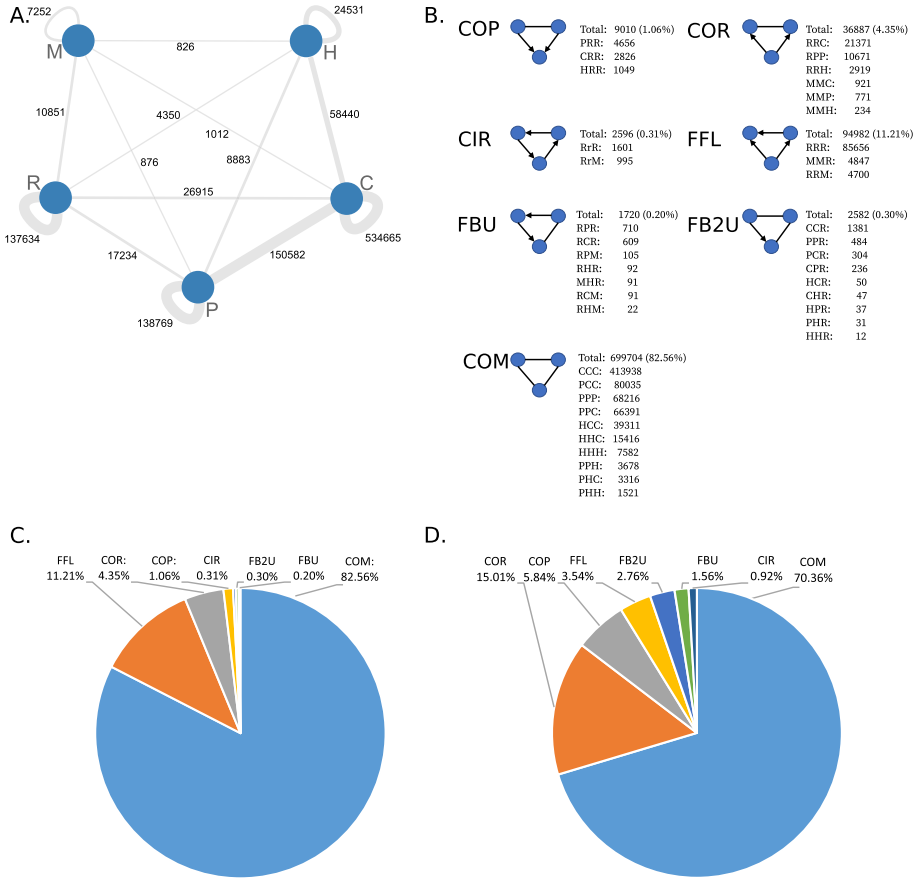

Fig. 2 **A**: Interconnection between different input networks (nodes) at the number of composite subgraphs. Each edge connecting two nodes represents how many subgraphs contain at least one edge from each network. Most subgraphs contain at least two edges from the same input networks. **B**: Overview of the counts and fractions of all detected subgraphs. **C**: Overview of different subgraph types detected in the human multi-edge network. **D**: Overview of different module types detected in the human multi-edge network. We omitted the ALL networks in **C** and **D** for visualization purposes since the ALL modules contain overlapping subgraphs with other module types

network and the regulatory TF-gene interaction network are counted with 26,915 occurrences. The smallest amount of subgraphs with two edges of the same type is 7252 and comes from miRNA-mRNA interactions, which also possess the lowest amount of shared subgraphs with the homologous network (826). While the interaction types and quantities used in Defoort et al. for *A. thaliana* and *C. elegans* were slightly different, we obtained comparable results with complex subgraphs being most abundant and subgraphs containing only protein–protein and homologous interactions having the highest subgraph counts [28].

Next, SUBATOMIC uses SCHype to cluster the composite subgraphs into 7 module types. For our *H. sapiens* composite network, this resulted in a total number of 5586 modules (2762 ALL, 1987 COM, 424 COR, 165 COP, 100 FFL, 78 FB2U, 44 FBU, 26 CIR). While COM, COR, COP, FFL, FB2U, FBU and CIR are generated on mutually exclusive subgraph types, ALL contains a joint clustering of all types together and allows to find interactions between different topological modules. Hence, ALL modules were most abundant, followed by COM and COR modules. CIR modules are the least present.

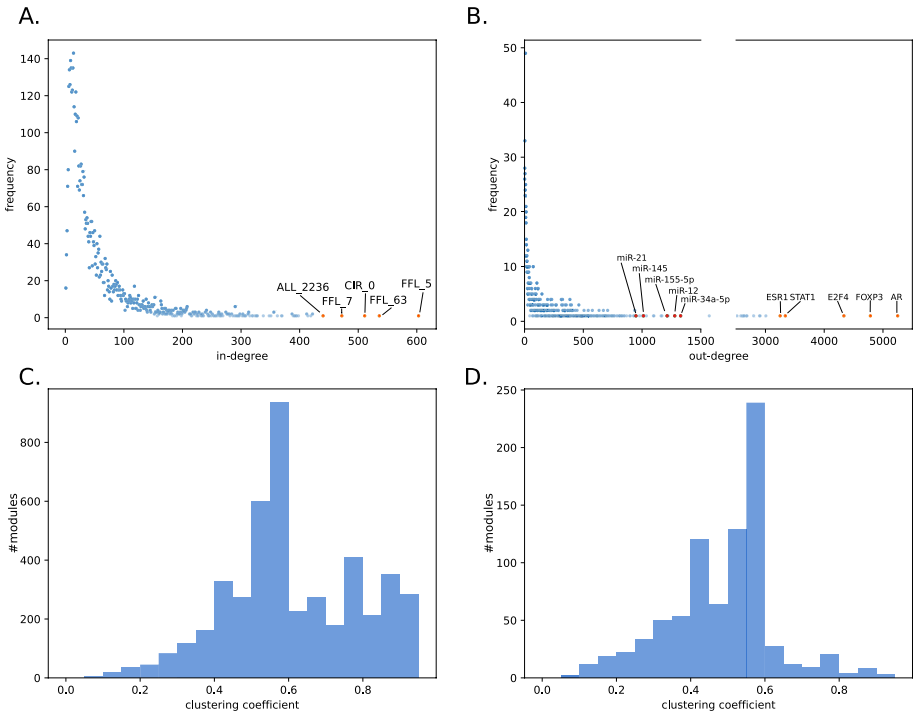

Fig. 3 Specific network properties of the regulator-modules network. **A** In-degree and **B** Out-degree distribution of the interactions between regulators and modules. The top 5 modules, transcription factors and miRNAs with the highest degree are shown with labels. In **B** we disrupted the axis between 1500 and 2500. **C** The clustering coefficient distribution of all modules. The mode was located between 0.55 and 0.6. The majority of modules had a higher connectivity than the mode. **D** The clustering coefficient distribution of all modules excluding COM and ALL modules. The majority of modules had a lower connectivity than the mode. The visualization of the clustering coefficient for each independent module is given in Additional file 6

With regard to the context of gene regulation, COR, COP, CIR and FFL modules are the most interesting ones, containing directed regulatory interactions.

Most modules are densely connected and regulated by several transcription factors and miRNAs

While calculating the modules, we kept track of their larger network context in the so-called 'superview' analysis. This includes the interactions of modules with each other as well as with regulators such as miRNAs and TFs. We first analyzed the specificity of regulators by looking at how many modules they target. This gives insights into whether regulators can be considered master regulators or specific regulators. In our analysis, we included a total of 526 TFs and 850 miRNAs.

On average, a TF targeted 6% and a miRNA targeted 2% of all modules. A total of 25 TFs regulated 5 or less modules, while five TFs were specific for only one module. Among the miRNA regulators, 90 regulated 5 or less modules while 19 only targeted one module. An average module is targeted by 33 TFs and 18 miRNAs. We then draw the degree distribution for in-degree and out-degree of the regulator-module interactions (Fig. 3A and B). Most modules and regulators had a low degree. The five highest degree modules are of FFL, CIR and ALL type. For regulators, a high degree can be interpreted as an indication for master regulators. Furthermore, we draw the distribution of the

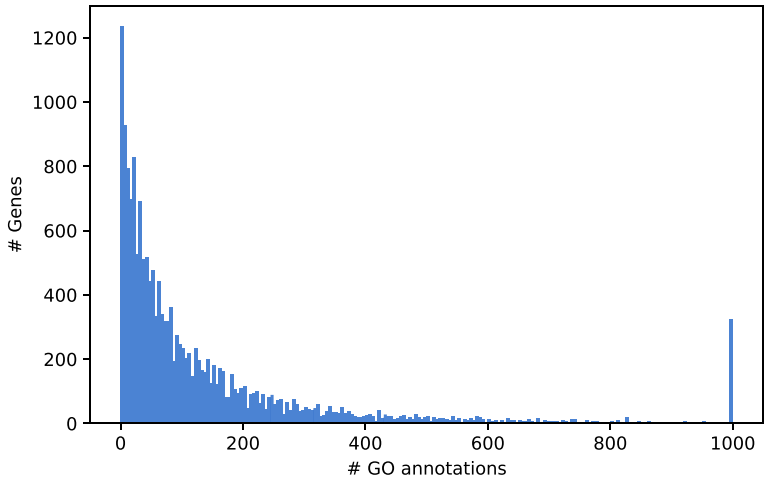

Fig. 4 Distribution of GO annotations for all genes included in the composite *H. sapiens* network. While the majority of genes is well-annotated, 345, 245, 181, 210, 256 and 167 genes have zero, one, two three, four and five GO terms, respectively. A total of 319 genes have more than 1000 GO-annotations and were stacked at the last bin. The histogram was generated with a bin size of 5

clustering coefficient for all modules (Fig. 3C). We detected a mode of 0.55–0.6 in the clustering coefficient distribution, with the majority of modules showing a higher connectivity than the mode. Since the modules were strongly dominated by COM modules, excluding COM and ALL modules, now the majority of modules had a lower connectivity than the mode (Fig. 3D). As the co-functional and protein–protein interactions networks are highly interconnected, their clustering coefficient is also very high.

