ChEA-KG: Human Transcription Factor Regulatory Network with a Knowledge Graph Interactive User Interface

Anna I. Byrd[1], John Erol Evangelista[1], Alexander Lachmann[1], Ho-Young Chung[1], Sherry L. Jenkins[1], Avi Ma'ayan[1,*]

[1]Department of Pharmacological Sciences, Department of Artificial Intelligence and Human Health, Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA

* To whom correspondence should be addressed. Tel: +1 (212) 241-1153; Fax: +1 (212) 849-2456; Email: avi.maayan@mssm.edu

Present Address: Avi Ma'ayan, Department of Pharmacological Sciences, Department of Artificial Intelligence and Human Health, Mount Sinai Center for Bioinformatics, One Gustave L. Levy Place, Box 1603, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA

ABSTRACT

Gene expression is controlled by transcription factors (TFs) that selectively bind and unbind to DNA to regulate mRNA expression of all human genes. TFs control the expression of other TFs, forming a complex gene regulatory network (GRN) with switches, feedback loops, and other regulatory motifs. Many experimental and computational methods have been developed to reconstruct the human intracellular GRN. Here we present a different approach. By submitting thousands of "up" and "down" gene sets from the RummaGEO resource for TF enrichment analysis with ChEA3, we distill signed and directed edges that connect human TFs to construct a high quality human GRN. The GRN has 131,581 signed and directed edges connecting 701 source TF nodes to 1,559 target TF nodes. The GRN is accessible via the ChEA-KG web server application, which provides interactive network visualization and analysis tools. Users may query the GRN for single or pairs of TFs or submit gene sets to perform TF enrichment analysis with ChEA3, placing the enriched TFs within the GRN. To demonstrate the utility of ChEA-KG, several TF-centric atlases are also made available via the ChEA-KG website. These atlases host TF subnetworks that regulate 131 major normal human cell-types (Cell Type Atlas); 69 tumour subtypes from 10 cancers (Cancer Atlas); 30 consensus perturbation response signatures for common mechanisms of action (MoA Atlas); and 24 aging signatures from tissues profiled by GTEx. Overall, ChEA-KG is an interactive web-server application that presents to users a new method of exploring the human gene regulatory network through both network visualization and transcription factor enrichment analysis. The ChEA-KG application is available from: https://chea-kg.maayanlab.cloud/.

KEYWORDS

gene regulatory network | transcription factors | knowledge graph | enrichment analysis | network visualization | cell types | cancer subtypes | lung squamous cell carcinoma | LINCS L1000 | Mechanisms of Action | Aging

INTRODUCTION

Unravelling the regulatory mechanisms underlying genome-wide gene expression has been a major goal of experimental and computational biologists for decades. Transcription factors (TFs) are key controllers of gene expression. TF activity directly and indirectly responds to extracellular cues to determine the behaviour of the cell by binding and unbinding to specific DNA sequences to increase (upregulate) or decrease (downregulate) the production of transcripts. TFs regulate the expression of other TFs forming an intricate and complex regulatory network commonly termed gene regulatory network (GRN) [1, 2]. Understanding the connectivity and dynamics of GRNs is critical to explaining disease mechanisms and other key intracellular physiological processes. GRNs can be represented *in-silico* as directed and signed graphs with nodes representing TFs and edges representing their regulatory relationships. To infer these associations, experimental data that measure TF activity directly or indirectly can inform computational methods. For example, various experimental methods have been developed to identify the binding locations of TFs on the DNA. Such experimental techniques include chromatin immunoprecipitation followed by sequencing (ChIP-seq) [3] or microarray (ChIP-chip) [4]; open chromatin assays, including DNase sequencing (DNase-seq) [5], formaldehyde-assisted identification of regulatory elements followed by sequencing (FAIRE-seq) [6], and DNase footprinting [7]; and chromatin interaction assays such as chromatin interaction analysis with paired end tag sequencing (ChIA-PET) [8, 9]. Aggregating the results from many studies that utilized these assays can be used to reconstruct GRNs [10-13]. However, typically such GRNs only have a subset of the entire repertoire of TFs and contain directed but unsigned edges. Alternative and complementary computational methods to reconstruct GRNs include text-mining [14], position weight matrices (PWM) analysis [15], and TF-gene co-expression analysis [16]. However, these methods may suffer from literature biases, lack of specificity, and reliance on indirect evidence. Regardless of the advantages and disadvantages of the computational and experimental methods used to reconstruct GRNs, once these GRNs are formed, they can be analyzed for their topological features such as network motifs [17-19], hub analysis [20], positive and negative feedback loops [21] and other topological features [22, 23]. Moreover, a useful application of *in-silico* GRNs in biomedical research is their use as background knowledge for transcription factor enrichment analysis (TFEA) [10, 11, 24].

Several web server and command line tools are available for performing TFEA. For example, Genomic Regions Enrichment of Annotations Tool (GREAT) [24] takes as input genomics locations to predict enriched pathway and biological processes based on knowledge about the binding sites of TFs in the nearby regions. ChIP-x enrichment analysis (ChEA) [11] aggregates gene sets from ChIP-seq and ChIP-chip publications and serves these sets for standard enrichment analysis with a proportion test. The updated version of ChEA, ChEA3 [10] aggregates TF-target associations from other sources such as co-expression, protein-protein interactions, and PWMs to identify consensus enriched TFs. Other tools such as oPOSSUM [25], Pscan [26], HOMER [27], and i-cisTarget [28] rely only on PWMs to infer the most likely TFs given a set of genes or a file with genomics regions annotations. Several other tools developed to perform TFEA are available offering various features and input types (Table 1, Supplemental Table S1). Here, we introduce a different method to reconstruct the human TF GRN. By combining TFEA with ChEA3 [10] with a massive collection of differentially expressed gene sets created for establishing RummaGEO [29], we infer the most likely regulatory relationships between all human TFs. The inferred GRN is stored in a knowledge graph (KG) database and served via a web server application called ChEA-KG. ChEA-KG is an interactive TFEA application that returns subnetworks of the connected top enriched TFs based on the topology of the reconstructed GRN. ChEA-KG also provides interactive network visualization and allows querying for relationships within the full GRN. To demonstrate the utility of ChEA-KG for context-specific analyses, we created four atlases featuring curated collections of TF regulatory subnetworks predicted to control marker gene sets from 131 human cell types; subnetworks regulating gene expression in 69 cancer subtypes from 10

cancer types; subnetworks of TFs regulating changes induced by drugs with 30 major mechanism of actions (MoAs); and TF subnetwoks responsible for observed changes due to aging in 24 human tissues.

METHODS

Construction of the TF-TF GRN with TFEA

The GRN construction method assumes that the up- or down-regulation of genes resulting from a single perturbation are controlled by TFs that directly regulate the transcription of those genes. These upstream TFs can be identified by submitting many differentially expressed up- and down-regulated gene sets for TFEA with ChEA3 [10]. Differentially expressed gene sets were downloaded from RummaGEO [29] on August 22nd, 2024. Gene sets were kept if they had at least five genes. Only studies with clear control and perturbation groups were included to ensure that differential expression resulted in a clear direction of transcription of the target genes. These studies were identified using regular expressions for terms that describe the experimental conditions, for example, “ctrl” and “wt”; and by manual classification of samples by DiSignAtlas [30]. TFEA was performed on each gene set using a local version of ChEA3 [10]. The 10 top-ranked TFs for each gene set were identified using the MeanRank method, and source-target edges were counted between the top 10 enriched TFs (sources) and TF-encoding genes from the input set. The sign of the regulatory relationship was determined based on the direction of differential expression comparing the perturbation group to the control group ("up" or "down" gene sets). Edges with a count lower than 10 were discarded.

Filtering the network to preserve significant edges

To remove edges that may be included by random chance, we determined the significance of each edge by comparing its observed frequency to its expected frequency in shuffled networks. Two methods were developed for generating expected edge counts. Both methods shuffle the original network. The “target set swap” (TSS) method swaps the source TF randomly, preserving the degree of the target nodes. The “node draw” (ND) method randomly swaps pairs of source and target nodes. Nodes are selected by the weighted frequency of occurrence as either a source or target role. This is repeated until the same number of edges have been sampled as were counted in the unfiltered network, preserving the relative connectivity distribution of the source and target nodes. For each shuffling method, 50 shuffled networks are produced. Then, the average and standard deviation of each edge is calculated. Next, the observed edge counts are compared to the average and standard deviations of the counts of each set of shuffled networks to calculate a z-score for the edge. The z-score was used to calculate a right-tailed p-value. Edges with a p-value >= 0.01 were discarded. Due to the large number of initial edges, some source-target node pairs were connected by both upregulated and downregulated edges. This conflict was resolved by preserving only the sign of the more significant edge. This produced two filtered GRNs: TSS-filtered and ND-filtered. Each of these GRNs was evaluated for accuracy by comparing the edges within these networks to edges from other sources of TF-TF associations not used to construct the GRNs.

Benchmarking the filtered networks against other GRNs

Each filtered network was independently benchmarked against two reference GRNs to determine if the GRN reconstruction process produces known TF-target associations, as well as which network pruning method produces the most reliable network. The two reference GRNs used to benchmark the ChEA-KG GRN are from text mining of TF-target interactions from TRRUST [14], and position weight matrices (PWMs) from TRANSFAC [31] and JASPAR [15]. These networks were selected because they are not used by ChEA3 [10] to perform enrichment analysis and were processed and available for download from Enrichr [32]. The TRRUST GRN has 379 source TFs, 553 target TFs, and 2,597 edges. The TRANSFAC and JASPAR PWMs GRN contains 246 source TFs, 1,652 target TFs, and 32,067 edges. The

statistical properties of the ChEA-KG filtered and unfiltered networks are summarized in Supplementary Table S2. The filtered and unfiltered ChEA-KG GRNs were compared to the benchmarking GRNs to determine the number of overlapping edges, and whether the number of overlapping edges is greater than what would be expected in shuffled networks with the similar size and structure. To generate the shuffled networks, source-target pairs in the original ChEA-KG filtered and unfiltered networks were randomly sampled from the source and target node distributions. This process ensures that each shuffled network has the same number of nodes and edges and similar connectivity distribution. The number of overlapping edges were normalized to the size of the filtered network. To determine statistical significance of the expected vs. observed overlapping edges between the ChEA-KG original and shuffled networks the Z test was applied. 100 shuffled networks were generated to produce an expected overlap distribution. The benchmarking procedure is outlined in a diagram (Supplementary Figure 1). This procedure was performed twice for each comparison, once with directed edges, and once with undirected edges, for a total of eight comparisons.

Visualizing the GRN

To better understand the global structure of the GRN, several dimensionality reduction methods were applied to visualize and cluster the GRN's TFs based on their connectivity similarity. First, we created a uniform manifold approximation projection (UMAP) [33] of the source TFs based on similarity between their target TF sets. Term frequency-inverse document frequency (TF-IDF) for each TF were computed [34]. Then, the Leiden algorithm was applied to identify clusters of similar TFs [35]. Next, we visualized the network edges as a clustered heatmap. To create the heatmap, first the network was converted to a directed, signed adjacency matrix where rows represent sources, and columns represent targets. Each (source, target) entry in the matrix was assigned a 0 if no edge exists, or a 1 or -1 for upregulated or downregulated edges, respectively. In addition, a heatmap of the Jaccard similarity scores between all TF pairs was generated. We calculated the Jaccard similarity between each pair of source TFs based on their shared targets. Both matrices were plotted as hierarchically clustered heatmaps using the Seaborn Python package [36].

Developing the ChEA-KG web server application

The ChEA-KG web-server application enables users to interact with and query the GRN. The application was built using the customizable Knowledge Graph User Interface (KG-UI) that uses Cytoscape.js [37] to visualize Cypher query results from a Neo4j database [38]. The GRN was ingested into a Neo4j database by serializing the GRN into node and edge lists. The user interface (UI) provides the ability to perform queries for finding neighbours of single TFs, finding shortest paths between pairs of TFs, displaying the subnetworks returned from the enrichment analysis with ChEA3 [10], displaying the Cell Atlas subnetworks using various layouts, expanding and shrinking the size of the displayed subnetwork, viewing properties of nodes and links, and downloading the displayed associations in tabular format.

Building the TF Cell Atlas

Cell types for the TF subnetwork Cell Atlas were identified manually based on literature curation. Blood cell types appearing in multiple tissues were removed. To construct the subnetworks for each cell type, cell-type specific marker gene sets were extracted from CellMarker [39], ASCT+B [40], PanglaoDB [41], Tabula Sapiens [42], Descartes [43], Azimuth [40], and TISSUES [44]. These marker genes were identified by manually mapping cell types to the terms that represent gene sets in gene set libraries served by Enrichr [32]. One gene set was selected for each cell type. Once the gene sets were present for each cell type, the gene sets were submitted to ChEA3 [10] for enrichment analysis and the networks were plotted using the KG-UI. We further analysed six representative subnetworks within the Cell Atlas to identify literature evidence linking each cell type to the nodes and link in each subnetwork of inferred TFs. Relevant citations were identified by manually searching PubMed

for the cell type, tissue type, and TF. For TFs with common synonyms documented in the NCBI database, for example, SPI1 and PU.1, both names were considered in the search. Relationships between the TFs within the subnetworks, including structural similarity, co-expression, known physical interactions, and transcription regulation were recorded.

Building the TF Cancer Atlas and analysing the subnetworks with GSFM

The ChEA-KG Cancer Atlas was created by submitting to ChEA-KG gene sets that are identified as upregulated in the transcriptomics clusters of the 10 tumour types from the Clinical Proteomic Tumour Atlas Consortium (CPTAC3) cohort [45]. For each cluster, identified subtypes were detected with the Leiden clustering algorithm [35] as described in the Multiomics2Targets workflow [46]. The upregulated marker gene sets for each subtype were extracted from the output of Multiomics2Targets, which uses the limma-voom to compute differential expression [47]. The visualization and integration of the ChEA-KG Cancer Atlas follows the same steps described for setting up the ChEA-KG Cell Atlas. To analyse the collective functions of each subnetwork produced by ChEA-KG, we identified consensus function predictions using the Gene Set Foundation Model (GSFM) [48] the KOMP2 [49] Mouse Phenotypes 2022 library from Enrichr [32] for the TFs in each subnetwork. The top 10 predictions with the highest score for each subnetwork of TF were aggregated and then sorted based on z-score. The top 10 aggregated phenotypes were assigned as the consensus phenotype for each subnetwork.

Building the TF MoA Atlas

The ChEA-KG MoA Atlas was created using precomputed differentially expressed gene sets from the LINCS L1000 Fireworks Display (L1000FWD) application [50]. Gene sets were downloaded directly from the L1000FWD download page. The set of unique drugs were identified and then cross referenced with the Drug Repurposing Hub [51] to assign known MoAs to each drug. All drugs with no known MoA were labelled as “unknown” and were discarded. The top 30 most common MoAs were selected as the set of terms for the gene set library. Up-regulated and down-regulated gene sets were produced for each MoA for a total of 60 gene sets. To create the consensus gene sets for each term, first, all the up- or all down-regulated genes for each MoA drug were counted. A consensus cutoff was then iteratively reduced until both up and down gene set size averages were between 150 and 200 genes. To quantify the overlap between each gene set, the Fisher’s exact test. Three comparisons were performed: upregulated vs. upregulated, upregulated vs. downregulated, and downregulated vs. downregulated gene sets. The p-values were recorded for each comparison and visualized as hierarchically clustered heatmaps using the Seaborn Python package [36]. The visualization and integration of the MoA Atlas follows the same procedure as the previously described Atlases. To identify shared modules of TFs between different MoAs, an adjacency matrix representing all edges across the 60 subnetworks was constructed. First, unique edges across all networks were combined into a single network. A directed binary adjacency matrix was created where 1 indicates a source-target edge and 0 indicates no edge. The resulting binary matrix was visualized as a hierarchically clustered heatmap using the Seaborn Python package [36]. Modules of similar TFs were identified and extracted. Their corresponding subnetworks were identified through manual inspection of the heatmap.

Building the TF Aging Atlas

The ChEA-KG Aging Atlas was created using differentially expressed gene sets pre-computed for the CFDE Gene Set Cart project [52] using the GTEx [53] Aging Signatures dataset. The 20-29 vs. 70-79 age group comparison was used as the representative set of signatures, up/down gene sets, for each tissue. The visualization and integration of the Aging Atlas follows the same procedure as the described for the other Atlases. The resulting subnetworks were tested for significant overlap to identify tissues with similar TF subnetworks. Significance was

assigned via permutation testing. Given a pair of subnetworks A and B with sizes M and N, respectively, and overlapping edges S_0, 1000 pairs of networks with sizes M and N and overlap S_1 were generated by randomly sampling edges from the full ChEA-KG background GRN. The p-value was calculated by dividing the number of times the randomized overlap, S_0, is greater than or equal to the observed overlap, S_1, divided by the number of permutations (1000). A pan-tissue regulatory TF-TF signed and directed network was constructed by aggregating all edges present in the Aging Atlas subnetworks. The resulting global network was visualized using Cytoscape [37]. The frequency of observing each node and link in the Aging Atlas subnetworks informs the size of the nodes and links in the global network.

RESULTS

The TF-TF GRN

The ChEA-KG human GRN of TF-TF interactions was created by applying TFEA with ChEA3 [10] to a massive collection of gene sets created by automated differential gene expression analysis applied to thousands of RNA-seq studies deposited in GEO to create RummaGEO [29]. By combining TF binding site knowledge with differential gene expression data, we can capture the most likely direct effects of one TF over another to construct a high fidelity GRN (Figure 1). The RummaGEO resource serves 171,441 human and 195,265 mouse gene sets spanning 29,294 GEO studies. For establishing ChEA-KG, only a subset of 29,328 gene sets extracted from 10,901 unique studies were used because the sets from these studies came from signatures where there is a clear separation between a control and a perturbation group. These gene sets were created by first downloading the uniformly aligned sample vectors of expression from ARCHS4 [54]. Then, samples from studies with control and perturbation conditions were analysed for differential expression with the limma-voom method [47]. TFEA was performed using ChEA3 [10] which covers 1,632 human TFs. During the construction process 10,405,833 signed edges were established between TFs. 373,250 of these edges are unique, connecting 727 source TFs to regulate the expression 1,560 target TFs. The network has self-loops and TFs can serve as both sources and targets. The GRN was filtered to exclude edges that did not occur more than expected when comparing the GRN to shuffled networks. After filtering, the GRN was reduced to 131,181 edges (ND filtering method) and 109,813 edges (TSS filtering method). The connectivity distributions of each of the filtered networks and the unfiltered network are visually compared (Figure 2). Other network properties such as number of edges by direction, average links per node, self-loops, and two-node feedback loops were recorded (Table 2).

