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

Background: 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.

Results: 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, we extend the enrichment analysis feature to generate cell-type and cancer TF atlases. The atlases display systematically generated master regulator subnetworks of TFs for marker gene sets from 131 major normal human cell-types and 69 tumour subtypes from 10 cancers, respectively.

Conclusions: 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 factor | Knowledge graph | Enrichment analysis | Network visualization | Cell types | Cancer subtypes | Lung Squamous Cell Carcinoma

BACKGROUND

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 [10, 11, 24].

Several web server and command line tools are available for performing transcription factor enrichment analysis (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], 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. The ChEA-KG web server also hosts a Cell Atlas and Cancer Atlas made of subnetworks that are found to regulate the expression of marker genes from 131 major human cell types and 69 tumor subtypes from 10 tumor types.

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 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 PrismEXP

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 phenotype predictions using the MGI Mammalian Phenotype Level 4 2021 library from Enrichr [32] and used the predictions made by PrismEXP about the phenotypes of 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.

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 [48]. 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 2). 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 (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 3). 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 [49]; TAL1 [50], GATA1 [51], IKZF1 [52], SPI1 [53], and LYL1 [54] 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 [55]. GATA1 also physically interacts with SPI1 [56] and controls the transcription of TAL1 [57], 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 [58], whereas HNF4A, SPDEF, and FOXA1 specifically regulate goblet cells by controlling maturation and differentiation [59-61]. ELF5 is a regulator of epithelial cell differentiation but has not been specifically linked to the intestine [62]. The T-Box family of genes plays a significant role in early embryonic development, but the exact role of TBX10 is unknown [63]. SPDEF and ATOH1 are transcriptional co-regulators [64] (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 [65]. 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 [66]. MYF5, MYF6, MYOG, and MYOD1 are well-documented to establish the skeletal muscle phenotype [67]. These four genes share structural similarity and are collectively known as the Myogenic regulatory factors (MRFs) [68]. TBX20 and TBX5 are T-box proteins that are required for cardiomyocyte proliferation and homeostasis [69] and maturation [70], respectively. LBX1 controls gene expression relating to migration of muscle precursors [71]. 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 [72]. 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 [73, 74]. TEAD1 expression is lowered in diabetic podocyte injury [74]. In the kidneys more broadly, GLI2 is upregulated after adriamycin neuropathy [75] whereas GLIS2 is known to maintain renal architecture [76] and GLIS3 is associated with renal anomalies [77]. Finally, PRRX1 is suppressed during regulation of the renal epithelial-to-mesenchymal transition (EMT) [78], a process linked to podocyte dysfunction [79].

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 [80]. PRRX1 is similarly known to induce EMT in embryonic stem and cancer cells [81]. AHR is known to upregulates the expression of mucin 5AC [82], a major component of mucus in the lung [83]. High expression of FOSL2 has been linked to poor prognosis in lung cancer patients [84]. Lastly, HNF1B regulates the expression of alpha(1)-antitrypsin (AAT; SERPINA1), which may defend the lung against antiproteases [85]. CDX2, which is similarly enriched in intestinal epithelial cells, autoregulates its own expression and as well as the expression of HNF1B

346 within intestinal epithelial cells, suggesting that it plays a similar regulatory role in lung and
347 intestinal epithelial cells [86].

348 *Bone marrow: MSCs*. TFs enriched for MSCs are known to be involved in bone development
349 and maintenance of processes related to cell fate (Figure 6B). SOX18 is known to be involved
350 in the differentiation of MSCs to endothelial cells [87]. SOX17 is known to control the
351 expression of fetal hematopoietic poetic stem cells [88], whereas MEOX1 is a key regulator of
352 proliferation in mesenchymal-like cells [89]. Differentiation of bone-marrow-derived MSCs is
353 regulated by GATA2 [90]. ERG and ELK3 are both members of the ETS family of TFs [91] and
354 involved in EMT [92]. ERG is directly linked to endo-cardial mesenchymal transformation
355 (EnMT) and EMT [93, 94]. Finally, it is known that TWIST1 suppresses the expression of
356 RUNX2 to suppress bone formation [95] while FOXO1 defends against oxidative damage in
357 the bone [96]. In conclusion, the subnetworks for each cell type in the ChEA-KG Cell Atlas
358 confirm many master regulators as well as predict novel key regulators.

359 **The TF subnetworks Cancer Atlas**

360 Besides the ChEA-KG Cell Atlas, ChEA-KG also has a Cancer Atlas. To create the ChEA-KG
361 Cancer Atlas we obtained 69 up-regulated gene sets extracted from 10 cancer types profiled
362 by the National Institutes of Health (NIH) CPTAC3 program [45]. Each of these gene sets
363 represents a subtype based on applying the Leiden clustering algorithm [35] to the RNA-seq
364 profiling of over a thousand tumours by Multiomics2Targets [46]. These up-regulated genes
365 from these tumour subtypes were submitted for analysis with ChEA-KG. As an example, we
366 applied this analysis to identify TF subnetworks enriched for lung squamous cell carcinoma
367 (LSCC) (Fig. 7). To characterize these subnetworks, we next identified key regulatory
368 mechanisms for each LSCC subnetwork by counting collective enriched terms in each
369 subnetwork (Supporting Table 4). We find that cluster 0 is enriched with terms related to T-
370 cell regulation, cluster 1 is enriched for midbrain morphology and sclerocornea phenotypes,
371 cluster 2 is enriched for immunity and NK cells, cluster 3 has term related to lung development,
372 and cluster 5 has term related to thyroid regulation. These enriched terms immediately suggest
373 key mechanisms of each subtype and provide clear mechanisms that may be correlated with
374 subtype survival and likelihood of the responsiveness to therapy of each subtype.

375 **DISCUSSION**

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

391 The ChEA-KG web server application facilitates the querying of the GRN for knowledge
392 discovery. The enrichment analysis feature of ChEA-KG can be used to uncover how the most
393 enriched TF master regulators are forming subnetworks of tight regulatory interactions. The
394 cell type and Cancer Atlas feature of ChEA-KG demonstrates how the GRN can be used to
395 accurately identify known master regulators subnetworks. The cell type specific subnetworks

also contain TF unknown to be relevant to the specific cell types, potentially leading to new discoveries. 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. In the future the GRN can be expanded to include non-TF-encoding genes. The subnetworks can additionally be dynamically simulated to further understand how groups of TFs work together as a functional regulatory module. 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.

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 direct 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 two atlases that visualize the master regulatory subnetworks that are master regulators of 131 human cell types and 69 tumour subtypes from 10 cancers. 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
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 [OT2OD030160].