Functional enrichment analysis of modules reveals unknown gene functions

Upon functional GO ontology enrichment analysis, 3805 modules have a list of enriched GO terms. We can capitalize on these functional annotations by generating hypotheses on gene functions for genes that are not well-characterized based on the guilt-by-association or guild-by-rewiring principle [51, 52]. Although most genes in our *H. sapiens* network are well-annotated and connected to many GO terms, 1404 genes have less than five GO terms and 345 have no GO term at all (Fig. 4). Limiting ourselves to protein-coding genes, we found 53 genes in the modules that were annotated with merely two or less GO terms, further referenced as 'weakly characterized genes', 25 of which had no GO term at all. (See supplement for a complete list of un-annotated genes and their module context). Several of these genes were present in well-annotated modules and we could predict their biological function in relation to the GO annotation and structure of the module. We selected five genes without current GO annotation for further analysis: the proline and serine-rich protein 1 (PROSER1), the cyclic nucleotide-binding domain-containing protein 1 (CNBD1), the leucine, glutamate and lysine rich 1 gene (LEKR1), the RIIa domain-containing protein 1 (RIIAD1), and the glutamate-rich protein 6B (ERICH6B) (Fig. 5). The gene PROSER1 (ENSG00000120685) appears in the modules ALL_1135, ALL_1880 and ALL_2888. The latter consists of a protein–protein complex of eight genes. The top 5 enriched terms are MLL3/4 complex, histone methyltransferase activity (H3-K4 specific), Set1C/COMPASS complex, histone H3-K4 methylation

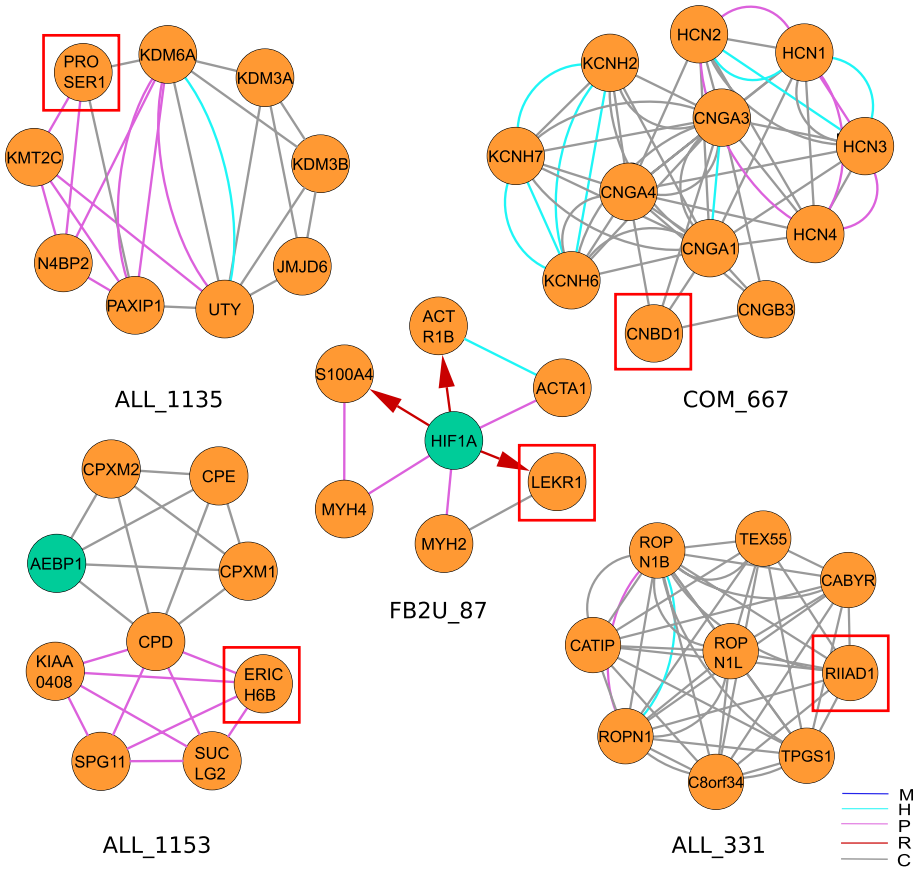

Fig. 5 Gene function prediction for poorly functionally characterized genes based on their module context. Based on the guilt-by-association principle, we created hypotheses on the biological function of these genes based on their module context. We derived function predictions for PROSER1, CNBD1, LEKR1, ERICH6B and RIIAD1

and histone methyltransferase complex. The histone methyltransferase complex GO term is shared by 6 out of 8 module genes. PROSER1 is directly connected to the histone methyltransferase KMT2 as well as to the PAXIP1 known to be involved in histone H3-K4 methylation [53]. Thus, we hypothesized that PROSER1 is part of a histone methylation complex. After our analysis, a recently published work confirmed the PROSER1s involvement in the regulation of various chromatin-associated proteins [54]. Next, we analyzed CNBD1 (ENSG00000176571). This gene appeared in the modules ALL_654 and COM_667. The top 5 enriched terms in COM_667 are HCN channel complex, intracellular cAMP-activated cation channel activity, intracellular cyclic nucleotide activated cation channel complex, intracellular cyclic nucleotide activated cation channel activity and cyclic nucleotide-gated ion channel activity. It is connected via co-functional edges to CNGB1, CNGA1 CNGA3, and CNGA4. All four are subunits of the cyclic nucleotide gated channel (CNGA) and appear in all of the top 5 enriched terms except for the HCN channel complex. Thus, we hypothesized that CNBD1 is also part of a cation channel complex and possesses cation channel activity. The LEKR1 (ENSG00000197980) gene appears in the modules ALL_3385 and FB2U_87. While ALL_3385 had no significant enrichment, the top 5 enriched terms in FBU_87 were cellular components muscle

myosin complex, dynactin complex, myosin filament, myosin II complex and sarcomere. It is connected with MYH2, which is also annotated with all significant terms except the dynactin complex. While we cannot derive a specific function for LEKR1, we can hypothesize that it plays a role in the myosin complex. The ERICH6B (ENSG00000165837) gene appears in ALL_1153 and COM_1252. The top 5 enriched terms are the molecular functions metallocarboxypeptidase activity, carboxypeptidase activity, metalloexopeptidase activity, exopeptidase activity, as well as the biological process protein processing. It has protein–protein interactions with four proteins, of which the carboxypeptidase D (CPD) is part of all enriched GO terms in this module, and the succinate–CoA ligase SUCLG2 is part of the cellular amide metabolic process. Although the function remains rather broad, we can hypothesize that this gene is involved in the amide metabolic process. Finally, the RIIAD1 (ENSG00000178796) gene appears in the modules ALL_331 and COM_380. The top 5 enriched terms in ALL_331 are sperm capacitation, sperm motility, flagellated sperm motility, cilium movement involved in cell motility and cilium or flagellum-dependent cell motility. Five out of nine genes are annotated with a cellular component of the motile cilium. We can hypothesize that RIIAD1 is involved in sperm motility. In fact, a recent paper mentioned RIIAD1 as co-expressed with the a-kinase anchor protein 3AKAP3, a gene whose knockdown was shown to induce infertility in male mice [55, 56]. A visualization of the top 30 enriched terms and an overview of GO terms is available in Additional files 2 and 6.

Dynamic modules are associated with hypoxia in three cancer cell lines

Hypoxia can lead to a variety of different responses in which cells can develop tolerance to severe tissue damage and might in turn promote aggressive cancer phenotypes [57, 58]. Such damaged cells can be embedded in the tumor microenvironment and influence treatment effectiveness [58, 59]. To contextualize the results obtained with SUBATOMIC, we chose a study that investigated the influence of cycling and chronic hypoxia on gene expression in melanoma (WM793B), ovarian cancer (SK-OV-3), and prostate cancer cell lines (PC-3) [58]. Chronic hypoxia is characterized by a permanent oxygen depletion and was modeled in the study with a permanent ambient oxygen concentration of 1%. In cyclic hypoxia, the availability of oxygen varied between 1 and 21% oxygen with a switch at six different time points. A permanent oxygen concentration of 21% was used as a control condition. We chose two main prioritization schemes to select modules potentially involved in hypoxia: one based on enriched GO terms and the other based on expression data.

We first filtered all modules based on at least one enrichment for a GO term containing the 'hypoxia' key word (Additional file 3). Hence, we identified 78 modules, further referenced as the 'hypoxia GO set'. We then investigated how tightly connected this set of modules was in the superview analysis (Table 2). We used the number of interactions between modules in our selected set and compared it against a background of 1000 randomly selected sets of modules of the same set size. This allowed us to see whether modules in the 'hypoxia GO set' are more connected than expected by chance. We calculated the upper boundary of a 95% confidence interval on the mean and standard deviation of the random set and a fold change comparing the interactions from the 'hypoxia GO set' with this upper boundary. We observed that modules of the 'hypoxia GO set' have

Table 2 Modules in the 'hypoxia GO set' show higher interconnectivity than expected by chance

Edge type	Hypoxia GO set	Random mean	Random std	Confidence down	Confidence up	Fold Change
M-type	490	22	24	21	24	21
R-type	27,476	1425	780	1377	1474	19
H-type	428	17	85	12	23	19
C-type	2772	202	216	189	216	13
P-type	872	52	70	48	57	15
Total	32,038	1720	880	1665	1774	18

We calculated mean, standard deviation and a 95% confidence interval based on a distribution of 1000 randomly sampled sets of modules. Confidence up = upper boundary of the confidence interval, Confidence down = lower boundary of the confidence interval, Fold change = comparison of the number of superview interactions in the 'hypoxia GO set' with the upper boundary of the random distribution. We observed a high fold change for all edge types in the 'hypoxia GO set'

about 18 times more edges between pairs of modules inside this set than random. Especially regulatory edges such as TF-target interactions and miRNA-mRNA interactions between these modules are respectively 21 and 19 times more often observed than expected by chance. This revealed a strong connection between those modules, indicating complex regulatory mechanisms behind hypoxia. Next, we investigated whether the 'hypoxia GO set' compared to other modules was enriched for significantly differentially expressed (DE) genes in chronic hypoxia with a minimum fold change greater than two. We did not apply this to cyclic hypoxia because the number of DE genes was quite low. A hypergeometric test was used comparing DE genes appearing inside the 'hypoxia GO set' with DE genes appearing in all other modules. We found a significant overexpression of DE genes in all the cell lines: WB793B (fold change 16.99, *p*-value 3.54E-38), PC3 (fold change 9.58, *p*-value 1.47E-36) and SK-OV-3 (fold change 7.74, *p*-value 1.43E-30). These results highlight that with a combination of superview analysis and functional annotation, we can already filter for condition-specific modules even without expression data integration.