Benchmarking the ND and TSS filtering methods

Since the TSS- and ND-filtered GRNs have different nodes and edges, the two methods were benchmarked by comparing the edges in those networks to edges from other TF-TF GRNs that are not included in ChEA3 [10] and RummaGEO [29], namely, literature curated TF-TF interactions from TRRUST [14], and TF-TF interactions based on position weight matrices (PWMs) from TRANSFAC [31] and JASPAR [15]. Each comparison was made with and without considering edge direction, for a total of four comparisons per network (Figure 3). Both filtered GRNs showed significantly higher overlap with interactions from TRRUST and TRANSFAC/JASPAR compared to edges in the shuffled networks of the same size and similar connectivity. The ND-filtered GRN showed the most significant results in three out of the four comparisons. It was only outperformed by the TSS-filtered GRN when compared against the TRANSFAC/JASPAR PWMs library (Figure 3, Supplementary Table S2). Based on the benchmarking result the ND-filtered GRN was selected for the ChEA-KG webserver application.

Identifying clusters within the GRN

Next, we visualised the entire GRN to examine its topology and identify TF modules. First, the network was clustered with the Leiden algorithm [35] and then visualized with a UMAP [33] (Figure 4). The sets of TFs from the 26 identified clusters were uploaded to Enrichr [32] to characterize the collective functions of each cluster of TFs. 13 clusters had clearly defined collective functions such as “patterning”, “cell cycle”, “hair and skin regulation”, and “stem cell maintenance”. Next, the GRN adjacency matrix was visualized as a clustered heatmap (Supplementary Figure S2). Only the 700 TFs that act as both sources and targets are included in the heatmap. The densely coloured areas in the heatmap show several regions of TF modules that regulate similar targets. Lastly, the Jaccard similarity scores between GRN TFs were visualized as a clustered distance table heatmap (Figure 5). Only 381 highly connected TFs were visualised because this subset of TFs had clearly visible defined clusters. The dendrogram from the distance table was extracted, and clusters were determined by cutting the dendrogram at 65% of the maximum tree height to produce 12 unique clusters. These clusters were annotated with enriched terms identified with Enrichr. The heatmap clusters were enriched for terms such as “protein folding”, “immune response”, and “cell cycle”.

The ChEA-KG web server application

ChEA-KG is an interactive web server application that enables users to interact with the GRN in multiple ways. The human GRN is stored in a Neo4j database as a KG and the ChEA-KG UI is querying the KG database to extract and display GRN subnetworks. Users of ChEA-KG can query the KG database for creating subnetworks centred on a single TF, and subnetworks that connect two-TFs based on the shortest paths between them. The results page has a toolbar that enables users to customize the subnetworks by limiting the number of nodes and edges, viewing the subnetwork as a table, adding a legend, changing the subnetwork layout, and displaying a tooltip with information associated with each TF. Users can also save the subnetwork as an image, or as node and edge lists in CSV files. ChEA-KG also has an enrichment analysis feature that enables users to associate their input gene sets with subnetworks of regulatory interacting TFs. The input gene sets are queried using the ChEA3 API [10] to identify the top enriched TFs for each gene set. These TFs are then visualized as a subnetwork. Users can adjust the number of edges in the displayed subnetwork by filtering by significance. The number of TFs can be adjusted to return between 5-25 top enriched TFs. Because the number of TFs in the GRN (1,559) is less than the number of TFs in the ChEA3 database (1,632), it is possible that some top-ranked TFs are not connected to any other TF in the GRN. Even for TFs that are in the GRN, the enrichment results may return TFs that are not directly connected to each other. If an enriched TF does not share edges with any other enriched TFs, including self-loops, they will be displayed as a single unconnected node.

The TF subnetworks cell atlas

Another feature of ChEA-KG is the TF subnetworks Cell Atlas. To create the Cell Atlas, we manually identified 131 major human cell types that can be found in 14 major human tissues. We then identified marker gene sets associated with each cell type by processing data from CellMarker [39], ASCT+B [40], PanglaoDB [41], Tabula Sapiens [42], Descartes [43], Azimuth [40], and TISSUES [44]. Then, we submitted these marker gene sets for enrichment analysis with ChEA3 [10] to create a subnetwork for each cell type. The TF Cell Atlas subnetworks are available through the ChEA-KG application. Users can select a subnetwork to view and interact with by using a drop-down menu of cell types organized by tissue. The full listing of cell types and marker genes are organized into a gene matrix transpose (GMT) file (Supplementary Table S3). From the collection of 131 cell types, we investigated the literature evidence of six subnetworks: erythrocytes in blood, cardiac muscle cells in heart, goblet cells in the intestine, mesenchymal stromal cells (MSCs) in bone marrow, epithelial cells in lung, and podocytes in the kidney (Figure 6).

Blood: Erythrocytes. TFs and interactions in the erythrocyte regulatory subnetwork were found to be involved in several erythrocyte-specific processes. NFE2 is known to control the

expression of globin genes in erythrocytes [55]; TAL1 [56], GATA1 [57], IKZF1 [58], SPI1 [59], and LYL1 [60] are known regulators of erythrocyte differentiation. TFs within the subnetwork are directly connected via direct transcriptional regulation, physical interactions, and structural similarity. LYL1 is known to bind to GATA1 to maintain erythropoiesis [61]. GATA1 also physically interacts with SPI1 [62] and controls the transcription of TAL1 [63], an interaction that is supported by the subnetwork (Figure 6A). The TFs AKNA, CEBPE, and TFEC were not found to be directly associated with erythrocytes in the literature and thus may represent undiscovered regulators of erythropoiesis.

Intestine: Goblet cells. The goblet cell regulatory subnetwork features several TFs known to control development of the intestine. CDX2, CDX1, and ATOH1 are known to regulate intestinal epithelial development [64], whereas HNF4A, SPDEF, and FOXA1 specifically regulate goblet cells by controlling maturation and differentiation [65-67]. ELF5 is a regulator of epithelial cell differentiation but has not been specifically linked to the intestine [68]. The T-Box family of genes plays a significant role in early embryonic development, but the exact role of TBX10 is unknown [69]. SPDEF and ATOH1 are transcriptional co-regulators [70] (Figure 6D). Overall, this subnetwork contains some known and some unknown master regulators of goblet cells.

Heart: Cardiac muscle cells. The cardiac muscle cell subnetwork includes TFs associated with both cardiac and skeletal muscle cells. RXRG2, an isoform of RXRG, is known to be highly expressed in cardiac muscle [71]. Several TFs in the cardiac muscle cell subnetwork are known regulators of skeletal muscle. For example, PITX3 is involved in myogenesis and has sustained expression in skeletal muscle cells [72]. MYF5, MYF6, MYOG, and MYOD1 are well-documented to establish the skeletal muscle phenotype [73]. These four genes share structural similarity and are collectively known as the Myogenic regulatory factors (MRFs) [74]. TBX20 and TBX5 are T-box proteins that are required for cardiomyocyte proliferation and homeostasis [75] and maturation [76], respectively. LBX1 controls gene expression relating to migration of muscle precursors [77]. Within the subnetwork, the skeletal muscle TFs, the MRFs and PITX3, regulate each other but no other network nodes, except for MYOD1, which regulates both RXRG and TBX20 (Figure 6C).

Kidney: Podocytes. TFs involved in the podocyte subnetwork are related to differentiation and gene expression during injury (Figure 6E). ATOH8 is a known regulator of kidney development and podocyte differentiation [78]. Several TFs in the subnetwork are associated with disease or injury of the kidney. ZHX3 is known to regulate podocyte gene expression during kidney injury [79, 80]. TEAD1 expression is lowered in diabetic podocyte injury [80]. In the kidneys more broadly, GLI2 is upregulated after adriamycin neuropathy [81] whereas GLIS2 is known to maintain renal architecture [82] and GLIS3 is associated with renal anomalies [83]. Finally, PRRX1 is suppressed during regulation of the renal epithelial-to-mesenchymal transition (EMT) [84], a process linked to podocyte dysfunction [85].

Lung: Epithelial cells. The epithelial cell subnetwork contains TFs involved in mucin production as well as EMT and lung cancer (Figure 6F). The TF SNAI2 plays a central role in EMT [86]. PRRX1 is similarly known to induce EMT in embryonic stem and cancer cells [87]. AHR is known to upregulates the expression of mucin 5AC [88], a major component of mucus in the lung [89]. High expression of FOSL2 has been linked to poor prognosis in lung cancer patients [90]. Lastly, HNF1B regulates the expression of alpha(1)-antitrypsin (AAT; SERPINA1), which may defend the lung against antiproteases [91]. CDX2, which is similarly enriched in intestinal epithelial cells, autoregulates its own expression and as well as the expression of HNF1B within intestinal epithelial cells, suggesting that it plays a similar regulatory role in lung and intestinal epithelial cells [92].

Bone marrow: MSCs. TFs enriched for MSCs are known to be involved in bone development and maintenance of processes related to cell fate (Figure 6B). SOX18 is known to be involved in the differentiation of MSCs to endothelial cells [93]. SOX17 is known to control the

395 expression of fetal hematopoietic poetic stem cells [94], whereas MEOX1 is a key regulator of
396 proliferation in mesenchymal-like cells [95]. Differentiation of bone-marrow-derived MSCs is
397 regulated by GATA2 [96]. ERG and ELK3 are both members of the ETS family of TFs [97] and
398 involved in EMT [98]. ERG is directly linked to endo-cardial mesenchymal transformation
399 (EnMT) and EMT [99, 100]. Finally, it is known that TWIST1 suppresses the expression of
400 RUNX2 to suppress bone formation [101] while FOXO1 defends against oxidative damage in
401 the bone [102]. In conclusion, the subnetworks for each cell type in the ChEA-KG Cell Atlas
402 confirm many master regulators as well as predict novel key regulators.

403 **The TF subnetworks cancer atlas**

404 Besides the ChEA-KG Cell Atlas, ChEA-KG also has a Cancer Atlas. To create the ChEA-KG
405 Cancer Atlas we obtained 69 up-regulated gene sets extracted from 10 cancer types profiled
406 by the National Institutes of Health (NIH) CPTAC3 program [45]. Each of these gene sets
407 represents a subtype based on applying the Leiden clustering algorithm [35] to the RNA-seq
408 profiling of over a thousand tumours by Multiomics2Targets [46]. These up-regulated genes
409 from these tumour subtypes were submitted for analysis with ChEA-KG. As an example, we
410 applied this analysis to identify TF subnetworks enriched for lung squamous cell carcinoma
411 (LSCC) (Figure 7). To characterize these subnetworks, we next identified key regulatory
412 mechanisms for each LSCC subnetwork by predicting gene functions for each TF in each
413 subnetwork using GSFM [48] and the KOMP2 [49] Mouse Phenotypes 2022 gene set library.
414 (Supplementary Table S5). We find that cluster 0 is enriched in functions related to T-cell
415 regulation, cluster 1 is enriched for midbrain morphology and sclerocornea phenotypes,
416 cluster 2 is shows enriched terms for immunity and NK cells, cluster 3 appears to be related
417 to lung development, and cluster 5 is related to thyroid regulation. These predicted functional
418 terms immediately suggest key mechanisms of each cancer subtype, and for classifying
419 tumours as hot or cold. The analysis provides clear mechanisms that may be correlated with
420 subtype survival and likelihood of the responsiveness to therapy for each subtype.

421 **The TF subnetworks MoA atlas**

422 The third ChEA-KG atlas is the MoA atlas. To create this atlas, we identified the 30 most
423 common MoAs in the LINCS L1000 chemical perturbations dataset [103]. We then created
424 consensus up and down gene sets for each MoA by aggregating genes that are up- or down-
425 regulated in at least 17% of the signatures associated with each MoA (Supplementary Table
426 S6). The gene sets were analysed for significant overlap using Fisher's exact test
427 (Supplementary Figure S3). A summary of each gene set, including the number of signatures
428 and unique drugs used to compute the gene set was computed (Table 3). To identify key TF
429 modules involved in drug response, we aggregated all unique edges from the MoA atlas
430 subnetworks and generated a hierarchically clustered binary adjacency matrix heatmap. We
431 identified five unique clusters of TFs with similar regulatory patterns (Figure 8A). For each
432 cluster, we used the member TFs to assign one or two representative subnetworks from the
433 MoA atlas (Figure 8B-F). We investigated the literature evidence connecting the TFs in
434 clusters 1, 2, 4, and 5 to each of their MoAs as follows.

435 <u>Cluster 1: Adrenergic Receptor Antagonists (down)</u>

436 The subnetwork representing Cluster 1 is composed of TFs enriched for genes downregulated
437 by sodium channel blockers (TP63, IRF6, EHF, ELF3, TFCP2L1) and adrenergic receptor
438 antagonists (GRHL3, EHF, GRHL1, FOXQ1, ELF3) (Figure 8B, 8C). These overlapping
439 modules are generally associated with epithelial-associated TFs [104-107]. The TFs enriched
440 in response to adrenergic receptor antagonists are associated with tumour suppressive TFs
441 known to control the epithelial-mesenchymal transition (EMT). To test the role of these TFs in
442 EMT regulation, 215 genes from an EMT-specific signature constructed via antibody-based
443 profiling of 736 cancer cell lines [108] were submitted to ChEA3 [10] to identify likely TF
444 regulators. All five TFs in the adrenergic receptor antagonist module were ranked within the

top 12 most likely regulators of the EMT-specific genes (Supplementary Table S6). Beta-adrenergic signalling has been linked to the progression of cancer, in part through activation of the EMT [109]. Beta-blockers, a type of adrenergic receptor antagonist, have thus been explored as a potential cancer therapeutics, particularly triple-negative breast cancer [110, 111]. This suggests that one mechanism by which beta-blockers can produce an anti-cancer response is through activation of the tumour-suppressive, EMT-regulating known module of GRHL3/EHF/GRHL1/FOXQ1/ELF3.

Cluster 2: HDAC Inhibitors (up)

The subnetwork representing Cluster 2 includes six TFs from the subnetwork enriched for genes upregulated by HDAC inhibitors (EGR1, JUNB, FOS, NR4A1, JUN, FOSB) (Figure 8D). Four of these TFs, namely FOS, FOSB, JUN, and JUNB, are considered immediate early genes (IEG) [112]. IEGs are characterized by having low basal expression levels and rapid upregulation following extracellular stimuli. These genes are tightly regulated, in part because of their global impact on transcription. Their enrichment downstream of HDAC inhibitors suggests that IEGs the TFs and their targets are upregulated in response to HDAC inhibitors. This was confirmed by a study that found that administration of HDAC inhibitors in aging mice increases levels of fos-positive cells [113]. This suggests a possible mechanism by which HDAC inhibitors upregulate IEGs by reversing histone hypoacetylation at their promoter.

Cluster 4: PI3K Inhibitors (down)

The TFs enriched for genes downregulated by PI3K inhibitors and captured by Cluster 4 – HMGA1, FOXM1, MYBL2, ZNF367, and E2F7 – form a highly connected positive feedback module. These TFs are implicated in cell cycle control and cell proliferation [114-117]. Additionally, four of the five TFs, all except ZNF367, have been associated with the PI3K/AKT/mTOR signalling pathway [118-121]. This suggests that PI3K inhibitors suppress expression of TFs that belong to this subnetwork of nested positive feedback loops. ZNF367 is an understudied TF that has previously been tied to the progression of colon cancer via YAP1 signalling [122]. Its involvement in this tightly connected subnetwork module downstream of the PI3K signalling pathway is highly likely but needs to be confirmed experimentally (Figure 8F).

Cluster 5: Glucocorticoid Receptor Agonists (down)

The genes downregulated by glucocorticoid receptor agonists are enriched for targets of TFs tied to developmental programming related to the Wnt cell signalling pathway [123-125]. The gene encoding for the transcription factor HEY1 is a known to be upregulated by Notch signalling [126]. The TF CXXC4 is a known repressor of the Wnt pathway and a highly connected member within this module [123]. Glucocorticoid receptors have also been shown to repress Wnt signaling in human osteoblasts [127]. This suggests that glucocorticoid receptor agonists activate this module of TFs to suppress Wnt-related developmental genes that may have additional roles in normal physiology and disease (Figure 8G).

The TF subnetworks of the aging atlas

The ChEA-KG aging atlas consists of aging-specific up and down gene sets for 24 tissues computed from the GTEx v8 publicly available RNA-seq dataset [128] (Supplementary Table S8). To explore how the various tissue-specific aging TF subnetworks may share components, a composite network was created by combining all edges and nodes from all the aging atlas subnetworks (Figure 9A). This network has visible clusters of TFs associated with specific tissues sharing similar aging specific regulatory modules (Figure 9B; Supplementary Figure S4). Clusters of co-regulating TFs were also identified by hierarchically clustering the adjacency matrix of the composite network (Supplementary Figure S5). One distinct network module was identified to regulate upregulated aging genes in breast, lung, blood vessel,

thyroid, liver, and skin tissues with the TF Intestine-Specific Homeobox (ISX) appearing as a central node in this module (Figure 9C).

ISX is an intestinal tissue master regulator that is also responsible for controlling vitamin A production [129]. ISX has been also suggested as a potential prognostic and therapeutic target for hepatocellular carcinoma (HCC) [130]. The ISX associated subnetwork includes three hepatocyte nuclear factors (HNFs), namely HNF4α, HNF4γ, and HNF4β. These TFs are primarily expressed in the liver, pancreas, and kidney and their dysregulation is implicated in diabetes [131]. Other TFs found in the subnetwork are also involved in liver and pancreatic functions including NR1H4, a regulator of bile acid synthesis [132]; MLXIPL, a master regulator of lipogenesis [133]; and RFX6, which regulates pancreatic islet cell differentiation [134]. The remaining TFs in the subnetwork, CDX1, CDX2, ELF3, GATA5, TBOX10, SPDEF, and AOTH1, are cellular differentiation regulators in both embryonic and adult tissues [69, 70, 135-137]. Dysregulation of developmental programs in adult tissues has recently been suggested as a cause of aging and age-related disease [138]. This module therefore implicates metabolic function and tissue development in the process of aging.

