BMC Bioinformatics

RESEARCH

Open Access

Elucidating gene expression patterns across multiple biological contexts through a large-scale investigation of transcriptomic datasets

Rebeca Queiroz Figueiredo[1,2], Sara Díaz del Ser[1,2], Tamara Raschka[1,2,3], Martin Hofmann-Apitius[1,2], Alpha Tom Kodamullil[1], Sarah Mubeen[1,2,3†] and Daniel Domingo-Fernández[1,3,4*†]

†Sarah Mubeen and Daniel Domingo-Fernández contributed equally to this work

*Correspondence: daniel.domingo.fernandez@scai.fraunhofer.de

[1] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, 53757 Sankt Augustin, Germany
[2] Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, 53115 Bonn, Germany
[3] Fraunhofer Center for Machine Learning, Sankt Augustin, Germany
[4] Enveda Biosciences, Boulder, CO 80301, USA

Abstract

Distinct gene expression patterns within cells are foundational for the diversity of functions and unique characteristics observed in specific contexts, such as human tissues and cell types. Though some biological processes commonly occur across contexts, by harnessing the vast amounts of available gene expression data, we can decipher the processes that are unique to a specific context. Therefore, with the goal of developing a portrait of context-specific patterns to better elucidate how they govern distinct biological processes, this work presents a large-scale exploration of transcriptomic signatures across three different contexts (i.e., tissues, cell types, and cell lines) by leveraging over 600 gene expression datasets categorized into 98 subcontexts. The strongest pairwise correlations between genes from these subcontexts are used for the construction of co-expression networks. Using a network-based approach, we then pinpoint patterns that are unique and common across these subcontexts. First, we focused on patterns at the level of individual nodes and evaluated their functional roles using a human protein–protein interactome as a referential network. Next, within each context, we systematically overlaid the co-expression networks to identify specific and shared correlations as well as relations already described in scientific literature. Additionally, in a pathway-level analysis, we overlaid node and edge sets from co-expression networks against pathway knowledge to identify biological processes that are related to specific subcontexts or groups of them. Finally, we have released our data and scripts at https://zenodo.org/record/5831786 and https://github.com/ContNeXt/, respectively and developed ContNeXt (https://contnext.scai.fraunhofer.de/), a web application to explore the networks generated in this work.

Keywords: Transcriptomic, Biological context, Co-expression networks, Gene expression, Network biology

Introduction

While gene expression profiling has markedly improved our understanding of the molecular underpinnings of biological processes, the knowledge we acquire from a particular study performed within a given context may not generalize to another. For instance, accumulating evidence shows that average gene expression varies extensively across cell lines or tissues of the same organism [38, 43] as well as across species [32]. Context-specificity has also been noted when investigating the reproducibility of protein–protein interactions (PPIs) across conditions in literature-curated PPI databases in Stacey et al. [39], finding no evidence for the occurrence of anywhere from 19 to 55% of interactions reported in these databases. These findings, however, are not altogether surprising given that PPI databases often store interactions that occur across various experimental conditions and contexts which may fail to be observed if either of these were to vary. Crucially, it is often these context-specific differences which are responsible for the variability of functions and unique characteristics of diverse cell types and tissues and their investigation is thus fundamental in understanding human biology.

Gene expression patterns that are specific to certain cell types or tissues can help us to better understand normal human physiology (e.g., which biological processes occur in specific cell types or tissues) as well as development biology (e.g., which genes are expressed in specific cell types or tissues at various developmental stages), and several studies have investigated differences in these two contexts. Specifically, Pierson et al. [30] and Dobrin et al. [5] analyzed gene expression patterns at the tissue-level, revealing function-specific patterns and subnetworks associated with obesity. Similarly, McKenzie et al. [24] analyzed co-expression changes in different cell types of the brain, discovering significant cell type-specific expression signatures, while also finding well-known cell type marker genes among the most enriched genes across cell types.

Another relevant context is cell line information, as these are widely used for the study of biological processes. In particular, cancer cell lines, such as HeLa, are frequently employed, having had many interactions characterized on them and representing the foremost models for the study of cancer biology as well as numerous other disease and normal conditions. Nonetheless, even cell lines classified to the same tissue can exhibit significant differences in gene expression [19]. For example, a study by Yu et al. [46] found that certain cell lines may not resemble the primary cells from which they originated. The discrepancies in regulation patterns across specific cell lines deem it necessary to employ tools such as the CellExpress system (developed by Lee et al. [19] which enables the analysis of over 4000 cancer cell lines for differences in gene expression levels) and resources such as the TCGA-110-CL cell line panel [46] to identify which cell lines are more suitable for a given study.

Biological networks of different types can be used to represent patterns characteristic to a particular context. These context-specific networks can be categorized based on whether they are directly derived from knowledge or data. Rachlin et al. [31] and Stacey et al. [39] are two illustrations of knowledge-driven approaches where authors generated context-specific PPI networks by leveraging information about biological processes from GO (The Gene Ontology Consortium et al. [41]) and co-occurrence literature, respectively. Similarly, the analysis of transcriptomic data through the construction of gene co-expression networks (Langfelder et al. [18]) can also serve to better understand

context-specific patterns within datasets [28]. Finally, hybrid approaches, as demonstrated by Kitsak et al. [16], have leveraged gene expression data from 64 different tissues and mapped genes expressed in specific tissues to a protein–protein interactome, revealing that these disease context-specific genes tend to be located in close proximity within the interactome. It is important to note that while transcriptomic experiments are often used as a proxy to reflect protein expression, the correlation between the two is often below 0.5 on average [26, 40]. Nevertheless, correlations between genes whose mRNA is differentially expressed and their protein products have been shown to be significantly higher than genes whose mRNA is not differentially expressed, lending support to the use of differential mRNA expression to infer changes at the protein level [17].

One of the challenges in conducting these hybrid approaches (i.e., approaches that combine data- and knowledge- derived networks) is the limited availability of context-specific resources on a large-scale (e.g., hundreds of experiments conducted within the same or similar conditions or context-specific interactomes). While there are several co-expression databases dedicated to storing context-specific information, such as species [27] and [20], the vast majority of transcriptomic datasets are not annotated with context information and thus, cannot be systematically leveraged to conduct contextualized analyses on a large-scale. Nonetheless, the Gemma system [21] has been made available to provide thousands of curated datasets,thus, more easily enabling data reuse and secondary analyses.

In this work, we apply a network-based approach to investigate transcriptomic patterns observed in a variety of subcontexts classified under three major biological contexts (i.e., tissues, cell types, and cell lines) by leveraging over 600 gene expression datasets (Fig. 1A). To do so, we first construct co-expression networks that capture the strongest gene expression correlations observed in each subcontext (Fig. 1B). Subsequently, a series of network-based analyses are conducted to enable the exploration of the similarities and differences across co-expression networks and provide insights on gene co-expression patterns across contexts (Fig. 1C). Furthermore, we study the consensus between patterns identified in the co-expression network and a human protein–protein interactome as well as pathways knowledge. Finally, we present ContNeXt, a web application we have developed to enable researchers to explore and reuse our work.

Methodology

Gene expression datasets

We identified publicly available transcriptomic datasets from each of the three contexts evaluated (i.e., tissues, cell types, and cell lines) using Gemma, a manually curated database containing metadata for over 10,000 datasets [21, 48] (Fig. 1A). This metadata is programmatically accessible through Gemma's API (https://gemma.msl.ubc.ca/resources/restapidocs) and is annotated using different ontologies. Specifically, for each of the three contexts of interest, the following ontologies were used: (i) UBERON for tissues [25], (ii) Cell Ontology (CL) for cell types [4], and (iii) Cell Line Ontology (CLO) for cell lines [34].