Subsequently, we took a closer look at three selected modules in the 'hypoxia GO set' (Fig. 6). For example, the module COM_256 resembles a co-functional protein complex that has 68 enriched GO terms including the most enriched hydroxylysine metabolic process, peptidyl-proline 4-dioxygenase activity as well as L-ascorbic acid binding (see supplement for full list of enriched GO terms). Five out of twelve genes are involved in response to hypoxia as well as response to decreased oxygen levels. The module is mostly regulated by three TFs: androgen receptor (AR), hypoxia inducible factor 1 subunit alpha (HIF1A) and endothelial PAS domain protein 1 (EPAS1), also known as hypoxia-inducible factor 2-alpha. HIF1A and EPAS1 are known to facilitate cellular adaptation to hypoxia and regulate many hypoxia-related genes in a variety of tissues [60–64]. Also, AR was shown to act as ligand-dependent TFs that confer to resistance against AR-targeted cancer therapies under hypoxic conditions [65, 66]. Many module genes were differentially expressed, namely the procollagen-lysine,2-oxoglutarate 5-dioxygenase 1 and 2 (PLOD1, PLOD2), the prolyl hydroxylase (EGLN3), as well as prolyl 4-hydroxylase subunits alpha 1 and 2 (P4HA1) and (P4HA2). PLOD1 and PLOD2 were shown to be involved in hypoxia-induced metastasis and glioblastoma tumor progression [63, 67]. EGLN3, which was upregulated in the chronic state in the SK-OV-3 cell line, catalyzes oxygen-dependent hydroxylation of the hypoxia induced factor (HIF)

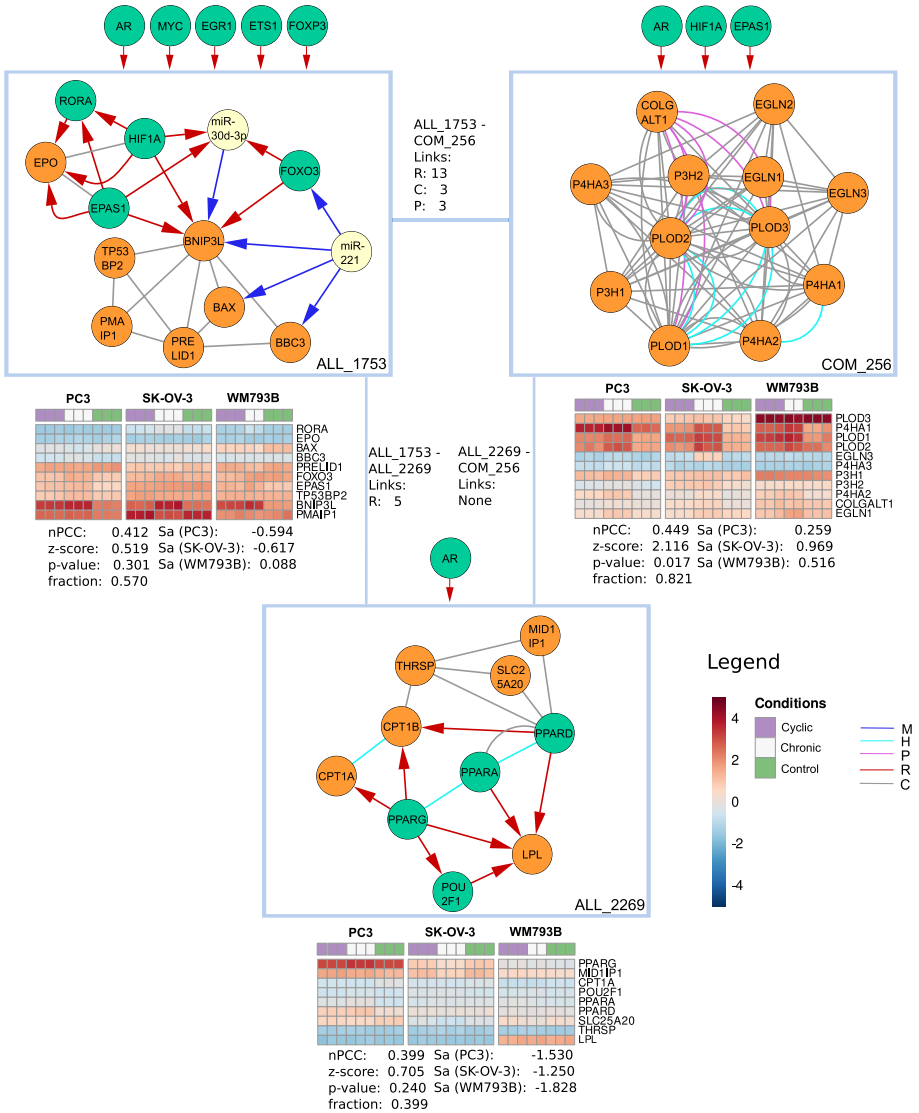

Fig. 6 Visualization of three superview-connected modules from the 'hypoxia GO set'. 'Links' indicate how many superview interactions exist per type between two modules. Regulators targeting at least 5 genes that were not present in a specific module are shown outside the boxes. nPCC indicates the correlation of genes within a module in all hypoxia samples together with a *z*-score derived by comparing this value with random modules of the same size as well as an *p*-value and the fraction of edges in a module for which expression data was available for both genes. The activity score Sa was displayed for each of the three cell lines PC3, SK-OV-3 and WM793B. Expression is shown for all module genes for which expression values were available based on the hypoxia dataset (GEO: GSE53012)

[68, 69]. Other genes such as the prolyl 3-hydroxylases, P4HA1 and P4HA2 were known to hydroxylate the 564-proline residue in the α-subunit of HIF [70]. Hence, we concluded that COM_256 is strongly involved in the reaction to hypoxia and shows a coherent but slightly different expression across the three cell lines.

We then investigated the ALL_1753 module. The module contained 177 significant GO term enrichments including cellular response to hypoxia and cellular response to decreased oxygen levels for six genes. It is centered around the BCL2 Interacting Protein 3 Like gene (BNIP3L), which is differentially expressed and targeted

by HIF1A and EPAS1. The module contained three interesting feed forward loops, where HIF1A, EPAS and the forkhead box O3a FOXO3 target BNIPL3 and miR-30d-3p that in turn also regulates BNIP3L [73]. Moreover, miR-30d-3p is known to be involved in hypoxia and directly regulate AR [71]. FOXO3 is activated in response to hypoxic stress [72]. Another DE gene in this module is the retinoic acid receptor-related orphan receptor (RORA), regulated by HIF1A and EPAS1. RORA is known to be induced by HIF1A and it plays a role in the nuclear accumulation of HIF1A [74]. Finally, miR-221 regulated FOXO3, BNIPL3, the bcl-2-binding component 3 (BBC3) and the apoptosis regulator BAX and exerted cytoprotective effects in hypoxia-reoxygenation injury [75]. Hence, we concluded a strong involvement of ALL_1753 in response to hypoxia and that the hypoxia response of BNIPL3 might be driven by the involvement of three regulatory feed forward loops.

At last, we investigated the ALL_2269 module. The module was enriched for 85 GO terms including the most enriched terms carnitine shuttle and carnitine O-palmitoyltransferase activity for two and three genes as well as positive regulation of fatty acid metabolic process that involved half of the module genes. The response to hypoxia and decreased oxygen levels was enriched due to the presence of three module genes. The module is centered by three homologous peroxisome proliferator-activated receptors, PPARG, PPARA and PPARD. Especially PPARG was shown to be activated under hypoxic conditions in correlation with HIF1A in lung cancer and hepatocellular carcinoma [76, 77]. It regulated the differentially expressed gene carnitine palmitoyltransferase 1A (CPT1A) and its homolog CPT1B, shown to regulate prostate cancer growth under hypoxic conditions [78]. While this module does not show strong dysregulation in the three analyzed cancer types, it contained genes and interactions highly relevant in reaction to hypoxia, as demonstrated in other studies.

All three modules were connected by many edges in the superview and shared a similar set of regulators. We demonstrated that the modules found by our GO term based approach are highly relevant in the hypoxia context, which is supported by the increased amount of hypoxia-specific DE genes. In the modules we identified regulatory structures such as feed forward loops involving interactions from complementary omics layers that help to explain and interpret the observed expression and allow to generate mechanistic hypotheses for hypoxia-induced mechanisms.

In a second prioritization approach, we wanted to use the dynamic response of genes towards a stimulus or condition as selection criteria (Additional file 4). We implemented a 'module activity' S_a approach that can capture the response of modules to a changing condition, for example based on differential expression data between two conditions [79] (see also methods). To find a set of highly hypoxia-related modules, we filtered for modules with a positive activity score S_a in all three cell lines. This resulted in a set of 52 modules that we further refer to as the 'hypoxia activity set'. Next, we analyzed the superview connections within the ALL modules in the same way as for the 'hypoxia GO set' (Table 3). We observed a 28 times higher connectivity between the activity modules as compared to random sets of the same size. While the number of interactions for M-type and R-type interactions was similar to the 'hypoxia GO set', the undirected interaction types H, C and P were much more enriched with

Table 3 'Modules in the 'hypoxia activity set' show higher interconnectivity than expected by chance

Edge type	Hypoxia	Random mean	Random std	Confidence down	Confidence up	Fold change
M-type	134	7	9	6	7	18
R-type	8494	458	330	437	478	18
H-type	624	5	25	4	7	90
C-type	5008	66	112	59	73	69
P-type	1676	17	35	15	19	88
Total	15,936	552	370	529	575	28

We calculated mean, standard deviation and a 95% confidence interval based on a distribution of 1000 randomly sampled sets of modules. Confidence up = upper boundary of the confidence interval, Confidence down = lower boundary of the confidence interval, Fold change = comparison of the number of superview interactions in the 'hypoxia GO set' with the upper boundary of the random distribution. We observed a high fold change for all edge types in the 'hypoxia GO set'

a 90, 69 and 88 times higher number of connections, respectively. Furthermore, the 'hypoxia GO set', and the 'hypoxia activity set' have 12 modules in common.

We then inspected three modules identified in the activity hypoxia set with regard to their relevance for hypoxia (Fig. 7).

The COR_347 module is dominated by HIF1A and MYC, each regulating 6 targets. The module contained 159 significant GO term enrichments including cellular response to hypoxia and cellular response to decreased oxygen levels for four and three genes, respectively. Besides HIF1A and the MYC proto-oncogene (MYC), the module is regulated by the *fos* proto-oncogene (FOS) with six regulatory interactions. MYC is one of the master regulators targeting 46% of all modules. It plays an important role in the development of cancer and regulates members of the hypoxia inducing factor protein family [80]. The hexokinase 2 gene (HK2) is regulated by MYC and HIF1A and shows dysregulation in the expression data. It was recently shown to be an important target of HIF1A in an oxygen-reduced environment [81].