DISCUSSION

The formation of the ChEA-KG GRN presents a unique new method to reconstruct in-silico gene regulatory networks. By combining a massive collection of diverse gene expression signatures with a comprehensive TF enrichment analysis, a reliable data driven GRN was constructed. The GRN contains almost all known human TFs, expanding on the size and coverage of most current GRN reconstruction techniques. For example, it includes almost all known human TFs (n=1,559) in its background GRN. In contrast, TRRUST [14], a highly used human GRN reconstructed from text mining, is considerably smaller (n=800). The accuracy of the GRN edges is supported by the benchmarking results. However, testing it against other sources of known TF-TF interactions would strengthen the confidence of the links. Ultimately, testing such regulatory interactions in-vitro and in-vivo would fully confirm the inferred regulatory relationship. This can be achieved by knockdown, knockout, and over-expression of the source TF followed by gene and protein expression of the target TF. Examining the topology of the GRN identified clear clusters with known shared functions for each module. For example, we identified a module of TFs that is related to the cell cycle, immune response, differentiation, and development.

The ChEA-KG web server application facilitates the querying of the GRN for knowledge discovery. The enrichment analysis feature of ChEA-KG can be used to uncover how the most enriched TF master regulators are forming subnetworks of tight regulatory interactions. In the future, such subnetworks could provide the basis for dynamical simulations to further understand how groups of TFs work together as a functional regulatory unit. In addition, the GRN can be expanded to include non-TF-encoding genes such as kinases, chromatin modifiers, and other relevant components that can enrich the contents of each subnetwork.

The several atlases of ChEA-KG demonstrate how the GRN can be used to systematically identify known master regulator subnetworks and propose new regulatory mechanisms in a variety of contexts. For example, the cell type specific subnetworks also contain TF unknown to be relevant to the specific cell types, potentially leading to new discoveries. This includes, for example, the role of MRFs in regulation of cardiac muscle cells, or the role of ELK3, an understudied member of the ETS family, in mesenchymal stromal cells regulation. The cancer specific subnetworks demonstrate how ChEA-KG can be integrated as part of a transcriptomics workflow to better understand the distinguishing factors between cancer subtypes. The MoA atlas subnetworks reveals connections between known regulatory modules and common drug MoAs. For example, a connection between HDAC inhibitors and IEGs is evident. Also, the connection between adrenergic receptor agonists and EMT

regulators is interesting. Finally, the global network that is made of all the upregulated aging atlas tissue subnetworks reveals a possible new master regulatory module of aging, featuring metabolic and developmental TFs that are cantered around ISX.

CONCLUSIONS

Overall, ChEA-KG presents several features not available in existing tools. In addition, the ChEA-KG network visualization interface enables users to easily see regulatory subnetworks that connect TFs with signed and directed links. Currently, we are not aware of a dynamic network visualization feature in any of the TF enrichment analysis servers/apps. Importantly, ChEA-KG has four atlases that visualize the master regulatory subnetworks for 131 human cell types, 69 tumour subtypes from 10 cancers, 30 drug MoA signatures, and 24 aging tissue signatures. By using the ChEA-KG GRN as a background regulatory network, ChEA-KG can provide insights to biologists about not only the TFs that are most relevant to their data, but also how these TFs form a regulatory module that works in concert to induce the observed changes in gene expression, adding more context to TF enrichment analysis results.

LIST OF ABBREVIATIONS

TF: Transcription Factor
GRN: Gene regulatory network
ChIP-seq: Chromatin immunoprecipitation followed by sequencing
ChIP-chip: Chromatin immunoprecipitation followed by microarray
DNase-seq: DNAse sequencing
FAIRE-seq: Formaldehyde-assisted identification of regulatory elements followed by sequencing
ChIA-PET: Chromatin interaction analysis with paired end tag sequencing
PWM: Positional weight matrices
TFEA: Transcription factor enrichment analysis
KG: Knowledge graph
TSS: Target set swap
ND: Node draw
UMAP: Uniform manifold approximation projection
TF-IDF: Term frequency-inverse document frequency
KG-UI: Knowledge Graph User Interface
UI: User interface
MoA: Mechanism of Action
LINCS: Library of Integrated Network-based Cellular Signatures
L1000 FWD: LINCS L1000 Fireworks Display
GTEx: Genotype-Tissue Expression Project
CPTAC3: Clinical Proteomic Tumor Atlas Consortium
GMT: Gene matrix transpose
MSCs: Mesenchymal stromal cells
MRFs: Myogenic regulatory factors
EMT: Epithelial-to-mesenchymal transition
EnMT: Endo-cardial mesenchymal transition
NIH: National Institutes of Health
LSCC: Lung squamous cell carcinoma

AVAILABILITY OF DATA AND MATERIALS

The ChEA-KG web-server application is available at: https://chea-kg.maayanlab.cloud/. The node and edge lists for the ND-filtered, TSS-filtered, and unfiltered GRNs, as well as the marker gene sets for both the Cell Atlas and Cancer Atlas, are available on the ChEA-KG site, https://chea-kg.maayanlab.cloud/download_files. The ChEA-KG source code is available from GitHub at: https://github.com/MaayanLab/ChEA-KG-UI.

Project name: ChEA-KG
Project home page: https://chea-kg.maayanlab.cloud/
Operating systems: Platform independent
Programming language: Python, TypeScript, React
License: GPL 3.0

FUNDING

This work was supported by the National Institutes of Health [OT2OD036435, U24CA264250, OT2OD030160, U24CA271114]. Funding for open access charge: National Institutes of Health [OT2OD036435].

AUTHOR CONTRIBUTIONS

AB and JEE: data curation, formal analysis, software, validation, visualization, writing – original draft; AL: methodology, software; HYC: initial data curation, formal analysis, software; SJ: supervision, project administration; AM: conceptualization, funding acquisition, investigation, methodology, supervision, project administration, writing – original draft, writing – review & editing.

FIGURES AND TABLES LEGENDS

Figure 1. Workflow for constructing the GRN.

Figure 2. Network connectivity distributions of the unfiltered and ND- and TSS- filtered GRNs. In-degree and out-degree refer to target and source nodes, respectively. A, B: before filtering, C, D: ND-filtered; E, F: TSS-filtered.

Figure 3. Benchmarking edge pruning methods (filtering). The expected edge overlap, calculated using randomly shuffled networks (shown with boxplots), is compared to the observed edges in the GRN (black dot). Comparisons are made with and without considering edge direction. The boxplots show the distribution of random trials. The black dots are the observed recovery for each comparison.

Figure 4. UMAP visualization of source TFs based on their target set similarity. Clusters are annotated with enriched terms.

Figure 5. Clustermap of TFs based on their Jaccard similarity score computed based on their edge similarity. The dendrogram at the top is color-coded by cluster. Regions of high similarity are annotated with their enrichment terms.

Figure 6. Six selected cell type specific TF regulatory subnetworks from the ChEA-KG TF Cell Atlas. A: Erythrocytes (blood). B: Mesenchymal stem cells (bone marrow). C: Cardiac muscle cells (heart). D: Goblet cells (intestine). E: Podocytes (kidney). F: Epithelial cells (lung). Edges and nodes are highlighted in yellow if they are confirmed by a manual literature search.

Figure 7. TF regulatory subnetworks identified by ChEA-KG from the six transcriptomics subclusters of LSCC from the ChEA-KG TF Cancer Atlas. A: Cluster 0 (n=29

642 patients/tumors). B: Cluster 1 (n=27). C: Cluster 2 (n=20). D: Cluster 3 (n=16). E: Cluster 4
643 (n=14). F: Cluster 5 (n=8). Red nodes represent TFs that are enriched by ChEA3 but also
644 present in the input set of upregulated genes in the specific cluster.

645 **Figure 8.** A: Hierarchically clustered binary adjacency matrix and selected co-regulating TF
646 modules. B-F: Representative subnetworks for each identified module. Subnetwork TFs
647 corresponding to module TFs are highlighted in pink. Module 1 consists of two MoAs and is
648 represented by both B and C. Module 2 is represented by D. Module 3 is represented by E.
649 Module 4 is represented by F. Module 5 is represented by G.

650 **Figure 9.** A: Cytoscape visualization of the Aging Atlas composite network. Node and edge
651 sizes are proportional to the number of networks in which each appears. TF submodules
652 corresponding to specific tissues are highlighted in purple and labeled. B: Heatmap of
653 hierarchically clustered p-values representing similarity between regulatory subnetworks for
654 genes upregulated in each tissue in response to aging. B: Subnetwork from (A) consisting of
655 ISX and its immediate neighbors, based on gene sets upregulated in response to aging.

656

657 **Table 1.** Comparison the features of ChEA-KG with existing relevant published tools to
658 predict TF regulatory interactions. A: Accessible in browser. B: Scripting library/package or
659 requires download. C: Interactive visualization of results. D: Network of TFs. E: Infers sign of
660 TF-target regulation. F: Background includes all human TFs.
661

662 **Table 2.** Network statistics. The nodes, edges, links per node, and self- and two-node
663 feedback loops are of the unfiltered GRN and ND- and TSS-filtered GRNs. Edges in positive
664 feedback loops have the same sign. Edges in negative feedback loops have opposite signs.
665

666 **Table 3.** Summary of each consensus gene set in the MoA Atlas, including the number of
667 signatures used to compute each gene set, the number of unique drug perturbations
668 included in the relevant signatures, and the final up and down gene set sizes.

669
670
671
672

673 **Table 1**

674

675

Name	PMID	Input	Output	A	B	C	D	E	F
ChEA-KG		Gene set	Subnetwork of enriched TFs and their connections	✓	×	✓	✓	✓	✓
AME	25953851	Genomic regions	Enriched motifs and TF binding sites	✓	×	×	×	×	×
BART	29608647	Gene set	Enrichment scores for TFs	✓	✓	×	×	×	×
ChEA3	31114921	Gene set	Ranked TFs predicted to regulate gene sets	✓	×	✓	×	×	×
Cluster Profiler	22455463	Gene set	Enrichment scores and visualization	×	✓	✓	✓	×	×
DoRothEA	31340985	Gene expression matrix	TF activity scores and regulatory networks	×	✓	✓	✓	✓	✓
GREAT	20436461	Genomic regions	Enriched terms across different gene set libraries	✓	×	×	×	×	×
GTRD	33231677	Genomic regions or gene lists	Enriched TF binding sites	✓	✓	✓	×	×	✓
HOMER	20513432	Genomic regions or gene lists	Enrichment of known or de novo motifs	×	✓	×	×	×	×
iRegulon	25058159	Gene set	Predicted TFs and their targets	×	✓	✓	✓	✓	✓
PSCAN	19487240	Genomic regions or gene lists	TF motif enrichment scores	✓	×	×	×	×	×
RcisTarget	28991892	Gene set	Ranked TFs associated with gene sets	×	✓	×	×	×	×
TFEA	34079046	Genomic regions or gene lists	Enrichment scores of TF motifs	×	✓	×	×	×	×
TFEA.ChIP	31347689	ChIP-seq peak files	TF enrichment scores	✓	✓	✓	×	×	×
TFmotifView	32324215	Genomic regions	Enriched TF motifs and their regulatory potential	✓	×	✓	×	×	×
TRANSFAC	16381825	Genomic regions or gene lists	Enriched TF binding sites and regulatory networks	✓	✓	✓	×	×	×
TRRUST	26066708	Gene set	TF-target networks with evidence from scientific literature	✓	×	✓	✓	✓	×
Viper	27322546	Gene expression signature and cell context-specific regulatory network	Inferred activity for each regulatory gene in the input dataset	×	✓	×	✓	✓	✓

676

677

678

679

680

681

682

683 **Table 2**

684

	Unfiltered network	ND-filtered	TSS-filtered
Nodes	1,560	1,559	1,560
Source nodes	727	700	548
Target nodes	1,560	1,559	1,560
Edges	373,250	131,181	109,813
Upregulated edges	184,134	63,531	51,382
Downregulated edges	189,116	67,650	58,431
Out-links per node	239.263	84.144	70.393
In-links per node	513.411	187.401	200.389
Links per node	239.263	84.144	70.393
Self-loops	1030	549	379
Positive feedback loops	27,368	5,088	2,243
Negative feedback loops	27,185	4,514	2,075

685

686

687 **Table 3**

MoA	Signatures	Drugs, up	Up genes	Down genes
dopamine receptor antagonist	345	16	146	88
cyclooxygenase inhibitor	124	16	94	96
adrenergic receptor antagonist	77	16	104	116
glucocorticoid receptor agonist	136	15	153	177
calcium channel blocker	62	11	159	97
histamine receptor antagonist	54	11	126	116
tubulin polymerization inhibitor	193	10	231	189
phosphodiesterase inhibitor	65	10	113	129
mTOR inhibitor	703	9	146	135
serotonin receptor antagonist	45	9	181	148
bacterial cell wall synthesis inhibitor	25	9	191	167
estrogen receptor agonist	131	7	85	114
adrenergic receptor agonist	39	7	214	159
PARP inhibitor	107	6	83	135
acetylcholine receptor antagonist	23	6	279	313
selective serotonin reuptake inhibitor (SSRI)	69	6	211	128
HDAC inhibitor	1757	6	321	276
sodium channel blocker	47	5	206	150
bacterial DNA gyrase inhibitor	28	5	183	244
HMGCR inhibitor	84	5	166	144
PI3K inhibitor	727	5	160	106
potassium channel blocker	41	5	245	155
ATPase inhibitor	27	5	250	198
HIV protease inhibitor	23	5	315	257
adenosine receptor agonist	21	5	229	263
AKT inhibitor	78	4	180	95
bacterial 50S ribosomal subunit inhibitor	22	4	248	281
DNA synthesis inhibitor	38	4	195	190
potassium channel activator	15	4	267	309

690 **REFERENCES**

691 1. Davidson EH, Erwin DH: **Gene regulatory networks and the evolution of animal body**
692 **plans**. *Science (New York, NY)* 2006, **311**(5762):796-800.
693 2. Levine M, Tjian R: **Transcription regulation and animal diversity**. *Nature* 2003,
694 **424**(6945):147-151.
695 3. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo**
696 **protein-DNA interactions**. *Science (New York, NY)* 2007, **316**(5830):1497-1502.
697 4. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J,
698 Hannett N, Kanin E *et al*: **Genome-wide location and function of DNA binding**
699 **proteins**. *Science (New York, NY)* 2000, **290**(5500):2306-2309.
700 5. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford
701 GE: **High-resolution mapping and characterization of open chromatin across the**
702 **genome**. *Cell* 2008, **132**(2):311-322.
703 6. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD: **FAIRE (Formaldehyde-Assisted**
704 **Isolation of Regulatory Elements) isolates active regulatory elements from human**
705 **chromatin**. *Genome research* 2007, **17**(6):877-885.
706 7. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE,
707 Neph S, Kuehn MS, Noble WS *et al*: **Global mapping of protein-DNA interactions in**
708 **vivo by digital genomic footprinting**. *Nature methods* 2009, **6**(4):283-289.
709 8. Furey TS: **ChIP-seq and beyond: new and improved methodologies to detect and**
710 **characterize protein-DNA interactions**. *Nature reviews Genetics* 2012, **13**(12):840-
711 852.
712 9. Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, Ariyaratne PN, Mohamed YB,
713 Ooi HS, Tennakoon C *et al*: **ChIA-PET tool for comprehensive chromatin interaction**
714 **analysis with paired-end tag sequencing**. *Genome biology* 2010, **11**(2):R22.
715 10. Keenan AB, Torre D, Lachmann A, Leong AK, Wojciechowicz ML, Utti V, Jagodnik KM,
716 Kropiwnicki E, Wang Z, Ma'ayan A: **ChEA3: transcription factor enrichment analysis**
717 **by orthogonal omics integration**. *Nucleic acids research* 2019, **47**(W1):W212-W224.
718 11. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A: **ChEA:**
719 **transcription factor regulation inferred from integrating genome-wide ChIP-X**
720 **experiments**. *Bioinformatics* 2010, **26**(19):2438-2444.
721 12. Hammal F, de Langen P, Bergon A, Lopez F, Ballester B: **ReMap 2022: a database of**
722 **Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative**
723 **analysis of DNA-binding sequencing experiments**. *Nucleic acids research* 2022,
724 **50**(D1):D316-D325.
725 13. Chèneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B: **ReMap 2018: an updated**
726 **atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq**
727 **experiments**. *Nucleic acids research* 2018, **46**(D1):D267-D275.
728 14. Han H, Shim H, Shin D, Shim JE, Ko Y, Shin J, Kim H, Cho A, Kim E, Lee T *et al*:
729 **TRRUST: a reference database of human transcriptional regulatory interactions**.
730 *Scientific reports* 2015, **5**:11432.
731 15. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R, Castro-Mondragon JA, Ferenc K,
732 Kumar V, Lemma RB, Lucas J, Chèneby J, Baranasic D *et al*: **JASPAR 2024: 20th**
733 **anniversary of the open-access database of transcription factor binding profiles**.
734 *Nucleic acids research* 2024, **52**(D1):D174-D182.
735 16. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano
736 A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a**
737 **mammalian cellular context**. *BMC bioinformatics* 2006, **7 Suppl 1**(Suppl 1):S7.
738 17. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs:**
739 **simple building blocks of complex networks**. *Science (New York, NY)* 2002,
740 **298**(5594):824-827.
741 18. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional**
742 **regulation network of Escherichia coli**. *Nature genetics* 2002, **31**(1):64-68.
743 19. Alon U: **Network motifs: theory and experimental approaches**. *Nature reviews*
744 *Genetics* 2007, **8**(6):450-461.