Leveraging the metadata from Gemma, we were able to classify the samples from each dataset to their corresponding context(s). To guarantee the quality of the annotations, we conducted an additional manual curation step where we confirmed that the Gemma

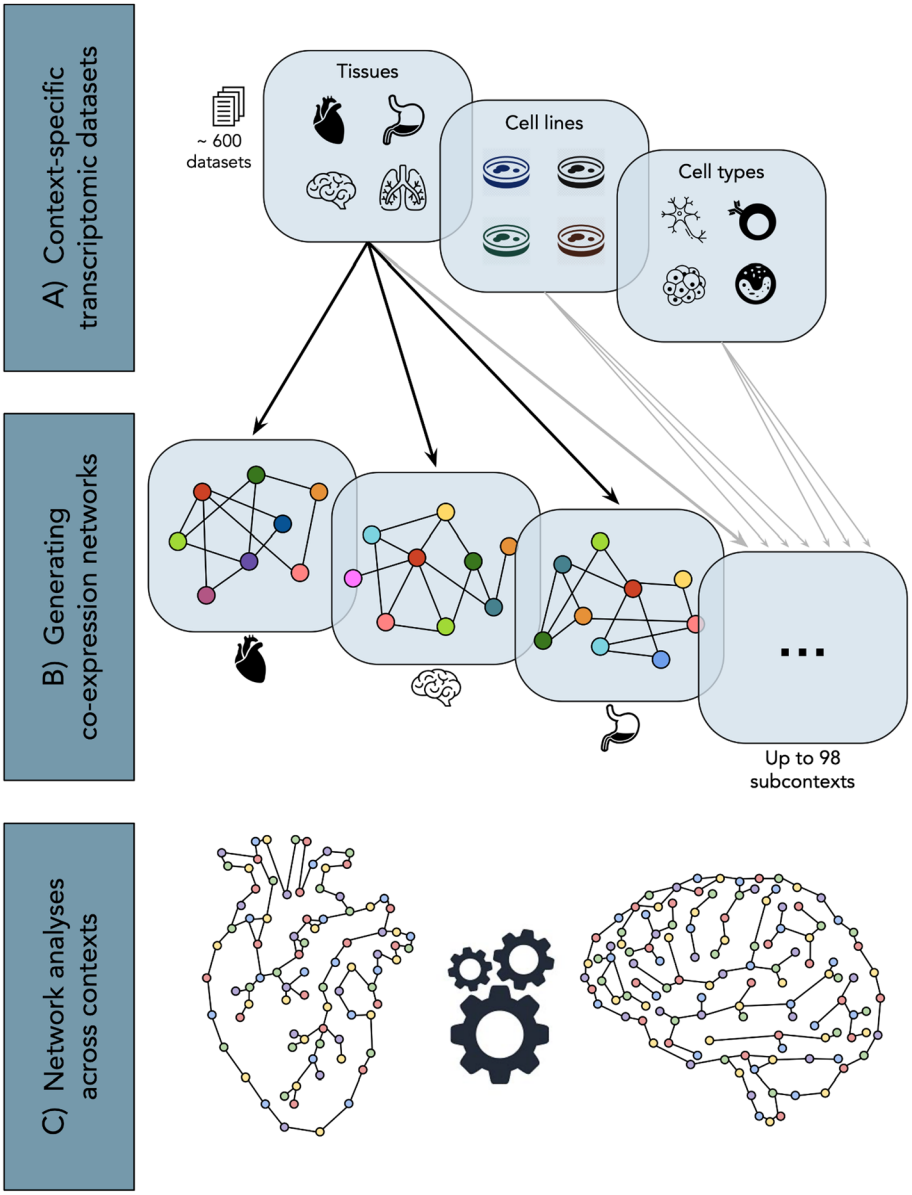

Fig. 1 Conceptualization of the presented study. **A** Over 600 context-specific transcriptomic datasets are collected and classified into 98 subcontexts (e.g., heart, astrocyte, and HeLa cell) under 3 major contexts (i.e., tissues, cell types, and cell lines), leveraging the Gemma database [21, 48] **B** Co-expression networks comprising the most strongly correlated edges observed in each subcontext are generated. **C** Network analyses provide insights on both common and unique patterns across the multiple contexts studied

sample annotations matched an ontology term for the given context present in the metadata, if available. Additionally, we filtered out samples that were not control or reference samples as our work focuses on comparing a normal physiological state in a variety of contexts. Finally, Gemma also includes annotations on dataset quality and samples that were annotated as unusable were excluded from our study.

After the initial annotation and curation steps, we implemented scripts for the downloading and processing of datasets found in Gene Expression Omnibus (GEO) [6]. While GEO incorporates several platforms, each measures different transcripts and requires a

dedicated pipeline, and merging data from several platforms is a complicated task which can introduce biases from probe sequences, arrays, or laboratory effects. Furthermore, conducting analyses combining raw data from multiple platforms can also introduce biases [33]. Thus, our work focuses on the most commonly used platform for humans, the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array platform (accession on GEO: GPL570). Out of 10,388 datasets in Gemma as of 22/04/2021, 9778 were filtered out while 610 remained for any one of the three contexts. In total, the tissue context was divided into 46 subcontexts, while the cell line and cell type contexts each contained 22 and 30 subcontexts, respectively (see Additional file 1: Tables S1–S3).

Generating co-expression networks from gene expression data

Co-expression networks were constructed using the WGCNA package in R (Langfelder et al. [18]). We followed the same procedure outlined in our previous work [9] to define the co-expression networks (Fig. 1B). This procedure focuses on the 1% highest similarity in the topological overlap matrix (TOM) to define the co-expression network for each subcontext; thus, facilitating the comparison of networks of the same size using a conservative cut-off in benchmark studies [29]. Given the platform used in this study, the most similar 1% in the TOM corresponds to 2,036,667 edges. We would like to note that the 1% cut-off is required as otherwise the networks would be fully connected, while we intend to focus only on the edges representing the most relevant transcriptomic patterns observed within each context. As edges representing a high topological overlap are also highly correlated in the TOM, we interchangeably refer to these edges as correlations for simplicity. Although this is not precise, the TOM value is based on the signed correlation but also takes the connectedness of nodes into account.

To run WGCNA, we used the raw expression data in the form of.CEL-files. Each dataset was individually pre-processed with the RMA function of the *oligo* R package to conduct background subtraction and quantile normalization. Next, we merged all samples from different datasets that belong to the same subcontext and applied batch correction using ComBat [14]. Regarding the mapping of the probes to genes, if there were multiple probes mapping to the same gene, we kept the most variable probe.

Protein—protein interaction network

We built a human protein–protein interactome as described in our previous work [9] as a knowledge template to compare against the co-expression networks generated. The interactome comprises interactions from well-established databases, including KEGG [15] and Reactome [13]. This network aims at representing the set of interactions that can occur in a physiological context, though it is worth mentioning that each of these interactions may not necessarily be occurring in a particular context at any given time.

Analyses

Controllability analysis

One of the more advanced techniques in analyzing networks is examining its controllability. We employed an algorithm developed by Liu et al. [22] which explores control theory to study the controllability of a directed network and thus identify driver nodes (i.e., the set of nodes that can offer control over the whole network) in order to classify each node and

edge in a network as indispensable, dispensable, or neutral. Ideally, minimizing the number of driver nodes offers adequate control over the network regarding the given biological system's dynamics. Using this algorithm, both nodes and edges can be classified as indispensable, dispensable, or neutral if their removal creates the need to increase, decrease, or cause no change in the number of driver nodes, respectively, so that controllability is maintained.

Pairwise co-expression network similarity

To evaluate similarity across co-expression networks, we calculated the overlap of edges across each pair of co-expression networks within a given context. Since all co-expression networks have the same number of edges, the number of shared edges between networks is readily comparable without the need to normalize values.

Similarity between co-expression networks and the interactome

We assessed the similarity of each co-expression network to the human interactome by calculating the number of shared edges. Here, it is important to note that edge directionality is ignored in the interactome since co-expression networks are inherently undirected. Furthermore, we evaluated the significance of the overlap by comparing the interactome to 1000 permuted co-expression networks. Permuted versions of the co-expression networks were created using the XSwap algorithm [12] (source code available at https://github.com/hetio/xswap), which ensures that the permuted versions preserve the structure of the original network (i.e., all edges are shuffled while maintaining the degree of each node).

Pathway—co-expression network similarity

To investigate the correspondence of transcriptomic signatures from co-expression networks with pathway knowledge, each of the context-specific co-expression networks were overlaid with pathways from KEGG [15]. The KEGG database was exclusively employed as it contains a feasible number of pathways for analysis (i.e., less than 350). For each gene set of a given pathway P from KEGG, we calculate every pairwise combination of nodes ($\boldsymbol{C_n}$) in P to determine the fraction of node combination pairs in $\boldsymbol{C_n}$ that exist as an edge in a given co-expression network $N = (n', E_N)$ where n' is the set of nodes in the co-expression network and E_N is the set of edges which connect the nodes n'. We term this the edge overlap, where $edge\ overlap = \left|\left\{\forall e_{u,v} s.t. (u,v) \in C_n \wedge u, v \in n' \wedge e_{u,v} \in E_N\right\}\right|$. The proportion of $\boldsymbol{C_n}$ that is in the edge overlap is the pathway-network similarity (Eq. 1). Using the pathway-network similarity, we create a similarity matrix with each network of a given context against every pathway from KEGG. This matrix is subsequently used to create a heatmap and hierarchical clustering of the co-expression networks is performed using Euclidean distances of their similarities to pathways.