ALL_2896 mostly contained COR and FFL subgraphs dominated by HIF1A. It was enriched for 11 GO annotations including procollagen-proline 4-dioxygenase activity and peptidyl-proline 4-dioxygenase activity. HIF1A regulated the differentially expressed gene P4HA1, which also occurred in COM_256 in the 'hypoxia GO set' in a co-regulatory manner with the prolyl 4-hydroxylase beta polypeptide gene (P4HB). Another regulator in this module was the signal transducer and activator of transcription 1 gene (STAT1). It regulated the sperm associated antigen 4 (SPAG4) together with HIF1A. While SPAG4 was not differentially expressed with a fold change greater than two, we still observed a consistent reduction in expression across all three cell lines.

ALL_2093 is enriched for 331 GO terms. The most enriched terms included regulation of metanephric cap mesenchymal cell proliferation and negative regulation of leukocyte adhesion to arterial endothelial cells, however the most enriched 15 terms only contained one gene. The response to hypoxia as well as to decreased oxygen levels included six genes. The module mostly consisted of COR and FFL subgraphs. Many of its genes connected to hypoxia already appeared in the modules described above, such as AR, MYC, HK2, STAT1 and PPARA. Another interesting gene is the krüppel-like factor 4 (KLF4). It was differentially expressed together with its target 6-Phosphofructo-2-Kinase/Fructose-2,6-Biphosphatase 3 (PFKFB3), which in turn is also regulated by

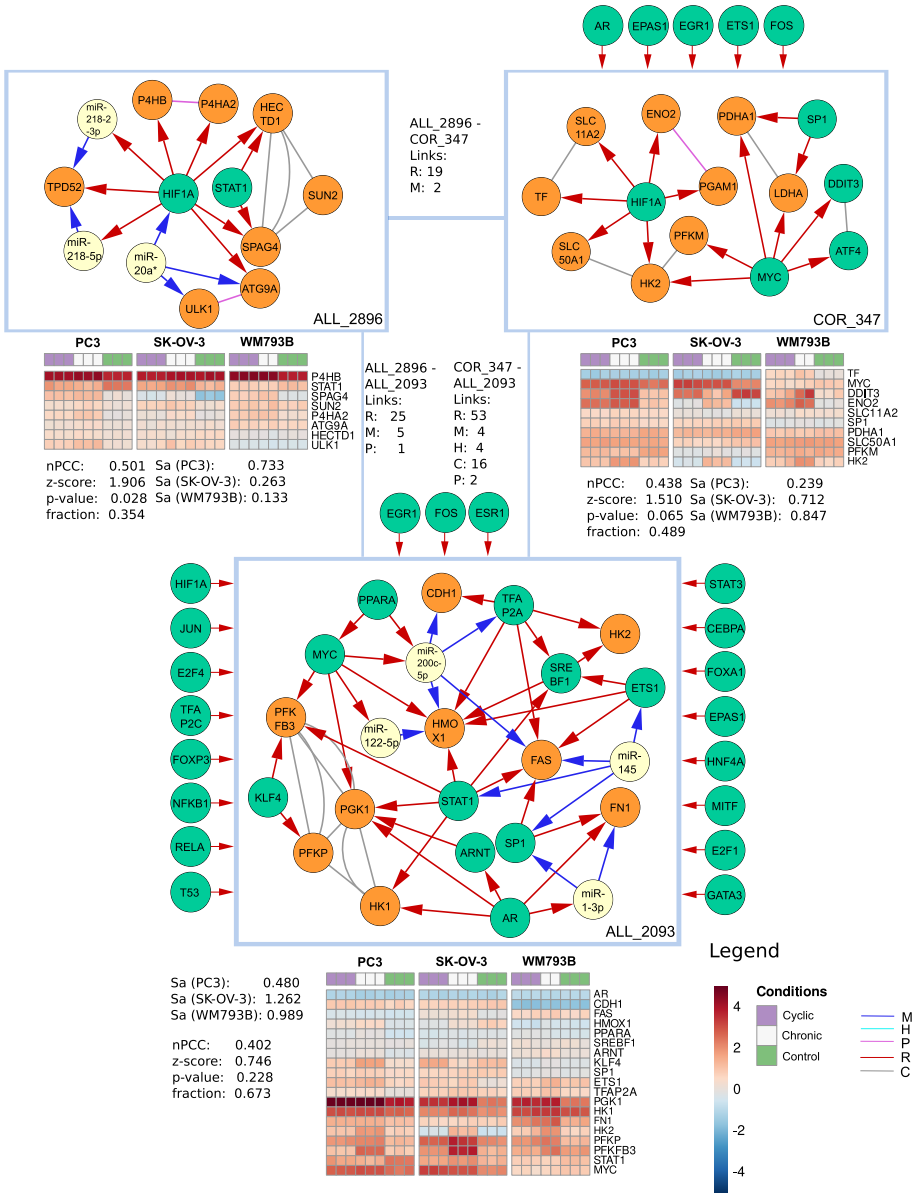

Fig. 7 Visualization of three superview connected modules from the 'hypoxia activity set'. 'Links' indicate how many superview interactions exist per type between two modules. Regulators targeting at least 5 genes that were not clustered in a specific module are shown at its top. nPCC indicates the correlation of genes within a module in all hypoxia samples together with a z-score derived by comparing this value with random modules of the same size as well as an *p*-value and the fraction of edges in a module for which expression data was available for both genes. The activity score Sa was displayed for each of the three cell lines PC3, SK-OV-3 and WM793B. Expression is shown for all module genes for which expression values were available

STAT and MYC. KLF4 was shown to be involved in Hypoxia-induced vascular smooth muscle cell migration [82]. However, while PFKFB3 is differentially expressed in PC-3 and SK-OV-3 cell lines, KLF4 is only significantly overexpressed in PC-3. Besides regulation via STAT1 and MYC, the PFKFB3 promotor also contains HIF1A binding sites and might not be dependent on KLF4 as an activator to be overexpressed under hypoxic conditions [83]. Furthermore, fibronectin 1 (FN1) is overexpressed and regulated by

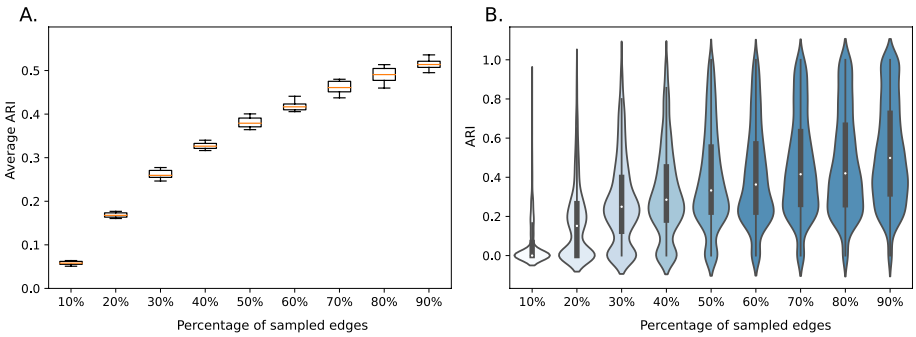

Fig. 8 **A**: Boxplots representing the average Adjusted Mutual Information score (AMI) for each subsampled set. Each box summarizes the results of 10 independent SUBATOMIC runs per 10%, 20%, …, 90% of interactions sampled from the full network. The orange line inside each box represents the average AMI for the 10 repetitions. **B**: Violin plot representing the distribution of one selected run. The width of each violin indicates how many values were present for a certain AMI value

two FFLs including AR, SP1 and miR-1-3p. The latter was show to be downregulated under hypoxic conditions in mouse lung tissues [84]. The specificity protein 1 (SP1) was described to also be regulated by HIF1A directly and is required for hypoxia-induced transcription of other downstream genes [85].

Module stability analysis

To assess the stability of the modules with regard to missing edges, we designed a re-sampling approach [86]. Given the full network as a base, we sampled 90%, 80%, …, 10% of edges without replacement from the set of interactions and re-run the SUBATOMIC pipeline on each of the reduced sets. We then removed the non-sampled edges from the modules of the full network to create a set of ground truth modules and calculated the module overlap based on the reduced networks. This allowed us to quantify the stability of the modules with regard to missing edges. To account for the randomness of the sampling process, we repeated each sampling ten times. We observed that with this amount of re-sampling, we get a stable output with regard to variance. We utilized three different metrics to compare the clusters to their ground truth: Jaccard Index (JI), Adjusted Rand Index (ARI) and the adjusted mutual information (AMI) [87–89]. For each module in our down-sampled set, we retrieved the maximum score in comparison to the ground truth and visualized the mean within a boxplot for each set (Fig. 8A, Additional file 8). In this way, when modules were merged or split up compared to the full network, only the module with the highest score was considered. While the scores were relatively stable between runs keeping 90%, 80% and 70% of the edges, the more edges were removed, the faster the scoring decreased. This trend was observed within all three scoring metrics. Since the variance was stable in-between runs, we then selected one of the 10 runs for each down sampling and visualized the distribution of scores in a violin plot (Fig. 8B). We observed a general increasing number of higher scores when adding more edges. Similar to the boxplots, we see that the violins of 90%, 80% and 70% of the edges were very similar to each other. We observed comparable results when using AMI or JI as metric in the comparison. In summary, our approach showed that modules are relatively stable when leaving out interactions, and similar modules are formed in the process. We

further analyzed the nPCC values calculated for the hypoxia dataset (Additional file 8). About 24% of all clusters had significant correlation of expression in-between module genes, further showing that derived SUBATOMIC modules are well-separated and supported by orthogonal expression data.

Discussion

We developed SUBATOMIC, an integration pipeline that decomposes multi-omics networks into topological modules and their interactions using composite subgraph clustering and statistical and functional analysis. The obtained modules are further embedded within their network and regulator context in a superview analysis as well as functionally annotated and visualized. In a post-processing step, we contextualized the obtained modules with condition-specific data in three hypoxia cell lines and calculated activity as well as expression correlation scores. Compared to our previous integration framework, SUBATOMIC contains major improvements [28]. Most importantly, we automated the workflow and integrated it into a general Snakemake pipeline. While the previous version was generated for one use case only, we now can perform the analysis workflow from decomposing a composite network up to the analysis of modules in one single execution. Moreover, it is applicable to multi-edge networks of any species. We adapted the superview analysis to output a summary of connections between modules and regulators. We also improved upon the functional characterization of modules by automating GO term enrichment. Since run-time can be an issue on large interaction networks, we parallelized time-critical steps in the pipeline and improved scalability and computability. Moreover, we added scripts to support the visualization of modules in Cytoscape. To calculate condition-specific activity of modules, we developed a post-processing step that calculates a biological activity score.