20. Barabási AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nature reviews Genetics* 2004, **5**(2):101-113.
21. Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, Eungdamrong NJ, Weng G, Ram PT, Rice JJ *et al*: **Formation of regulatory patterns during signal propagation in a Mammalian cellular network**. *Science (New York, NY)* 2005, **309**(5737):1078-1083.
22. Kadelka C, Butrie TM, Hilton E, Kinseth J, Schmidt A, Serdarevic H: **A meta-analysis of Boolean network models reveals design principles of gene regulatory networks**. *Science advances* 2024, **10**(2):eadj0822.
23. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA: **Structure and evolution of transcriptional regulatory networks**. *Current opinion in structural biology* 2004, **14**(3):283-291.
24. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G: **GREAT improves functional interpretation of cis-regulatory regions**. *Nature biotechnology* 2010, **28**(5):495-501.
25. Kwon AT, Arenillas DJ, Worsley Hunt R, Wasserman WW: **oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets**. *G3 (Bethesda, Md)* 2012, **2**(9):987-1002.
26. Zambelli F, Pesole G, Pavesi G: **Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes**. *Nucleic acids research* 2009, **37**(Web Server issue):W247-252.
27. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities**. *Molecular cell* 2010, **38**(4):576-589.
28. Herrmann C, Van de Sande B, Potier D, Aerts S: **i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules**. *Nucleic acids research* 2012, **40**(15):e114.
29. Marino GB, Clarke DJB, Lachmann A, Deng EZ, Ma'ayan A: **RummaGEO: Automatic mining of human and mouse gene sets from GEO**. *Patterns (New York, NY)* 2024, **5**(10):101072.
30. Zhai Z, Lin Z, Meng X, Zheng X, Du Y, Li Z, Zhang X, Liu C, Zhou L, Zhang X *et al*: **DiSignAtlas: an atlas of human and mouse disease signatures based on bulk and single-cell transcriptomics**. *Nucleic acids research* 2024, **52**(D1):D1236-D1245.
31. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K *et al*: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes**. *Nucleic acids research* 2006, **34**(Database issue):D108-110.
32. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A *et al*: **Enrichr: a comprehensive gene set enrichment analysis web server 2016 update**. *Nucleic acids research* 2016, **44**(W1):W90-97.
33. McInnes L, Healy J, Saul N, Großberger L: **UMAP: Uniform Manifold Approximation and Projection**. *J Open Source Softw* 2018, **3**(29):861.
34. Salton G, Buckley C: **Term-weighting approaches in automatic text retrieval**. *Inf Process Manag* 1988, **24**(5):513-523.
35. Traag VA, Waltman L, van Eck NJ: **From Louvain to Leiden: guaranteeing well-connected communities**. *Scientific reports* 2019, **9**(1):5233.
36. Waskom M: **seaborn: statistical data visualization**. *J Open Source Softw* 2021, **6**(60):3021.
37. Franz M, Lopes CT, Fong D, Kucera M, Cheung M, Siper MC, Huck G, Dong Y, Sumer O, Bader GD: **Cytoscape.js 2023 update: a graph theory library for visualization and analysis**. *Bioinformatics* 2023, **39**(1):ppublish.
38. Lyon W: **Fullstack GraphQL applications with GRANDstack**. New York, NY: Manning Publications; 2022.

39. Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, Luo T, Xu L, Liao G, Yan M *et al*: **CellMarker: a manually curated resource of cell markers in human and mouse**. *Nucleic acids research* 2019, **47**(D1):D721-D728.
40. Börner K, Teichmann SA, Quardokus EM, Gee JC, Browne K, Osumi-Sutherland D, Herr BW, 2nd, Bueckle A, Paul H, Haniffa M *et al*: **Anatomical structures, cell types and biomarkers of the Human Reference Atlas**. *Nature cell biology* 2021, **23**(11):1117-1128.
41. Franzén O, Gan LM, Björkegren JLM: **PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data**. *Database : the journal of biological databases and curation* 2019, **2019**:ppublish.
42. Jones RC, Karkanias J, Krasnow MA, Pisco AO, Quake SR, Salzman J, Yosef N, Bulthaup B, Brown P, Harper W *et al*: **The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans**. *Science (New York, NY)* 2022, **376**(6594):eabl4896.
43. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, Zager MA, Aldinger KA, Blecher-Gonen R, Zhang F *et al*: **A human cell atlas of fetal gene expression**. *Science (New York, NY)* 2020, **370**(6518):ppublish.
44. Palasca O, Santos A, Stolte C, Gorodkin J, Jensen LJ: **TISSUES 2.0: an integrative web resource on mammalian tissue expression**. *Database : the journal of biological databases and curation* 2018, **2018**:ppublish.
45. Li Y, Dou Y, Da Veiga Leprevost F, Geffen Y, Calinawan AP, Aguet F, Akiyama Y, Anand S, Birger C, Cao S *et al*: **Proteogenomic data and resources for pan-cancer analysis**. *Cancer cell* 2023, **41**(8):1397-1406.
46. Deng EZ, Marino GB, Clarke DJB, Diamant I, Resnick AC, Ma W, Wang P, Ma'ayan A: **Multiomics2Targets identifies targets from cancer cohorts profiled with transcriptomics, proteomics, and phosphoproteomics**. *Cell reports methods* 2024, **4**(8):100839.
47. Law CW, Chen Y, Shi W, Smyth GK: **voom: Precision weights unlock linear model analysis tools for RNA-seq read counts**. *Genome biology* 2014, **15**(2):R29.
48. Clarke DJB, Marino GB, Ma'ayan A: **A Gene Set Foundation Model Pre-Trained on a Massive Collection of Diverse Gene Sets**. *bioRxiv* 2025.
49. Groza T, Gomez FL, Mashhadi HH, Muñoz-Fuentes V, Gunes O, Wilson R, Cacheiro P, Frost A, Keskivali-Bond P, Vardal B *et al*: **The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease**. *Nucleic acids research* 2022, **51**(D1):D1038-D1045.
50. Wang Z, Lachmann A, Keenan AB, Ma'ayan A: **L1000FWD: fireworks visualization of drug-induced transcriptomic signatures**. *Bioinformatics* 2018, **34**(12):2150-2152.
51. Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, Johnston SE, Vrcic A, Wong B, Khan M *et al*: **The Drug Repurposing Hub: a next-generation drug library and information resource**. *Nat Med* 2017, **23**(4):405-408.
52. Marino GB, Olaiya S, Evangelista JE, Clarke DJB, Ma'ayan A: **GeneSetCart: assembling, augmenting, combining, visualizing, and analyzing gene sets**. *Gigascience* 2025, **14**.
53. **The Genotype-Tissue Expression (GTEx) project**. *Nature genetics* 2013, **45**(6):580-585.
54. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma'ayan A: **Massive mining of publicly available RNA-seq data from human and mouse**. *Nat Commun* 2018, **9**(1):1366.
55. Andrews NC, Erdjument-Bromage H, Davidson MB, Tempst P, Orkin SH: **Erythroid transcription factor NF-E2 is a haematopoietic-specific basic-leucine zipper protein**. *Nature* 1993, **362**(6422):722-728.
56. Kassouf MT, Hughes JR, Taylor S, McGowan SJ, Soneji S, Green AL, Vyas P, Porcher C: **Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells**. *Genome research* 2010, **20**(8):1064-1083.

855 57. Briegel K, Bartunek P, Stengl G, Lim KC, Beug H, Engel JD, Zenke M: **Regulation and**
856 **function of transcription factor GATA-1 during red blood cell differentiation**.
857 *Development (Cambridge, England)* 1996, **122**(12):3839-3850.
858 58. Dijon M, Bardin F, Murati A, Batoz M, Chabannon C, Tonnelle C: **The role of Ikaros in**
859 **human erythroid differentiation**. *Blood* 2008, **111**(3):1138-1146.
860 59. Back J, Dierich A, Bronn C, Kastner P, Chan S: **PU.1 determines the self-renewal**
861 **capacity of erythroid progenitor cells**. *Blood* 2004, **103**(10):3615-3623.
862 60. Capron C, Lécluse Y, Kaushik AL, Foudi A, Lacout C, Sekkai D, Godin I, Albagli O,
863 Poullion I, Svinartchouk F *et al*: **The SCL relative LYL-1 is required for fetal and adult**
864 **hematopoietic stem cell function and B-cell differentiation**. *Blood* 2006,
865 **107**(12):4678-4686.
866 61. Chiu SK, Saw J, Huang Y, Sonderegger SE, Wong NC, Powell DR, Beck D, Pimanda JE,
867 Tremblay CS, Curtis DJ: **A novel role for Lyl1 in primitive erythropoiesis**.
868 *Development (Cambridge, England)* 2018, **145**(19):epublish.
869 62. Cantor AB, Orkin SH: **Transcriptional regulation of erythropoiesis: an affair**
870 **involving multiple partners**. *Oncogene* 2002, **21**(21):3368-3376.
871 63. Mouthon MA, Bernard O, Mitjavila MT, Romeo PH, Vainchenker W, Mathieu-Mahul D:
872 **Expression of tal-1 and GATA-binding proteins during human hematopoiesis**.
873 *Blood* 1993, **81**(3):647-655.
874 64. Guo RJ, Suh ER, Lynch JP: **The role of Cdx proteins in intestinal development and**
875 **cancer**. *Cancer biology & therapy* 2004, **3**(7):593-601.
876 65. Ye DZ, Kaestner KH: **Foxa1 and Foxa2 control the differentiation of goblet and**
877 **enteroendocrine L- and D-cells in mice**. *Gastroenterology* 2009, **137**(6):2052-2062.
878 66. Noah TK, Kazanjian A, Whitsett J, Shroyer NF: **SAM pointed domain ETS factor**
879 **(SPDEF) regulates terminal differentiation and maturation of intestinal goblet cells**.
880 *Experimental cell research* 2010, **316**(3):452-465.
881 67. Garrison WD, Battle MA, Yang C, Kaestner KH, Sladek FM, Duncan SA: **Hepatocyte**
882 **nuclear factor 4alpha is essential for embryonic development of the mouse colon**.
883 *Gastroenterology* 2006, **130**(4):1207-1220.
884 68. Luk IY, Reehorst CM, Mariadason JM: **ELF3, ELF5, EHF and SPDEF Transcription**
885 **Factors in Tissue Homeostasis and Cancer**. *Molecules (Basel, Switzerland)* 2018,
886 **23**(9):epublish.
887 69. Naiche LA, Harrelson Z, Kelly RG, Papaioannou VE: **T-box genes in vertebrate**
888 **development**. *Annual review of genetics* 2005, **39**:219-239.
889 70. Lo YH, Chung E, Li Z, Wan YW, Mahe MM, Chen MS, Noah TK, Bell KN, Yalamanchili
890 HK, Klisch TJ *et al*: **Transcriptional Regulation by ATOH1 and its Target SPDEF**
891 **in the Intestine**. *Cellular and molecular gastroenterology and hepatology* 2017, **3**(1):51-
892 71.
893 71. Liu Q, Linney E: **The mouse retinoid-X receptor-gamma gene: genomic organization**
894 **and evidence for functional isoforms**. *Molecular endocrinology (Baltimore, Md)* 1993,
895 **7**(5):651-658.
896 72. L'Honoré A, Coulon V, Marcil A, Lebel M, Lafrance-Vanasse J, Gage P, Camper S,
897 Drouin J: **Sequential expression and redundancy of Pitx2 and Pitx3 genes during**
898 **muscle development**. *Developmental biology* 2007, **307**(2):421-433.
899 73. Olson EN, Brennan TJ, Chakraborty T, Cheng TC, Cserjesi P, Edmondson D, James G,
900 Li L: **Molecular control of myogenesis: antagonism between growth and**
901 **differentiation**. *Molecular and cellular biochemistry* 1991, **104**(1-2):7-13.
902 74. Zammit PS: **Function of the myogenic regulatory factors Myf5, MyoD, Myogenin**
903 **and MRF4 in skeletal muscle, satellite cells and regenerative myogenesis**. *Seminars*
904 *in cell & developmental biology* 2017, **72**:19-32.
905 75. Xiang FL, Guo M, Yutzey KE: **Overexpression of Tbx20 in Adult Cardiomyocytes**
906 **Promotes Proliferation and Improves Cardiac Function After Myocardial Infarction**.
907 *Circulation* 2016, **133**(11):1081-1092.
908 76. Steimle JD, Moskowitz IP: **TBX5: A Key Regulator of Heart Development**. *Current*
909 *topics in developmental biology* 2017, **122**:195-221.
910 77. Brohmann H, Jagla K, Birchmeier C: **The role of Lbx1 in migration of muscle**
911 **precursor cells**. *Development (Cambridge, England)* 2000, **127**(2):437-445.

78. Ross MD, Martinka S, Mukherjee A, Sedor JR, Vinson C, Bruggeman LA: **Math6 expression during kidney development and altered expression in a mouse model of glomerulosclerosis**. *Developmental dynamics : an official publication of the American Association of Anatomists* 2006, **235**(11):3102-3109.
79. Liu G, Clement LC, Kanwar YS, Avila-Casado C, Chugh SS: **ZHX proteins regulate podocyte gene expression during the development of nephrotic syndrome**. *The Journal of biological chemistry* 2006, **281**(51):39681-39692.
80. Chugh SS: **Transcriptional regulation of podocyte disease**. *Translational research : the journal of laboratory and clinical medicine* 2007, **149**(5):237-242.
81. Zhou D, Fu H, Han Y, Zhang L, Liu S, Lin L, Stolz DB, Liu Y: **Sonic hedgehog connects podocyte injury to mesangial activation and glomerulosclerosis**. *JCI insight* 2019, **4**(22):epublish.
82. Attanasio M, Uhlenhaut NH, Sousa VH, O'Toole JF, Otto E, Anlag K, Klugmann C, Treier AC, Helou J, Sayer JA *et al*: **Loss of GLIS2 causes nephronophthisis in humans and mice by increased apoptosis and fibrosis**. *Nature genetics* 2007, **39**(8):1018-1024.
83. Uhlenhaut NH, Treier M: **Transcriptional regulators in kidney disease: gatekeepers of renal homeostasis**. *Trends in genetics : TIG* 2008, **24**(7):361-371.
84. Du L, Lu Y, Wang J, Zheng Y, Li H, Liu Y, Wu X, Zhou J, Wang L, He L *et al*: **LncRNA KIFAP3-5:1 inhibits epithelial-mesenchymal transition of renal tubular cell through PRRX1 in diabetic nephropathy**. *Cell biology and toxicology* 2024, **40**(1):47.
85. Li Y, Kang YS, Dai C, Kiss LP, Wen X, Liu Y: **Epithelial-to-mesenchymal transition is a potential pathway leading to podocyte dysfunction and proteinuria**. *The American journal of pathology* 2008, **172**(2):299-308.
86. Sundararajan V, Tan M, Tan TZ, Ye J, Thiery JP, Huang RY: **SNAI1 recruits HDAC1 to suppress SNAI2 transcription during epithelial to mesenchymal transition**. *Scientific reports* 2019, **9**(1):8295.
87. Ocaña OH, Córcoles R, Fabra A, Moreno-Bueno G, Acloque H, Vega S, Barrallo-Gimeno A, Cano A, Nieto MA: **Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer Prrx1**. *Cancer cell* 2012, **22**(6):709-724.
88. Chiba T, Uchi H, Tsuji G, Gondo H, Moroi Y, Furue M: **Arylhydrocarbon receptor (AhR) activation in airway epithelial cells induces MUC5AC via reactive oxygen species (ROS) production**. *Pulmonary pharmacology & therapeutics* 2011, **24**(1):133-140.
89. Gendler SJ, Spicer AP: **Epithelial mucin genes**. *Annual review of physiology* 1995, **57**:607-634.
90. Sarode P, Zheng X, Giotopoulou GA, Weigert A, Kuenne C, Günther S, Friedrich A, Gattenlöhner S, Stiewe T, Brüne B *et al*: **Reprogramming of tumor-associated macrophages by targeting β-catenin/FOSL2/ARID5A signaling: A potential treatment of lung cancer**. *Science advances* 2020, **6**(23):eaaz6105.
91. Hu C, Perlmutter DH: **Cell-specific involvement of HNF-1beta in alpha(1)-antitrypsin gene expression in human respiratory epithelial cells**. *American journal of physiology Lung cellular and molecular physiology* 2002, **282**(4):L757-765.
92. Boyd M, Hansen M, Jensen TG, Perearnau A, Olsen AK, Bram LL, Bak M, Tommerup N, Olsen J, Troelsen JT: **Genome-wide analysis of CDX2 binding in intestinal epithelial cells (Caco-2)**. *The Journal of biological chemistry* 2010, **285**(33):25115-25125.
93. Ikhapoh IA, Pelham CJ, Agrawal DK: **Sry-type HMG box 18 contributes to the differentiation of bone marrow-derived mesenchymal stem cells to endothelial cells**. *Differentiation; research in biological diversity* 2015, **89**(3-4):87-96.
94. He S, Kim I, Lim MS, Morrison SJ: **Sox17 expression confers self-renewal potential and fetal stem cell characteristics upon adult hematopoietic progenitors**. *Genes & development* 2011, **25**(15):1613-1627.
95. Sun L, Burnett J, Gasparyan M, Xu F, Jiang H, Lin CC, Myers I, Korkaya H, Liu Y, Connarn J *et al*: **Novel cancer stem cell targets during epithelial to mesenchymal transition in PTEN-deficient trastuzumab-resistant breast cancer**. *Oncotarget* 2016, **7**(32):51408-51422.
96. Kamata M, Okitsu Y, Fujiwara T, Kanehira M, Nakajima S, Takahashi T, Inoue A, Fukuhara N, Onishi Y, Ishizawa K *et al*: **GATA2 regulates differentiation of bone marrow-derived mesenchymal stem cells**. *Haematologica* 2014, **99**(11):1686-1696.