Similarity between a pathway and co-expression network.

$$pathway - network\ similarity\ (P, N) = \frac{edge\ overlap}{|C_n|} \tag{1}$$

Implementation

Scripts to retrieve and process the datasets as well as to deploy the web application are available at https://github.com/ContNeXt. We have also provided comprehensive

documentation to modify the filtering steps and add extensions to the scripts. For network analysis and visualizations, we used the Python NetworkX library [11] (https://networkx.github.io/), and Matplotlib, and seaborn, respectively. The processed data used in this work is available at Zenodo at https://zenodo.org/record/5831786.

Results

In "Overview of co-expression networks and interactome" section, we provide an overview of the co-expression and PPI networks, while in "Analyses at the protein-level", "Analyses at the network-level" and "Mapping co-expression networks to pathway knowledge" sections, we outline each of the analyses conducted, specifically at the protein-, network-, and pathway- levels (Fig. 2). Finally, "ContNeXt—a web application to explore gene expression patterns across contexts" section presents ContNeXt, a web application developed to explore the results of this work.

Overview of co-expression networks and interactome

From 364, 222, and 103 (at times overlapping) datasets that were categorized into 46 distinct tissues, 30 distinct cell types, 22 distinct cell lines, respectively, we systematically constructed co-expression networks corresponding to each of these contexts. The exact breakdown of the number of datasets and samples for each subcontext can be found in Additional file 1: Tables S1–S3. Figure 3 summarizes the size of each corresponding co-expression network. We find that across different contexts, the collected data, which depends on the study objectives, is biased towards certain groups of related subcontexts. For instance, in the tissue context, a large number of subcontexts belong to tissues of the nervous system, while in the cell type context, the majority of subcontexts are related to the immune system. This bias can especially be seen in the cell line context, where nearly all cell lines are derived from cancer cells. Finally, we investigated the correlation between the number of samples or datasets used to generate the co-expression networks and the size of the networks as a potential source of bias. We found no such dependency

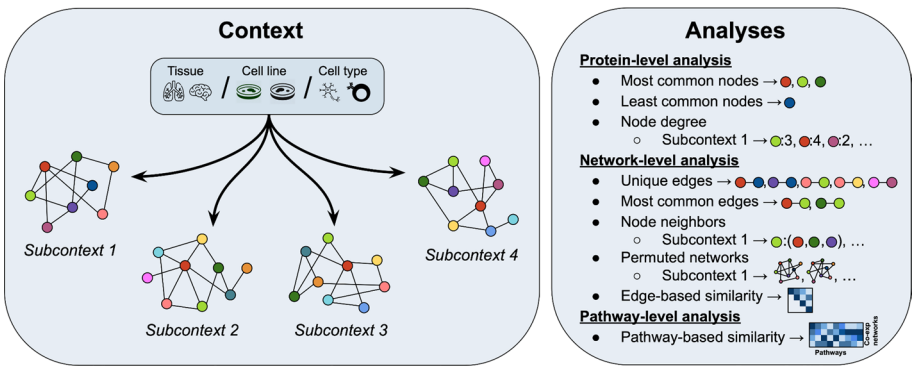

Fig. 2 Overview of analyses conducted across all subcontexts in three different contexts (i.e., tissues, cell lines, and cell types). At the protein-level, patterns surrounding each single node are investigated ("Analyses at the protein-level" section). The network-level analysis focuses on the relations between nodes (or node pairs) ("Analyses at the network-level" section) and the pathway-level analysis leverages defined node and edge sets to gain insights on context-specific co-expression networks ("Mapping co-expression networks to pathway knowledge" section)

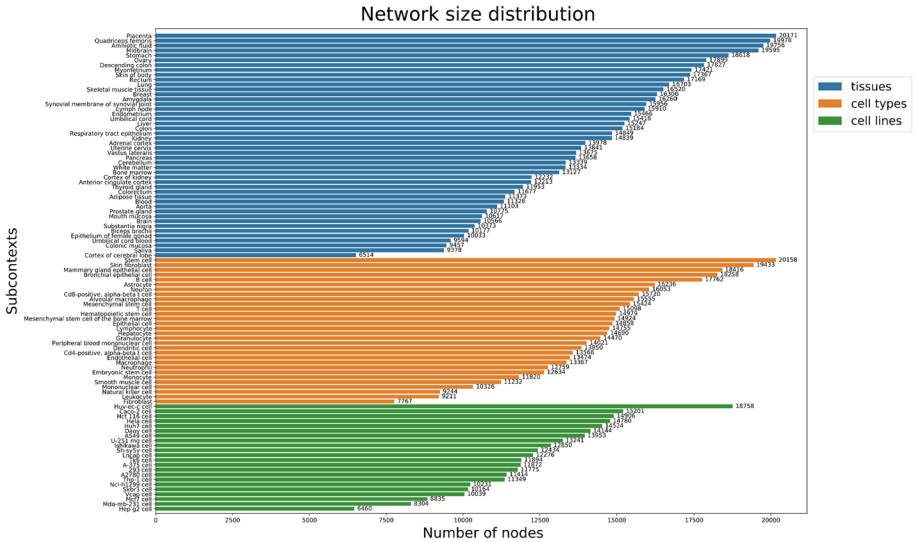

Fig. 3 Distribution of network size for each of the three contexts. Distributions of network size are given as the number of nodes in each subcontext. In the tissue context, the cortex of cerebral lobe network had the fewest number of nodes (i.e., 6514), while the placenta network had the largest number of nodes (i.e., 20,171) across not only all networks of the tissue context, but also across all other contexts. In the cell type context, the fibroblast network had the least number of nodes (i.e., 7767), while the stem cell network had the highest number of nodes (i.e., 20,158). In the cell line context, the HepG2 cell line network had the least number of nodes (i.e., 6460), while the Huv-ec-c cell line network had the largest number of nodes (i.e., 18,758). Generally, the networks within each context tended to vary greatly in size. For example, the tissue context includes networks ranging in size from 6514 to 20,171 nodes

between the number of samples or datasets and the network size (Additional file 2: Fig. S1).

The human interactome we employed (see Methods "Protein—protein interaction network" section), generated in our earlier work [9], contains 8601 nodes and 199,535 edges. These numbers place our interactome on the same scale as other, recently published human interactomes [23], Vinayagam et al. [42]. Nonetheless the size of the interactome, with regard to the number of nodes (proteins), is less than half of the largest co-expression network. This was to be expected, as the majority of proteins measured in transcriptomic experiments have not yet been investigated in the literature and little is known of their functionality. Nodes of the interactome can be visualized in the web application (see "ContNeXt—a web application to explore gene expression patterns across contexts" section) along with their neighbors, betweenness centrality, degree centrality, controllability classification, and information on whether the node is a housekeeping gene.

In order to discern unique features of context-specific co-expression networks which could be of biological significance, we first sought to identify genes known to arise from generic processes whose patterns are more likely to be stable and unaffected by any given context or condition. In particular, we investigated the presence of these, so called, housekeeping genes in each of the co-expression networks, noting that these genes are indicative of shared biology given their role in cell maintenance, and therefore, exhibit constant expression levels across all cells and conditions (Eisenberg and Levanon [7]). Thus, by better understanding which genes have critical roles in basic cell maintenance,

we could better direct our focus in determining genes of interest. The housekeeping genes dataset made available from Eisenberg and Levanon [7] consisted of 3804 genes (Additional file 1: Table S4), 1723 of which were present in the interactome (20% of the overall interactome).

To analyze the structural properties of the interactome, we employed an algorithm (see Method) that has been applied to identify the importance of nodes and edges in biological networks (Additional file 2: Text S1). The results of the controllability analysis indicate that the interactome has 1233 driver nodes with which the network can be controlled. Overall, 74.6% of the nodes were classified as neutral, 16.17% dispensable, and 9.2% indispensable. A list of the full classifications can be seen in Additional file 1: Table S5, with the indispensable nodes listed in Additional file 1: Table S6, and a summary of these nodes can be seen in Table 1. We observed that the indispensable nodes were highly connected, as expected, had the highest average betweenness centrality, and a significant portion (i.e., ~ 25%) were housekeeping genes. By comparison, neutral nodes were found to have half as many connections and an average betweenness centrality 10 times lower than indispensable nodes. However, the proportion of neutral nodes that were housekeeping genes were comparable to that of the indispensable nodes. By contrast, differences between the dispensable and indispensable nodes were far more pronounced; the average degree of dispensable nodes was only ~ 6, compared to ~ 107 for indispensable nodes, while the average betweenness centrality was more than 1000 times lower. Additionally, only ~ 8% of dispensable nodes were housekeeping genes, compared to roughly a quarter for both indispensable and neutral nodes.