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

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

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

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

511 **Table 1**

512

513

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	×	✓	×	✓	✓	✓

514

515

516

517

518

519

520

521 **Table 2**

522

	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

523

524

REFERENCES

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

642 43. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, Zager MA, Aldinger KA,
643 Blecher-Gonen R, Zhang F *et al*: **A human cell atlas of fetal gene expression**. *Science*
644 *(New York, NY)* 2020, **370**(6518):ppublish.
645 44. Palasca O, Santos A, Stolte C, Gorodkin J, Jensen LJ: **TISSUES 2.0: an integrative web**
646 **resource on mammalian tissue expression**. *Database : the journal of biological databases*
647 *and curation* 2018, **2018**:ppublish.
648 45. Li Y, Dou Y, Da Veiga Leprevost F, Geffen Y, Calinawan AP, Aguet F, Akiyama Y, Anand S,
649 Birger C, Cao S *et al*: **Proteogenomic data and resources for pan-cancer analysis**.
650 *Cancer cell* 2023, **41**(8):1397-1406.
651 46. Deng EZ, Marino GB, Clarke DJB, Diamant I, Resnick AC, Ma W, Wang P, Ma'ayan A:
652 **Multiomics2Targets identifies targets from cancer cohorts profiled with**
653 **transcriptomics, proteomics, and phosphoproteomics**. *Cell reports methods* 2024,
654 **4**(8):100839.
655 47. Law CW, Chen Y, Shi W, Smyth GK: **voom: Precision weights unlock linear model**
656 **analysis tools for RNA-seq read counts**. *Genome biology* 2014, **15**(2):R29.
657 48. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma'ayan
658 A: **Massive mining of publicly available RNA-seq data from human and mouse**. *Nat*
659 *Commun* 2018, **9**(1):1366.
660 49. Andrews NC, Erdjument-Bromage H, Davidson MB, Tempst P, Orkin SH: **Erythroid**
661 **transcription factor NF-E2 is a haematopoietic-specific basic-leucine zipper protein**.
662 *Nature* 1993, **362**(6422):722-728.
663 50. Kassouf MT, Hughes JR, Taylor S, McGowan SJ, Soneji S, Green AL, Vyas P, Porcher C:
664 **Genome-wide identification of TAL1's functional targets: insights into its mechanisms**
665 **of action in primary erythroid cells**. *Genome research* 2010, **20**(8):1064-1083.
666 51. Briegel K, Bartunek P, Stengl G, Lim KC, Beug H, Engel JD, Zenke M: **Regulation and**
667 **function of transcription factor GATA-1 during red blood cell differentiation**.
668 *Development (Cambridge, England)* 1996, **122**(12):3839-3850.
669 52. Dijon M, Bardin F, Murati A, Batoz M, Chabannon C, Tonnelle C: **The role of Ikaros in**
670 **human erythroid differentiation**. *Blood* 2008, **111**(3):1138-1146.
671 53. Back J, Dierich A, Bronn C, Kastner P, Chan S: **PU.1 determines the self-renewal capacity**
672 **of erythroid progenitor cells**. *Blood* 2004, **103**(10):3615-3623.
673 54. Capron C, Lécluse Y, Kaushik AL, Foudi A, Lacout C, Sekkai D, Godin I, Albagli O, Poullion I,
674 Svinartchouk F *et al*: **The SCL relative LYL-1 is required for fetal and adult hematopoietic**
675 **stem cell function and B-cell differentiation**. *Blood* 2006, **107**(12):4678-4686.
676 55. Chiu SK, Saw J, Huang Y, Sonderegger SE, Wong NC, Powell DR, Beck D, Pimanda JE,
677 Tremblay CS, Curtis DJ: **A novel role for Lyl1 in primitive erythropoiesis**. *Development*
678 *(Cambridge, England)* 2018, **145**(19):epublish.
679 56. Cantor AB, Orkin SH: **Transcriptional regulation of erythropoiesis: an affair involving**
680 **multiple partners**. *Oncogene* 2002, **21**(21):3368-3376.
681 57. Mouthon MA, Bernard O, Mitjavila MT, Romeo PH, Vainchenker W, Mathieu-Mahul D:
682 **Expression of tal-1 and GATA-binding proteins during human hematopoiesis**. *Blood*
683 1993, **81**(3):647-655.
684 58. Guo RJ, Suh ER, Lynch JP: **The role of Cdx proteins in intestinal development and**
685 **cancer**. *Cancer biology & therapy* 2004, **3**(7):593-601.
686 59. Ye DZ, Kaestner KH: **Foxa1 and Foxa2 control the differentiation of goblet and**
687 **enteroendocrine L- and D-cells in mice**. *Gastroenterology* 2009, **137**(6):2052-2062.
688 60. Noah TK, Kazanjian A, Whitsett J, Shroyer NF: **SAM pointed domain ETS factor (SPDEF)**
689 **regulates terminal differentiation and maturation of intestinal goblet cells**. *Experimental*
690 *cell research* 2010, **316**(3):452-465.
691 61. Garrison WD, Battle MA, Yang C, Kaestner KH, Sladek FM, Duncan SA: **Hepatocyte nuclear**
692 **factor 4alpha is essential for embryonic development of the mouse colon**.
693 *Gastroenterology* 2006, **130**(4):1207-1220.
694 62. Luk IY, Reehorst CM, Mariadason JM: **ELF3, ELF5, EHF and SPDEF Transcription Factors**
695 **in Tissue Homeostasis and Cancer**. *Molecules (Basel, Switzerland)* 2018, **23**(9):epublish.
696 63. Naiche LA, Harrelson Z, Kelly RG, Papaioannou VE: **T-box genes in vertebrate**
697 **development**. *Annual review of genetics* 2005, **39**:219-239.
698 64. Lo YH, Chung E, Li Z, Wan YW, Mahe MM, Chen MS, Noah TK, Bell KN, Yalamanchili HK,
699 Klisch TJ *et al*: **Transcriptional Regulation by ATOH1 and its Target SPDEF**
700 **in the Intestine**. *Cellular and molecular gastroenterology and hepatology* 2017, **3**(1):51-71.

65. Liu Q, Linney E: **The mouse retinoid-X receptor-gamma gene: genomic organization and evidence for functional isoforms**. *Molecular endocrinology (Baltimore, Md)* 1993, **7**(5):651-658.
66. L'Honoré A, Coulon V, Marcil A, Lebel M, Lafrance-Vanasse J, Gage P, Camper S, Drouin J: **Sequential expression and redundancy of Pitx2 and Pitx3 genes during muscle development**. *Developmental biology* 2007, **307**(2):421-433.
67. Olson EN, Brennan TJ, Chakraborty T, Cheng TC, Cserjesi P, Edmondson D, James G, Li L: **Molecular control of myogenesis: antagonism between growth and differentiation**. *Molecular and cellular biochemistry* 1991, **104**(1-2):7-13.
68. Zammit PS: **Function of the myogenic regulatory factors Myf5, MyoD, Myogenin and MRF4 in skeletal muscle, satellite cells and regenerative myogenesis**. *Seminars in cell & developmental biology* 2017, **72**:19-32.
69. Xiang FL, Guo M, Yutzey KE: **Overexpression of Tbx20 in Adult Cardiomyocytes Promotes Proliferation and Improves Cardiac Function After Myocardial Infarction**. *Circulation* 2016, **133**(11):1081-1092.
70. Steimle JD, Moskowitz IP: **TBX5: A Key Regulator of Heart Development**. *Current topics in developmental biology* 2017, **122**:195-221.
71. Brohmann H, Jagla K, Birchmeier C: **The role of Lbx1 in migration of muscle precursor cells**. *Development (Cambridge, England)* 2000, **127**(2):437-445.
72. 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.
73. 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.
74. Chugh SS: **Transcriptional regulation of podocyte disease**. *Translational research : the journal of laboratory and clinical medicine* 2007, **149**(5):237-242.
75. 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.
76. 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.
77. Uhlenhaut NH, Treier M: **Transcriptional regulators in kidney disease: gatekeepers of renal homeostasis**. *Trends in genetics : TIG* 2008, **24**(7):361-371.
78. 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.
79. 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.
80. 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.
81. 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.
82. 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.
83. Gendler SJ, Spicer AP: **Epithelial mucin genes**. *Annual review of physiology* 1995, **57**:607-634.
84. 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.
85. 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.