We applied SUBATOMIC to a composite network consisting of human TF-target gene, miRNA-mRNA, protein–protein, co-functional and homologous interactions. Our approach yielded 5586 modules. The majority of modules were enriched for GO terms, demonstrating that module topology and biological function are closely interrelated, since clustering was performed on topological features. Most modules obtained were COM modules, made up of undirected edges. This is expected because we included more undirected than directed interactions. Moreover, we observed that in most of the cases the interactions in a three-node subgraph came more often from the same input network rather than from different networks. This is due to the fact that the input networks do not possess exactly the same set of nodes and can also have a different number of interactions. Instead of using the full HumanNet database, we included only high-quality interactions with a log-likelihood score greater than three. This balances the number of interactions to a comparable amount by setting cut-offs on quality values.

Given the modules and their functional annotation, we demonstrated how SUBATOMIC can be used to predict the function of unannotated genes. Out of 53 weakly annotated genes that also appeared in the inferred modules, we selected PROSER1, CNBD1, LEK1, RIIAD1 and ERICH6B for an in-depth analysis. While some of our predictions, such as the chromatin modification role of PROSER1 are very novel, others were supported by recent publications such as the potential involvement of RIIAD1 in sperm motility. However, functional characterization based on guild-by-association

principle generates guiding functional hypotheses that still need experimental conformation.

In the case of contextualized modules, we put forward two prioritization strategies: either through enrichment of related GO terms or the module activity score. We showed that modules that share GO terms, such as hypoxia, are strongly interconnected and accumulate DE genes for related conditions. Also, genes not annotated with the specific GO terms but relevant for the condition can be detected due to close distances inside one module. However, when genes are not well characterized or do lack specific GO key words linking them to a condition, a GO term-based prioritization strategy might miss out on important modules. We further developed a more data-driven prioritization scheme by implementing module activity based on differential expression data. Using the activity score, we found a small set of modules relevant in the hypoxia context. This set was strongly interconnected and partially overlapped with the GO term based set of genes. While modules on itself show a static view and give insights on what is possible in an organism, contextualizing adds condition-specific dynamicity. For example, despite the fact that one of the main driving hypoxia genes HIF1A was not present in the most current probe annotation for the expression array from the hypoxia study, it was detected in many modules based on GO term annotation and activity. Especially genes targeted by HIF1A revealed high differential expression in HIF1A-containing modules. Our results indicate that we can prioritize the large amount of modules in different ways to end up with sets of modules highly relevant for a condition or disease context. With the activity prioritization method, we were able to find modules with different topologies strongly connected in the superview. We showed that many genes in these modules are already known to play important roles in hypoxia. Moreover, contextualizing the modules with expression data from three different cell lines revealed that the activation of response mechanisms can differ and that different parts of a module can be active in different tissues. For example, EGLN3 is a known hypoxia-induced factor and showed dysregulation only in the SK-OV-3 cell line (see COM_256). In ALL_2093, KLF4 is only weakly expressed in WM793B while its target PFKB3 is strongly expressed in all three cell lines under hypoxic conditions; thus other regulators such as MYC and STAT also targeting PFKB3 might have a stronger regulatory influence. Overall, the combination of annotated SUBATOMIC modules based on different topologies, their superview connections, their contextualization with expression data and their visualization among different conditions and cell lines delivers a versatile tool to deeply investigate multi-edge networks.

SUBTATOMIC is not limited to the data types described in this study and offers additional analysis opportunities. While we restricted our analysis to genes and miRNAs, it is possible to add any type of nodes and interactions such as metabolite interactions, lncRNA interactions, or siRNA interactions, as long as the input networks share a common set of nodes for intersection. Moreover, while we used a static prior network that was contextualized in a later step to add dynamicity, it is possible to directly analyze a dynamic composite network by including condition- or patient specific association networks inferred from context-specific high-throughput data such as transcriptomics or proteomics, e.g. co-expression networks at either bulk or single

cell level. Also, hybrid approaches are possible, combining public databases for some interaction types with condition-specific interactions for others.

To evaluate our modules from a clustering perspective, we designed a re-sampling approach where we randomly removed a fixed number of edges over ten repetitions for nine different sampling fractions (10–90%). We showed that modules remain stable when removing a small percentage of edges but become less and less stable when removing more. This trend was confirmed by using three different metrics when comparing modules of sampled networks with the full network. Since modules are strongly influenced by the size and completeness of the input prior network, the analysis also demonstrated that there is a saturation effect in stability when adding more edges, since the differences in average ARI between adjacent sampling points were decreasing while adding more edges. However, not every module could be found back in the sub-sampled modules. This is expected since some edges might be essential to connect two parts of a module, and when removed the module might split into two. Since clustering was done on two-node and three-node hypergraphs, nodes could appear in several modules. To avoid comparing non-related modules, we decided to only include the best pairwise ARI between the ground truth and subsampled modules to assess clustering stability. Furthermore, stability was supported by a large amount of modules with significant nPCC values as well as a high number of modules with significantly enriched GO terms. Thus, we can conclude that SUBATOMIC predicts modules in a stable manner.

Broadly, two main distinctions can be made when it comes to module inference: on the one hand, methods generate modules directly from experimental read-outs such as expression data, or use prior networks as a base for the inference. On the other hand, clustering can be based on only one data modality or include multiple ones. Methods such as WGCNA, lmQCM, MiBiOmics and TPSC are examples of methods that generate expression based co-expression clusters [38, 90–92]. While they are widely used and are shown to produce modules that correlate to many biological features, these modules are often very large, do not consider causal regulatory interactions, or have no multi-omics data integration. Other methods go one step further and additionally integrate protein–protein interactions [41, 93, 94]. The method of Dittrich et al. combines clustering on expression data with protein–protein interaction networks to derive modules that represent merged, overlapping and independent communities [95]. While the classification of modules in independent, overlapping and merging communities already give some network context to the clusters, it is not as flexible with the input data and they did not consider directionality of interactions and the clustering yielded only a small number of modules. The Multi-omics Module Analysis Method (MOMA) uses a deep learning approach to derive omics-specific module representations which are further integrated within an attention layer to find relevant modules for disease prediction [96]. Sparse Multiple Canonical Correlation Network Analysis (SmCCNet) derives a few large modules to connect omics measurements with specific phenotypes [97]. Another class of methods tries to use static prior networks as a base and uses expression data in a contextualization step to find active subnetworks [97–99]. For example, the connect separate connected components (C3) modularizes a network into disease-relevant modules by iteratively connecting sub-networks made of a small number of disease-associated proteins [100].

DeRegNet combines prior regulatory networks with omics abundance measurements to identify maximally deregulated subnetworks [101]. Some more of these methods and strategies have been further reviewed by other authors [39, 102–104]. Compared to existing methods, SUBATOMIC tries to address open issues and creates a comprehensive analysis framework covering various aspects of module inference (see Additional file 8 that further shows a tabular comparison of the features in-between the here mentioned module inference methods and SUBATOMIC.)

It is based on a topological clustering approach that allows to interpret clausal relationships between gene and emphasizes different regulatory mechanisms. It introduces flexibility allowing for operation on literature-defined prior networks as well as on omics-derived association networks. Moreover, it can include all types of nodes and biological interactions in an integrated manner. Networks are divided into a large number of small and interpretable modules with district topological properties, while still keeping track of their global network context and regulators. Furthermore, static network modules can be inferred and contextualized with expression data using ECD, nPCC and module activity scores. This combination makes it a unique and outstanding method in the field of composite network clustering.

While SUBATOMIC was shown to be able to answer many biological questions, it also comes with some limitations. Input prior networks are often incomplete and might complicate biological interpretation. However, we expect that interaction databases continuously grow, and thus multi-edge networks will become more and more complete, and our stability analysis further demonstrated that modules remain stable when confronted with missing interactions. Furthermore, if networks do not overlap for a certain quantity of nodes, the derived subgraphs will mostly assemble in modules from separated interaction types. Also this issue will be solved with a growing amount of databases. Another limit is computability. We used SCHype for our clustering algorithm and demonstrated that it was able to process more than 750,000 subgraphs in an adequate amount of time. However, the number of detected subgraphs grows super-linear with larger networks. Thus, there exists an upper limit in network size that is still computable. Furthermore, while we parallelized GOATOOLS to annotate several modules at a time, the gain of computational speed was accompanied by an increase in space consumption. This limits the number of cores that can be used for parallelization. This will be addressed in a future version integrating a more space-efficient annotation tool.

Conclusion

In conclusion, we developed an automated subgraph clustering framework that takes basic building blocks of interactions and clusters them into modules. The modules are further characterized and contextualized by superview calculation, regulator analysis, GO term enrichment, and module activity scoring. SUBATOMIC can be used to investigate conditions and diseases, find interactions between functionally related modules, and derive novel gene functions for uncharacterized genes. The main limiting factor is the availability of interconnected networks. We believe this issue will be solved in time with an ever-increasing number of interactions being discovered. Our approach distinguishes itself from other module inference methods by clustering based on topological

features to create a high number of small and easily interpretable modules with different regulatory properties, while still keeping the overall network context in mind.

Methods

The main analysis workflow has been integrated into one Snakemake pipeline. A Snakemake workflow diagram is available in the Additional file 7, and a schematic overview can be found in Fig. 1. All software, including some scripts for pre- and post-processing analysis, as well as a Docker version are made available on GitHub (https://github.com/CBIGR/SUBATOMIC).

Subgraph detection

For subgraph detection, we used the 'Index-based subgraph algorithm' (ISMAGS) [44, 105]. We followed the subgraph representation used in ISMAGS, where a three-node subgraph is represented as a three letter code, which specifies that a given edge originates from a certain set of input interactions. Interactions in one input network need to be either all directed or all undirected (e.g. all interactions in a TF-target network should be directed; all interactions in a protein–protein network should be un-directed, …). Each input network is characterized by a unique one-letter representation that can be freely chosen by the user. As an example, the subgraph RRP would contain one directed edge from network R, another directed edge from network R as well as an undirected edge from the protein–protein interaction network. A letter for a directed network can be set to lower case to indicate that the direction is reversed (see ISMAGS paper [44]). Due to symmetry, some subgraphs are redundant to one another (e.g., PPC, PCP and CPP represent the same subgraph). A custom-made script calculated a non-redundant set of subgraph representations based on a provided list of directed and undirected network letters. This set is then used by the pipeline as a guide to search for subgraphs and can be further fine-tuned by the user to remove additional unwanted subgraphs. The three-node subgraphs were then identified by ISMAGS [44]. ISMAGS takes for each iteration the three-letter subgraph representation and the parts of the composite network that contain interactions for this subgraph. It outputs all three-node subgraphs that satisfy the defined representation. Besides three-node subgraphs, we also searched for two-node subgraphs that possess special properties: i.e., all pairs of nodes where each node contains a directed edge pointing at the other node (DD-type) and all pair of nodes connected by one undirected and one directed edge (DU-type).

