Identifying Reproducible Transcription Regulator Coexpression Patterns with Single Cell Transcriptomics

Alexander Morin[1,2,3], C. Pan Chu[1,2,3], Paul Pavlidis[1,2,]*

1. Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada

2. Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada

3. Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC, Canada

* Corresponding author

Paul Pavlidis

177 Michael Smith Laboratories

2185 East Mall

University of British Columbia

Vancouver BC V6T1Z4

Canada

604 827 4157

paul@msl.ubc.ca

Abstract

The proliferation of single cell transcriptomics has potentiated our ability to unveil patterns that reflect dynamic cellular processes such as the regulation of gene transcription. In this study, we leverage a broad collection of single cell RNA-seq data to identify the gene partners whose expression is most coordinated with each human and mouse transcription regulator (TR). We assembled 120 human and 103 mouse scRNA-seq datasets from the literature (>28 million cells), constructing a single cell coexpression network for each. We aimed to understand the consistency of TR coexpression profiles across a broad sampling of biological contexts, rather than examine the preservation of context-specific signals. Our workflow therefore explicitly prioritizes the patterns that are most reproducible across cell types. Towards this goal, we characterize the similarity of each TR's coexpression within and across species. We create single cell coexpression rankings for each TR, demonstrating that this aggregated information recovers literature curated targets on par with ChIP-seq data. We then combine the coexpression and ChIP-seq information to identify candidate regulatory interactions supported across methods and species. Finally, we highlight interactions for the important neural TR ASCL1 to demonstrate how our compiled information can be adopted for community use.

Author Summary

A common way to analyze gene expression (transcriptomics) data is to correlate gene transcript levels across samples for every pair of genes (coexpression). Coordinated expression between genes may imply a shared biological function, though this warrants cautious interpretation given assumptions about cellular processes inferred from RNA abundances alone. Still, coexpression inference is often used to nominate genes whose expression may be controlled by transcription regulators (TRs). The rapid generation of diverse single cell transcriptomics data has unlocked our ability to discover coexpression patterns across individual cells — though these signals are often noisy. Reproducible patterns across studies can help distinguish meaningful biological relationships from spurious correlations. We used this study to analyze a broad collection of single cell data spanning numerous tissues in human and mouse to infer global TR coexpression patterns. We aimed to learn which interactions were generally observable, to better potentiate future examinations of reproducible coexpression in specific contexts. We evaluate the predictive performance of these global single cell coexpression rankings using independent gene regulation evidence, and highlight TR-gene pairs that are supported across data modalities as well as species. By disseminating these rankings, we hope that other researchers can extract insight for their own TRs of interest.

Introduction

The widespread adoption of single cell genomic methodologies, particularly single cell/nucleus RNA sequencing (herein, scRNA-seq), has significantly advanced our ability to characterize dynamic cellular processes. The scale with which scRNA-seq data has been generated has created an unprecedented opportunity to understand the reproducibility of these cellular patterns. This is important because, despite its power, scRNA-seq results in sparse gene transcript counts due to both biological and technical factors (Crow et al., 2016; Heumos et al., 2023).

Gene regulation is a field that stands to greatly benefit from the single cell era. A primary objective is to map the temporal and context-specific interactions between transcription regulators (TRs) and their target genes. However, understanding the sets of genes regulated by each TR — regardless of context — remains a challenge. Despite the availability of genetic tools, linking TRs to direct gene targets is hindered by multiple factors. These include the cost and difficulty of collecting experimental data implicating direct regulation, such as TR binding information from chromatin immunoprecipitation sequencing (ChIP-seq), and the inherent complexity of the underlying biology (Lambert et al., 2018, Rothenberg 2019).

Gene coexpression is a traditional and widely adopted approach for predicting TR-target relationships. This analysis is often cast as generating a predicted gene regulatory network, where the strength of covariation between gene transcript levels serves as edge weights (Sonawane et al., 2019). The fundamental assumption is that if a TR protein influences a gene's transcription, the TR gene itself must also be expressed. However, this assumption may be compromised when the dynamic expression of TRs and their targets are uncoupled. Further, this covariation does not implicate a causative directionality (i.e., regulatory influence) between gene pairs. Despite these limitations, coexpression analysis has been extensively applied as a cost-effective and genome-wide strategy to investigate gene regulation and is commonly integrated with other data modalities (Aibar et al., 2017; Bravo González-Blas et al., 2023).

The emergence of scRNA-seq has made it possible to study coexpression at a finer level of granularity than afforded by bulk tissue, mitigating cell type compositional effects that impact bulk tissue interpretation (McCall et al., 2016; Farahbod and Pavlidis, 2019; Farahbod and Pavlidis, 2020; Zhang et al., 2021). However, cautious interpretation is still warranted due to the sparsity of scRNA-seq data. Correspondingly, the benefits of a meta-analytic framework (Lee et al., 2004; Mistry et al., 2013; Ballouz et al., 2015) have been extended to single cell coexpression to tasks such as gene function prediction (Crow et al., 2016; Crow and Gillis, 2018) and understanding reproducible patterns in the brain (Harris et al. 2021; Suresh et al. 2023; Werner and Gillis 2023). Importantly, these studies typically focused on the preservation of the global coexpression network structure, rather than any specific gene profile.

We drew inspiration from these works and our experience in aggregating ChIP-seq and TR perturbation studies to identify reproducible TR-target interactions (Morin et al., 2023). This stemmed from the recognition that the evidence from various lines of gene regulation methods often do not intersect, necessitating comprehensive data compilation (Hu et al., 2007; Gitter et al., 2009; Cusanovich et al., 2014; Garcia-Alonso et al., 2019; Kang et al., 2020). In this study, we adopt a “TR-centric” approach towards aggregating single cell coexpression networks, with the primary goal of learning reproducible TR interactions. Specifically, our focus was to assemble a diverse range of scRNA-seq data to better understand the coexpression range of all measurable TRs in mouse and human. Our key aim was to prioritize the genes that are most frequently coexpressed with each TR, hypothesizing that this prioritization can facilitate the identification of direct TR-target interactions. We further reasoned that this information would help establish expectations for more focused data aggregations.

Methods

All analyses were performed in the R statistical computing environment (R version 4.2.1 http://www.r-project.org). The associated scripts can be found at https://github.com/PavlidisLab/TR_singlecell.

Genomic tables

Gene annotations were based on NCBI RefSeq Select (mm10 and hg38) (https://www.ncbi.nlm.nih.gov/refseq/refseq_select/). High-confidence one-to-one orthologous genes were accessed via the DIOPT resource (V9; Hu et al. 2011; Hu et al., 2025). We kept only genes with a score of at least five that were also reciprocally the best score between mouse and human and excluded genes with more than one match. This resulted in 16,981 orthologous genes. Cytosolic L and S ribosomal genes were obtained from Human Genome Organization (groups 728 and 729; https://www.genenames.org/data/genegroup/#!/group/). This encompassed 89 human genes, which we subset to the 82 genes with a one-to-one mouse ortholog. Transcription regulator identities were acquired from Animal TFDB (V4; Shen et al., 2023).

scRNA-seq data acquisition and preprocessing

We focused on datasets with count matrices that had cell identifiers readily matched to author-annotated cell types. This was primarily sourced through two means: 1) From the “Cell x Gene” database (https://cellxgene.cziscience.com/), which has pre-processed and annotated data. When a single submission (“collection”) contained multiple downloads (for example, different tissue lineages), we downloaded all and combined them into a single dataset keeping only unique cells. 2) Automated screening followed by human curation of the Gene Expression Omnibus (GEO) database (Barrett et al., 2013). Here, we preserved the author-annotated cell types, save for when a biologically uninformative delimiter was used (e.g., “Neuron-1” and “Neuron-2”), in which case we collapsed these cell types into one to prevent overly sparse cell-type populations. We further acquired two tissue-panel datasets. The first was downloaded from the Human Protein Atlas (Uhlén et al., 2015; https://www.proteinatlas.org/download/rna_single_cell_read_count.zip, June 2023), covering 31 tissue-specific datasets which we collapsed into a single dataset and thus treated as a single network. Similarly, we downloaded each of 20 tissue datasets from the Tabula Muris Consortium (2018; https://figshare.com/articles/dataset/Robject_files_for_tissues_processed_by_Seurat/5821263; July 2023), which were also combined as one dataset.