761 86. Boyd M, Hansen M, Jensen TG, Perearnau A, Olsen AK, Bram LL, Bak M, Tommerup N,
762 Olsen J, Troelsen JT: **Genome-wide analysis of CDX2 binding in intestinal epithelial**
763 **cells (Caco-2)**. *The Journal of biological chemistry* 2010, **285**(33):25115-25125.
764 87. Ikhapoh IA, Pelham CJ, Agrawal DK: **Sry-type HMG box 18 contributes to the**
765 **differentiation of bone marrow-derived mesenchymal stem cells to endothelial cells**.
766 *Differentiation; research in biological diversity* 2015, **89**(3-4):87-96.
767 88. He S, Kim I, Lim MS, Morrison SJ: **Sox17 expression confers self-renewal potential and**
768 **fetal stem cell characteristics upon adult hematopoietic progenitors**. *Genes &*
769 *development* 2011, **25**(15):1613-1627.
770 89. Sun L, Burnett J, Gasparyan M, Xu F, Jiang H, Lin CC, Myers I, Korkaya H, Liu Y, Connarn J
771 *et al*: **Novel cancer stem cell targets during epithelial to mesenchymal transition in**
772 **PTEN-deficient trastuzumab-resistant breast cancer**. *Oncotarget* 2016, **7**(32):51408-
773 51422.
774 90. Kamata M, Okitsu Y, Fujiwara T, Kanehira M, Nakajima S, Takahashi T, Inoue A, Fukuhara
775 N, Onishi Y, Ishizawa K *et al*: **GATA2 regulates differentiation of bone marrow-derived**
776 **mesenchymal stem cells**. *Haematologica* 2014, **99**(11):1686-1696.
777 91. Sharrocks AD: **The ETS-domain transcription factor family**. *Nature reviews Molecular cell*
778 *biology* 2001, **2**(11):827-837.
779 92. Maroulakou IG, Bowe DB: **Expression and function of Ets transcription factors in**
780 **mammalian development: a regulatory network**. *Oncogene* 2000, **19**(55):6432-6442.
781 93. Leshem O, Madar S, Kogan-Sakin I, Kamer I, Goldstein I, Brosh R, Cohen Y, Jacob-Hirsch J,
782 Ehrlich M, Ben-Sasson S *et al*: **TMPRSS2/ERG promotes epithelial to mesenchymal**
783 **transition through the ZEB1/ZEB2 axis in a prostate cancer model**. *PloS one* 2011,
784 **6**(7):e21650.
785 94. Vijayaraj P, Le Bras A, Mitchell N, Kondo M, Juliao S, Wasserman M, Beeler D, Spokes K,
786 Aird WC, Baldwin HS *et al*: **Erg is a crucial regulator of endocardial-mesenchymal**
787 **transformation during cardiac valve morphogenesis**. *Development (Cambridge, England)*
788 2012, **139**(21):3973-3985.
789 95. Bialek P, Kern B, Yang X, Schrock M, Sosic D, Hong N, Wu H, Yu K, Ornitz DM, Olson EN *et*
790 *al*: **A twist code determines the onset of osteoblast differentiation**. *Developmental cell*
791 2004, **6**(3):423-435.
792 96. Liao L, Su X, Yang X, Hu C, Li B, Lv Y, Shuai Y, Jing H, Deng Z, Jin Y: **TNF-α Inhibits**
793 **FoxO1 by Upregulating miR-705 to Aggravate Oxidative Damage in Bone Marrow-**
794 **Derived Mesenchymal Stem Cells during Osteoporosis**. *Stem cells (Dayton, Ohio)* 2016,
795 **34**(4):1054-1067.