Subgraph clustering and module inference

The subgraphs produced by ISMAGS were subsequently grouped into one of the following subgraph types: complex subgraphs (COM), feed forward loop (FFL), co-pointing subgraphs (COP), co-regulated subgraphs (COR), circular feedback subgraph (CIR), feedback undirected subgraph (FBU) and feedback 2 undirected subgraph (FB2U) and two-node feedback subgraph (2FB) [28]. Each of these subgraph types is characterized by a unique combination of directed and undirected edges as

visualized in Fig. 1. For example, the COM type contains all subgraphs that exclusively consist of undirected edges. Given the undirected network letters C, P and H, any combination of these three letters that led to a set of non-redundant subgraphs was grouped together into the COM type. Each module type consists of specific subgraphs types and served as the input for the following clustering.

We inferred clusters for each of the above defined subgraph types separately as well as on the union of all classes (ALL) using the SCHype algorithm [45]. This algorithm is based on the Perron-Frobenius theorem, and clusters a hypergraph solving an optimization problem by maximizing the edge-to-node ratio in each cluster for a network [106]. The input is a hypergraph, were each hypernode represents a three-node subgraph calculated by ISMAGS or a two-node subgraph. SCHype was run with default settings ($p=1$) and output several modules for each of the eight classes of modules and for clustering all subgraphs together. These modules were further filtered for subsequent analysis: clusters containing 5–50 genes were kept and modules containing more than 90% of homologous edges were excluded.

Superview calculation

The superview step characterizes interactions between modules as well as between regulators and modules. Each module was compared to every other module by counting how many edges per input network were shared between those two modules. This value was compared against the shared edge count in a random sampling. In the sampling, we generated 1000 times two random modules, both having an equivalent amount of nodes as the two modules under investigation. The derived distribution was used for a z-score transformation with $z = \frac{x-\mu}{\sigma}$ given the mean μ and the standard deviation σ from the random distribution. The z-score was further evaluated by calculating a p-value (significance cut-off: 0.05) for that distribution with a right tailed-test $1 - CDF(z)$ where CDF is the cumulative distribution function. The output was composed of one file per module type (ALL, COM, COR, ...) containing interactions between every module of this type and all other modules for each input network and displayed the count of shared interactions, z-score, and p-value. If no interactions exist between the two modules, the z-score was set to 0 and the p-value was set to 1.

The superview calculated three more outputs that characterized the relationship between modules and regulators. For each regulator (TF or miRNA), we calculated the RF-module connection strength for each RF-module pair and each interaction type with $\frac{1}{|N|}\sum_{n_i \in N} n_i$ with N being all module genes and $n_i=1$ if an edge exist, else $x = 0$. This gave a fraction on how many interactions between a TF and a module existed and was used to find regulators that are strongly connected to a module or a set of modules. Another analysis displayed the fraction how many distinct TFs or miRNAs target one particular module for each module with $\frac{no.regulator-moduleinteractions}{totalnumberregulators}$. This gave a module-specificity and allowed to investigate whether a module was targeted by a few or many regulators. Another analysis displayed how many modules are targeted by a certain regulator and shows a fraction of how this compared to the total number of derived modules. For each regulator, we calculated the regulator specificity

by $\frac{no.targetedmodule}{totalnumberofmodules}$, which allowed to investigate which TFs and miRNAs targeted a wide range of modules, as opposed to some that were specific for a single module or a small number of modules.

Functional enrichment analysis

For each module, SUBATOMIC performed a functional enrichment to gain insights into its biological relevance. We used the Python implementation of GOATOOLS to calculate the enrichment of GO terms for each module [46]. We provided three options as enrichment background: all genes present in one specific type of modules, all genes present in the input networks, or all genes written in a user-specified file. For our analyses we used a user-defined input containing all annotated human genes according to Ensembl as input. Results were summarized in one file per module. Only results with a corrected p-value > 0.05 were kept based on Benjamini and Hochberg FDR correction [107]. Additionally to the standard GOATOOLS output, we provided a rank for the ascendingly sorted p-values for each module, since p-values can strongly differ between modules and enriched functionality depends on which processes are well annotated as well as how many GO terms are available. The rank allowed to filter for the top n entries per module. We further reported the log2 fold-change for each significant GO term.

Visualization

To visualize the modules in Cytoscape, we provide a number of files that can be imported. The most important file is a nnf file containing the network representation of the modules. Additionally, we generated a noa file to annotate each node with its type (TF, gene or miRNA), its gene name, and a short optional functional description. Each run of the pipeline also resulted in a Cytoscape style sheet in xml format to format the network in a way consistent with the provided information in nnf and noa file format, that can as well be imported. The xml file can also be adapted for more customized style choices.

Run time considerations

Operations on graphs often come with a high computational cost. Several steps in the pipeline were parallelized, but some bottlenecks remain. The subgraph detection algorithm is highly efficient and can find millions of subgraphs in less than a minute. SCHype can cluster hundreds of thousands of subgraphs, but its run time increases super-linear. Furthermore, the superview calculation comparing all modules against each other and the functional annotation are the most time-consuming steps. Since these steps process one module at a time, we parallelized them in a way that each module can be processed by a different core. In principle, as many modules can be processed in parallel as there are cores available. However, since each separate process needs a certain amount of memory, a careful balance between the number of cores available and the amount of memory must be taken in order to find a suited number of cores for parallelization. For our application on a *H. sapiens* composite network, we ran SUBATOMIC limited to eight cores and 70 GB RAM on a 2 × 18-core Intel Xeon Gold 6240 (Cascade Lake @ 2.6 GHz) processor, which resulted in two days of run time.

Construction of the composite H. sapiens network

We unified five different types of interactions from different sources into one composite network representation (Table 1). From OmniPath, we included 53,232 TF-target gene interactions formed by 526 regulators and 14,488 target genes (access 17.01.2022) [19, 20]. We included all interactions in the evidence classes A, B and C from DoRothEA as well as the TF-target gene and TF-miRNA interactions. From the same database, we included 11,085 miRNA-target interactions between 850 miRNAs and 3210 target genes (access 17.01.2022). The homology between the genes was retrieved from Ensembl based on the GeneTree pipeline [49]. This pipeline takes a reciprocal best BLAST approach in a simple case, but also considers more complex ontologies by resolving one-to-many and many-to-many relations. We applied a minimum reciprocal sequence identity of 50% as a threshold to include homologous interactions between a pair of genes. We only considered genes involved in at least one of the other data sets included in this analysis. Moreover, we added homologous miRNAs with identical seed sequences. This summed up to a total of 10,847 interactions for 4862 genes or miRNAs. We obtained protein–protein interactions and co-functional interactions from HumanNet v2 [21, 50]. We used the log-likely-hood score (LLS) provided by HumanNet to filter for interactions with $LLS \leq 3.0$. Since the number of interactions in HumanNet was magnitudes bigger than the number of regulatory interactions, these filtering steps tried to balance the number of interactions without losing to many included genes. We selected 24,773 protein–protein interactions formed by 10,950 genes and 66,373 co-functional interactions formed by 10,683 genes. For the undirected interaction sets, duplicates were merged (e.g., A–B and B–A is equivalent for undirected interactions). This removed a total of 6 protein–protein interactions, 28 co-functional interactions and 39 homologous interactions. We mapped all genes to Ensembl identifiers to make them comparable between the networks. For miRNAs, we kept the standard naming convention (e.g. hsa-miR-600e), which avoided potential overlaps with Ensembl gene identifiers. We omitted genes that could not be mapped to an Ensembl ID. The genes included in the analysis were based on the human genome version 38. We included annotated genes on chromosomes 1–22 as well as X and Y. All interactions were merged into one file to create the composite network. In this file, each interaction was represented by the two interacting nodes as well as their edge color. The edge color was represented by a network specific letter (TF-target gene interactions: R, miRNA-target gene interactions: M, homologous interactions: H, protein–protein interactions: P, Co-functional interactions: C, Additional file 1).

The hypoxia expression data set

We used expression data from three different cell lines under cyclic and chronic hypoxia conditions to contextualize the modules (GEO: GSE53012) [58]. The Affymetrix microarray data were processed and normalized using the Single Channel Array Normalization (SCAN) [108]. It corrected the effect of technical bias, such as GC content by applying a mixture-modeling approach [108]. To calculate biological activity, we used the *p*-values from the differential expression analysis from the original publication.

Contextualization

We implemented several methods to contextualize modules with expression data from perturbation experiments or experiments with case and control samples. We calculated the average Pearson correlation coefficient nPCC between each pair of genes in a module and compared it against the nPCC of a sampling of 1000 modules to measure co-expression within a module. Next, we derived a z-score for each module nPCC and transformed it into a p-value via the CDF of the standard normal distribution. We also added an implementation of the Expression Correlation Differential Score (ECD score), which highlights modules specific for an experimental condition as compared to the control condition [28]. There we subtracted for each edge in a module the Pearson correlation of case samples from the Pearson correlation of condition samples and averaged this over all edges per module. We repeated this for 1000 randomly generated modules to obtain a background distribution, which is then used to calculate a z-score that was consequently transformed into p-values using the standard normal CDS. Given enough expression values and conditions, this allowed us to address the dynamicity of edges using the guilt by association and guilt by rewiring principles.

We implemented one additional metric compared to the previous framework to capture modules responding to a change in condition. This module activity score (s_a) used p-values of a comparison between different conditions (e.g., from differential expression between case and control samples) and transformed them into a z-score $z_i = \theta^{-1}(1 - p_i)$ with θ^{-1} being the inverse normal CDF [79]. For each module, we calculated the aggregate z-score $z_a = \frac{1}{\sqrt{k}} \sum_{allmodulegenes} z_i$ for each module with size k. For each module size, we drew 1000 random modules of the same size and used this as a background distribution to calculate the normalized activity score $s_a = \frac{(z_a - \mu_k)}{\sigma_k}$. High values of s_a indicated that the module could be interpreted as an active subnetwork in the specific experimental condition.

Abbreviations

2FB	Two-node feedback subgraph
CIR	Circular feedback subgraph
COM	Complex subgraph
COP	Co-pointing subgraph
COR	Co-regulated subgraph
FB2U	Feedback 2 undirected subgraph
FBU	Feedback undirected subgraph
FFL	Feed forward loop
RF	Regulatory factor
TF	Transcription factor
COF	Co-functional interaction
DE	Differentially expressed

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04908-3.