969 97. Sharrocks AD: **The ETS-domain transcription factor family**. *Nature reviews Molecular*
970 *cell biology* 2001, **2**(11):827-837.
971 98. Maroulakou IG, Bowe DB: **Expression and function of Ets transcription factors in**
972 **mammalian development: a regulatory network**. *Oncogene* 2000, **19**(55):6432-6442.
973 99. Leshem O, Madar S, Kogan-Sakin I, Kamer I, Goldstein I, Brosh R, Cohen Y, Jacob-
974 Hirsch J, Ehrlich M, Ben-Sasson S *et al*: **TMPRSS2/ERG promotes epithelial to**
975 **mesenchymal transition through the ZEB1/ZEB2 axis in a prostate cancer model**.
976 *PloS one* 2011, **6**(7):e21650.
977 100. Vijayaraj P, Le Bras A, Mitchell N, Kondo M, Juliao S, Wasserman M, Beeler D, Spokes
978 K, Aird WC, Baldwin HS *et al*: **Erg is a crucial regulator of endocardial-mesenchymal**
979 **transformation during cardiac valve morphogenesis**. *Development (Cambridge,*
980 *England)* 2012, **139**(21):3973-3985.
981 101. Bialek P, Kern B, Yang X, Schrock M, Sosic D, Hong N, Wu H, Yu K, Ornitz DM, Olson
982 EN *et al*: **A twist code determines the onset of osteoblast differentiation**.
983 *Developmental cell* 2004, **6**(3):423-435.
984 102. Liao L, Su X, Yang X, Hu C, Li B, Lv Y, Shuai Y, Jing H, Deng Z, Jin Y: **TNF-α Inhibits**
985 **FoxO1 by Upregulating miR-705 to Aggravate Oxidative Damage in Bone Marrow-**
986 **Derived Mesenchymal Stem Cells during Osteoporosis**. *Stem cells (Dayton, Ohio)*
987 2016, **34**(4):1054-1067.
988 103. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF,
989 Tubelli AA, Asiedu JK *et al*: **A Next Generation Connectivity Map: L1000 Platform and**
990 **the First 1,000,000 Profiles**. *Cell* 2017, **171**(6):1437-1452.e1417.
991 104. Zheng L, Xu M, Xu J, Wu K, Fang Q, Liang Y, Zhou S, Cen D, Ji L, Han W *et al*: **ELF3**
992 **promotes epithelial-mesenchymal transition by protecting ZEB1 from miR-141-3p-**
993 **mediated silencing in hepatocellular carcinoma**. *Cell Death Dis* 2018, **9**(3):387.
994 105. Albino D, Longoni N, Curti L, Mello-Grand M, Pinton S, Civenni G, Thalmann G,
995 D'Ambrosio G, Sarti M, Sessa F *et al*: **ESE3/EHF controls epithelial cell differentiation**
996 **and its loss leads to prostate tumors with mesenchymal and stem-like features**.
997 *Cancer Res* 2012, **72**(11):2889-2900.
998 106. Ke CY, Xiao WL, Chen CM, Lo LJ, Wong FH: **IRF6 is the mediator of TGFβ3 during**
999 **regulation of the epithelial mesenchymal transition and palatal fusion**. *Scientific*
1000 *reports* 2015, **5**:12791.
1001 107. Qiao Y, Jiang X, Lee ST, Karuturi RK, Hooi SC, Yu Q: **FOXQ1 regulates epithelial-**
1002 **mesenchymal transition in human cancers**. *Cancer Res* 2011, **71**(8):3076-3086.
1003 108. Koplev S, Lin K, Dohlman AB, Ma'ayan A: **Integration of pan-cancer transcriptomics**
1004 **with RPPA proteomics reveals mechanisms of epithelial-mesenchymal transition**.
1005 *PLoS Comput Biol* 2018, **14**(1):e1005911.
1006 109. Cole SW, Sood AK: **Molecular pathways: beta-adrenergic signaling in cancer**. *Clin*
1007 *Cancer Res* 2012, **18**(5):1201-1206.
1008 110. Chang A, Botteri E, Gillis RD, Löfling L, Le CP, Ziegler AI, Chung NC, Rowe MC, Fabb
1009 SA, Hartley BJ *et al*: **Beta-blockade enhances anthracycline control of metastasis in**
1010 **triple-negative breast cancer**. *Sci Transl Med* 2023, **15**(693):eadf1147.
1011 111. Løfling LL, Støer NC, Sloan EK, Chang A, Gandini S, Ursin G, Botteri E: **β-blockers and**
1012 **breast cancer survival by molecular subtypes: a population-based cohort study**
1013 **and meta-analysis**. *Br J Cancer* 2022, **127**(6):1086-1096.
1014 112. Vacca A, Itoh M, Kawaji H, Arner E, Lassmann T, Daub CO, Carninci P, Forrest ARR,
1015 Hayashizaki Y, Aitken S *et al*: **Conserved temporal ordering of promoter activation**
1016 **implicates common mechanisms governing the immediate early response across**
1017 **cell types and stimuli**. *Open Biol* 2018, **8**(8).
1018 113. Montalvo-Ortiz JL, Keegan J, Gallardo C, Gerst N, Tetsuka K, Tucker C, Matsumoto M,
1019 Fang D, Csernansky JG, Dong H: **HDAC inhibitors restore the capacity of aged mice**
1020 **to respond to haloperidol through modulation of histone acetylation**.
1021 *Neuropsychopharmacology* 2014, **39**(6):1469-1478.
1022 114. Akaboshi S, Watanabe S, Hino Y, Sekita Y, Xi Y, Araki K, Yamamura K, Oshima M, Ito T,
1023 Baba H *et al*: **HMGA1 is induced by Wnt/beta-catenin pathway and maintains cell**
1024 **proliferation in gastric cancer**. *The American journal of pathology* 2009, **175**(4):1675-
1025 1685.

115. Wierstra I, Alves J: **FOXM1, a typical proliferation-associated transcription factor**. *Biol Chem* 2007, **388**(12):1257-1274.
116. Musa J, Aynaud MM, Mirabeau O, Delattre O, Grünewald TG: **MYBL2 (B-Myb): a central regulator of cell proliferation, cell survival and differentiation involved in tumorigenesis**. *Cell Death Dis* 2017, **8**(6):e2895.
117. de Bruin A, Maiti B, Jakoi L, Timmers C, Buerki R, Leone G: **Identification and characterization of E2F7, a novel mammalian E2F family member capable of blocking cellular proliferation**. *The Journal of biological chemistry* 2003, **278**(43):42041-42049.
118. Yang R, Wang M, Zhang G, Bao Y, Wu Y, Li X, Yang W, Cui H: **E2F7-EZH2 axis regulates PTEN/AKT/mTOR signalling and glioblastoma progression**. *Br J Cancer* 2020, **123**(9):1445-1455.
119. Liao J, Jiang L, Wang C, Zhao D, He W, Zhou K, Liang Y: **FoxM1 Regulates Proliferation and Apoptosis of Human Neuroblastoma Cell through PI3K/AKT Pathway**. *Fetal Pediatr Pathol* 2022, **41**(3):355-370.
120. Li Z, Zhou H, Xia Z, Xia T, Du G, Franziska SD, Li X, Zhai X, Jin B: **HMGA1 augments palbociclib efficacy via PI3K/mTOR signaling in intrahepatic cholangiocarcinoma**. *Biomark Res* 2023, **11**(1):33.
121. Deng Q, Wu L, Li Y, Zou L: **MYBL2 in synergy with CDC20 promotes the proliferation and inhibits apoptosis of gastric cancer cells**. *Adv Clin Exp Med* 2021, **30**(9):957-966.
122. Lei T, Gao Y, Duan Y, Cui C, Zhang L, Si M: **Inhibition of zinc finger protein 367 exerts a tumor suppressive role in colorectal cancer by affecting the activation of oncogenic YAP1 signaling**. *Environ Toxicol* 2021, **36**(11):2278-2290.
123. Kojima T, Shimazui T, Hinotsu S, Joraku A, Oikawa T, Kawai K, Horie R, Suzuki H, Nagashima R, Yoshikawa K *et al*: **Decreased expression of CXXC4 promotes a malignant phenotype in renal cell carcinoma by activating Wnt signaling**. *Oncogene* 2009, **28**(2):297-305.
124. Li X, Xu Y, Chen Y, Chen S, Jia X, Sun T, Liu Y, Li X, Xiang R, Li N: **SOX2 promotes tumor metastasis by stimulating epithelial-to-mesenchymal transition via regulation of WNT/β-catenin signal network**. *Cancer Lett* 2013, **336**(2):379-389.
125. Guan L, Li T, Ai N, Wang W, He B, Bai Y, Yu Z, Li M, Dong S, Zhu Q *et al*: **MEIS2C and MEIS2D promote tumor progression via Wnt/β-catenin and hippo/YAP signaling in hepatocellular carcinoma**. *J Exp Clin Cancer Res* 2019, **38**(1):417.
126. Fischer A, Schumacher N, Maier M, Sendtner M, Gessler M: **The Notch target genes Hey1 and Hey2 are required for embryonic vascular development**. *Genes & development* 2004, **18**(8):901-911.
127. Ohnaka K, Tanabe M, Kawate H, Nawata H, Takayanagi R: **Glucocorticoid suppresses the canonical Wnt signal in cultured human osteoblasts**. *Biochem Biophys Res Commun* 2005, **329**(1):177-181.
128. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N: **The genotype-tissue expression (GTEx) project**. *Nature genetics* 2013, **45**(6):580-585.
129. Lobo GP, Amengual J, Baus D, Shivdasani RA, Taylor D, von Lintig J: **Genetics and diet regulate vitamin A production via the homeobox transcription factor ISX**. *The Journal of biological chemistry* 2013, **288**(13):9017-9027.
130. Wang LT, Liu KY, Chiou SS, Huang SK, Hsu SH, Wang SN: **Phosphorylation of intestine-specific homeobox by ERK1 modulates oncogenic activity and sorafenib resistance**. *Cancer Lett* 2021, **520**:160-171.
131. Lau HH, Ng NHJ, Loo LSW, Jasmen JB, Teo AKK: **The molecular functions of hepatocyte nuclear factors - In and beyond the liver**. *J Hepatol* 2018, **68**(5):1033-1048.
132. de Aguiar Vallim TQ, Tarling EJ, Edwards PA: **Pleiotropic roles of bile acids in metabolism**. *Cell Metab* 2013, **17**(5):657-669.
133. Chang X, Tian C, Jia Y, Cai Y, Yan P: **MLXIPL promotes the migration, invasion, and glycolysis of hepatocellular carcinoma cells by phosphorylation of mTOR**. *BMC Cancer* 2023, **23**(1):176.

1082 134. Smith SB, Qu HQ, Taleb N, Kishimoto NY, Scheel DW, Lu Y, Patch AM, Grabs R, Wang
1083 J, Lynn FC *et al*: **Rfx6 directs islet formation and insulin production in mice and**
1084 **humans**. *Nature* 2010, **463**(7282):775-780.
1085 135. Yu J, Liu D, Sun X, Yang K, Yao J, Cheng C, Wang C, Zheng J: **CDX2 inhibits the**
1086 **proliferation and tumor formation of colon cancer cells by suppressing Wnt/β-**
1087 **catenin signaling via transactivation of GSK-3β and Axin2 expression**. *Cell Death*
1088 *Dis* 2019, **10**(1):26.
1089 136. Ju Y, Fang S, Liu L, Ma H, Zheng L: **The function of the ELF3 gene and its**
1090 **mechanism in cancers**. *Life Sci* 2024, **346**:122637.
1091 137. Morrisey EE, Ip HS, Tang Z, Lu MM, Parmacek MS: **GATA-5: a transcriptional**
1092 **activator expressed in a novel temporally and spatially-restricted pattern during**
1093 **embryonic development**. *Developmental biology* 1997, **183**(1):21-36.
1094 138. Silva-García CG: **Devo-Aging: Intersections Between Development and Aging**.
1095 *Geroscience* 2023, **45**(4):2145-2159.

1096

Fig. 1

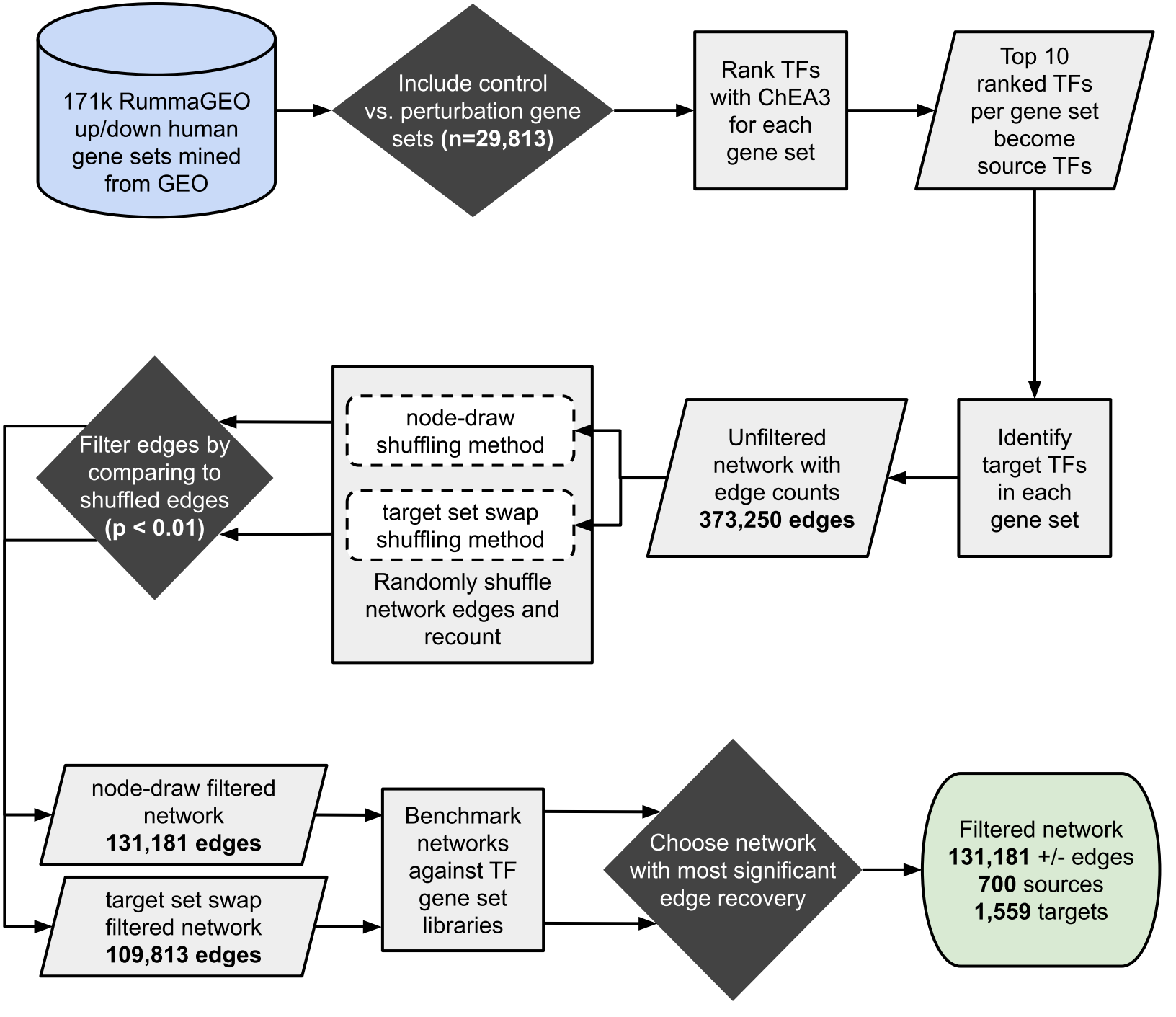

171k RummaGEO up/down human gene sets mined from GEO
Include control vs. perturbation gene sets (n=29,813)
Rank TFs with ChEA3 for each gene set
Top 10 ranked TFs per gene set become source TFs
Identify target TFs in each gene set
Unfiltered network with edge counts 373,250 edges
node-draw shuffling method
target set swap shuffling method
Randomly shuffle network edges and recount
Filter edges by comparing to shuffled edges (p < 0.01)
node-draw filtered network 131,181 edges
target set swap filtered network 109,813 edges
Benchmark networks against TF gene set libraries
Choose network with most significant edge recovery
Filtered network 131,181 +/- edges 700 sources 1,559 targets

Fig. 2

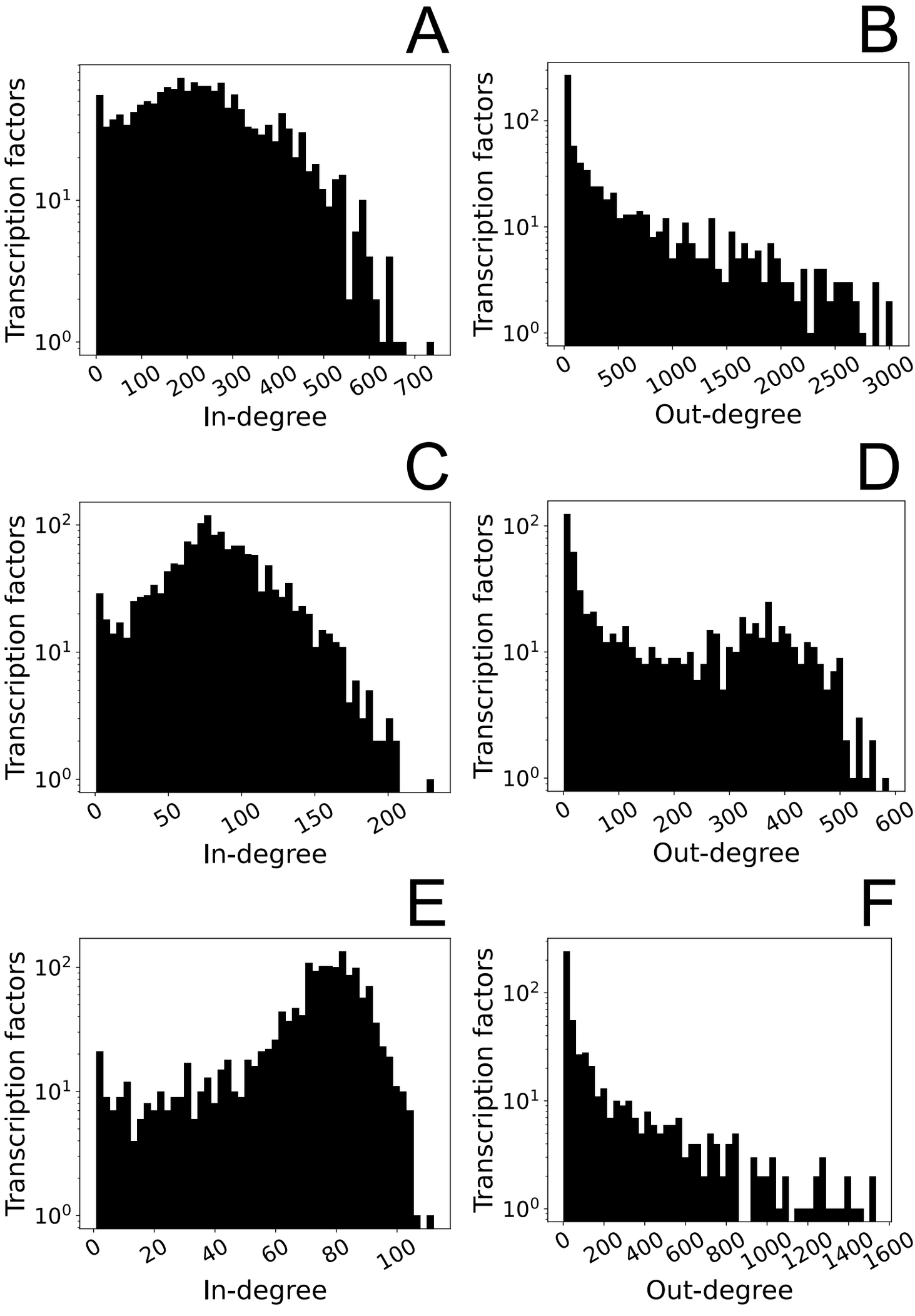

A
B
C
D
E
F
Transcription factors
In-degree
Out-degree
10^0
10^1
10^2
0
100
200
300
400
500
600
700
500
1000
1500
2000
2500
3000
50
150
20
40
60
80
1200
1400
1600