796

Fig. 1

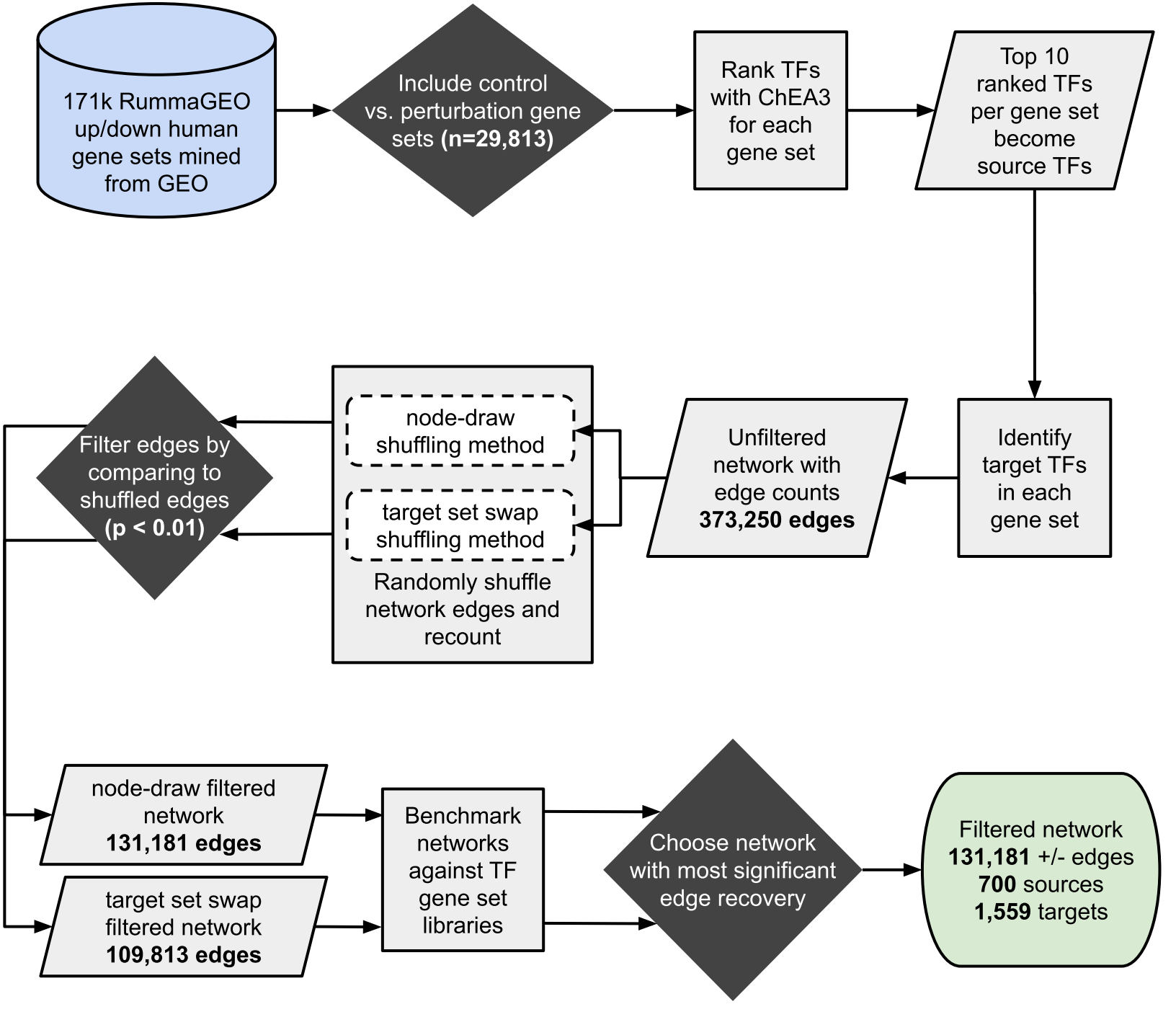

171k RummaGEO up/down human gene sets mined from GEO
Include control vs. perturbation gene sets (n=29,813)
Rank TFs with ChEA3 for each gene set
Top 10 ranked TFs per gene set become source TFs
Identify target TFs in each gene set
Unfiltered network with edge counts 373,250 edges
node-draw shuffling 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

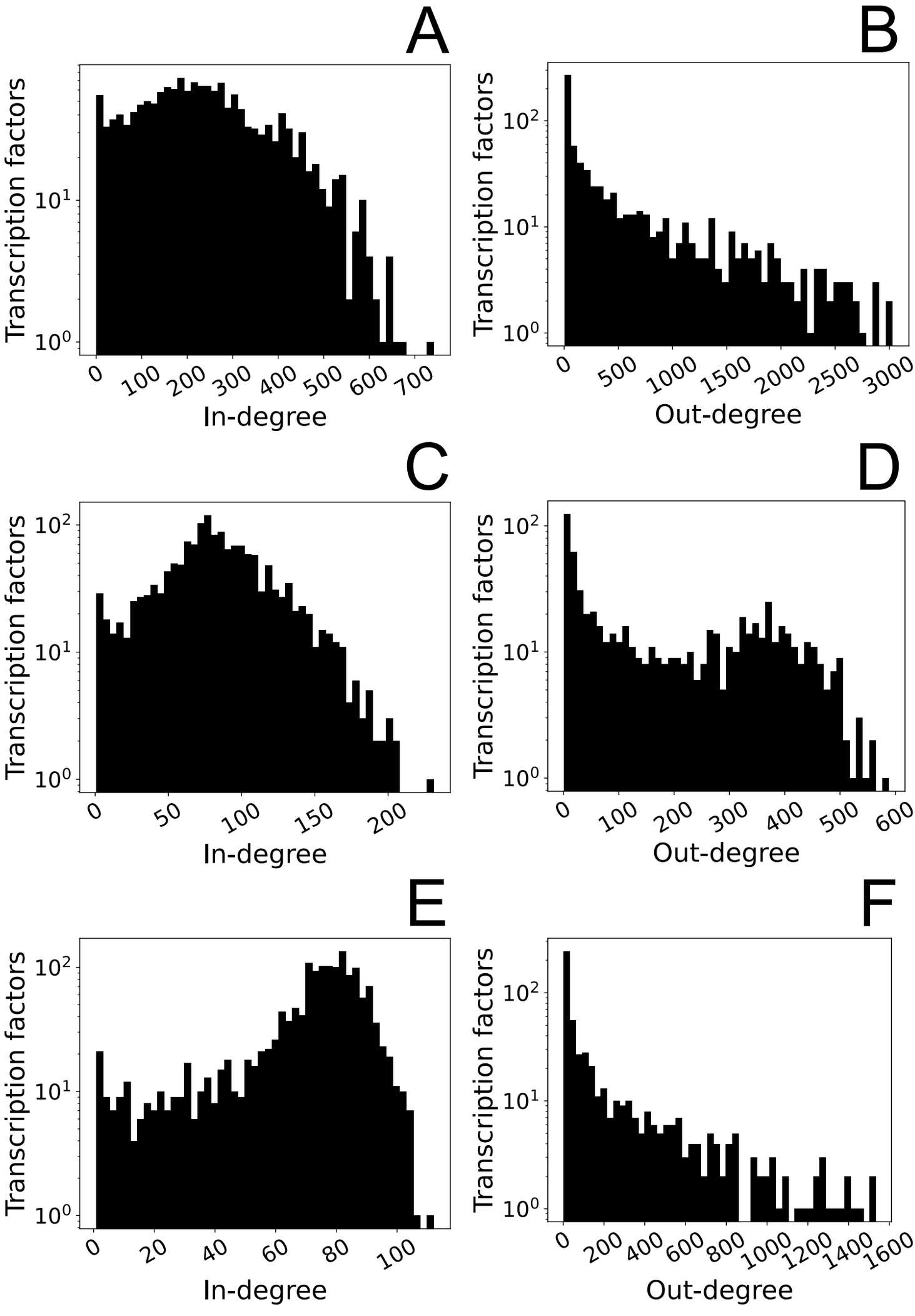

A
B
C
D
E
F
Transcription factors
In-degree
Out-degree
10^0
10^1
10^2
0
100
200
300
400
500
600
700
500
1000
1500
2000
2500
3000
50
150
20
40
60
80
1200
1400
1600