Analyses at the protein-level

We begin by exploring general trends for all co-expression networks of each context at the protein-level by focusing on the most and least common proteins (i.e., present in all or exactly one network within a context). We first used the results of the previously-mentioned controllability analysis of the interactome as well as housekeeping

Table 1 Regarding the interactome controllability, 6417 of the total nodes (74.6%) were classified as neutral; i.e., removing them will have no effect on the number of driver nodes in the network, representing the largest proportion of nodes in the interactome. 1391 (16.17% of the interactome) nodes were dispensable, meaning their removal would decrease the number of driver nodes in the network. Lastly, 793 nodes (9.2% of the interactome) were determined to be indispensable, which caused an increase in the need for driver nodes at their removal. In all three categories (i.e., betweenness centrality, degree, and housekeeping gene proportion), indispensable nodes had the highest value, followed by neutral, and dispensable with the lowest values

	Total number	Scaled betweenness centrality mean	Scaled betweenness centrality median	Scaled betweenness centrality mode	Degree mean	Degree median	Degree mode	Proportion housekeeping gene (%)
Indispensable	793	0.024519	0.006825	0.002642	107.08	60	29	24.59
Dispensable	1391	0.000019	0.00000	0.00000	6.44	4	1	7.84
Neutral	6417	0.004090	0.001101	0.00000	47.56	31	13	22.11

The indispensable nodes are listed in Additional file 1: Table S6. Betweenness centrality scores were scaled between 0 and 1 to facilitate comparability

genes and overlapped them with the most and least common proteins in each context, shown in Additional file 1: Table S7. As summarized in Table 2, of the most common nodes (i.e., proteins that could be found in each network within a given context), we found that the cell type context had the largest number of proteins across all networks (301 proteins), while the tissue network had the fewest (22 proteins). Among the most common nodes, the ratio of housekeeping genes was greater than the proportion of housekeeping genes present in the interactome (i.e., 20%), comprising nearly 50% of the most common nodes in each of the contexts.

Overlap of co-expression networks with the interactome

While only considering the proteins present in the interactome as well as at least one co-expression network, we conducted an in-depth investigation of whether proteins in the co-expression networks of a given context could consistently be identified in the human interactome network. We first noted trends at the protein-level by comparing the most and least common proteins across co-expression networks within a context against the most and least connected proteins of the interactome. As the co-expression network and interactome sizes vastly differed, we studied this overlap considering the top or bottom most proteins in proportions roughly equivalent in size. We selected various cut-offs for each context, corresponding to the number of co-expression networks (see Additional file 2: Text S2 for details on the cut-offs for each context). This ensured the inclusion of either the maximal or minimal possible overlap of the common proteins of the co-expression networks and connected proteins of the interactome, depending on whether our investigation focused on the most commonly or most uniquely occurring proteins, respectively. A detailed list of the resulting overlaps can be seen in Additional file 1: Table S8.

Table 2 Most and least common proteins per context. The most and least common proteins of the co-expression networks (i.e., in all or exactly one network within a context) were overlapped with proteins given distinct classifications from the controllability analysis of the interactome as well as with housekeeping genes. 22 proteins were identified as the most common proteins, that is, found in all 46 co-expression networks of the tissue context. Of the 30 co-expression networks of the cell type context, 301 proteins were found in all of them, while among 22 co-expression networks in the cell line context, 185 proteins were identified in each network By comparison, no proteins were found to be unique to a single co-expression network in the tissue context, while only one was found in the cell type context belonging to the stem cell co-expression network. On the other hand, 106 least common proteins were found in the cell line context, only one of which is a housekeeping gene and none of which are indispensable

	Tissue context		Cell type context		Cell line context	
Proteins in all co-expression networks	22	2 indispensable 11 housekeeping	301	21 indispensable 180 housekeeping	185	15 indispensable 81 housekeeping
Proteins unique to one co-expression network	0	0 indispensable 0 housekeeping	1	0 indispensable 0 housekeeping	106	0 indispensable 1 housekeeping

A full list of the proteins found in all or in a single network per context can be seen in Additional file 1: Table S7

Most common proteins

First, we focus on the most common proteins. Among the most commonly occurring proteins in the tissue context that overlapped with proteins from the interactome, a number of proteins belonged to the MAPK protein family (Additional file 1: Table S8). Proteins in this family are instrumental in transduction of extracellular signals to cellular responses and complex cellular processes such as apoptosis, development, differentiation, proliferation, and transformation [47]. While only the larger two comparisons in the tissue context (Additional file 2: Fig. S2; lower two diagrams) resulted in an overlap, a significant portion of these overlapping proteins were also indispensable, or housekeeping. Within the large overlaps between the common cell type proteins and most connected interactome proteins (Additional file 2: Fig. S3), a larger proportion of housekeeping genes was found than in any of the contexts studied, with more than half of each overlap being a housekeeping gene (i.e., 50–67%), and more of the proteins are also indispensable.

In cell lines, we observed a substantial overlap of most common proteins that are also found in the interactome overall, including when using the strictest cut-offs, however, significantly less were found to be indispensable or housekeeping than in the tissue and cell type contexts (Additional file 2: Fig. S4). We select a proportional set from each context (400 of the most common proteins per context) to compare their overlaps with the interactome (Additional file 1: Table S9A). The overlaps all had a similar number of proteins in them, between 30 and 37 proteins. Across contexts, there was a similar proportion of the overlap which are indispensable nodes of the interactome (~ 32% in tissues, 40% in cell types, and ~ 43% in cell lines). On the other hand, the proportion of housekeeping genes varied more, with 43% of the proteins from the cell line overlap, while tissues and cell types both had more than 60%. Overall, housekeeping genes seem to be best represented in the co-expression networks. We observed a number of proteins in all of the context's overlaps belonging to the Ribosomal protein (RP) family (Additional file 1: Table S9A), from both small and large subunits. RPs are essential in protein synthesis [45]. The tissue overlap had one from large and one from small subunit, the cell type overlap had four from large and one from small subunit, and the cell line overlap had one small subunit RP. We also found that the average number of relations for the proteins in the interactome that overlapped with the approximately top 400 most common proteins in the tissue and cell line networks (~ 73 and ~ 72 relations, respectively), was much higher than the average number of relations overall in the interactome (~ 46 relations). This suggests that the common tissue- and cell line-wide proteins across the co-expression networks are better represented in the scientific literature. In the cell type networks, this average was less high, ~ 60 relations, but still more than overall in the interactome.

Least common proteins

Next, we investigated the least common proteins in the co-expression networks and their overlap with the least connected proteins in the interactome. This time, the tissue context presented a more consistent overlap while increasing the protein pool, but still a minimal overlap (Additional file 2: Fig. S5). The overlap with the interactome and the cell

type context was about the same as in the tissue context (Additional file 2: Fig. S6). In the cell line context, we found a small, steadily increasing overlap with each interval comparison, which was not the case in the most common proteins (Additional file 2: Fig. S7). The overlap with the interactome in the larger comparisons was roughly the same as in every other context. The minimal overlaps suggest that little is currently known of these proteins. Additionally, we also selected proportional sets of the 400 least common proteins in each context, also occurring the interactome overall against the 400 least connected nodes of the interactome (Additional file 1: Table S9B). The sizes of the overlap didn't vary as much as in the most common and connected comparison, with each context having around 30 proteins in the overlap. As expected, with these overlaps, either one or no proteins are also indispensable or housekeeping. We observe an overwhelming number of proteins belonging to the ZNF protein family in each of the overlaps (i.e., 10/34 (29%) in tissues, 11/33 (33%) in cell types, and 4/27 (15%) in cell lines) (Additional file 1: Table S9B). While ZNFs are widely found in the organism, they play critical roles in specific tissues, and in the development of many diseases [2].

Analyses at the network-level

We first focused on analyzing edges of the co-expression networks, including the unique and most commonly occurring edges within contexts. Additionally, we leveraged prior knowledge from a referential human interactome and studied the correspondence of edges from this network against the strongest pairwise correlations of the co-expression networks. Subsequently, we validated these findings by conducting an equivalent comparison against randomly generated versions of the co-expression networks. Finally, we conducted a similarity analysis on the network edges within each context.