Fig. 3

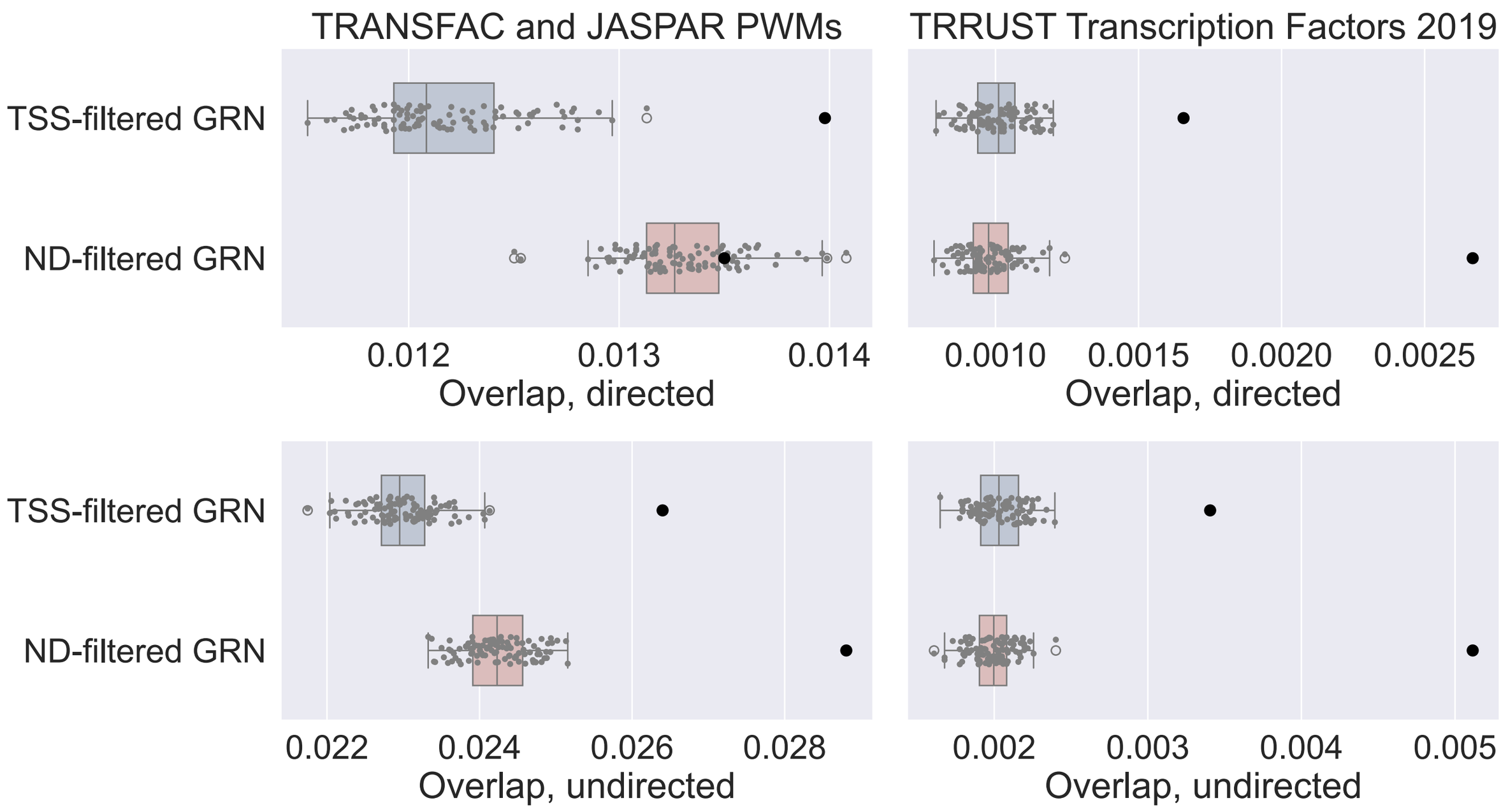

TRANSFAC and JASPAR PWMs
TRRUST Transcription Factors 2019
TSS-filtered GRN
ND-filtered GRN
0.012
0.013
0.014
Overlap, directed
0.0010
0.0015
0.0020
0.0025
Overlap, directed
TSS-filtered GRN
ND-filtered GRN
0.022
0.024
0.026
0.028
Overlap, undirected
0.002
0.003
0.004
0.005
Overlap, undirected

Fig. 4

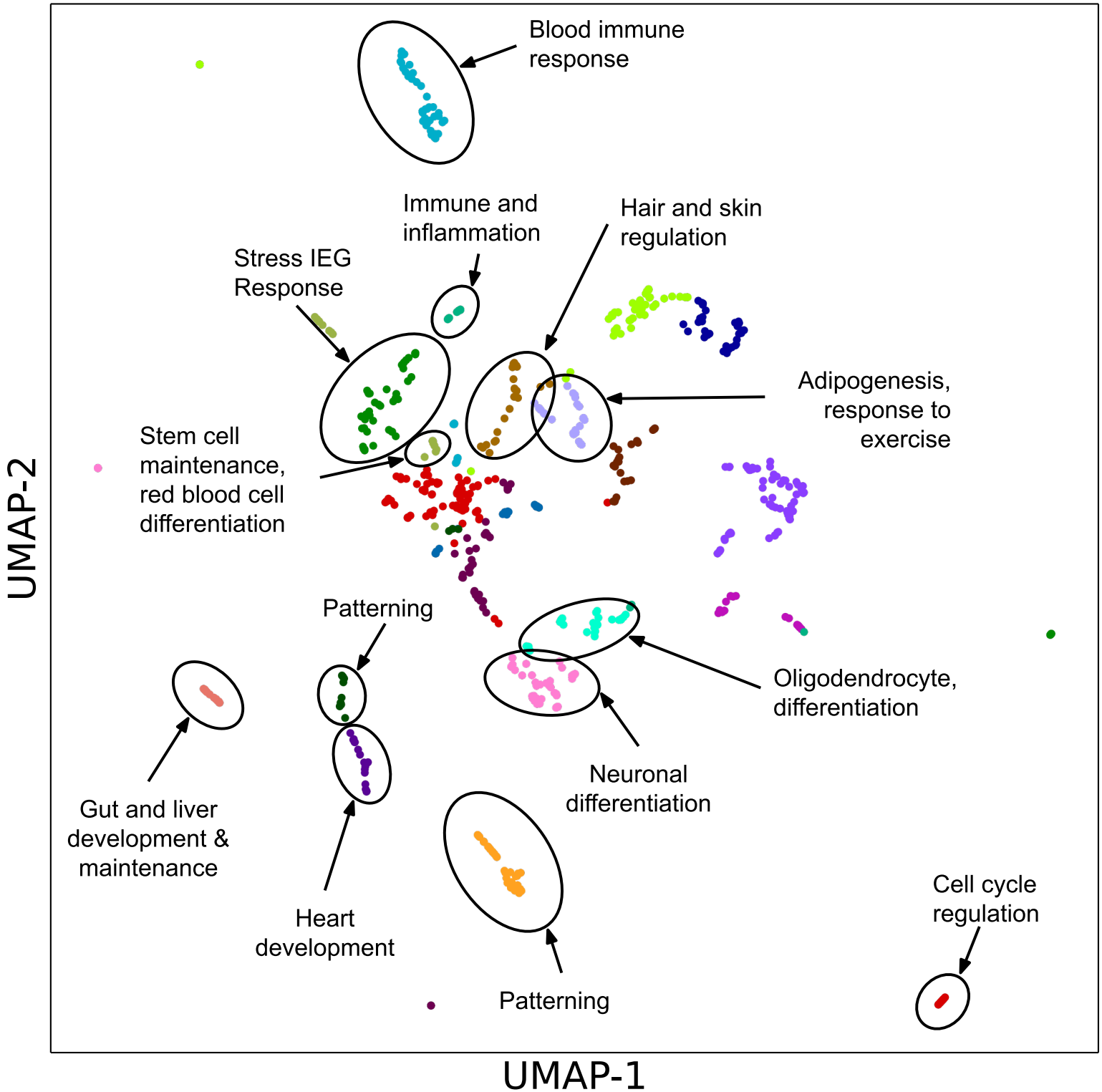

Blood immune response
Immune and inflammation
Hair and skin regulation
Stress IEG Response
Adipogenesis, response to exercise
Stem cell maintenance, red blood cell differentiation
Patterning
Oligodendrocyte, differentiation
Neuronal differentiation
Gut and liver development & maintenance
Heart development
Cell cycle regulation
Patterning
UMAP-2
UMAP-1

Fig. 5

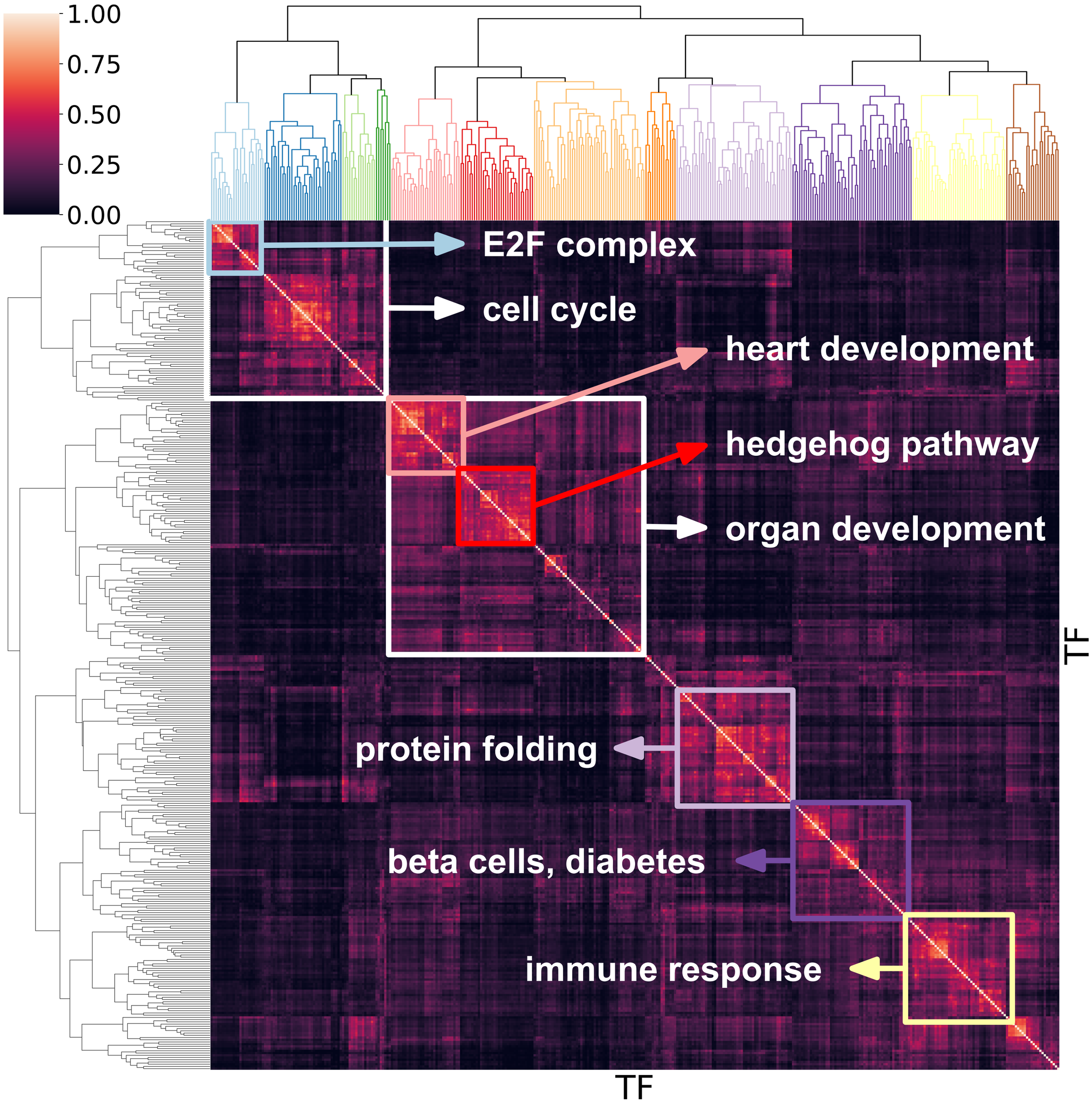

1.00
0.75
0.50
0.25
0.00
E2F complex
cell cycle
heart development
hedgehog pathway
organ development
protein folding
beta cells, diabetes
immune response
TF
TF