Fig. 3

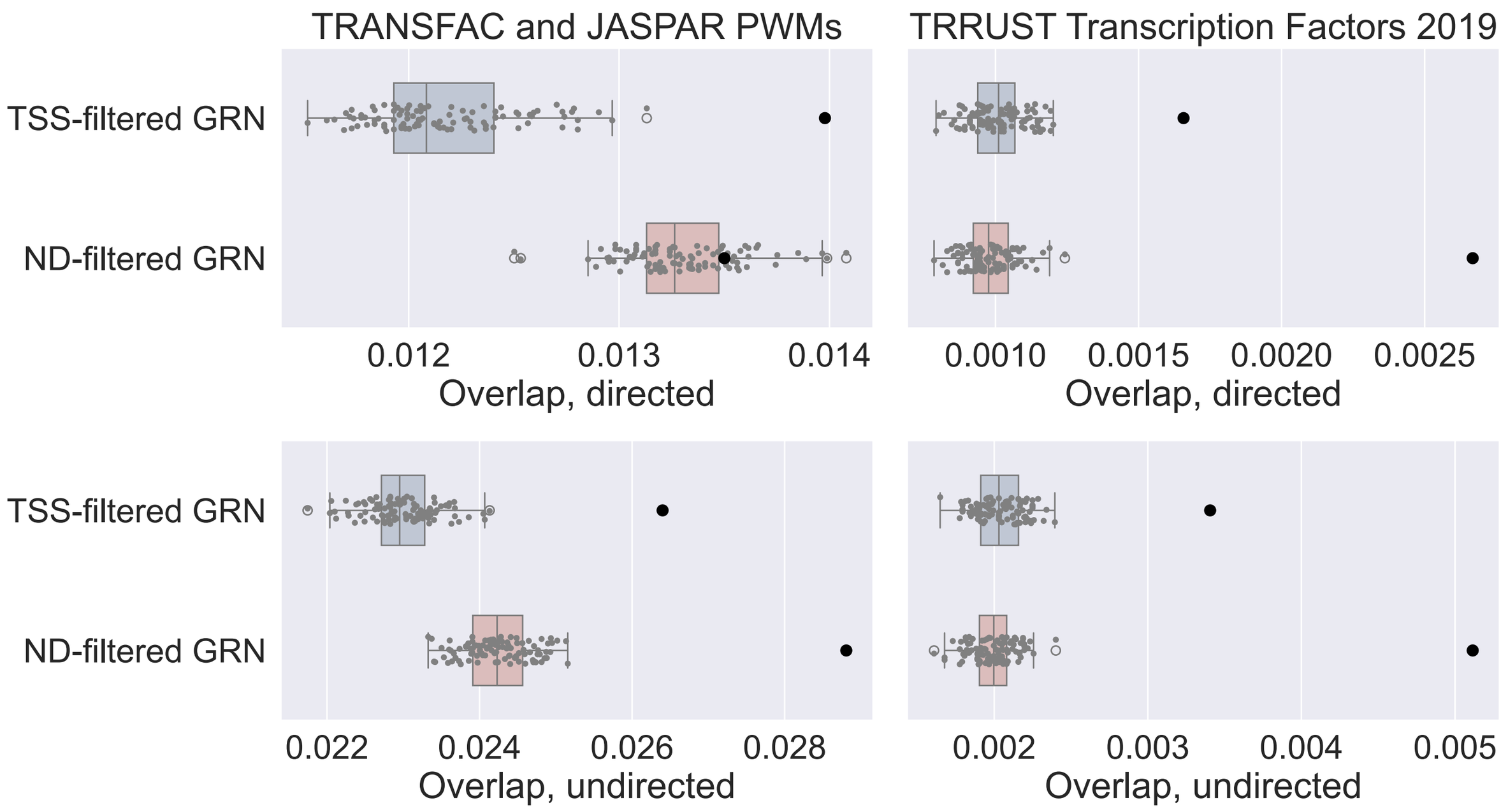

TRANSFAC and JASPAR PWMs
TRRUST Transcription Factors 2019
TSS-filtered GRN
ND-filtered GRN
0.012
0.013
0.014
Overlap, directed
0.0010
0.0015
0.0020
0.0025
Overlap, directed
TSS-filtered GRN
ND-filtered GRN
0.022
0.024
0.026
0.028
Overlap, undirected
0.002
0.003
0.004
0.005
Overlap, undirected

Fig. 4

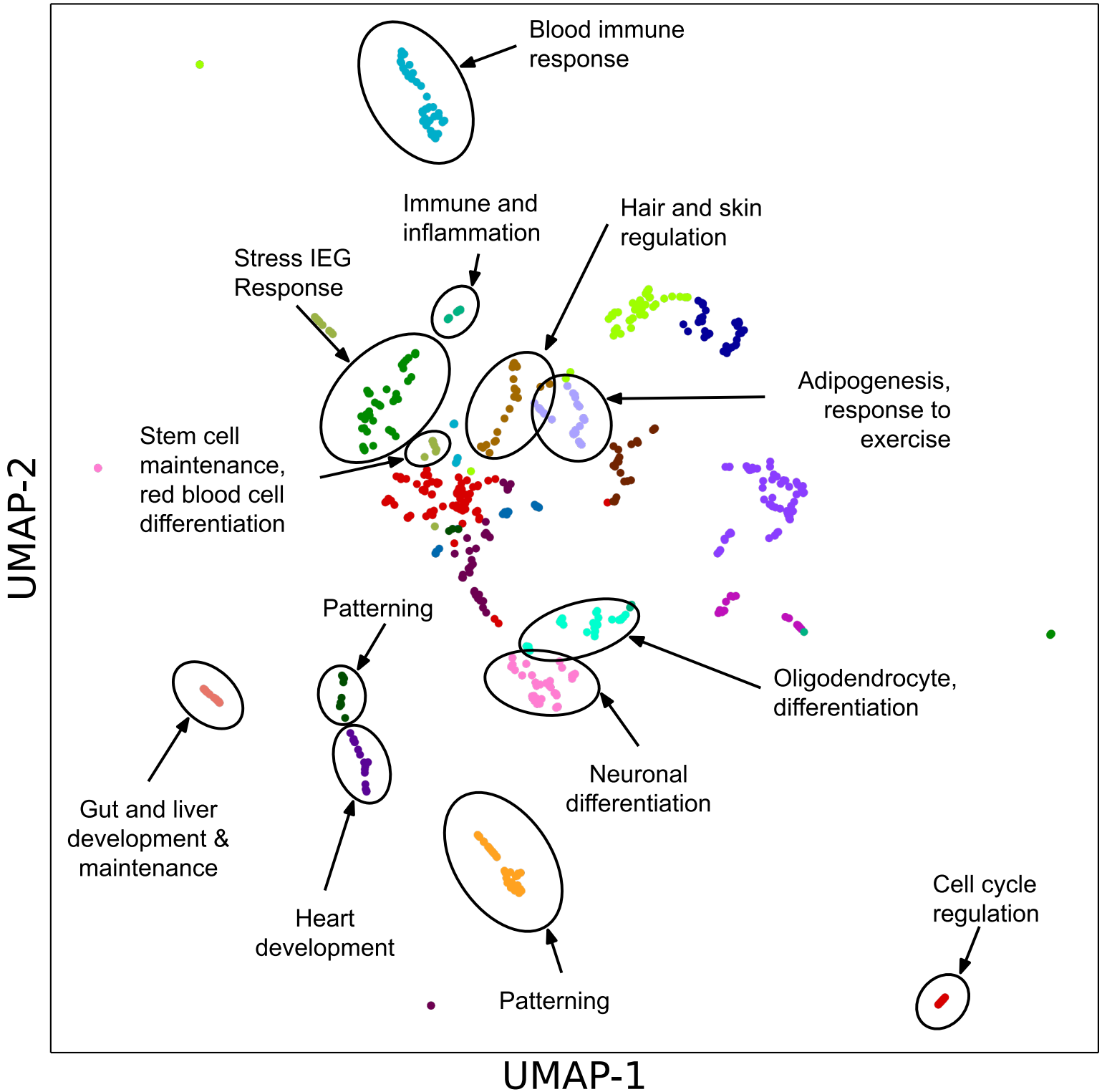

Blood immune response
Immune and inflammation
Hair and skin regulation
Stress IEG Response
Adipogenesis, response to exercise
Stem cell maintenance, red blood cell differentiation
Patterning
Oligodendrocyte, differentiation
Neuronal differentiation
Gut and liver development & maintenance
Heart development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