Additional file 1: General description of the input networks and inferred modules.

Additional file 2: Analysis of weakly characterized genes.

Additional file 3: GO term based module analysis.

Additional file 4: Activity based module analysis.

Additional file 5: Abbreviations of gene names.

Additional file 6: Supplemental figures.

Additional file 7: Pipeline description and module stability.

Additional file 8: Comparsion of SUBATOMIC with other methods.

Acknowledgements
We would like to thank Kenneth Stoop and Pieter Audenaert for their help with running the ISMAGS algorithm. Moreover, we would like to thank Hayoung Kim, Heesoo Song and Jietse Verweirder for their support in prototyping the Snakemake pipeline.

Author contributions
Contributions according to the CRediT system (https://casrai.org/credit/, JL = Jens Uwe Loers, VV = Vanessa Vermeirssen): Conceptualization: JL, VV, Data curation: JL, Formal analysis: JL, Funding acquisition: JL, VV, Investigation: JL, Methodology: JL, VV, Project administration: VV, Software: JL, Supervision: VV, Validation: JL, Visualization: JL, Writing—original draft: JL, VV, Writing—rewriting and editing JL, VV. Both authors read and approved the final manuscript.

Funding
J.L. is supported by a BOF PhD scholarship from Ghent University, and this work was also funded by a BOF Starting Grant BOF/STA/201909/030 'Multi-omics data integration to elucidate the causes of complex diseases'.

Availability of data and materials
GitHub: https://github.com/CBIGR/SUBATOMIC, code of the pipeline as well as link to the Docker version. Zenodo: https://doi.org/10.5281/zenodo.6556413, raw data of the input networks and SUBATOMIC output.

Declarations

Ethics approval and consent to participate
The data analyzed in this study were from public databases, so ethical approval and consent participation were not required.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 17 May 2022 Accepted: 24 August 2022
Published online: 05 September 2022

References

1. Orphanides G, Reinberg D. A unified theory of gene expression. Cell. 2002;108:439–51.
2. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13:613–26.
3. Venters BJ, Pugh BF. How eukaryotic genes are transcribed. Crit Rev Biochem Mol Biol. 2009;44:117–41.
4. Reiter F, Wienerroither S, Stark A. Combinatorial function of transcription factors and cofactors. Curr Opin Genet Dev. 2017;43:73–81.
5. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet. 2019;20:207–20.
6. Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell. 2010;6:479–91.
7. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136:215–33.
8. Morozova N, Zinovyev A, Nonne N, Pritchard L-L, Gorban AN, Harel-Bellan A. Kinetic signatures of microRNA modes of action. RNA. 2012;18:1635–55.
9. Yao Q, Chen Y, Zhou X. The roles of microRNAs in epigenetic regulation. Curr Opin Chem Biol. 2019;51:11–7.
10. Zhang Z, Cao Y, Zhai Y, Ma X, An X, Zhang S, et al. MicroRNA-29b regulates DNA methylation by targeting Dnmt3a/3b and Tet1/2/3 in porcine early embryo development. Dev Growth Differ. 2018;60:197–204.
11. Bhat SA, Ahmad SM, Mumtaz PT, Malik AA, Dar MA, Urwat U, et al. Long non-coding RNAs: mechanism of action and functional utility. Non-Coding RNA Res. 2016;1:43–50.
12. Gao N, Li Y, Li J, Gao Z, Yang Z, Li Y, et al. Long non-coding RNAs: the regulatory mechanisms, research strategies, and future directions in cancers. Front Oncol. 2020;10:2903.
13. Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, et al. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell. 2011;147:358–69.
14. Zhou Y, Meng X, Chen S, Li W, Li D, Singer R, et al. IMP1 regulates UCA1-mediated cell invasion through facilitating UCA1 decay and decreasing the sponge effect of UCA1 for miR-122-5p. Breast Cancer Res. 2018;20:32.
15. Conrad B, Antonarakis SE. Gene duplication: a drive for phenotypic diversity and cause of human disease. Annu Rev Genomics Hum Genet. 2007;8:17–35.
16. Lan X, Pritchard JK. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. Science. 2016;352:1009–13.

17. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019;29:1363–75.
18. Garcia-Alonso L, Iorio F, Matchan A, Fonseca N, Jaaks P, Peat G, et al. Transcription factor activities enhance markers of drug sensitivity in cancer. Cancer Res. 2018;78:769–80.
19. Türei D, Valdeolivas A, Gul L, Palacio-Escat N, Klein M, Ivanova O, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol Syst Biol. 2021;17: e9923.
20. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods. 2016;13:966–7.
21. Hwang S, Kim CY, Yang S, Kim E, Hart T, Marcotte EM, et al. HumanNet v2: human gene networks for disease research. Nucleic Acids Res. 2019;47:D573–80.
22. Manke T, Bringas R, Vingron M. Correlating protein-DNA and protein-protein interaction networks. J Mol Biol. 2003;333:75–85.
23. Zhang C, Lee S, Mardinoglu A, Hua Q. Investigating the combinatory effects of biological networks on gene co-expression. Front Physiol. 2016;7:160.
24. Martinez NJ, Ow MC, Barrasa MI, Hammell M, Sequerra R, Doucette-Stamm L, et al. A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity. Genes Dev. 2008;22:2535–49.
25. Guo Y, Alexander K, Clark AG, Grimson A, Yu H. Integrated network analysis reveals distinct regulatory roles of transcription factors and microRNAs. RNA. 2016;22:1663–72.
26. Baur B, Shin J, Zhang S, Roy S. Data integration for inferring context-specific gene regulatory networks. Current Opin Syst Biol. 2020;23:38–46.
27. Williams RM, Candido-Ferreira I, Repapi E, Gavriouchkina D, Senanayake U, Ling ITC, et al. Reconstruction of the global neural crest gene regulatory network in vivo. Dev Cell. 2019;51:255-276.e7.
28. Defoort J, Van de Peer Y, Vermeirssen V. Function, dynamics and evolution of network motif modules in integrated gene regulatory networks of worm and plant. Nucleic Acids Res. 2018;46:6480–503.
29. Dolinski K, Chatr-aryamontri A, Tyers M. Systematic curation of protein and genetic interaction data for computable biology. BMC Biol. 2013;11:43.
30. Lander AD. The edges of understanding. BMC Biol. 2010;8:40.
31. Azad AKM. Integrating heterogeneous datasets for cancer module identification. In: Keith JM, editor. Bioinformatics: volume II: structure, function, and applications. New York: Springer; 2017. p. 119–37.
32. Azad AKM, Lee H. Voting-based cancer module identification by combining topological and data-driven properties. PLoS ONE. 2013;8: e70498.
33. Bennett L, Kittas A, Muirhead G, Papageorgiou LG, Tsoka S. Detection of composite communities in multiplex biological networks. Sci Rep. 2015;5:10345.
34. Bodein A, Scott-Boyer M-P, Perin O, Lê Cao K-A, Droit A. Interpretation of network-based integration from multi-omics longitudinal data. Nucleic Acids Res. 2021;18(9):551.
35. Bodein A, Chapleur O, Droit A, Lê Cao K-A. A generic multivariate framework for the integration of microbiome longitudinal studies with other data types. Front Genet. 2019;10:963.
36. Bonnet E, Calzone L, Michoel T. Integrative multi-omics module network inference with lemon-tree. PLoS Comput Biol. 2015;11: e1003983.
37. Durmaz A, Henderson TAD, Bebek G. Frequent subgraph mining of functional interaction patterns across multiple cancers. Pac Symp Biocomput. 2021;26:261–72.
38. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform. 2008;9:559.
39. Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. Nat Commun. 2018;9:1090.
40. Silverbush D, Cristea S, Yanovich-Arad G, Geiger T, Beerenwinkel N, Sharan R. Simultaneous integration of multi-omics data improves the identification of cancer driver modules. Cell Syst. 2019;8:456-466.e5.
41. Wu C, Zhang F, Li X, Zhang S, Li J, Su F, et al. Composite functional module inference: detecting cooperation between transcriptional regulation and protein interaction by mantel test. BMC Syst Biol. 2010;4:82.
42. Hiraga T. Hypoxic microenvironment and metastatic bone disease. Int J Mol Sci. 2018;19:E3523.
43. Todd VM, Vecchi LA, Clements ME, Snow KP, Ontko CD, Himmel L, et al. Hypoxia inducible factor signaling in breast tumors controls spontaneous tumor dissemination in a site-specific manner. Commun Biol. 2021;4:1–18.
44. Houbraken M, Demeyer S, Michoel T, Audenaert P, Colle D, Pickavet M. The index-based subgraph matching algorithm with general symmetries (ISMAGS): exploiting symmetry for faster subgraph enumeration. PLoS ONE. 2014;9: e97896.
45. Michoel T, Nachtergaele B. Alignment and integration of complex networks by hypergraph-based spectral clustering. Phys Rev E. 2012;86: 056111.
46. Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, et al. GOATOOLS: a python library for gene ontology analyses. Sci Rep. 2018;8:10872.
47. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
48. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47:D155–62.
49. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara genetrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2009;19:327–35.
50. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011;21:1109–21.
51. Hou L, Chen M, Zhang CK, Cho J, Zhao H. Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. Hum Mol Genet. 2014;23:2780–90.
52. Tian W, Zhang LV, Taşan M, Gibbons FD, King OD, Park J, et al. Combining guilt-by-association and guilt-by-profiling to predict saccharomyces cerevisiaegene function. Genome Biol. 2008;9:S7.