Following the advice of the Harvard Chan Bioinformatics Core (https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html), we uniformly applied relatively lenient filtering rules for all datasets. We required a minimum cell count of 500 UMI (or equivalent) and 250 expressed genes, and a ratio of the $\log_{10}$ count of genes over $\log_{10}$ UMI counts greater than 0.8 for all experiments, save for SMART-seq assays, where the cutoff was relaxed to 0.6 as this technology can result in greater read depth for select genes (Wang et al., 2021). We applied standard CPM library normalization on the raw counts of all datasets (Seurat V4.1.1 NormalizeData “RC”), having observed that the log transformation in other normalization schemes resulted in elevated correlation reproducibility in our null comparisons.

scRNA-seq network construction

Aggregate single cell coexpression networks were constructed as described by Crow et al. (2016). Every dataset consists of a gene by cell normalized counts matrix, where each cell is associated to an annotated cell type. We fix genes to the RefSeq Select protein coding genes, setting unreported genes to counts of 0. This was done so that every resulting network had equal dimensionality.

For a given dataset, we performed the following steps for each cell type:

1. Subset the count matrix to only cells of the current cell type.
2. Set genes with non-zero counts in fewer than 20 cells to NA.
3. Calculate the gene-gene Pearson’s correlation matrix.
4. Set NA correlations resulting from NA counts to 0.
5. Make the correlation matrix triangular to prevent double ranking symmetric elements.
6. Rank the entire correlation matrix jointly, using the minimum ties method.

The resulting rank matrices across cell types were then summed into one matrix that was re-ranked and standardized into the range [0, 1] by dividing each element by the maximum rank. Higher values correspond to consistently positive coexpressed gene pairs, and values closer to 0 represent more consistently negative pairs. Step 2 is applied to ensure noisy coexpression values are not calculated from overly sparse populations, as recommended by Ballouz et al. (2015). The zero imputation in Step 4 is to ensure the ranking procedure includes non-measured genes, placing them in between positive and negative correlations. Thus, each dataset is represented as a single gene by gene matrix of coexpression scores aggregated across all labeled cell types. A gene profile refers to a single gene vector (such as a TR gene) from a single matrix; a set of profiles is the collection of profiles extracted from the experiments that measured the given gene.

Gene profile similarity

Coexpression profiles from any one dataset may not have a full complement of measured genes and thus contain tied ranks corresponding to missing values in our framework. Consequently, metrics of similarity that compare all of two lists, such as Spearman's correlation, are inappropriate and so we focused on the agreement of the Top and Bottom K genes between profiles. We calculated various set overlap metrics between lists and, finding our conclusions to be consistent, opted for the interpretability of reporting the size of the Top_K and Bottom_K overlaps. We restrict our reporting to TRs that were measured in at least five datasets.

For each TR, we calculated the pairwise similarities among its set of profiles. Averaging these similarity metrics was used to infer the consistency of a TR's coexpression profile across datasets. This process was also applied to each of the 82 ribosomal genes to provide a comparison with a set of genes known to be coexpressed. To generate a null comparison, a random TR was selected from each network to create a set of shuffled profiles, and pairwise similarities were calculated and averaged as above. This process was repeated 1000 times, generating a null distribution of average pairwise similarities. A TR with an average similarity greater than any of the 1000 nulls has an empirical *p-value* < 0.001.

Aggregating TR profiles and the effect of gene measurement sparsity

To prioritize the gene partners most commonly coexpressed with each gene, we averaged the set of rank-standardized profiles for the given gene into one aggregate profile. As each dataset-level profile had variable gene measurement, there was variable delineation between the positive coexpression values, the non-measured gene pair ties, and negative coexpression values. Therefore, for a given gene's set of profiles, we imputed all tied values to the median value before averaging, standardizing the values of non-measured gene pairs. A schematic is shown in Supplemental Fig. 1C.

Gene set enrichment

For each aggregate profile, we performed GO enrichment analysis of "biological process" terms with the "precRecall" R implementation of ermineJ (https://github.com/PavlidisLab/ermineR; Ballouz et al., 2016), using the aggregate values as scores. This approach considers the full scored list to find enriched terms but places greater emphasis on the top of the gene list. We analyzed 3,284 terms that had 20-200 genes and set the false discovery rate at 0.05 for considering terms significant. For the orthologous coexpression rankings we used human genes to map GO annotations.

ChIP-seq data acquisition and summarization

All ChIP-seq data was downloaded from the Unibind database (Puig et al., 2021; https://unibind.uio.no/downloads/; September 2022). For every TR experiment, we scored gene binding intensity using the same approach as in Morin et al., 2023, using a continuous scoring function (Ouyang et al., 2009; detailed in the Supplement). To generate an aggregate binding profile, we averaged the gene binding vectors specific to each TR. A “consensus” list of ASCL1 bound regions consisted of the union of all its peaks across ASCL1 Chip-Seq datasets (detailed in Supplement).

Literature curation evaluation

TR-target interactions supported by low-throughput experimental evidence were collected from our prior study (Chu et al., 2021), which compiled information from other resources (see Supplement for details) and then significantly expanded upon neurologically-relevant TRs. Since Chu et al. (2021) was published, we have further expanded this collection, to a total of 27,629 experiments encompassing 772 TRs and 5,899 gene targets. We then used each TR’s aggregate profile’s ranking as a score and its curated targets as labels, calculating AUC metrics (AUPRC: area under the precision recall curve and AUROC: area under the receiver operator curve) using the ROCR package (Sing et al., 2005; V1.0-11). To generate a null comparison for each TR, we randomly sampled from the entire literature curation corpus a number of targets equal to the count of curated targets for the given TR, and calculated AUCs using the TR’s aggregate profile as a score and the shuffled targets as labels. This process was repeated 1000 times to generate a null distribution of AUC values. The observed AUCs (using the TR’s true curated targets) were then compared to this distribution of null AUCs. A quantile of 1 means that the observed AUC was better than every single null AUC (empirical *p-value* < 0.001). We restrict our reporting to TRs that had at least five curated targets.

Cross-species coexpression profile comparison

There were 1,246 TRs with a one-to-one orthologous match between mouse and human that were also measured in at least 5 datasets in each species. For each of these TRs, we subset their aggregate profiles to the 16, 981 orthologous genes. Each orthologous TR thus has a mouse and human aggregate profile, generated separately across the respective species’ datasets. To generate a consensus orthologous profile for each TR, we took the rank product between its human and mouse aggregate profiles. To compare ortholog aggregate profiles, we calculated Spearman’s correlation and TopK and BottomK overlaps. Null comparisons were generated in a manner consistent with the individual profile comparison: similarities were calculated between randomly shuffled aggregate profiles between species over 1000 iterations.