Fig. 5

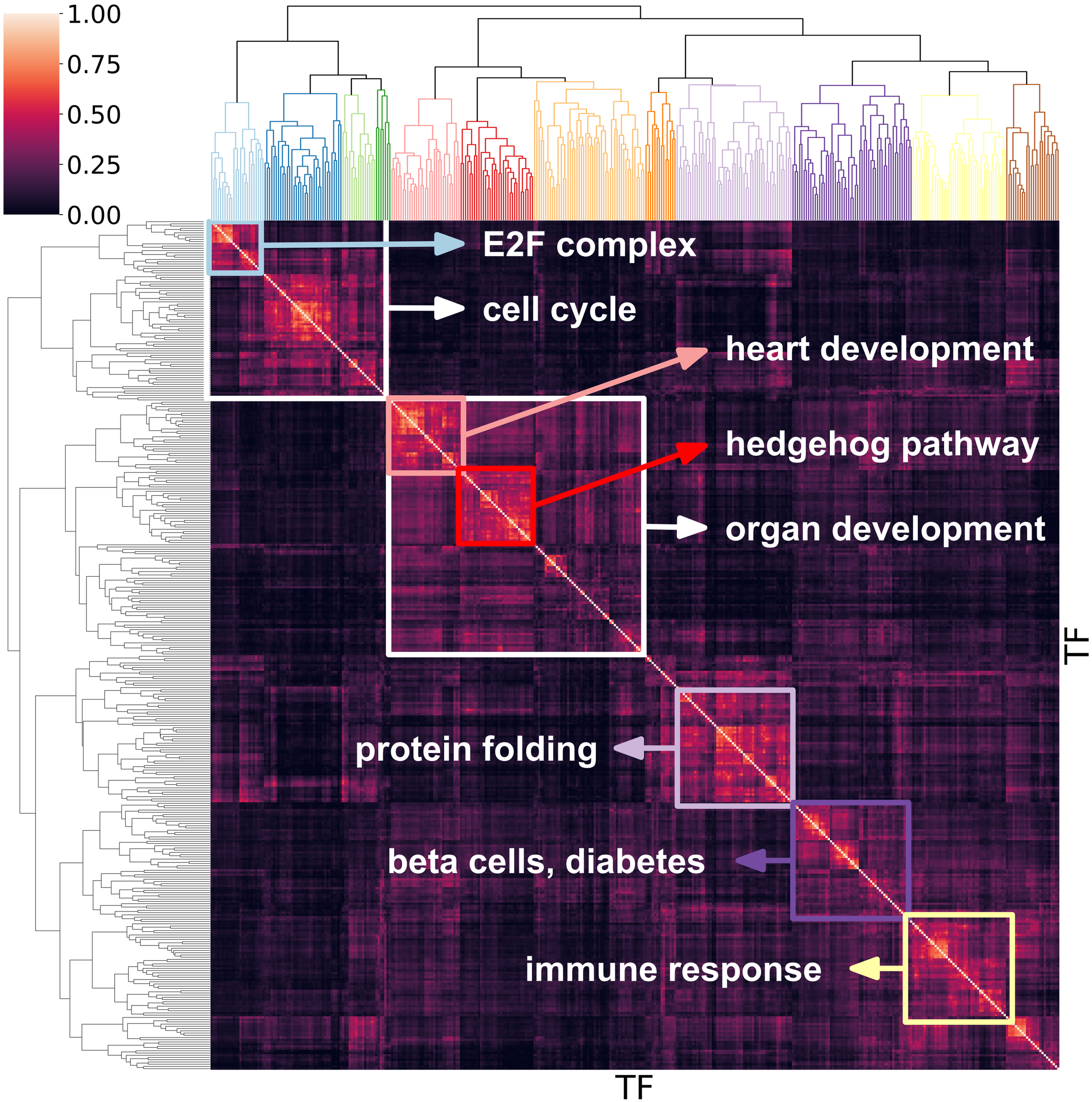

1.00
0.75
0.50
0.25
0.00
E2F complex
cell cycle
heart development
hedgehog pathway
organ development
protein folding
beta cells, diabetes
immune response
TF
TF