Fig. 6

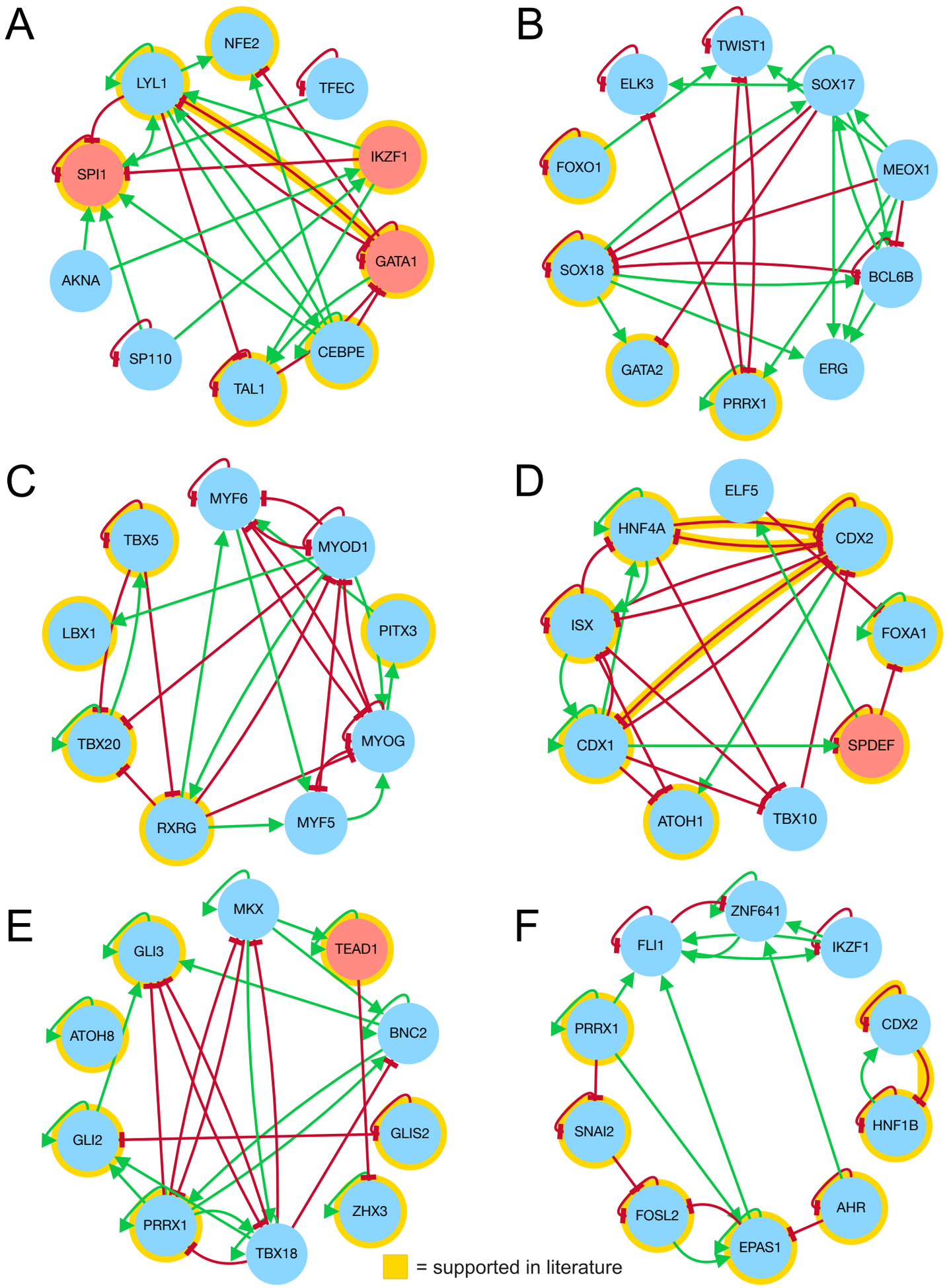
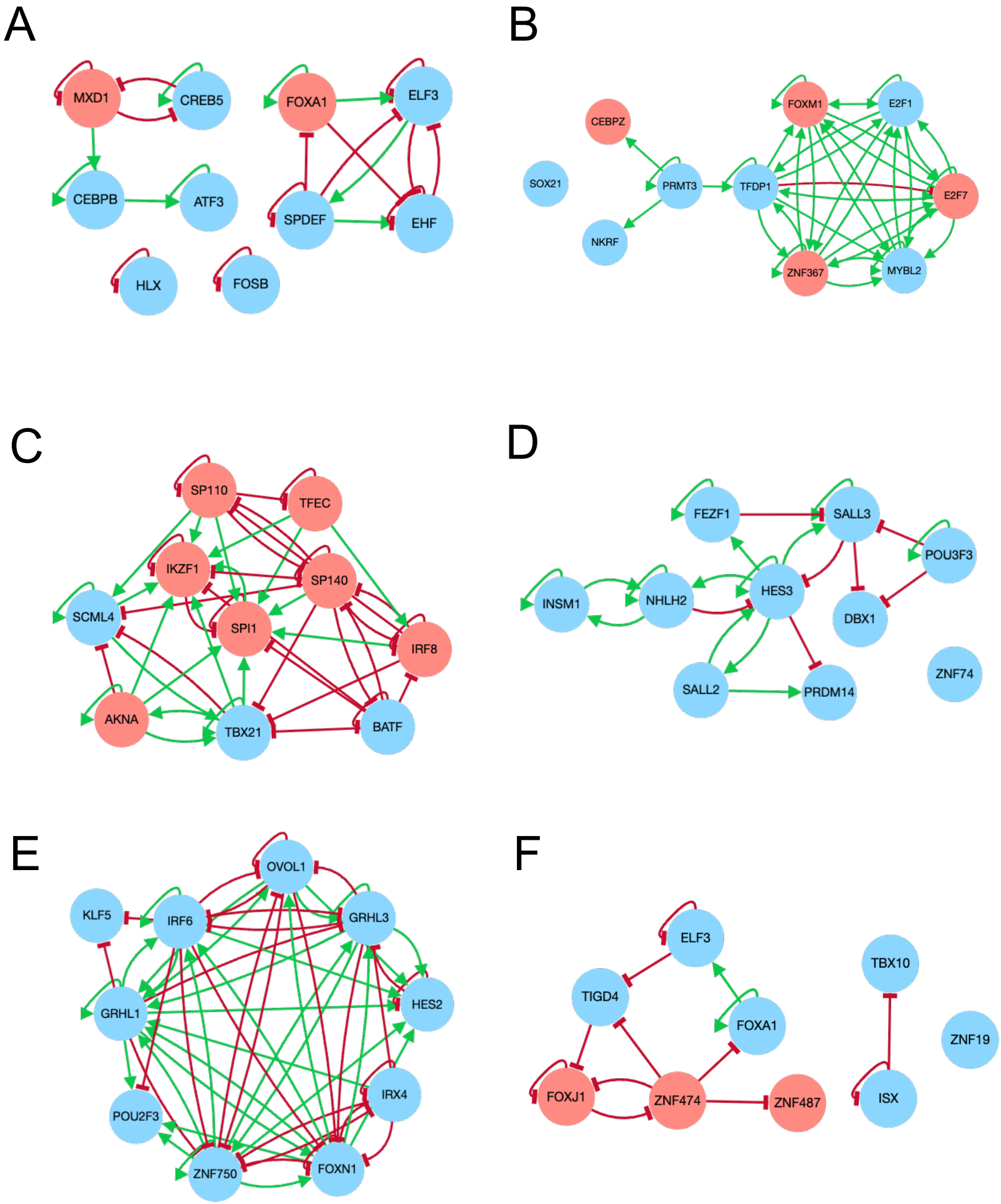
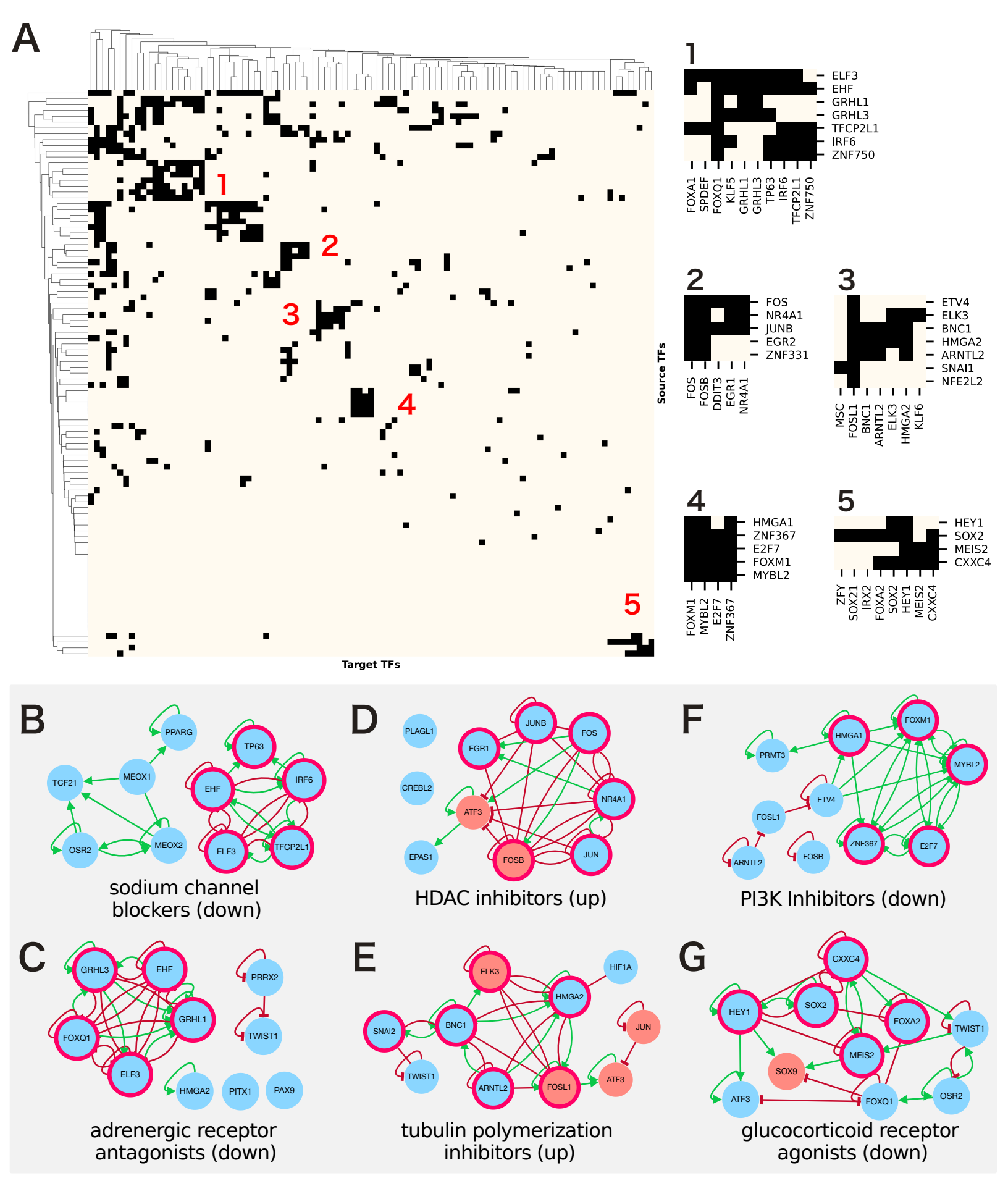

A
NFE2
LYL1
TFEC
IKZF1
SPI1
GATA1
AKNA
SP110
CEBPE
TAL1
B
TWIST1
ELK3
SOX17
FOXO1
MEOX1
SOX18
BCL6B
GATA2
ERG
PRRX1
C
MYF6
TBX5
MYOD1
LBX1
PITX3
TBX20
MYOG
RXRG
MYF5
D
ELF5
HNF4A
CDX2
ISX
FOXA1
CDX1
SPDEF
ATOH1
TBX10
E
MKX
GLI3
TEAD1
ATOH8
BNC2
GLI2
GLIS2
PRRX1
ZHX3
TBX18
F
ZNF641
FLI1
IKZF1
PRRX1
CDX2
SNAI2
HNF1B
FOSL2
AHR
EPAS1
= supported in literature

Fig. 7


A
MXD1
CREB5
CEBPB
ATF3
HLX
FOSB
FOXA1
ELF3
SPDEF
EHF
B
CEBPZ
SOX21
PRMT3
NKRF
TFDP1
FOXM1
E2F1
E2F7
ZNF367
MYBL2
C
SP110
TFEC
IKZF1
SP140
SCML4
SPI1
IRF8
AKNA
TBX21
BATF
D
FEZF1
SALL3
POU3F3
INSM1
NHLH2
HES3
DBX1
SALL2
PRDM14
ZNF74
E
OVOL1
KLF5
IRF6
GRHL3
GRHL1
HES2
IRX4
POU2F3
ZNF750
FOXN1
F
ELF3
TIGD4
FOXA1
TBX10
ZNF19
FOXJ1
ZNF474
ZNF487
ISX


A
1
2
3
4
5
Target TFs
Source TFs
ELF3
EHF
GRHL1
GRHL3
TFCP2L1
IRF6
ZNF750
FOXA1
SPDEF
FOXQ1
KLF5
GRHL1
GRHL3
TP63
IRF6
TFCP2L1
ZNF750
FOS
NR4A1
JUNB
EGR2
ZNF331
FOS
FOSB
DDIT3
EGR1
NR4A1
ETV4
ELK3
BNC1
HMGA2
ARNTL2
SNAI1
NFE2L2
MSC
FOSL1
BNC1
ARNTL2
ELK3
HMGA2
KLF6
HMGA1
ZNF367
E2F7
FOXM1
MYBL2
FOXM1
MYBL2
E2F7
ZNF367
HEY1
SOX2
MEIS2
CXXC4
ZFY
SOX21
IRX2
FOXA2
SOX2
HEY1
MEIS2
CXXC4
B
PPARG
TP63
TCF21
MEOX1
EHF
IRF6
OSR2
MEOX2
ELF3
TFCP2L1
sodium channel blockers (down)
D
PLAGL1
JUNB
FOS
EGR1
CREBL2
NR4A1
ATF3
EPAS1
FOSB
JUN
HDAC inhibitors (up)
F
FOXM1
HMGA1
PRMT3
MYBL2
ETV4
FOSL1
ZNF367
E2F7
ARNTL2
FOSB
PI3K Inhibitors (down)
C
GRHL3
EHF
PRRX2
GRHL1
FOXQ1
TWIST1
ELF3
HMGA2
PITX1
PAX9
adrenergic receptor antagonists (down)
E
ELK3
HIF1A
HMGA2
SNAI2
BNC1
JUN
TWIST1
ARNTL2
FOSL1
ATF3
tubulin polymerization inhibitors (up)
G
CXXC4
HEY1
SOX2
FOXA2
TWIST1
MEIS2
SOX9
ATF3
FOXQ1
OSR2
glucocorticoid receptor agonists (down)

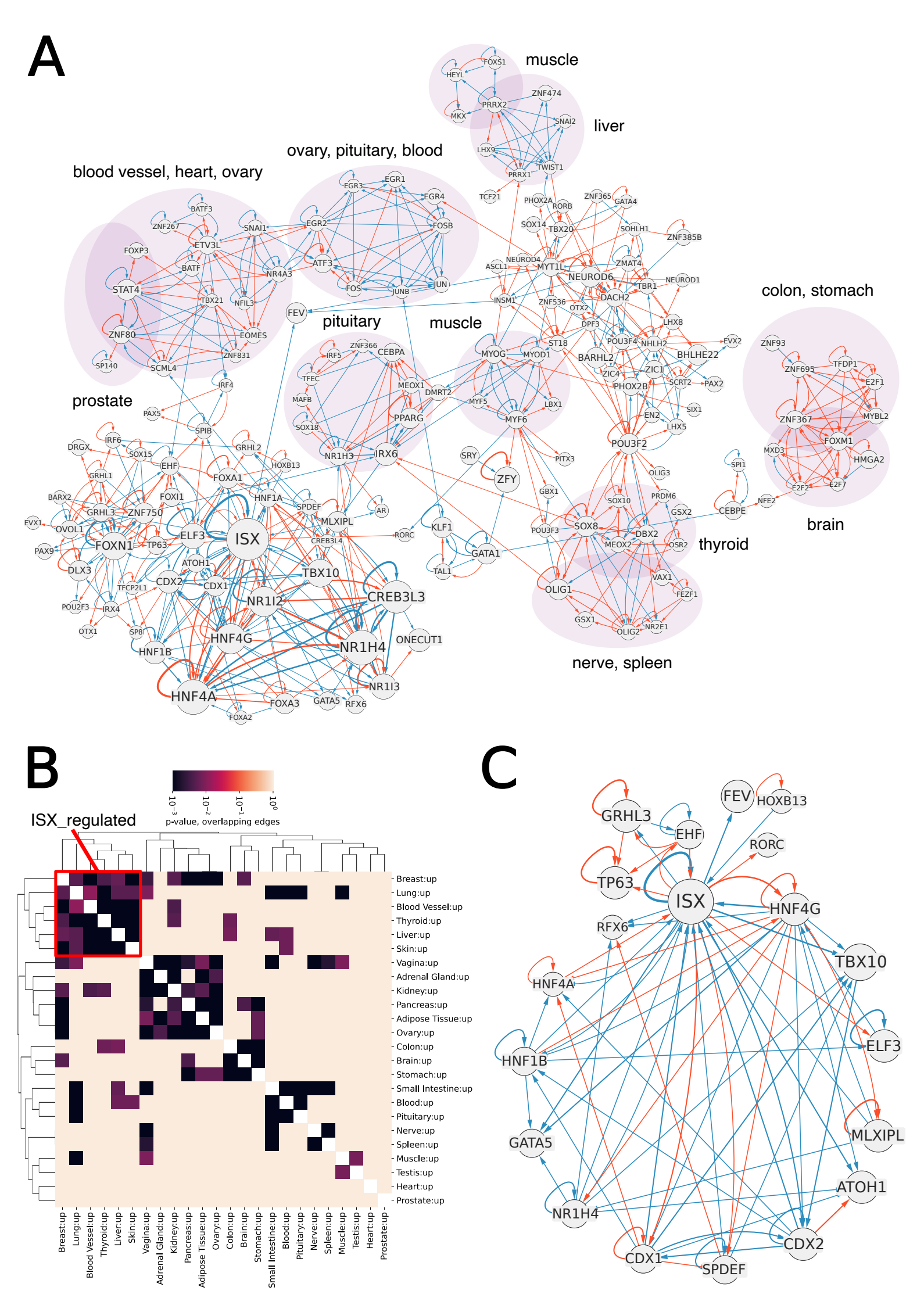

A
muscle
liver
ovary, pituitary, blood
blood vessel, heart, ovary
pituitary
muscle
colon, stomach
prostate
brain
thyroid
nerve, spleen
ISX
B
ISX_regulated
p-value, overlapping edges
Breast:up
Lung:up
Blood Vessel:up
Thyroid:up
Liver:up
Skin:up
Vagina:up
Adrenal Gland:up
Kidney:up
Pancreas:up
Adipose Tissue:up
Ovary:up
Colon:up
Brain:up
Stomach:up
Small Intestine:up
Blood:up
Pituitary:up
Nerve:up
Spleen:up
Muscle:up
Testis:up
Heart:up
Prostate:up
C
ISX
FEV
HOXB13
GRHL3
EHF
RORC
TP63
HNF4G
RFX6
TBX10
HNF4A
ELF3
HNF1B
MLXIPL
GATA5
ATOH1
NR1H4
CDX2
CDX1
SPDEF