To learn the specificity of a TR’s aggregate coexpression profile with its matched ortholog in the reciprocal species, we combined the framework applied in this study with prior studies examining the conservation of coexpression (Patel et al., 2012; Suresh et al., 2023). For each TR in a species, we selected the given TR’s top 200 coexpressed partners (Top_{200}). We next calculated the overlap of this gene set with the Top_{200} gene sets of each of the 1,246 TRs in the other species. We then treated the mismatched (non-orthologous) overlaps as a distribution and represented the matched (ortholog) TR’s Top_{200} as a quantile with respect to this distribution. We refer to this quantile as the *Ortholog retrieval score*. A score of 1 means that the given TR’s ortholog shared more top coexpressed partners than any other TR in the other species. This procedure was then repeated for the reciprocal species. The result is a pair of *Ortholog retrieval scores* for each TR: how well a human TR’s aggregate profile recovered its mouse ortholog relative to all other mouse TRs (human in mouse), and the recovery of the mouse ranking across human TRs (mouse in human).

Integrating coexpression and ChIP-seq profiles

For TRs with ChIP-seq data, we took the rank product of the TR's aggregate coexpression profile and its aggregate binding profile, re-ranking the result (Breitling et al., 2004; Wang et al., 2013; Morin et al., 2023). This orders genes by placing equal weight on their (positive) coexpression evidence and their binding evidence. We further report a second prioritization scheme for each TR, categorizing genes based on a cut-off of the rankings for both data types and species:

1. Stringent: Required a gene's presence in the Top 500 of both coexpression and binding in both species (orthologous genes only).

2. Elevated: Genes needed to make the Top 500 cut-off for both data types in one species and in one data type for the other species (orthologous genes only).

3. Species-specific: Top 500 cut-off for both data types in one species only. Notably, this may include genes absent from the one-to-one orthologous set, or TRs that had ChIP-seq data in one species only. Consequently, this tier had greater coverage than the others.

4. Mixed-species: Allowed genes ranked in the Top 500 in both data types, but each in only one species (orthologous genes only).

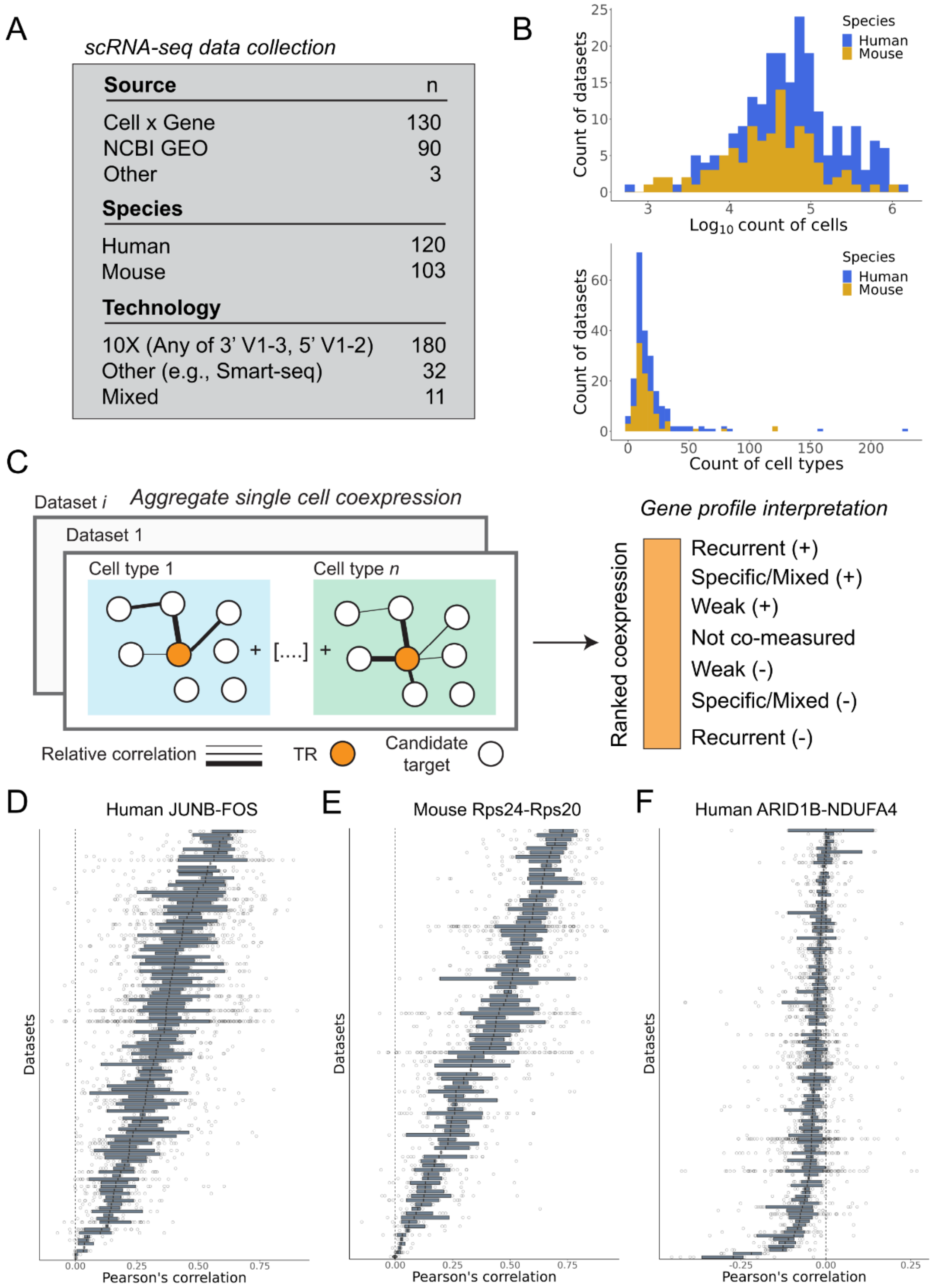
Results

Assembling a broad corpus of single cell RNA-seq data

To establish a diverse range of biological contexts for constructing single cell coexpression networks, we acquired scRNA-seq data from public resources (Methods). Our focus was strictly on datasets that included author-annotated cell type labels in the metadata, and all identified datasets underwent consistent preprocessing. In total, we analyzed 120 human datasets and 103 mouse datasets (Fig. 1A; Metadata in Supplemental Table 1). This corpus spans a wide range of biological contexts, scRNA-seq technologies, and counts of assayed cells. After filtering, the median human dataset measured 15,341 protein coding genes across 74,148 cells and 14 cell types; in mouse 13,996 genes across 36,755 cells and 12 cell types (Fig. 1B; Supplemental Figs. 1A,B). There was appreciable spread in these counts, with tissue atlas studies typically exhibiting the broadest coverage. The complete dataset is over 2.8 x10^7 cells.

Constructing single cell coexpression networks

We constructed aggregated single cell coexpression networks for each dataset using the approach outlined by Crow et al., 2016 (Methods). In brief, this entails generating a gene-by-gene correlation matrix for each cell type within a dataset, ranking each cell type correlation matrix, and consolidating them into a single network per dataset (Fig. 1C). Notably, unlike in Harris et al. (2021), where information was consolidated across



scRNA-seq data collection

Source	n
Cell x Gene	130
NCBI GEO	90
Other	3
Species	
Human	120
Mouse	103
Technology	
10X (Any of 3' V1-3, 5' V1-2)	180
Other (e.g., Smart-seq)	32
Mixed	11

Figure 1. Overview of study design. (A) Counts of datasets by source, technology, and species. (B) Top panel: Counts of cells across the dataset corpus. Bottom panel: Counts of cell types. (C) Schematic of the single cell coexpression aggregation framework and the interpretation of an individual gene coexpression profile. (D, E) Examples of the most reproducible positively coexpressed gene pairs. Each bar represents a dataset/network, and each point represents the gene pair's correlation in a cell type within the dataset. (F) Example of one of the most reproducibly negative coexpression gene pairs.

datasets for a single cell type, we first aggregate across cell types within a dataset before aggregating across datasets. In doing so, we explicitly prioritize signals shared across cell types. This strategy also minimizes effects due to expression differences between cell types, which we consider a separate question from “within cell” regulatory interactions (Farahbod and Pavlidis, 2020).