Unique and most commonly occurring edges

We first assessed whether there were any edges specific to particular tissue networks, identifying 45,963,343 unique edges in total (i.e., 49% of all edges). We also identified 34,584,720 unique edges in the cell type context (i.e., 57% of all edges) and 31,941,789 unique edges in the cell line context (i.e., 71% of all edges). These proportions are similar to findings by Stacey et al. [39] who found that over half of edges in several PPI databases are context-specific. Figure 4 illustrates the frequency of unique and common edges in all networks within a context. We find that edges which are common to at least 25% of networks within a context are rare (i.e., between 0.07 and 0.16%), while those which are in at least 75% of networks are nearly negligible (i.e., 33 edges in total for tissues, 9 for cell types, and 4 for cell lines). As only the 1% strongest correlations were selected for each network, it was foreseen that a large number of edges in our resulting co-expression networks would be specific to a single subcontext. Although these unique edges are interesting to explore for a given subcontext (green portions in Fig. 4), given the sheer volume of unique edges, their investigation was outside of the scope of this work.

We hypothesize that these common edges correspond to basal correlations that are not specific as they appear in the majority of networks within one or more contexts. Thus, we analyze the most frequently occurring edges in each of the three contexts. Unsurprisingly, the two housekeeping genes of the tubulin alpha families (i.e., TUBA1C and TUBA1B) are nearly always found to be connected to each other (in 83 out of 98

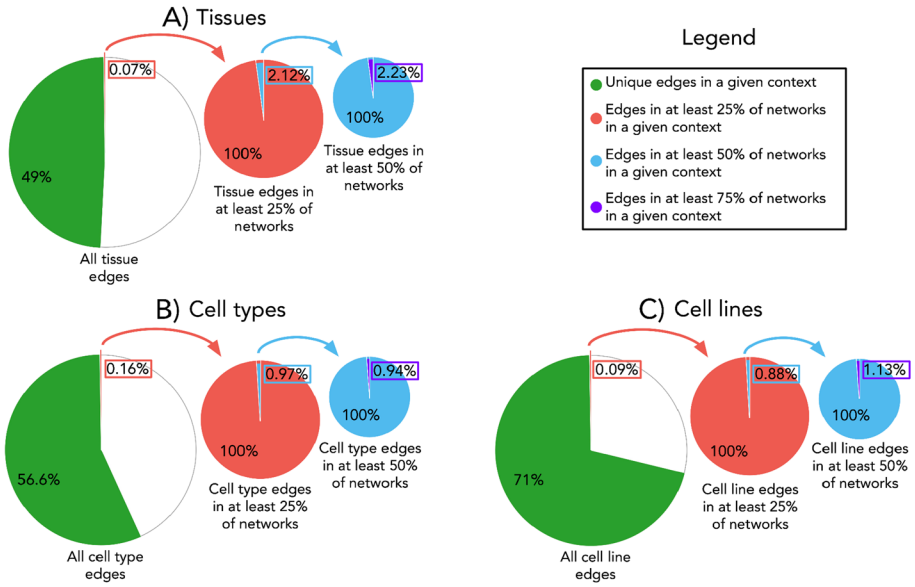

Fig. 4 Frequency of edge occurrence across networks within a context. Proportions of edges are given as those that are unique, or common to varying degrees, in networks within the **A** tissue, **B** cell type, and **C** cell line context. From the total set of edges that occur across all networks within each context, the fraction of edges that are unique (i.e., appear in at most one network within a given context) are shown in green. From this total set of edges, the fraction of those which appear in at least 25% of networks within a given context are magnified in a consecutively smaller pie chart (i.e., predominantly in red). Similarly, those which appear in at least 50% of networks within a given context are magnified and illustrated in a pie chart predominantly in blue. Finally, of this latter group of edges, the fraction of edges that are most common (i.e., appear in at least 75% of all networks within a given context) are highlighted in purple

networks), regardless of context. Additionally, IFITM2 and IFITM3, proteins of the interferon-induced transmembrane family, which play a key role in immune system functions, are also often seen connected to each other in 84 out of 98 networks. Members of the human leukocyte antigens (HLA) protein family are also often interconnected across the cell type and cell line contexts. This is in line with Crow et al. [3] who found that certain gene modules are predictably found across biological conditions, such as those of the immune response. In our previous paper [9], we found that of the most common edges among 63 major diseases, members of the Metallothionein (MT) family of proteins, were in nearly half of these edges. Similarly, here again we observed that a large number of MT proteins share neighbors across networks in every context.

Of the most common edges throughout all contexts (see Additional file 2: Text S3), none were indispensable within the interactome. When widening our search to the top 100,000, we found only seven, three, and one edge in the tissue, cell type, and cell line contexts to be indispensable in the interactome, respectively. Next, these most common edges found in the majority of networks of a given context were compared to the interactome network to identify concordance between the two. We performed a range of comparisons on the most common edges by focusing only on the top 1000 to 10,000 edges, in increments of 1000. Then, the most common edges in each co-expression network were compared to the interactome. Overall, we found little overlap in the most common edges. In the tissue context, we found an overlap of only 5% in the top 1000 most common edges against the interactome, with this overlap decreasing to 4% when considering

the top 10,000 most common edges. In comparison these proportions ranged from ~ 7 to 3% in the cell type context between the top 1000 and 10,000 most common edges, and 4% to 2% in the cell line context.

The strongest correlations tend to correspond with protein–protein interactions more than expected by chance

In this section, we investigate whether the strongest correlations present in the co-expression networks correspond to PPIs more often than what would be expected by chance. For this purpose, we permuted each co-expression network for each context 1000 times while maintaining the original graph structure (see Methods). We next compared the overlap of edges between these permuted co-expression networks with the human interactome (the results of the first 100 permutations can be seen in Additional file 1: Table S10). Our results show that, on average, the original co-expression networks have 1.55 times as many edges in common with the human interactome as compared to the permuted networks, which exhibited a comparatively low variability in their overlap within a subcontext. Across all contexts, the maximum difference in overlap was for the ovary subcontext, where the original ovary co-expression network had 3.3 times as many edges in common with the interactome as compared to the permuted versions. In comparison, the saliva co-expression network showed the smallest difference in edge overlap between the original and permuted co-expression networks, with the overlap of the interactome with the original co-expression network having only 1.01 times as many edges as the permuted versions on average. Thus, we find that co-expression patterns correspond with PPIs more than expected by chance.

Edge-based similarity across co-expression networks

Next, we investigated edge similarity across networks within a given context. By comparing the co-expression networks to each other rather than just the interactome, we could identify the networks that were most similar edgewise. In the tissue context, two pairs of networks displayed the highest degree of similarity, namely the brain and the cortex of the cerebral lobe, and the colon and the rectum (Fig. 5A). This finding was not surprising given that these pairs of tissues are anatomically related (i.e., both are of the brain or the colorectum). The cell line context had a few standout pairs of networks which had the highest degree of similarity (Fig. 5B). Specifically, the highest similarity was between two different human breast cancer cell lines: MDA-MB-231 and MCF7. Additionally, the MCF7 cell line again had a high similarity with a human colon cancer cell line, HCT 116. On the other hand, in the cell type context, rather than specific pairs showing the highest similarity with each other, a few selected subcontexts had a high similarity with most of the other networks overall (Fig. 5C). In particular, the peripheral blood mononuclear cell network showed high similarity with its more specific cell type networks, including monocytes, T cells, and lymphocytes. Overall, these results lend support to how network similarity can reflect similarity across related cell types, tissues, or cell lines.

Mapping co-expression networks to pathway knowledge

Lastly, we attempted to establish patterns across co-expression networks at a pathway-level by overlaying pathway knowledge with the co-expression networks. If a given