Fig. 6

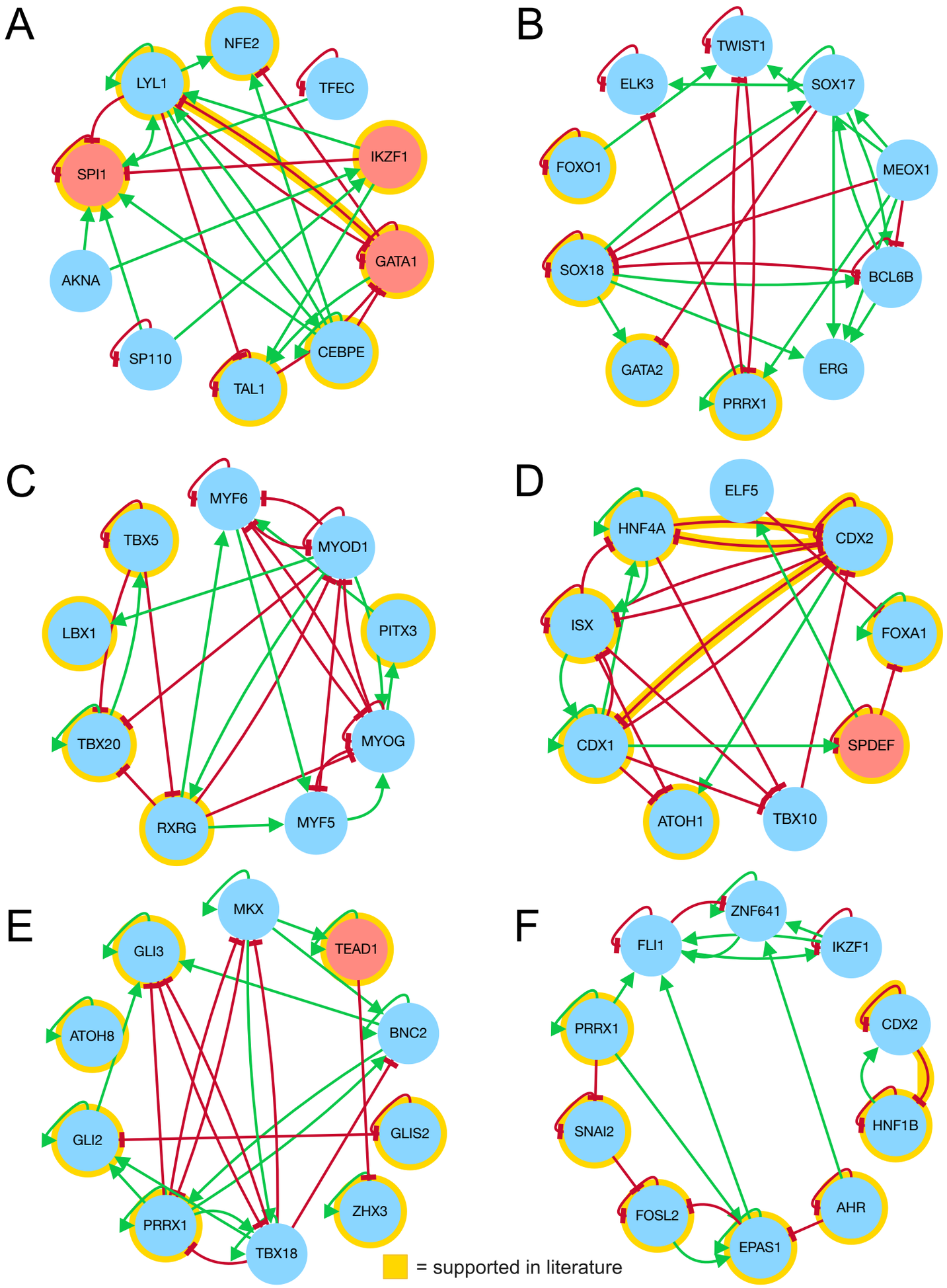

A
NFE2
LYL1
TFEC
IKZF1
SPI1
GATA1
AKNA
SP110
CEBPE
TAL1
B
TWIST1
ELK3
SOX17
FOXO1
MEOX1
SOX18
BCL6B
GATA2
ERG
PRRX1
C
MYF6
TBX5
MYOD1
LBX1
PITX3
TBX20
MYOG
RXRG
MYF5
D
ELF5
HNF4A
CDX2
ISX
FOXA1
CDX1
SPDEF
ATOH1
TBX10
E
MKX
GLI3
TEAD1
ATOH8
BNC2
GLI2
GLIS2
PRRX1
ZHX3
TBX18
F
ZNF641
FLI1
IKZF1
PRRX1
CDX2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

Fig. 7

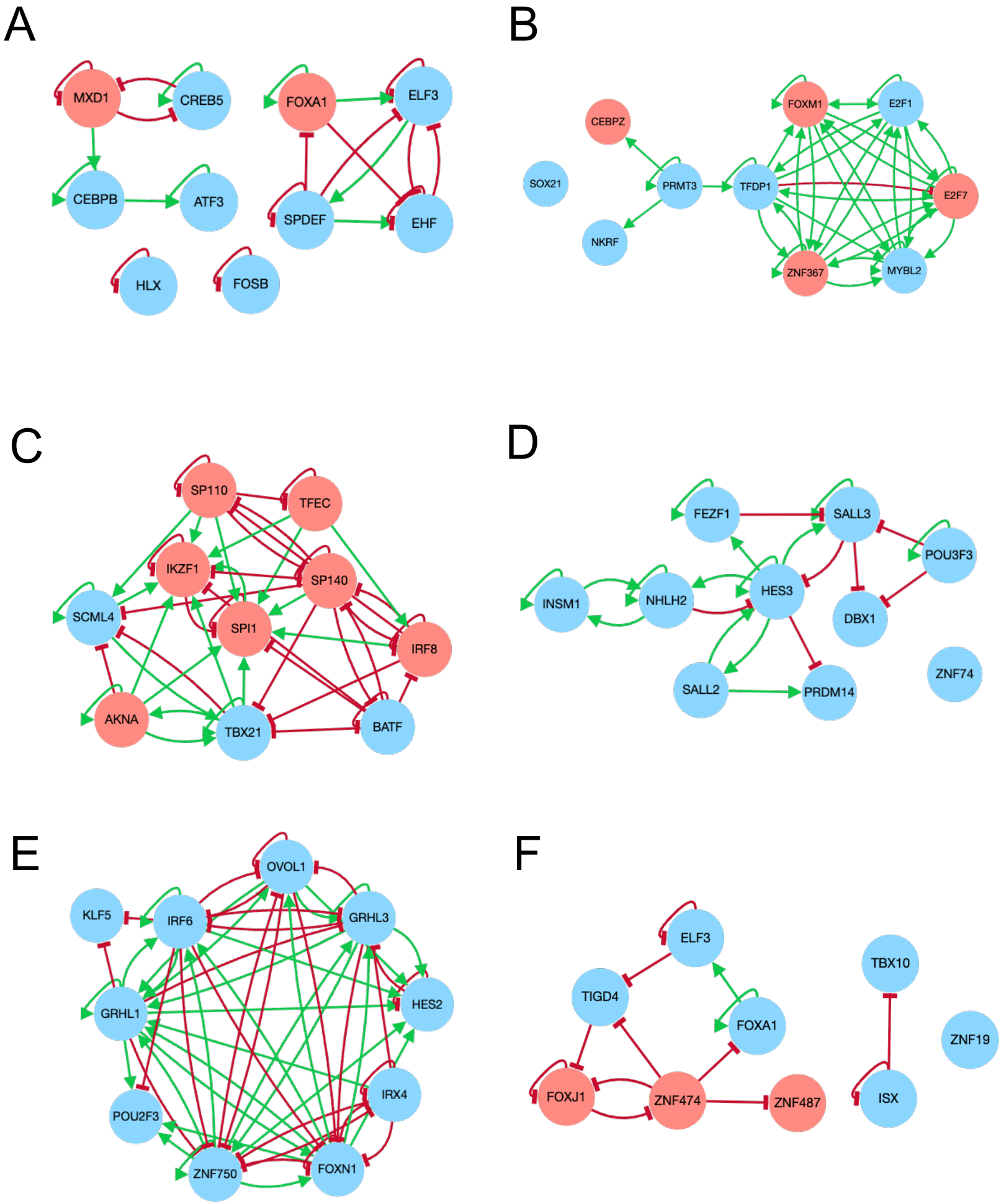
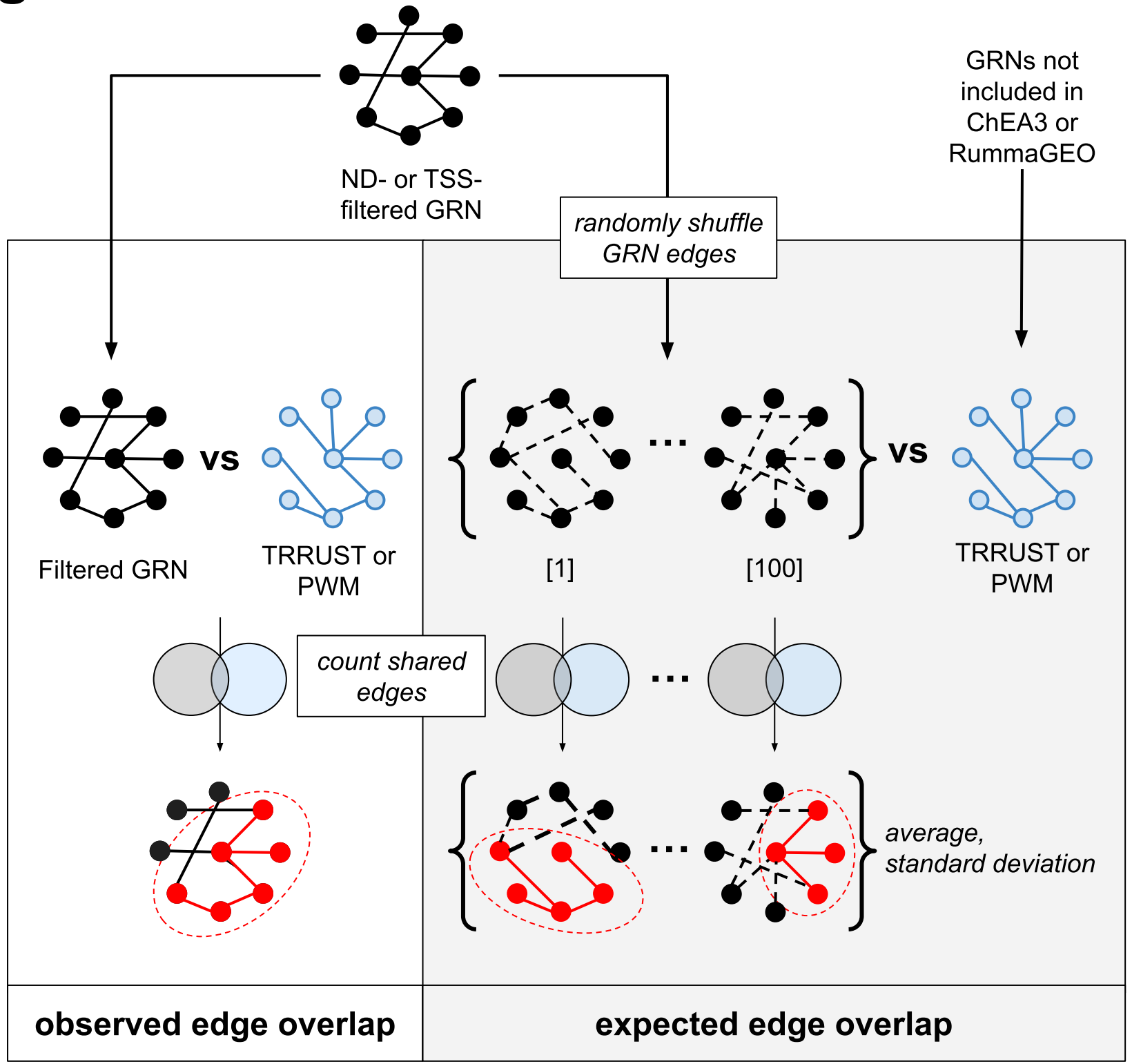