This procedure aims to rank coexpression partners, as illustrated in Fig. 1C, by ordering from “top” to “bottom”: consistently high positive interactions across cell types; mixed/specific positive interactions; weak-to-no coexpression; non-measured gene pairs; and then the increasingly most reproducibly negative coexpressed pairs. From this network, it is possible to extract a single gene column (herein, gene profile), such as for a TR, with the relative ordering reflecting the strength of its aggregate transcript covariation with all other genes.

While the focus of this study is on TRs, we first examined the globally most consistent coexpressed gene pairs (Figs. 1D-F). Top examples include TRs that dimerize to form the pleiotropic AP-1 complex, such as JUNB and FOS, as well as members of the ribosomal complex. Given the known biological coexpression of ribosomal genes (Li et al., 2016), we use a set of 82 large (L) and small (S) ribosomal genes that are highly conserved between mouse and human as a positive control when examining TR-gene coexpression in the following analyses (Methods). We also show one of the most consistently negative coexpressed TR-gene pairs in human. Aligning with our prior observations (Lee et al., 2004), the magnitudes of these values are smaller and less consistent than the positive coexpression profiles, contributing to the complexity in identifying repressive interactions (discussed in the Supplemental Material).

Similarity of TR-target profiles

Before prioritizing reproducible TR-gene interactions, we examined the concordance of the TR coexpression profiles between datasets. We expected that distinct profiles generated for the same TR and similar contexts would have elevated similarity relative to mismatched contexts or gene profiles. At the same time, the underlying data we used was from differing cell types, as datasets could be from different tissues. While we expected this would affect the degree of similarity, a total absence of overlap between profiles would raise questions about the efficacy of our framework in finding reproducible interactions.

We report here on the size of the overlap (*K)* of the top positively coexpressed (Top_K) genes between each pair of gene profiles (negative coexpression is discussed in the Supplemental Material). We examined a range of *K,* from 200 — approximately the top 1% of protein coding genes — to 1000, finding that our main conclusions were robust to this cut-off. To contextualize the similarity between TR profiles, we generated null similarities, iteratively sampling TRs across datasets and calculating the overlap of the shuffled TR profiles. We also report the similarity of the set of 82 L/S ribosomal genes.

First, for each TR we pairwise compared its profiles across studies. As expected, the most similar pairs were supported by datasets investigating similar biological contexts. For example, the best pairing in human (Top_{200} = 163/200) was between *FOXM1* profiles from two studies that both assayed the developing human intestine (Fawkner-Corbett et al., 2021; Elmentaite et al., 2021). The highest Mouse Top$_{200}$ (150/200) was associated with E2f8, derived from two studies of the blood-brain barrier (Posner et al., 2022; van Lengerich et al., 2023). The magnitude of the best ribosomal gene pairs was comparable: the best global human ribosomal pairing (Top$_{200}$ = 161/200) belonged to *RPS13*, originating from two immune cell studies (Liu et al., 2021; Domínguez Conde et al., 2022).

While these observations support the ability to find consistent coexpression patterns within pairs of similar contexts, our ultimate aim was to combine information across contexts. Seeking a more global summary of TR profile overlap, we calculated the mean Top$_{200}$ overlap for each TR profile across all unique pairs of networks measuring the TR. We again use the similarities from the pairs of randomly sampled TRs and the 82 ribosomal genes as reference.

In Figs. 2A,B, we show the average Top$_{200}$ of shuffled TR pairs across 1000 iterations. The typical null sample had an average Top$_{200}$ value of 2.7/200 in human and 2.6/200 in mouse. The ribosomal genes, approximating an empirical “upper bound,” averaged 61/200 in human and 44/200 in mouse. The distribution of average Top$_{200}$ values was highly skewed for TRs, with 67% of human TRs and 68% of mouse TRs having an average Top$_{200}$ value greater than the maximum value achieved across all of the null samples (empirical *p-value* < 0.001; represented as red lines in Figs 2A, B). And while the best individual ribosomal data pairs were equivalent in overlap size compared to the best individual TR pairs, ribosomal genes typically had a much greater average Top$_{200}$ than even the best TR. This underscores the unusual uniformity of ribosomal protein gene coexpression across distinct cellular contexts — it is an outlier. A similar comparison for the Bottom$_{200}$ is provided in Supplemental Figs. 2A-D.

TRs with the highest mean Top$_{200}$ values, indicative of the most consistent positive coexpression profiles across studies, were often associated with fundamental cellular housekeeping processes. For example, *E2F8* led in human (mean Top$_{200}$ 40.4/200), with mouse *E2f8* similarly having one of the most consistent profiles (Figs. 2A,B). The E2F family are well characterized regulators of the cell cycle (Emanuele et al., 2020), and other E2F members also ranked high in both species, as did regulators involved in early transcriptional response to environmental signals, such as AP-1 complex members *FOS* and *JUN*. In mouse, the highest mean Top$_{200}$ belonged to *Mxd3*, a MYC-antagonist whose human ortholog also had elevated similarity. More broadly, there was appreciable correlation between human and mouse orthologous TRs (Methods) in the consistency of their positive and negative coexpression profiles (Supplemental Figs. 3A,B).

TRs with context-restricted activity might be expected to exhibit relatively low cross-dataset similarity in our broad corpus. However, this is not necessarily the case. For example, the neural regulator NEUROD6 (Tutukova et al., 2021) had one of the most consistent TR profiles in human (mean Top$_{200}$ rank 44th out of 1,605 TRs), despite being only measured in 22 of 120 datasets. This shows that restricted expression does not preclude the identification of reproducible patterns. In contrast, human PAX6 —