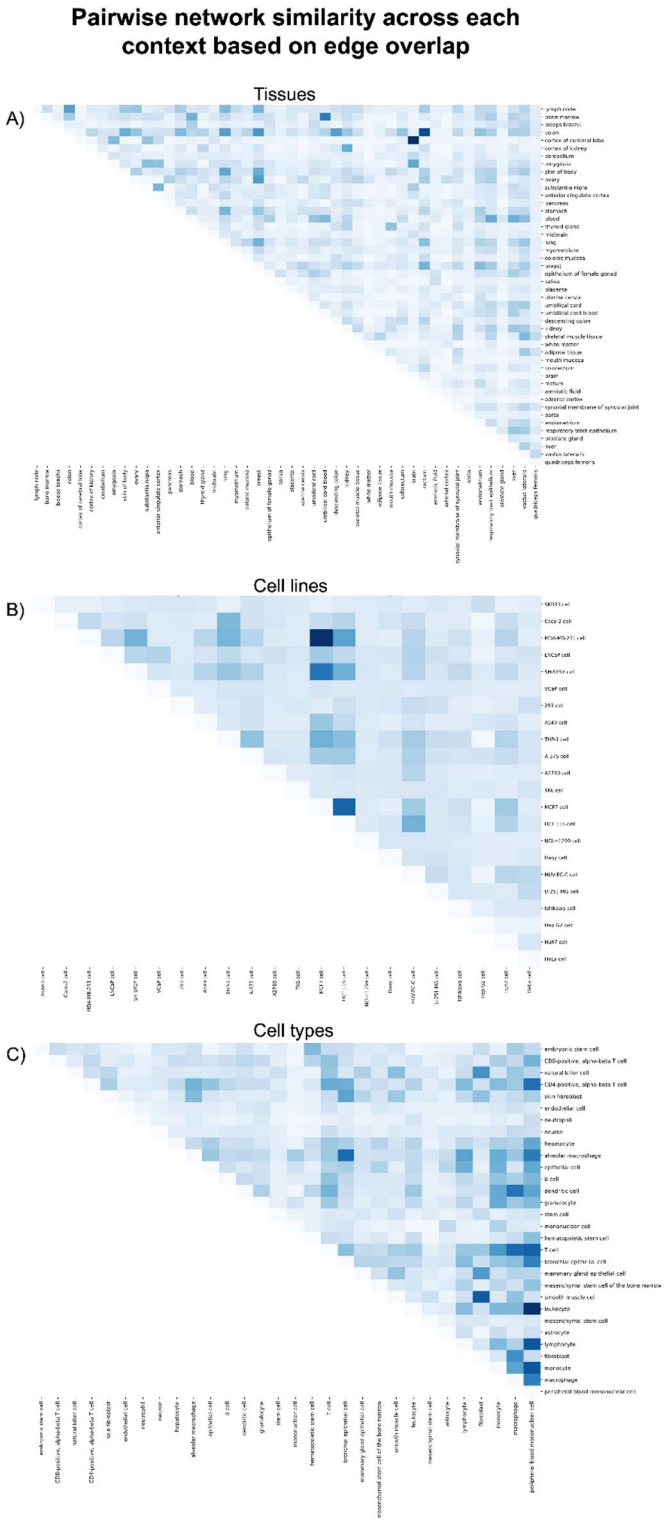

Fig. 5 Pairwise co-expression network similarity across contexts. For each pair of co-expression networks within a given context, edge overlap was calculated as a measure of similarity between networks for the **A** tissue, **B** cell line, and **C** cell type contexts. A high quality version of the figure is available at https://github.com/ContNeXt/scripts/blob/main/figures/figure5.pdf

pathway is related to a specific network (e.g., fatty acid metabolism pathway and the liver co-expression network), we would expect that the proteins in the pathway would be strongly correlated in the co-expression network. Furthermore, we assume that, given a set of highly co-expressed genes of which a majority are involved in a particular pathway, the remaining genes may be functionally relevant to the pathway as well. We therefore seek to identify the pathways associated with networks from each of the investigated contexts. Using the KEGG database [15], we mapped pathway knowledge to co-expression networks according to Eq. 1 (see Methods).

We found several groups of tissues that had high similarities with pathways related to the given tissues (Fig. 6). For instance, the two tissue networks corresponding to cortex of cerebral lobe and brain shared a large group of pathways exhibiting a high degree of similarity, including nine synaptic pathways (Fig. 6; green oval) (Additional file 1: Table S11). Furthermore, the three networks for liver, cortex of kidney, and kidney also had the highest level of similarity with numerous pathways, including eight involving the regulation of fatty acids as well as 11 involving amino acid metabolism and degradation (Fig. 6; red oval) (Additional file 1: Table S12). Not surprisingly, the adipose tissue network also showed the highest similarity with adipose-related pathways, such as adipocytokine signaling pathway and regulation of lipolysis in adipocytes pathway.

In the cell type context, while no groups of network shared distinct pathways among them, we found three cell types having distinct groups of pathways with very high similarity unique to a single network. For example, a number of pathways showed a high degree of similarity to the neutrophil co-expression network (Additional file 2: Fig. S8; red oval), namely, 11 that regulate the immune response (Additional file 1: Table S13). Additionally, the co-expression network for hepatocytes, the primary cell type of the liver, had the highest level of similarity with many pathways (Additional file 2: Fig. S8; yellow oval), including six involving basic liver function as well as many metabolic pathways, particularly 10 pertaining to amino acids metabolism and seven for other specific molecules (Additional file 1: Table S14). Lastly, we found an additional group of pathways that were exclusively similar to one network, namely the neuron (Additional file 2: Fig. S8; green oval). Specifically, this included five pathways related to neurotransmitter systems, long-term depression, and pathways related to addiction (Additional file 1: Table S15).

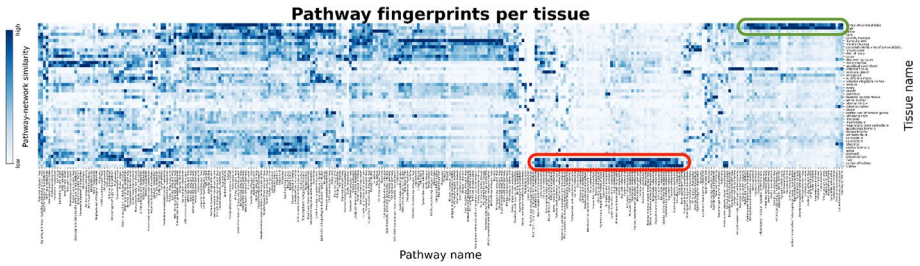

Fig. 6 Similarity between tissue-specific co-expression networks and KEGG pathways. The similarity between a particular pathway and a co-expression network is defined as the percentage of pairwise combinations of proteins of a given KEGG pathway that can be found in a co-expression network as edges. Light blue corresponds to a lower similarity, while dark blue corresponds to a high similarity. A high quality version of this figure is available at https://github.com/ContNeXt/scripts/blob/main/figures/figure6_highquality.pdf and can also be visualized in the web application

Analogous to the cell type context, while related groups of networks from the cell line context were not found to be similar to related groups of pathways (Additional file 2: Fig. S9), several individual cell lines were observed to be highly similar to a group of pathways. However, these pathways were not necessarily unique to the cell line, showing some similarity with other cell lines as well. Interestingly, we found a large group of pathways (i.e., 70 in total) with consistently high similarity with nearly all cell lines, with the exception of the THP-1 cell line (Additional file 2: Fig. S9; green rectangle). These include 24 different signaling pathways and 16 different cancer pathways (Additional file 1: Table S16). Notably, we found a group of pathways that were distinctly similar to two cell lines (i.e., A549 and TK6). Specifically, 14 pathways showed a high degree of similarity to the A549 cell line co-expression network (Additional file 2: Fig. S9; yellow oval). This cell line originated from adenocarcinomic human alveolar basal epithelial cells from lung cancer and is used as a model for drug metabolism [10]. Of these 14 pathways that, on average, showed the highest similarity to this cell line relative to the others, eight were pathways involving metabolism and three were pathways related to compound biosynthesis (Additional file 1: Table S17). Similarly, we identified a group of pathways which showed a higher similarity to the TK6 cell line, originating from a human B lymphoblastoid cell [36], over all other cell lines (Additional file 2: Fig. S9; red oval), including five signaling pathways (Additional file 1: Table S18).

ContNeXt—a web application to explore gene expression patterns across contexts

To provide access to the co-expression networks and analyses presented in this work, we have developed ContNeXt, a web application that facilitates the large-scale exploration and analysis of transcriptomic patterns across multiple contexts. The main page of the web application allows users to search co-expression patterns for a given node in a particular context or browse and query specific nodes in a certain subcontext (Fig. 7A). With interactive network visualizations, users can explore these patterns and employ functionalities such as filtering or search boxes (Fig. 7B). Similarly, the heatmaps presented in this work can be interactively explored through the web application (Fig. 7C). Finally, both the processed data and networks can be downloaded directly from the web application.

Discussion

We have presented a large-scale network-based approach that aims at revealing common and specific biological processes and mechanisms across contexts by identifying transcriptional patterns that are unique to various cell types, tissues, and cell lines, as well as patterns which are consistent across them. In order to do so, we constructed co-expression networks to capture the strongest correlations observed in 98 specific subcontexts belonging to these three biological contexts (i.e., tissues, cell types, and cell lines) and conducted a series of analyses at the protein, network, and pathway levels. Finally, we developed a web application to enable users to query and display these networks and ultimately, explore shared and distinct co-expression patterns for multiple contexts.