53. Shinsky SA, Monteith KE, Viggiano S, Cosgrove MS. Biochemical reconstitution and phylogenetic comparison of human SET1 family core complexes involved in histone methylation. J Biol Chem. 2015;290:6361–75.
54. Wang X, Rosikiewicz W, Sedkov Y, Martinez T, Hansen BS, Schreiner P, et al. PROSER1 mediates TET2 O-GlcNAcylation to regulate DNA demethylation on UTX-dependent enhancers and CpG islands. Life Sci Alliance. 2021;5: e202101228.
55. Urizar-Arenaza I, Osinalde N, Akimov V, Puglia M, Candenas L, Pinto FM, et al. Phosphoproteomic and functional analyses reveal sperm-specific protein changes downstream of kappa opioid receptor in human spermatozoa*. Mol Cell Proteomics. 2019;18:S118–31.
56. Xu K, Yang L, Zhang L, Qi H. Lack of AKAP3 disrupts integrity of the subcellular structure and proteome of mouse sperm and causes male sterility. Development. 2020;147:dev181057.
57. Chen P-S, Chiu W-T, Hsu P-L, Lin S-C, Peng I-C, Wang C-Y, et al. Pathophysiological implications of hypoxia in human diseases. J Biomed Sci. 2020;27:63.
58. Olbryt M, Habryka A, Student S, Jarząb M, Tyszkiewicz T, Lisowska KM. Global gene expression profiling in three tumor cell lines subjected to experimental cycling and chronic hypoxia. PLoS ONE. 2014;9: e105104.
59. Watts ER, Walmsley SR. Inflammation and hypoxia: HIF and PHD isoform selectivity. Trends Mol Med. 2019;25:33–46.
60. Depoix CL, de Selliers I, Hubinont C, Debieve F. HIF1A and EPAS1 potentiate hypoxia-induced upregulation of inhibin alpha chain expression in human term cytotrophoblasts in vitro. Mol Hum Reprod. 2017;23:199–209.
61. Lee JW, Ko J, Ju C, Eltzschig HK. Hypoxia signaling in human diseases and therapeutic targets. Exp Mol Med. 2019;51:1–13.
62. Wang GL, Semenza GL. Purification and characterization of hypoxia-inducible factor 1 (∗). J Biol Chem. 1995;270:1230–7.
63. Wang Z, Shi Y, Ying C, Jiang Y, Hu J. Hypoxia-induced PLOD1 overexpression contributes to the malignant phenotype of glioblastoma via NF-κB signaling. Oncogene. 2021;40:1458–75.
64. Ziello JE, Jovin IS, Huang Y. Hypoxia-inducible factor (HIF)-1 regulatory pathway and its potential for therapeutic intervention in malignancy and ischemia. Yale J Biol Med. 2007;80:51–60.
65. Geng H, Xue C, Mendonca J, Sun X-X, Liu Q, Reardon PN, et al. Interplay between hypoxia and androgen controls a metabolic switch conferring resistance to androgen/AR-targeted therapy. Nat Commun. 2018;9:4972.
66. Mitani T, Yamaji R, Higashimura Y, Harada N, Nakano Y, Inui H. Hypoxia enhances transcriptional activity of androgen receptor through hypoxia-inducible factor-1α in a low androgen environment. J Steroid Biochem Mol Biol. 2011;123:58–64.
67. Gilkes DM, Bajpai S, Wong CC, Chaturvedi P, Hubbi ME, Wirtz D, et al. Procollagen Lysyl hydroxylase 2 Is essential for hypoxia-induced breast cancer metastasis. Mol Cancer Res. 2013;11:456–66.
68. Bruick RK, McKnight SL. A conserved family of Prolyl-4-hydroxylases that modify HIF. Science. 2001;294:1337–40.
69. To KKW, Huang LE. Suppression of hypoxia-inducible factor 1alpha (HIF-1alpha) transcriptional activity by the HIF prolyl hydroxylase EGLN1. J Biol Chem. 2005;280:38102–7.
70. Shah R, Smith P, Purdie C, Quinlan P, Baker L, Aman P, et al. The prolyl 3-hydroxylases P3H2 and P3H3 are novel targets for epigenetic silencing in breast cancer. Br J Cancer. 2009;100:1687–96.
71. Kumar B, Khaleghzadegan S, Mears B, Hatano K, Kudrolli TA, Chowdhury WH, et al. Identification of miR-30b-3p and miR-30d-5p as direct regulators of androgen receptor signaling in prostate cancer by complementary functional microRNA library screening. Oncotarget. 2016;7:72593–607.
72. Bakker WJ, Harris IS, Mak TW. FOXO3a is activated in response to hypoxic stress and inhibits HIF1-induced apoptosis via regulation of CITED2. Mol Cell. 2007;28:941–53.
73. Chaanine AH, Kohlbrenner E, Gamb SI, Guenzel AJ, Klaus K, Fayyaz AU, et al. FOXO3a regulates BNIP3 and modulates mitochondrial calcium, dynamics, and function in cardiac stress. Am J Physiol Heart Circ Physiol. 2016;311:H1540–59.
74. Li H, Zhou L, Dai J. Retinoic acid receptor-related orphan receptor RORα regulates differentiation and survival of keratinocytes during hypoxia. J Cell Physiol. 2018;233:641–50.
75. Chen Q, Zhou Y, Richards AM, Wang P. Up-regulation of miRNA-221 inhibits hypoxia/reoxygenation-induced autophagy through the DDIT4/mTORC1 and Tp53inp1/p62 pathways. Biochem Biophys Res Commun. 2016;474:168–74.
76. Xu R, Luo X, Ye X, Li H, Liu H, Du Q, et al. SIRT1/PGC-1α/PPAR-γ correlate with hypoxia-induced chemoresistance in non-small cell lung cancer. Front Oncol. 2021;11:2791.
77. Zhao Y-Z, Liu X-L, Shen G-M, Ma Y-N, Zhang F-L, Chen M-T, et al. Hypoxia induces peroxisome proliferator-activated receptor γ expression via HIF-1-dependent mechanisms in HepG2 cell line. Arch Biochem Biophys. 2014;543:40–7.
78. Rios-Colon L, Kumar P, Kim S, Sharma M, Su Y, Kumar A, et al. Carnitine palmitoyltransferase 1 regulates prostate cancer growth under hypoxia. Cancers (Basel). 2021;13:6302.
79. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002;18(suppl_1):S233–40.
80. Li Y, Sun X-X, Qian DZ, Dai M-S. Molecular crosstalk between MYC and HIF in cancer. Front Cell Develop Biol. 2020;8:590576.
81. Menendez MT, Teygong C, Wade K, Florimond C, Blader IJ. siRNA screening identifies the host hexokinase 2 (HK2) gene as an important hypoxia-inducible transcription factor 1 (HIF-1) target gene in toxoplasma gondii-infected cells. MBio. 2015;6:e00462.
82. Shan F, Huang Z, Xiong R, Huang Q-Y, Li J. HIF1α-induced upregulation of KLF4 promotes migration of human vascular smooth muscle cells under hypoxia. J Cell Physiol. 2020;235:141–50.
83. Obach M, Navarro-Sabaté A, Caro J, Kong X, Duran J, Gómez M, et al. 6-Phosphofructo-2-kinase (pfkfb3) gene promoter contains hypoxia-inducible factor-1 binding sites necessary for transactivation in response to hypoxia. J Biol Chem. 2004;279:53562–70.

84. Sysol JR, Chen J, Singla S, Zhao S, Comhair S, Natarajan V, et al. Micro-RNA-1 is decreased by hypoxia and contributes to the development of pulmonary vascular remodeling via regulation of sphingosine kinase 1. Am J Physiol Lung Cell Mol Physiol. 2018;314:L461–72.
85. Woo SK, Kwon MS, Geng Z, Chen Z, Ivanov A, Bhatta S, et al. Sequential activation of hypoxia-inducible factor 1 and specificity protein 1 is required for hypoxia-induced transcriptional stimulation of Abcc8. J Cereb Blood Flow Metab. 2012;32:525–36.
86. Levine E, Domany E. Resampling method for unsupervised estimation of cluster validity. Neural Comput. 2001;13:2573–93.
87. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218.
88. Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J Mach Learn Res. 2010;11:2837–54.
89. Jaccard P. The distribution of the flora in the alpine zone.1. New Phytol. 1912;11:37–50.
90. Liu Y, Ye X, Yu CY, Shao W, Hou J, Feng W, et al. TPSC: a module detection method based on topology potential and spectral clustering in weighted networks and its application in gene co-expression module discovery. BMC Bioinform. 2021;22:111.
91. Zhang J, Huang K. Normalized ImQCM: an algorithm for detecting weak quasi-cliques in weighted graph with applications in gene co-expression module discovery in cancers. Cancer Inform. 2014;13s3:CIN.S14021.
92. Zoppi J, Guillaume J-F, Neunlist M, Chaffron S. MiBiOmics: an interactive web application for multi-omics data exploration and integration. BMC Bioinform. 2021;22:6.
93. Lu X, Liu F, Miao Q, Liu P, Gao Y, He K. A novel method to identify gene interaction patterns. BMC Genomics. 2021;22:436.
94. Lu X, Zhu Z, Peng X, Miao Q, Luo Y, Chen X. InFun: a community detection method to detect overlapping gene communities in biological network. SIViP. 2021;15:681–6.
95. 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:i223-231.
96. Moon S, Lee H. MOMA: a multi-task attention learning algorithm for multi-omics data interpretation and classification. Bioinformatics. 2022;38:2287–96.
97. Shi WJ, Zhuang Y, Russell PH, Hobbs BD, Parker MM, Castaldi PJ, et al. Unsupervised discovery of phenotype-specific multi-omics networks. Bioinformatics. 2019;35:4336–43.
98. Ghiassian SD, Menche J, Barabási A-L. A DIseAse MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol. 2015;11:e1004120.
99. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. EnrichNet: network-based gene set enrichment analysis. Bioinformatics. 2012;28:i451–7.
100. Wang B, Hu J, Wang Y, Zhang C, Zhou Y, Yu L, et al. C3: connect separate connected components to form a succinct disease module. BMC Bioinform. 2020;21:433.
101. Winkler S, Winkler I, Figaschewski M, Tiede T, Nordheim A, Kohlbacher O. De novo identification of maximally deregulated subnetworks based on multi-omics data with DeRegNet. BMC Bioinform. 2022;23:139.
102. Nguyen H, Shrestha S, Tran D, Shafi A, Draghici S, Nguyen T. A comprehensive survey of tools and software for active subnetwork identification. Front Genet. 2019;10:155.
103. Wu S, Chen D, Snyder MP. Network biology bridges the gaps between quantitative genetics and multi-omics to map complex diseases. Curr Opin Chem Biol. 2022;66: 102101.
104. Alcalá-Corona SA, Sandoval-Motta S, Espinal-Enríquez J, Hernández-Lemus E. Modularity in biological networks. Front Genet. 2021; 12:701331.
105. Demeyer S, Michoel T, Fostier J, Audenaert P, Pickavet M, Demeester P. The index-based subgraph matching algorithm (ISMA): fast subgraph enumeration in large networks using optimized search trees. PLoS ONE. 2013;8:e61183.
106. Horn RA, Johnson CR. Matrix analysis. 2nd ed. Cambridge; New York: Cambridge University Press; 2012.
107. Benjamini Y, Hochberg Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. J R Stat Soc Series B. 1995;57:289–300.
108. Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics. 2012;100:337–44.

Publisher's Note