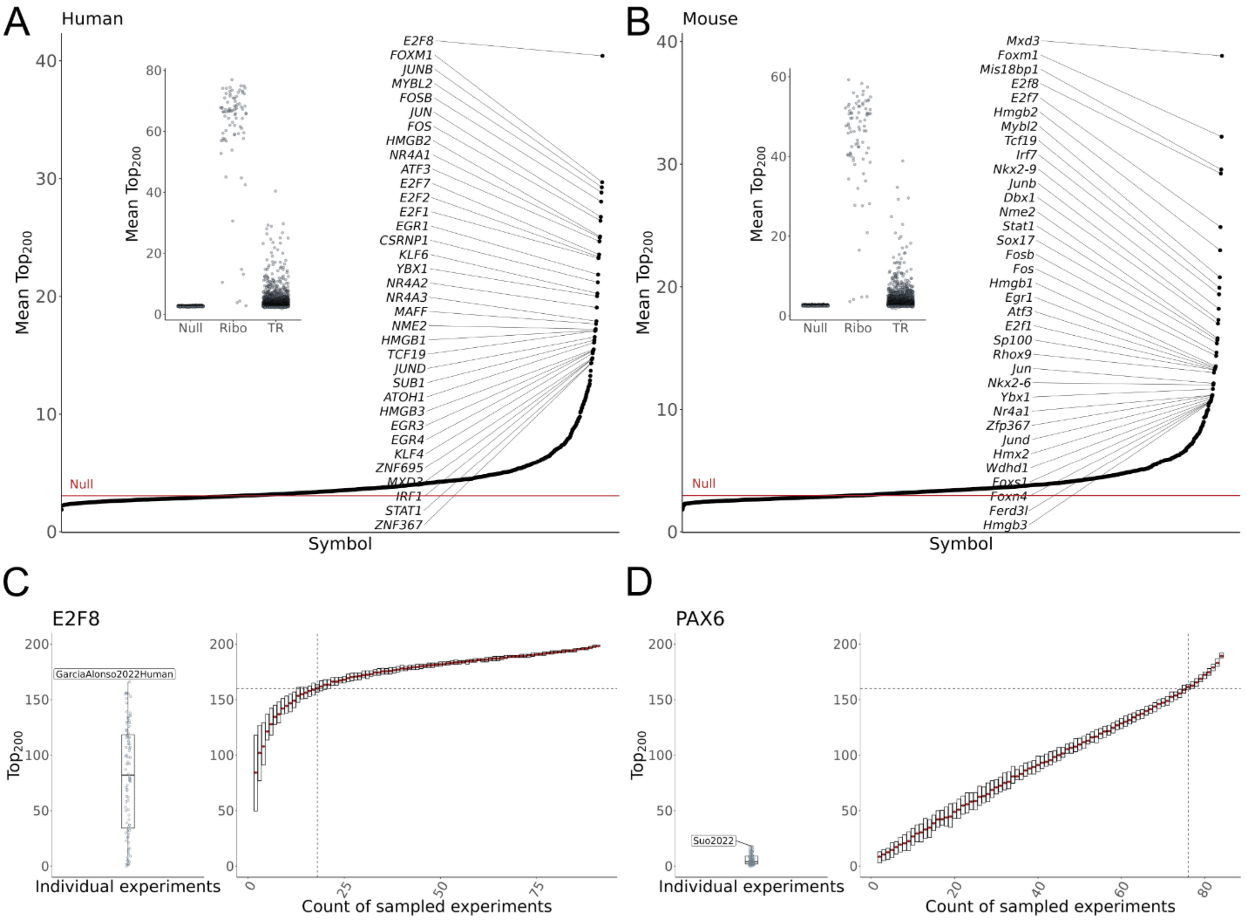


Figure 2. Similarity of TR profiles. (A) Inset: distribution of the mean Top_{200} overlaps for the null background, 82 ribosomal genes, and 1,605 human TRs. The null was generated through 1000 iterations of sampling one TR profile from each of 120 human datasets and calculating the average size of the Top_{200} overlap between every pair of sampled profiles. The ribosomal genes represent a “base case” scenario. Main: The average Top_{200} overlap of all human TRs, with the red line indicating the best null overlap. (B) Same as in A, save for 103 mouse experiments and 1,484 TRs. (C,D) Saturation analysis of global TR profiles for human (C) E2F8 and (D) PAX6. Left panels show the spread of Top_{200} overlaps between individual dataset profiles and the global E2F8 and PAX6 profiles. Right panels show the spread of overlaps when iteratively subsampling and aggregating datasets at increasing steps. Dotted lines indicate the average number of sampled datasets required to reach 80% of the global profile. E2F8 recovers its global profile with relatively fewer datasets than does PAX6.

necessary for the development and function of several nervous and pancreatic tissues (Wen et al., 2009; Yeung et al., 2016) — had a mean Top_{200} value marginally above the null, improving slightly at K=1000 (Supplemental Figs. 2E,F). Although PAX6 can also be described as a context-restricted, it was detected in 85 of 120 datasets, suggesting greater heterogeneity in its coexpression profiles compared to NEUROD6.

Ranking aggregated coexpression to prioritize TR-target candidates

The preceding section demonstrated that similar TR profiles could be identified across this biologically heterogeneous corpus, supporting the potential to find reproducibly coexpressed gene pairs. We thus turned to our primary aim of prioritizing these consistent interactions, generating a unified gene ranking for each TR using all compiled data. This process involves aggregating information at two levels: first, across cell types *within* a dataset (as in the previous section; Fig. 1C), and then, for each TR, aggregating their profiles *across* datasets (Methods, Supplemental Fig. 1C). This approach aims to maintain the interpretability of an aggregate profile relative to a profile from an individual network (Fig. 1C): the extremes represent the most consistent positive and negative correlations, while the middle of the list encompasses weak and non-measured coexpression gene pairs.

As before, we used the set of ribosomal genes to validate that our aggregation workflow prioritized known biological coexpression (Supplemental Material; Supplemental Figs. 4A,B). We next performed GO biological process enrichment on all aggregate profiles (Supplemental Fig. 4C), finding that most TRs (91% human, 86% mouse) were associated with at least one term (FDR 0.05). E2F8 coexpression partners were enriched for multiple terms relating to cytokinesis and chromosomal organization, as expected for its known role in these processes (Emanuele et al., 2020). We also frequently observed that terms affiliated with tissue-specific processes were enriched for TRs implicated in those tissues. Examples include glial development and myelination terms for the oligodendrocyte TRs OLIG1/2 (Szu et al., 2021), neuronal synaptic functionality for the aforementioned NEUROD6 (Tutukova et al., 2021), leukocyte and cytokine processes for IRF8 (Salem et al., 2020), and hematopoietic terms for the erythroid GATA1 (Ferreira et al., 2005). Some tissue-selective TRs were enriched for more general regulator terms (e.g., “cell fate commitment” for mouse Pax6) or had disparate tissue-specific terms (e.g., “regulation of osteoblast differentiation” and “regulation of neuron differentiation” for SOX4), potentially reflecting data heterogeneity. While GO is an imperfect resource, these results agree with our other observations that our analysis yields biologically-relevant signals.

We examined the relationship between the aggregated global TR profiles and the constituent datasets through two analyses. First, we assessed how well individual experiments aligned with the global profiles to identify potential biases (Supplemental Material). As shown in Supplemental Figs. 2H-L, datasets with the highest agreement were large studies of broad tissues using the 10X Chromium platform, though consistencies between platforms were still observed (Supplemental Fig. 2G).

Second, we performed a saturation analysis to determine how many datasets are needed, on average, to recover each TR’s global profile (≥80% overlap in Top_{200}

genes). By iteratively subsampling and aggregating datasets, we evaluated the convergence of sampled TR profiles to the global set. For example, E2F8 (Fig. 2C) required an average of 18 of 92 datasets to reach saturation, while PAX6 (Fig. 2D) showed a linear trend, indicating saturation has not yet been achieved. These results suggest future work is needed to explore not only replicable context-specific patterns for TRs such as PAX6, but also the extent to which globally consistent partners can be found when using more data.

Recovery of literature-curated TR-target interactions

Equipped with a unified single cell coexpression profile for each human and mouse TR, we aimed to assess their concordance with an orthogonal line of regulation evidence. While coexpression is expected to prioritize both direct and indirect regulatory interactions (the latter we would consider false positives), the rankings should still demonstrate a greater ability to recover true direct interactions relative to a null expectation.

In a previous study (Morin et al., 2023), we evaluated the utility of aggregating TR perturbation and ChIP-seq experiments, using literature-curated low-throughput interactions as positive labels and calculating area under the curve (AUC) metrics (Marbach et al., 2012; Garcia-Alonso et al., 2019). We applied the same framework here, using curated TR-target interactions we have collected (Chu et al., 2021, since expanded) and assembled from other resources (see Supplement for further discussion). We considered TRs that had a minimum of five curated targets, resulting in 451 TRs analyzed in human (median count of curated targets = 18) and 434 in mouse (median count = 17).