We believe that one strength of our work is its robustness, as we have systematically leveraged hundreds of curated datasets, thereby ensuring a diverse sample of experiments conducted in similar settings whilst applying a common preprocessing and

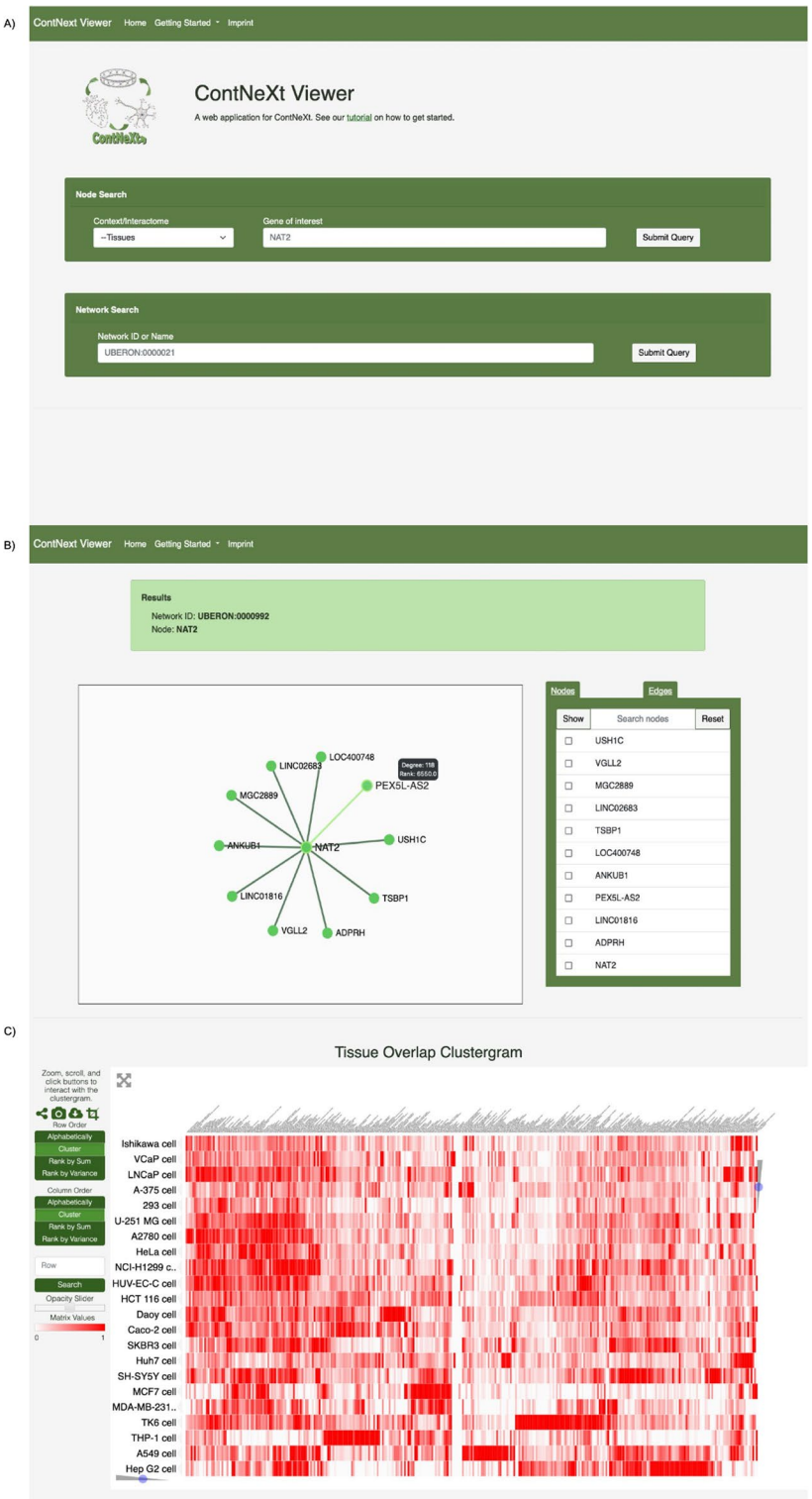

Fig. 7 ContNeXt web application. **A** Main page. Users can query for specific genes or directly explore the networks of a given context. **B** Network page. Users can explore and navigate through the neighbors of a specific gene for each network. **C** Heatmap visualization. Heatmaps presented in this work can be interactively viewed to investigate pairwise co-expression network -based similarity as well as pathway- co-expression network -based similarity

analysis pipeline. However, although we applied a conservative inclusion/exclusion criteria, we cannot assume that every dataset in the same (sub)context is equivalent and thus, some of the patterns identified may be dataset-specific. To account for this factor and reduce noise and variability across datasets, we focused on the 1% strongest correlations, keeping in mind that the choice of cut-off can influence the resulting co-expression network [44], and also constrained our analysis to subcontexts with a large number of samples. Still, independently of this minimum criteria, there are differences in the number of datasets per subcontext that could lead to variability for specific subcontexts with a small sample size. Another limitation is that we have exclusively relied on the platform with a large number of datasets in the Gemma database. Similarly, we also employed Gemma's context annotations to classify the datasets. While it is technically possible to include more platforms in our analysis as well as annotate datasets from other databases, each additional platform would require its own independent processing pipeline and a significant curation effort. Furthermore, in the cell line context, it is important to note that the majority of cell lines originate from widely used immortal cancer cell lines, which might differ from the normal human cells used for the cell type and tissue contexts. Finally, we would like to remark on two other limitations of our analysis. Firstly, while we employed a large and high-quality version of the protein–protein human interactome, some parts of the graphs are more dense than others as some proteins are under-studied [35]. Secondly, some of the analyses are influenced by the size of the co-expression networks (Fig. 3), as the fewer nodes a network has, the more dense it is due to the larger amount of connections between its nodes.

Lastly, we would like to mention some of the prospects we foresee for future work. Firstly, by further incorporating single-cell experiment datasets, we can potentially identify more granular patterns. Additional single-cell RNA-seq datasets can be included in our work to verify whether the observed tissue-specific transcriptional patterns are indeed characteristic to specific tissues, or are influenced by their cellular composition, as observed by Farahbod and Pavlidis [8]. While this large-scale exercise is not feasible at the moment due to the lack of available data of this kind, we expect that it could be conducted in future. Secondly, disease-specific gene expression datasets can be exploited to compare disease-specific signatures with the ones observed in a related normal tissue or cell type in order to identify the biological processes and pathways that are dysregulated in the disease context. Thirdly, as demonstrated by Azevedo et al. [1] and Sealfon et al. [37], machine learning models could be trained on the generated co-expression networks to classify signatures coming from new samples into a particular context given its specific characteristics.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04765-0.

Additional file 1 Supplementary Tables.

Additional file 2 Supplementary File including supplementary figure and tables.

Acknowledgements
We would like to thank the entire Gemma team, especially Paul Pavlidis, for their support using their tool. Furthermore, we would like to thank André Gemünd for his technical assistance.

Author contributions
DDF and SM conceived and designed the study. RQF and TR processed the transcriptomic datasets. RQF implemented the methodology and analyzed the results supervised by SM and DDF. SDS implemented the web application. ATK, MHA and DDF acquired the funding. RQF, SM, and DDF wrote the manuscript. ATK and MHA reviewed the manuscript. All authors read and approved final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was developed in the Fraunhofer Cluster of Excellence "Cognitive Internet Technologies". This work is supported by the German Federal Ministry of Education and Research (BMBF, grant 01ZX1904C).

Availability of data and materials
All data supporting the conclusions of this article are available at https://zenodo.org/record/5831786 and scripts can be found at https://github.com/ContNeXt/scripts. ContNeXt and its source code are available at https://contnext.scai.fraunhofer.de and https://github.com/ContNeXt/web_app, respectively.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
DDF received salary from Enveda Biosciences. Other authors do not declare any competing interests.