Fig. S1

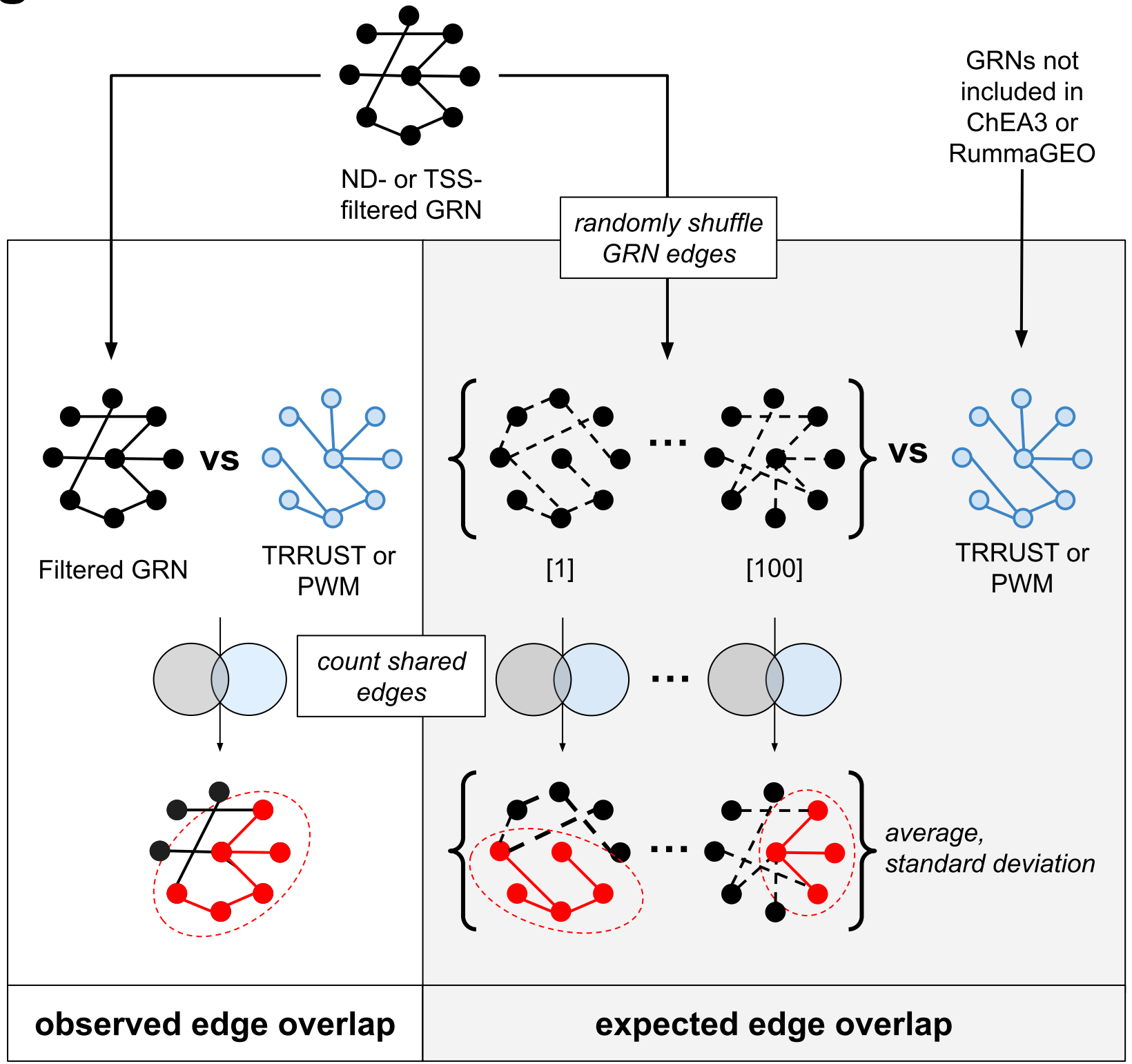
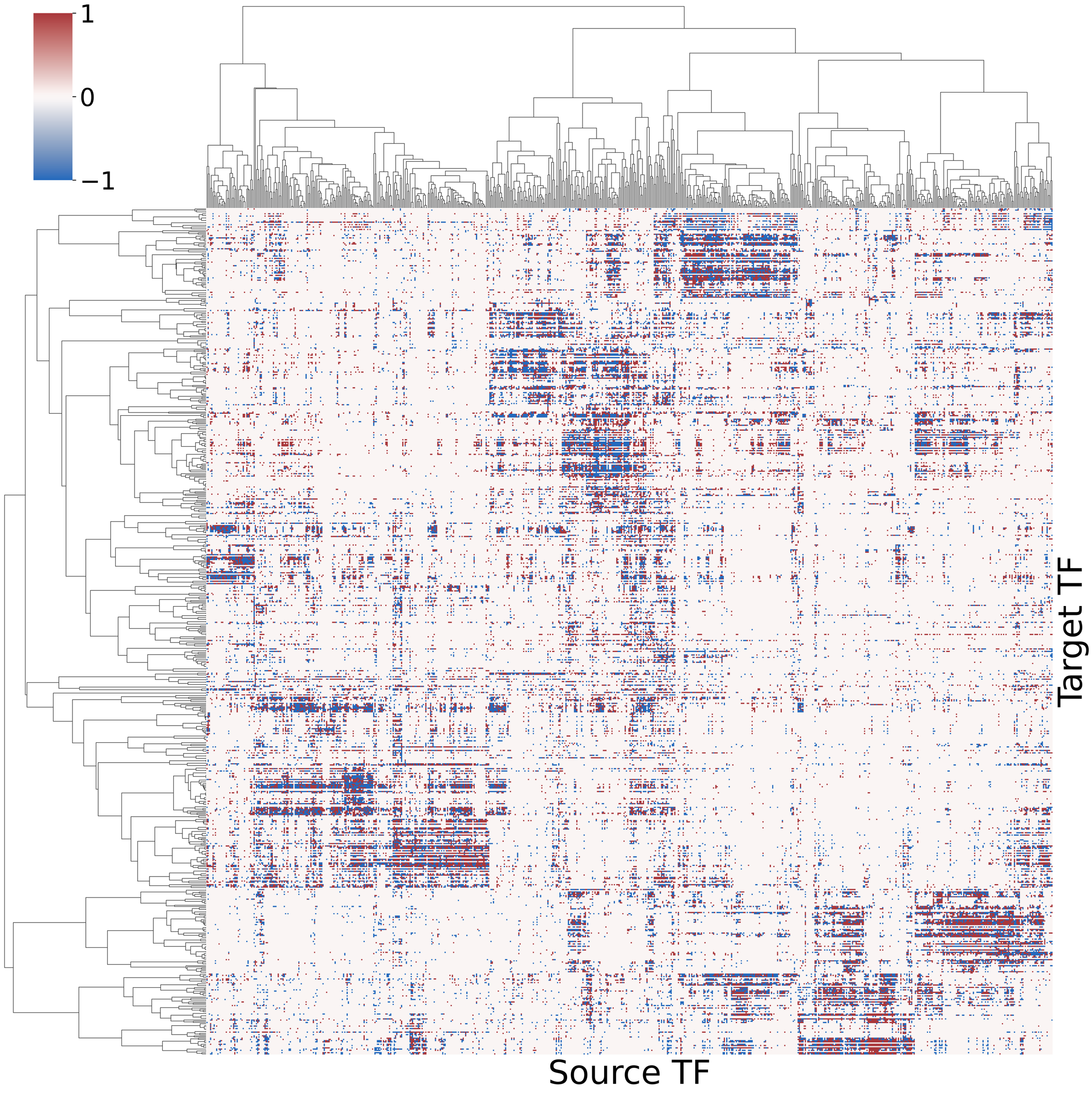
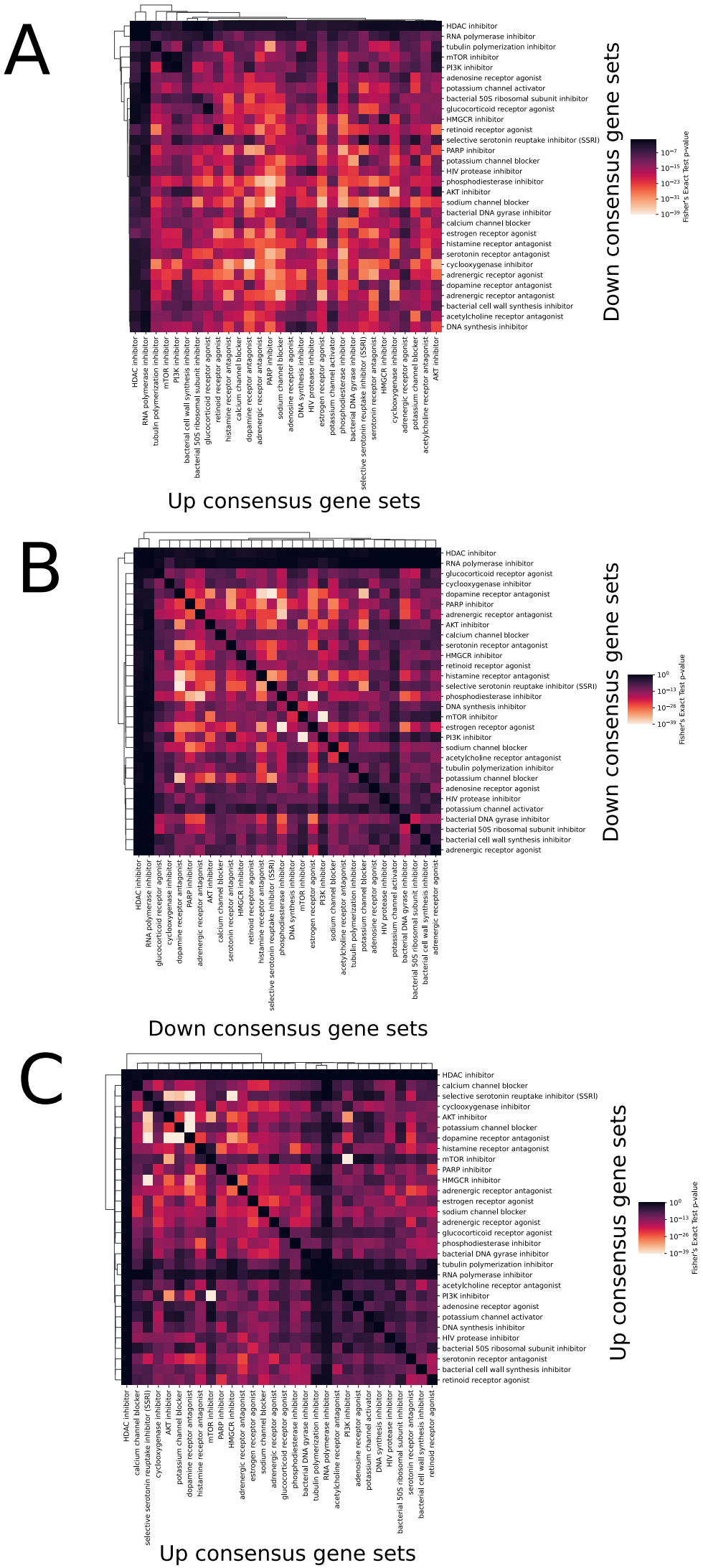
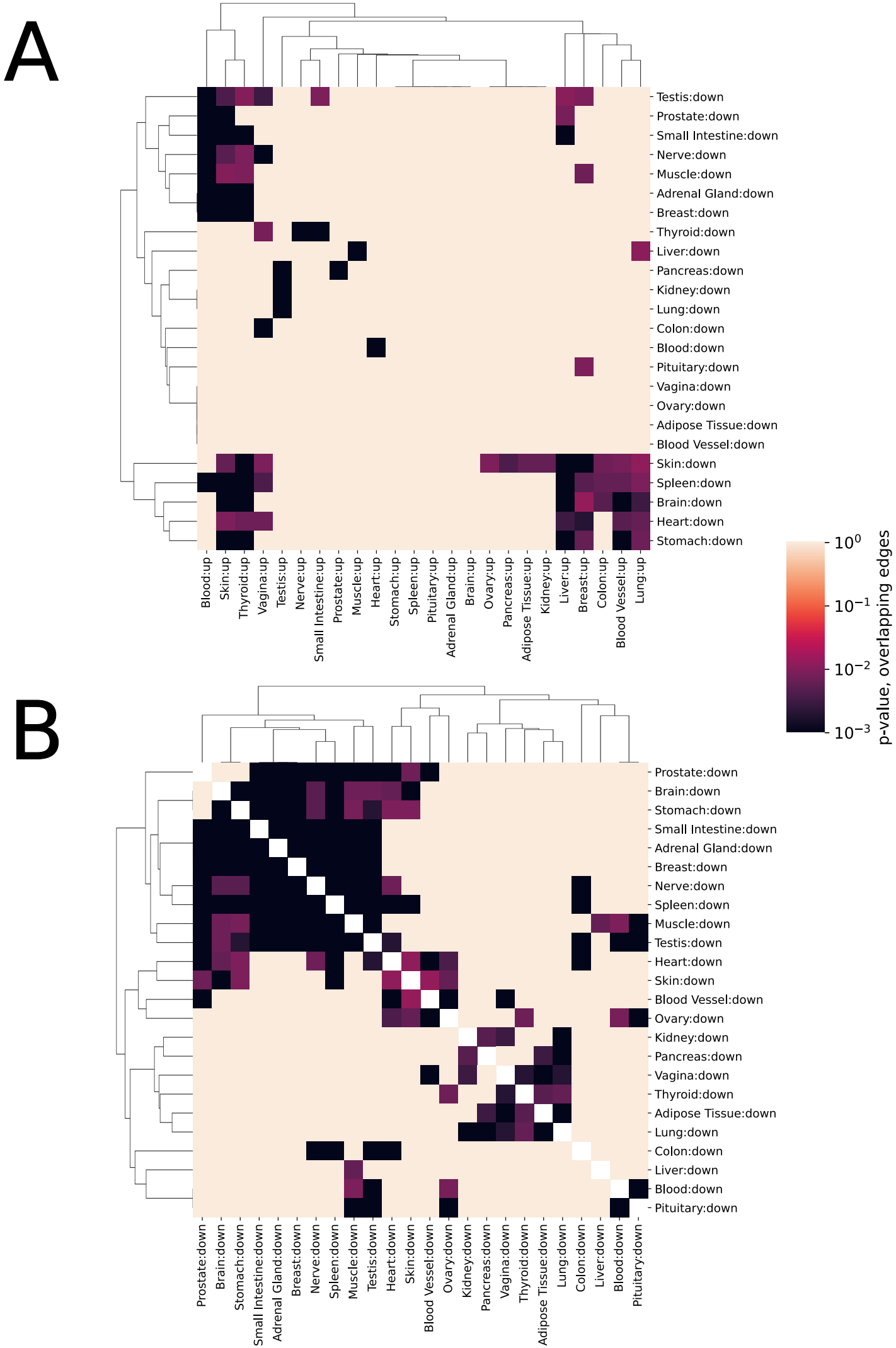
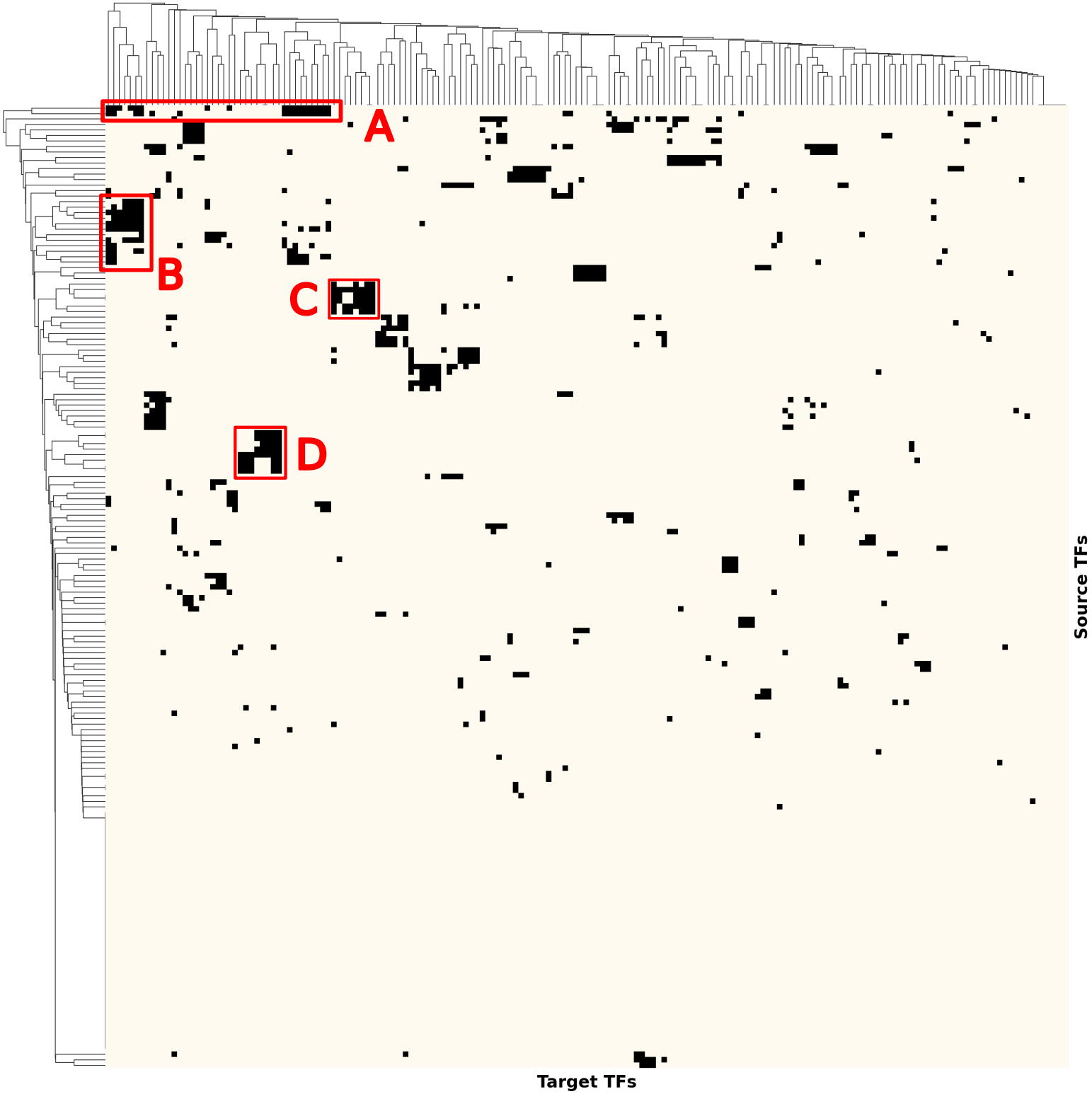

ND- or TSS-filtered GRN
randomly shuffle GRN edges
GRNs not included in ChEA3 or RummaGEO
Filtered GRN
vs
TRRUST or PWM
[1]
[100]
vs
TRRUST or PWM
count shared edges
average, standard deviation
observed edge overlap
expected edge overlap

Fig. S2




A
HDAC inhibitor
RNA polymerase inhibitor
tubulin polymerization inhibitor
mTOR inhibitor
PI3K inhibitor
adenosine receptor agonist
potassium channel activator
bacterial 50S ribosomal subunit inhibitor
glucocorticoid receptor agonist
HMGCR inhibitor
retinoid receptor agonist
selective serotonin reuptake inhibitor (SSRI)
PARP inhibitor
potassium channel blocker
HIV protease inhibitor
phosphodiesterase inhibitor
AKT inhibitor
sodium channel blocker
bacterial DNA gyrase inhibitor
calcium channel blocker
estrogen receptor agonist
histamine receptor antagonist
serotonin receptor antagonist
cyclooxygenase inhibitor
adrenergic receptor agonist
dopamine receptor antagonist
adrenergic receptor antagonist
bacterial cell wall synthesis inhibitor
acetylcholine receptor antagonist
DNA synthesis inhibitor
Down consensus gene sets
Up consensus gene sets
Fisher's Exact Test p-value
B
Down consensus gene sets
Down consensus gene sets
C
Up consensus gene sets
Up consensus gene sets


A
Testis:down
Prostate:down
Small Intestine:down
Nerve:down
Muscle:down
Adrenal Gland:down
Breast:down
Thyroid:down
Liver:down
Pancreas:down
Kidney:down
Lung:down
Colon:down
Blood:down
Pituitary:down
Vagina:down
Ovary:down
Adipose Tissue:down
Blood Vessel:down
Skin:down
Spleen:down
Brain:down
Heart:down
Stomach:down
Blood:up
Skin:up
Thyroid:up
Vagina:up
Testis:up
Nerve:up
Small Intestine:up
Prostate:up
Muscle:up
Heart:up
Stomach:up
Spleen:up
Pituitary:up
Adrenal Gland:up
Brain:up
Ovary:up
Pancreas:up
Adipose Tissue:up
Kidney:up
Liver:up
Breast:up
Colon:up
Blood Vessel:up
Lung:up
B
Prostate:down
Brain:down
Stomach:down
Small Intestine:down
Adrenal Gland:down
Breast:down
Nerve:down
Spleen:down
Muscle:down
Testis:down
Heart:down
Skin:down
Blood Vessel:down
Ovary:down
Kidney:down
Pancreas:down
Vagina:down
Thyroid:down
Adipose Tissue:down
Lung:down
Colon:down
Liver:down
Blood:down
Pituitary:down
10^0
10^{-1}
10^{-2}
10^{-3}
p-value, overlapping edges


A
B
C
D
Source TFs
Target TFs