We first examined the effectiveness of the aggregate profiles in recovering curated targets relative to the individual TR profiles that compose each aggregate. On average, the aggregate profiles outperformed (better prioritized curated targets) the expected AUC value from an individual profile (Supplemental Fig. S5A). Therefore, aggregating the coexpression networks typically maintains or improves performance on this benchmark.

Next, we evaluated the efficacy of the coexpression rankings in recovering curated targets relative to a null distribution of AUCs (*Quant_coexpression)*. While the raw AUC values were typically better than random (Fig. 3B, Supplemental Fig. 5B), we report the quantile of the observed value relative to a null to standardize the comparison across TRs (discussed in Supplemental Material). This null was created by size-matching and randomly sampling from the pool of curated targets from the entire literature-curation corpus. The latter helps account for biases in the coverage of targets in the low-throughput literature. A *Quant_coexpression* value of 1 indicates that an aggregate profile outperformed every null sample.

ASCL1 is provided as an example of this procedure for one TR in Fig. 3A. As illustrated in Fig. 3C, the coexpression aggregates consistently exceeded the null AUCs, reflected by a median AUROC *Quant_coexpression* of 0.95 in human and 0.93 in mouse. The pile-up of quantiles near or equal to 1 indicates that, while not universal, a majority of

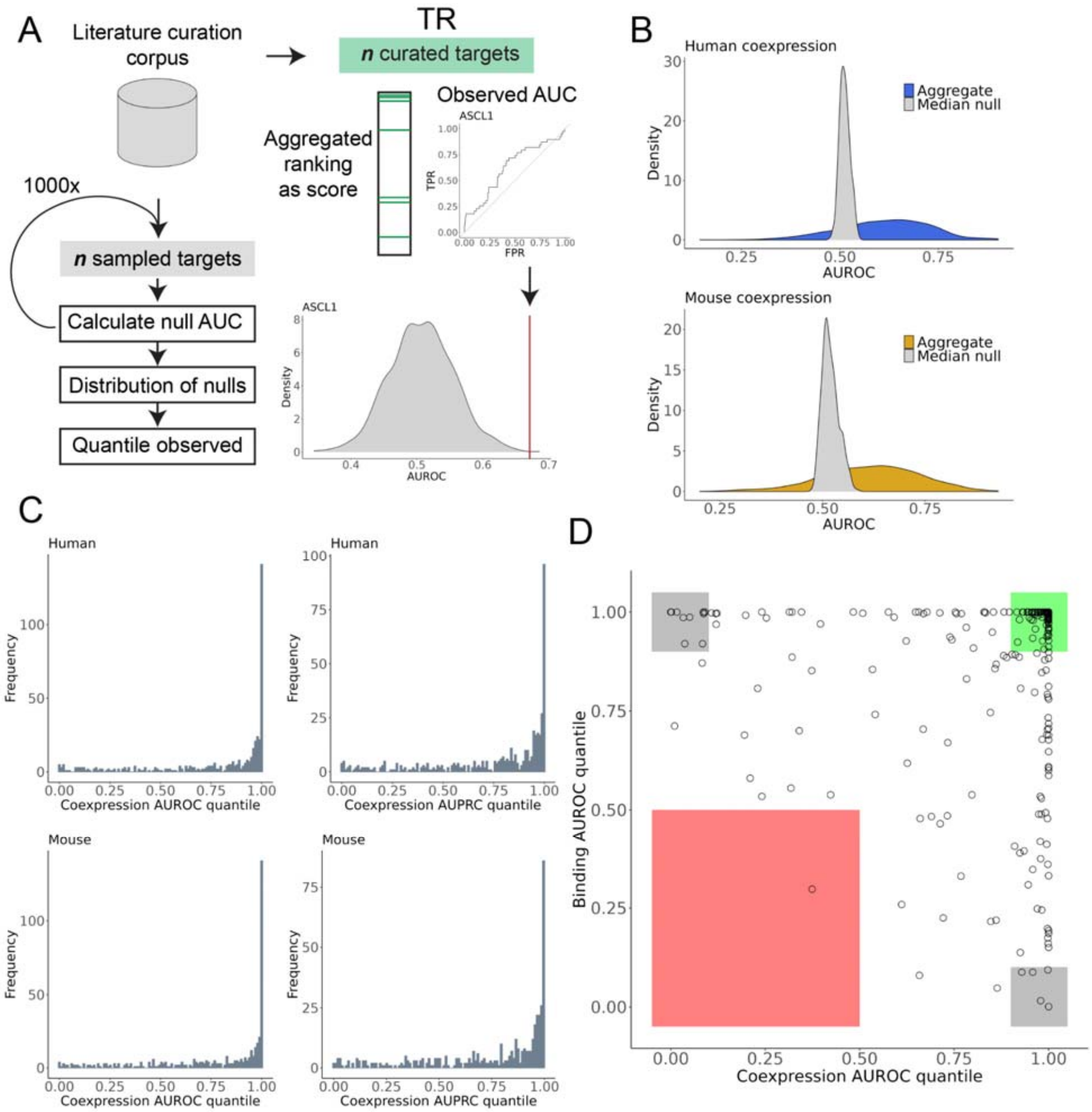


Figure 3. Recovery of literature curated targets by aggregate rankings. (A) Schematic of literature curation evaluation. (B) Distributions of the observed AUROCs for 451 human and 434 mouse aggregate TR coexpression profiles, along with the distribution of the median null AUROCs generated for each profile. (C) Histograms of the AUROC and AUPRC coexpression quantiles for human and mouse. (D) Scatter plot of the AUROC quantiles for the coexpression and binding profiles of 253 human TRs that had binding data and at least five curated targets. Green box indicates TRs for which both genomic methods were effective in the benchmark, grey box for only one method, and red box for neither method being effective.

TR single cell coexpression rankings excelled in prioritizing matched curated targets over randomly sampled targets. These observations strongly suggest that these aggregate rankings are capable of prioritizing regulatory interactions that were identified through targeted biochemical assays.

To further contextualize these performances, we conducted a similar null AUC analysis, this time using aggregate ChIP-seq signals. In brief, we applied the same approach as in Morin et al., 2023, scoring gene-level binding intensity for each ChIP-seq experiment, then averaging these signals within each TR's set of experiments to create a single unified ranking of gene binding for each TR. In total, we considered 4,115 human experiments for 253 TRs and 3,564 mouse experiments for 241 TRs from the Unibind database (Puig et al., 2021, Methods) that had at least five curated targets. As with the aggregate coexpression signal, we compared the unified binding ranking's ability to recover TR-specific curated targets relative to a null of sampled targets (*Quant_binding*).

We anticipated that TR ChIP-seq, as a more direct form of regulatory inference, might outperform coexpression (Garcia-Alonso et al., 2019). However, in our hands the aggregate binding evidence was on par with single cell coexpression in its ability to predict known targets (Supplemental Fig. 5C), further motivating integration of both data types. Supporting this, integrating the coexpression and binding rankings for available TRs typically led to elevated performance in the benchmark (Supplemental Fig. 5D).

Among TRs with both binding and coexpression data, many performed well in the benchmark for both data types separately, as demonstrated for human TRs in Fig. 3D. In human, 134 of 253 (53%) TRs had AUCs (AUPRC or AUROC) *Quant_binding* > 0.9 and *Quant_coexpression* > 0.9; in mouse 126 of 241 (52%). This signifies that, for these specific regulators, aggregated coexpression and binding profiles both effectively prioritize curated TR targets relative to sampled targets. This alignment highlights TRs whose activity may be more readily identified through distinct data modalities. Further, of the TRs performant in both lines of evidence, more than half did so in both species (human 83 of 134, mouse 83 of 126), suggesting convergence of evidence across not only experiments, but also species.