Received: 11 February 2022 Accepted: 3 June 2022
Published online: 15 June 2022

References

1. Azevedo T, Dimitri GM, Lió P, Gamazon ER. Multilayer modelling of the human transcriptome and biological mechanisms of complex diseases and traits. NPJ Sys Biol Appl. 2021;7(1):1–13. https://doi.org/10.1038/s41540-021-00186-6.
2. Cassandri M, Smirnov A, Novelli F, Pitolli C, Agostini M, Malewicz M, et al. Zinc-finger proteins in health and disease. Cell Death Discov. 2017;3(1):1–12. https://doi.org/10.1038/cddiscovery.2017.71.
3. Crow M, Lim N, Ballouz S, Pavlidis P, Gillis J. Predictability of human differential gene expression. Proc Natl Acad Sci. 2019;116(13):6491–500. https://doi.org/10.1073/pnas.1802973116.
4. Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, et al. The cell ontology 2016: enhanced content, modularization, and ontology interoperability. J Biomed Semant. 2016;7(1):1–10. https://doi.org/10.1186/s13326-016-0088-7.
5. Dobrin R, Zhu J, Molony C, Argman C, Parrish ML, Carlson S, Allan MF, Pomp D, Schadt EE. Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. Genome Biol. 2009;10(5):1–3. https://doi.org/10.1186/gb-2009-10-5-r55.
6. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10. https://doi.org/10.1093/nar/30.1.207.
7. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. 2013;29(10):569–74. https://doi.org/10.1016/j.tig.2013.05.010.
8. Farahbod M, Pavlidis P. Untangling the effects of cellular composition on coexpression analysis. Genome Res. 2020;30(6):849–59. https://doi.org/10.1101/gr.256735.119.
9. Figueiredo RQ, Raschka T, Kodamullil AT, Hofmann-Apitius M, Mubeen S, Domingo-Fernández D. Towards a global investigation of transcriptomic signatures through co-expression networks and pathway knowledge for the identification of disease mechanisms. Nucleic Acids Res. 2021;49(14):7939–53. https://doi.org/10.1093/nar/gkab556.
10. Foster KA, Oster CG, Mayer MM, Avery ML, Audus KL. Characterization of the A549 cell line as a type II pulmonary epithelial cell model for drug metabolism. Exp Cell Res. 1998;243(2):359–66. https://doi.org/10.1006/excr.1998.4172.
11. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th Python in Science Conference (SciPy2008); 2008. Pp. 11–5.
12. Hanhijärvi S, Garriga, GC, Puolamäki K. Randomization techniques for graphs. In: Proceedings of the 2009 SIAM International Conference on Data Mining; 2009. pp. 780–91. https://doi.org/10.1137/1.9781611972795.67
13. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, D'Eustachio P. The reactome pathway knowledgebase. Nucleic Acids Res. 2020;48(D1):D498-503. https://doi.org/10.1093/nar/gkz1031.
14. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27. https://doi.org/10.1093/biostatistics/kxj037.
15. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021;49(D1):D545–51. https://doi.org/10.1093/nar/gkaa970.
16. Kitsak M, Sharma A, Menche J, Guney E, Ghiassian SD, Loscalzo J, Barabási AL. Tissue specificity of human disease module. Sci Rep. 2016;6(1):1–12. https://doi.org/10.1038/srep35241.
17. Koussounadis A, Langdon SP, Um IH, Harrison DJ, Smith VA. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. Sci Rep. 2015;5(1):1–9. https://doi.org/10.1038/srep10775.
18. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform. 2008;9(1):1–13. https://doi.org/10.1186/1471-2105-9-559.

19. Lee YF, Lee CY, Lai LC, Tsai MH, Lu TP, Chuang EY. Cell Express: a comprehensive microarray-based cancer cell line and clinical sample gene expression analysis online system. Database. 2018. https://doi.org/10.1093/database/bax101.
20. Lee J, Shah M, Ballouz S, Crow M, Gillis J. CoCoCoNet: conserved and comparative co-expression across a diverse set of species. Nucleic Acids Res. 2020;48(W1):W566–71. https://doi.org/10.1093/nar/gkaa348.
21. Lim N, Tesar S, Belmadani M, Poirier-Morency G, Mancarci BO, Sicherman J, et al. Curation of over 10,000 transcriptomic studies to enable data reuse. Database. 2021. https://doi.org/10.1093/database/baab006.
22. Liu YY, Slotine JJ, Barabási AL. Controllability of complex networks. Nature. 2011;473(7346):167–73. https://doi.org/10.1038/nature10011.
23. Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human binary protein interactome. Nature. 2020;580(7803):402–8. https://doi.org/10.1038/s41586-020-2188-x.
24. McKenzie AT, Wang M, Hauberg ME, Fullard JF, Kozlenkov A, Keenan A, et al. Brain cell type specific gene expression and co-expression network architectures. Sci Rep. 2018;8(1):1–9. https://doi.org/10.1038/s41598-018-27293-5.
25. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012;13(1):1–20. https://doi.org/10.1186/gb-2012-13-1-r5.
26. Nusinow DP, Szpyt J, Ghandi M, Rose CM, McDonald ER III, Kalocsay M, et al. Quantitative proteomics of the cancer cell line encyclopedia. Cell. 2020;180(2):387–402. https://doi.org/10.1016/j.cell.2019.12.023.
27. Obayashi T, Kagaya Y, Aoki Y, Tadaka S, Kinoshita K. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. Nucleic Acids Res. 2019;47(D1):D55-62. https://doi.org/10.1093/nar/gky1155.
28. Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH. Functional organization of the transcriptome in human brain. Nat Neurosci. 2008;11(11):1271–82. https://doi.org/10.1038/nn.2207.
29. Perkins AD, Langston MA. Threshold selection in gene co-expression networks using spectral graph theory techniques. BMC Bioinform. 2009;10(11):1–11. https://doi.org/10.1186/1471-2105-10-S11-S4.
30. Pierson E, GTEx Consortium, Koller D, Battle A, Mostafavi S. Sharing and specificity of co-expression networks across 35 human tissues. PLoS Comput Biol. 2015;11(5):e1004220. https://doi.org/10.1371/journal.pcbi.1004220.
31. Rachlin J, Cohen DD, Cantor C, Kasif S. Biological context networks: a mosaic view of the interactome. Mol Syst Biol. 2006;2(1):66. https://doi.org/10.1038/msb4100103.
32. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 2012;13(7):505–16. https://doi.org/10.1038/nrg3229.
33. Rung J, Brazma A. Reuse of public genome-wide gene expression data. Nat Rev Genet. 2013;14(2):89–99. https://doi.org/10.1038/nrg3394.
34. Sarntivijai S, Lin Y, Xiang Z, Meehan TF, Diehl AD, Vempati UD, et al. CLO: the cell line ontology. J Biomed Semant. 2014;5(1):1–10. https://doi.org/10.1186/2041-1480-5-37.
35. Schaefer MH, Serrano L, Andrade-Navarro MA. Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types. Front Genet. 2015;6:260. https://doi.org/10.3389/fgene.2015.00260.
36. Schwartz JL, Jordan R, Evans HH, Lenarczyk M, Liber HL. Baseline levels of chromosome instability in the human lymphoblastoid cell TK6. Mutagenesis. 2004;19(6):477–82. https://doi.org/10.1093/mutage/geh060.
37. Sealfon RS, Wong AK, Troyanskaya OG. Machine learning methods to model multicellular complexity and tissue specificity. Nat Rev Mater. 2021. https://doi.org/10.1038/s41578-021-00339-3.
38. Sonawane AR, et al. Understanding tissue-specific gene regulation. Cell Rep. 2017;21(4):1077–88. https://doi.org/10.1016/j.celrep.2017.10.001.
39. Stacey RG, Skinnider MA, Chik JHL, Foster LJ. Context-specific interactions in literature-curated protein interaction databases. BMC Genom. 2018;19(1):1–10. https://doi.org/10.1186/s12864-018-5139-2.
40. Trapotsi MA, Hosseini-Gerami L, Bender A. Computational analyses of mechanism of action (MoA): data, methods and integration. RSC Chem Biol. 2022. https://doi.org/10.1039/D1CB00069A.
41. The Gene Ontology Consortium. The gene ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021;49(D1):D325–34. https://doi.org/10.1093/nar/gkaa1113.
42. Vinayagam A, Gibson TE, Lee HJ, Yilmazel B, Roesel C, Hu Y, et al. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. Proc Natl Acad Sci. 2016;113(18):4976–81. https://doi.org/10.1073/pnas.1603992113.
43. Whitehead A, Crawford DL. Variation in tissue-specific gene expression among natural populations. Genome Biol. 2005;6(2):1–14. https://doi.org/10.1186/gb-2005-6-2-r13.
44. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics. 2007;8(1):1–14. https://doi.org/10.1186/1471-2105-8-22.
45. Yoshihama M, Uechi T, Asakawa S, Kawasaki K, Kato S, Higa S, Maeda N, Minoshima S, Tanaka T, Shimizu N, Kenmochi N. The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. Genome Res. 2002;12(3):379–90. https://doi.org/10.1101/gr.214202.
46. Yu K, Chen B, Aran D, Charalel J, Yau C, Wolf DM, et al. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. Nat Commun. 2019;10(1):1–11. https://doi.org/10.1038/s41467-019-11415-2.
47. Zhang W, Liu HT. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. Cell Res. 2002;12(1):9–18. https://doi.org/10.1038/sj.cr.7290105.
48. Zoubarev A, Hamer KM, Keshav KD, McCarthy EL, Santos JRC, Van Rossum T, et al. Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. Bioinformatics. 2012;28(17):2272–3. https://doi.org/10.1093/bioinformatics/bts430.

Publisher's Note