A
MXD1
CREB5
CEBPB
ATF3
HLX
FOSB
FOXA1
ELF3
SPDEF
EHF
B
CEBPZ
SOX21
PRMT3
NKRF
TFDP1
FOXM1
E2F1
E2F7
ZNF367
MYBL2
C
SP110
TFEC
IKZF1
SP140
SCML4
SPI1
IRF8
AKNA
TBX21
BATF
D
FEZF1
SALL3
POU3F3
INSM1
NHLH2
HES3
DBX1
SALL2
PRDM14
ZNF74
E
OVOL1
KLF5
IRF6
GRHL3
GRHL1
HES2
IRX4
POU2F3
ZNF750
FOXN1
F
ELF3
TIGD4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

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

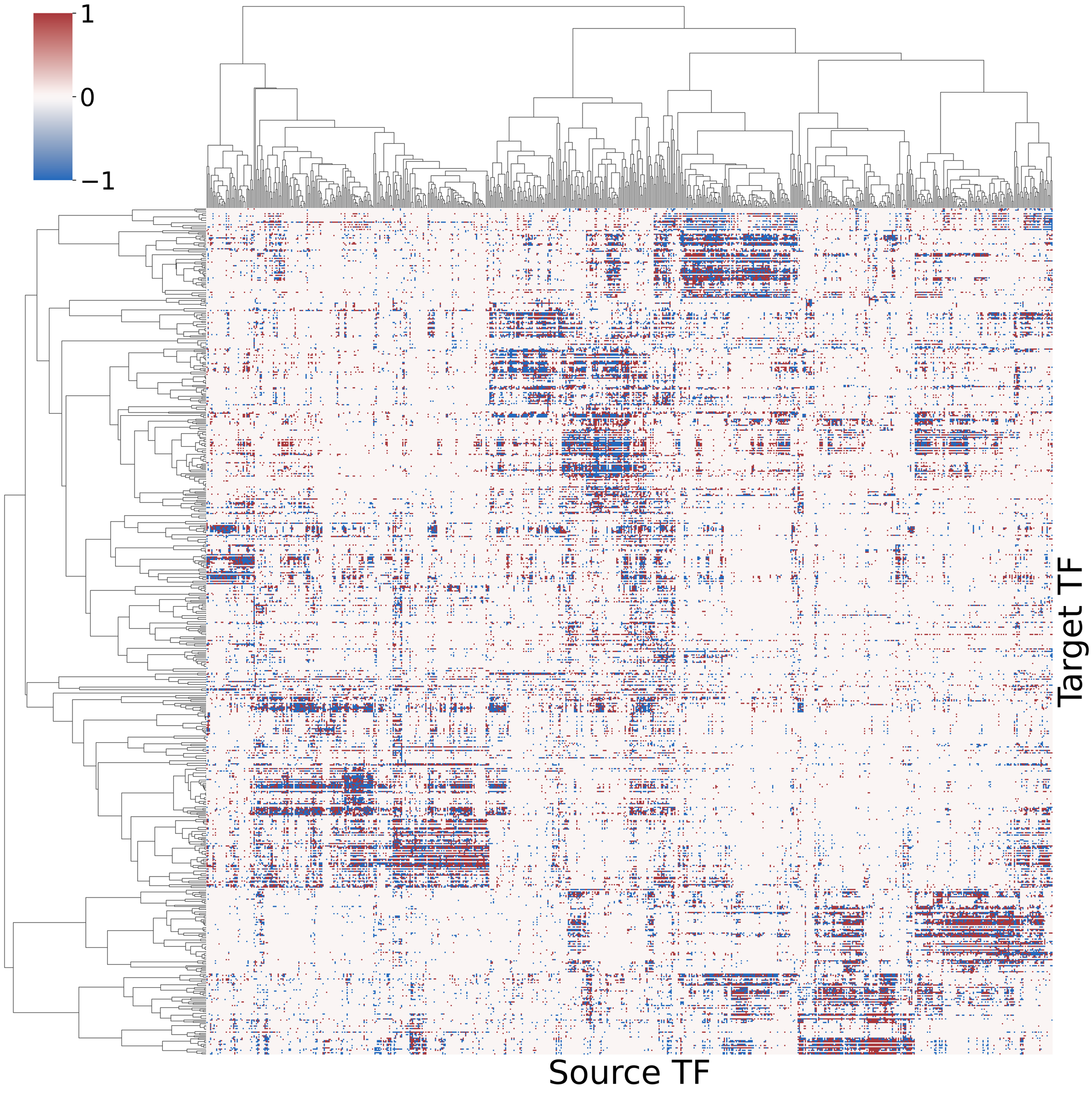



Fig. S2