This agreement of evidence encompassed broadly active TRs, such as those involved in the AP-1 complex. However, it also included more specialized factors, such as the neuronal-specifying ASCL1, and the aforementioned PAX6. This suggests that, even though the average overlap of PAX6 profiles was weak (Fig. 2D; Supplemental Figs. 2E,F), there was still a consensus of recurrent curated PAX6 targets within these smaller intersects. We also find cases where only one data type was performant. LEF1, for example, had an AUROC *Quant_coexpression* value of 1 in both species but a *Quant_binding* value of 0 and 0.22 in human and mouse, respectively.

Finally, because negative expression correlations might be of interest for identifying repressive interactions, we conducted an analysis of the reproducibility and performance of the relations predicted from the bottom of the rankings. We found that for some TRs, negative correlations performed better than positive correlations in the benchmark, though this was the exception (Supplemental Fig. 5B). This suggests that for some TRs, repressive activity might be inferable from coexpression (see Supplement for discussion; Supplemental Tables 2-3).

Identification of conserved interactions

It has been observed that, despite the high evolutionary turnover of regulatory DNA sequences, TR-target relations exhibit relatively high conservation (Yue et al., 2014), with coexpression providing an attractive means to nominate common and divergent interactions (Monaco et al., 2015; Lee et al., 2020; Suresh et al., 2023). Here, our aim was to identify the extent to which individual TR aggregate coexpression profiles were preserved between mouse and human, focusing on orthologous genes (Methods). A meta-analytic comparison of TR single cell coexpression profiles between these two species is lacking, and we reasoned that evidence of conservation using this global data corpus would provide future support for studies that focus on specific TR patterns in a more focused context.

Figure 4A demonstrates the similarity distributions between ortholog aggregate coexpression profiles, overlaid with the median observed and shuffled null values. Although there was appreciable spread in these similarity metrics, most orthologs shared more similarity in their profiles than would be expected from shuffled TRs, suggestive of conserved TR coexpression. While there are TRs that agree less well between species, we are cautious in interpreting this as species-specific regulatory rewiring, given the relatively modest effect size and the absence of an exact match in cellular contexts covered across both species.

Given our emphasis on reproducible interactions, we focused on the overlap at the extremes of these species rankings (Figs. 4B,C; Supplemental Fig. 6C). To quantify the specificity of this overlap, we applied a slightly modified framework of the Top$_K$ overlap used in this study, consistent with prior studies (Methods; Patel et al., 2012; Suresh et al., 2023) and illustrated in Fig. 4B. The result is a pair of ortholog retrieval scores for each TR: how well a human TR's ranking recovered its mouse ortholog relative to all other mouse TRs (human in mouse), and the recovery of the mouse ranking across human TRs (mouse in human), with a value of 1 indicating perfect retrieval.

As demonstrated in Fig. 4C, there was considerable preservation of single cell aggregate TR coexpression profiles between mouse and human. The median ortholog retrieval score for human was 0.969, with 175/1,246 (14%) TRs having a perfect value of 1; in mouse these values were 0.973 and 172/1,246 (14%), respectively. These relative values correspond to a median *Top$_{200}$* overlap of 14 genes, with FOXM1 and HMGB2 each having a maximal *Top$_{200}$* of 149 genes (Fig. 4A). While the most conserved TRs (by *Top$_{200}$* overlap) were led by regulators of housekeeping processes such as cell division, we also observed this preservation among more specific TRs, such as the aforementioned NEUROD6 (human in mouse and mouse in human = 1, *Top$_{200}$* = 50). Logically, many of these highly preserved TRs also had similar profiles across datasets within species (as shown in Figs. 2A,B), and those that were weakly preserved generally lacked consistency within species (Supplemental Figs. 6A,B). These findings collectively contribute to characterizing the extent to which each TR can

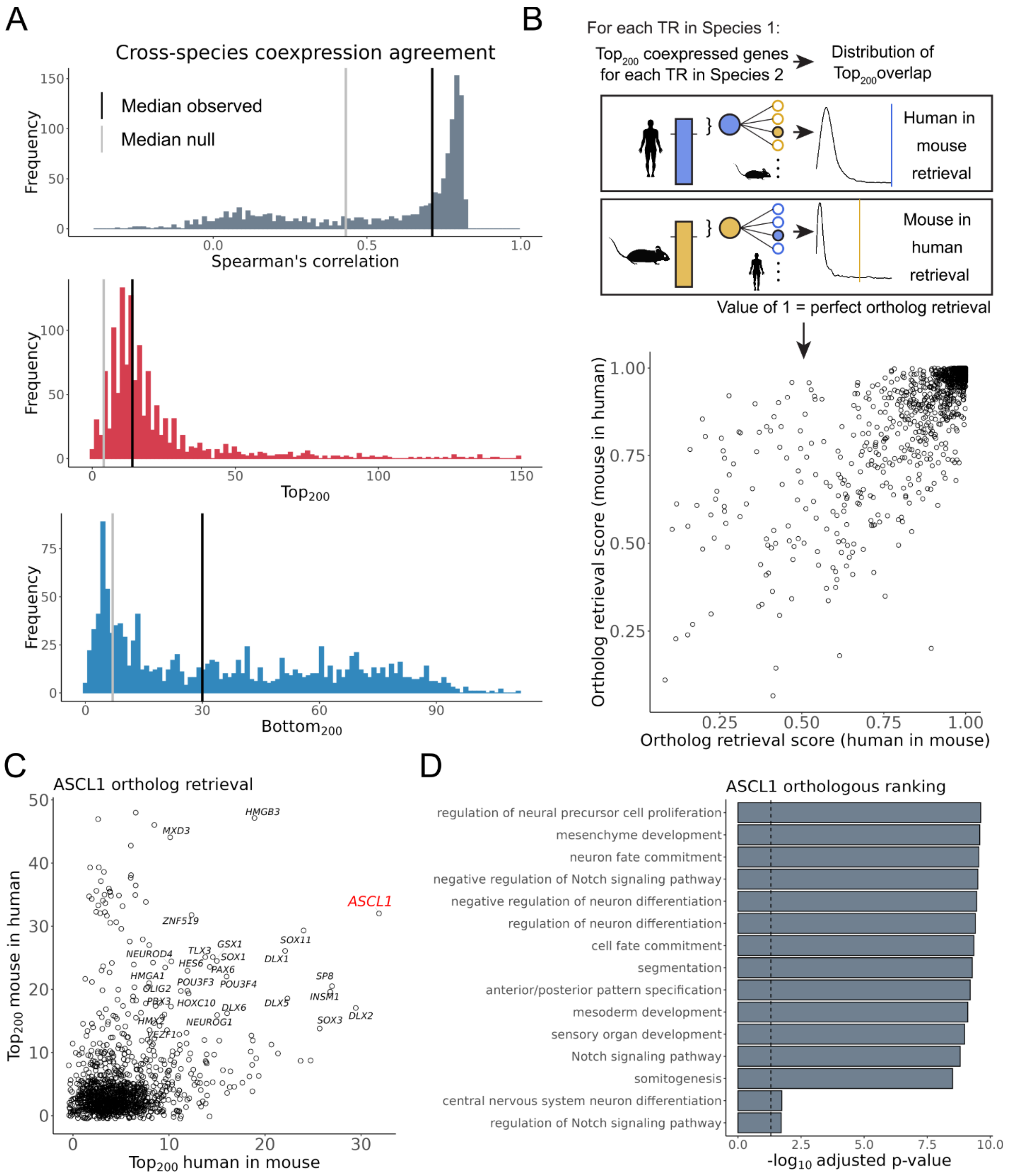


Figure 4. Preservation of mouse and human single cell coexpression profiles. (A) Distribution of coexpression agreement between the aggregate single cell coexpression profiles of 1,246 orthologous TRs. Black lines indicate the median value for the TRs, grey lines indicate the median of null values generated by shuffling pairs of orthologous TRs. (B) Top: Schematic of the ortholog retrieval workflow, adapted from Suresh et al., 2023. Bottom: Scatterplot of the resulting ortholog retrieval scores (C) Scatter plot of the ASCL1 *Top*$_{200}$ overlaps. (D) The top 15 GO terms when combining the human and mouse top ASCL1 coexpressed gene partners.

be defined by a set of coexpressed gene partners, facilitating inferences into their biological roles.

In Fig. 4C we illustrate this overlap procedure for ASCL1, an essential pioneer nervous system regulator that is also relevant to cancer (Castro et al., 2011). Of the 200 genes that were most consistently coexpressed with human ASCL1, 32 of their mouse orthologs were also in the mouse Ascl1 Top$_{200}$ set. This marked the largest overlap human ASCL1 had with any mouse TR (human in mouse = 1). In the reciprocal comparison, where mouse Ascl1 was queried against all human TRs, human ASCL1 ranked 30th (mouse in human = 0.98). The 29 human TRs with a greater overlap with mouse Ascl1 did not have a sizable overlap in the reciprocal comparison, save for HMGB3. Conversely, TRs other than ASCL1 with elevated overlap across species included the ASCL1 curated targets INSM1, HES6, and DLX5 (Castro et al., 2006; Nelson et al., 2009; Kito-Shingaki et al., 2014). Other TRs are well-characterized for operating in a regulatory network with ASCL1 — though not necessarily as direct downstream targets — such as DLX1/2/6, GSX1/2, SP8, and OLIG2 (Wang et al., 2013; Al-Jaberi et al., 2015; Liu et al., 2017; Aslanpour et al., Lunden et al., 2019; 2020). GO enrichment of the top ASCL1 coexpressed gene partners using information from both species returned numerous terms that are consistent with ASCL1's role in brain development (Fig. 4D).

Combining single cell coexpression and aggregated binding reveals numerous reproducible interactions

Up to this point, we have presented evidence supporting the existence of recurrent single cell TR-gene coexpression patterns within (Fig. 2) and across species (Fig. 4), demonstrating that this information can prioritize curated experimental interactions (Fig. 3). One of our primary motivations is to prioritize the direct gene targets of TRs (Morin et al., 2023). However, the correlation of TR-gene transcripts serves as an indirect form of gene regulation evidence — it does not confer information about the causative directionality of this covariation. We thus now turn to identifying interactions corroborated by TR binding evidence, using the same aggregated Unibind ChIP-seq data examined in the literature curation evaluation. We reasoned that, as in our earlier work, knowledge of binding can help focus attention on expression patterns more likely to reflect direct regulatory relations.

We present two straightforward strategies for prioritizing reproducible interactions, acknowledging the use of relatively arbitrary cut-offs for the sake of reporting. All summarized rankings are made available for researchers interested in conducting their own exploration. We first combined the single cell coexpression and binding profiles into a final ordered ranking for TRs with ChIP-seq data, using the common rank product summary (Breitling et al., 2004; Wang et al., 2013; Morin et al., 2023). This was done separately for each species (317 TRs in human, 305 in mouse), as well as across species for orthologous TRs with available data (216 TRs). This establishes convenient lists that order the protein coding genes most associated with each TR based on their aggregated single cell coexpression and binding profiles.

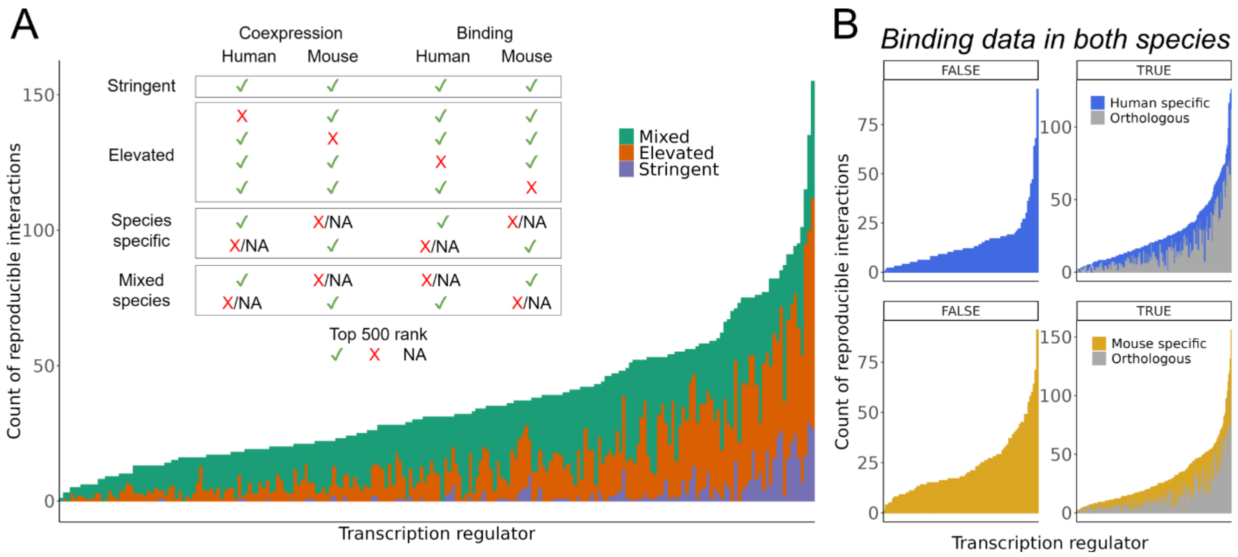


Figure 5. Count of interactions supported across methods and species. (A) Inset: criteria used to group interactions into tiers. Bar chart: Count of unique interactions gained in each orthologous tier (Stringent, Elevated, and Mixed-Species) for the 216 TRs with binding data in both species. (B) Count of Species-Specific interactions for 317 TRs in human (top) and 305 TRs in mouse (bottom). TRs are split by those with ChIP-seq data in one species only (left) and thus are ineligible for consideration in the orthologous interactions, and those with ChIP-seq data in both species (right). Grey bars indicate the count of interactions already found in the Stringent and Elevated sets, coloured bars indicate the count of Species-Specific interactions that were gained due to lacking orthologs or because they had elevated ChIP-seq signal in one species and not the other.

Recognizing that a gene may be prioritized (have a better rank product) if ranked exceptionally well in one data type or species only, we introduce a second scheme for more balanced consideration across lines of evidence. For each TR, genes are categorized into tiers by their status across the rankings, as illustrated in the inset of Figure 5A. This collection provides examples of regulatory interactions supported by both binding and single cell coexpression evidence.

Fig. 5A shows the counts of unique orthologous interactions gained in each tier of evidence for the available TRs. The Stringent level, representing the most reproducible interactions across both species and genomic methods, contains 545 TR-gene pairs corresponding to 101 TRs and 357 unique genes. The TRs with the largest Stringent collection featured multiple AP-1 members, led by FOSL1 with 29 genes, along with immunity TRs such as STAT1, STAT2, and IRF1. More specialized TRs also had among the largest Stringent sets, such as the hematopoietic factors SPI1 (n = 27), GATA1 (n = 16) and GATA2 (n = 11), and the hepatic HNF4A (n = 8). This once again suggests conservation of many regulatory interactions, although it is essential to note that this observation is influenced by the limited coverage of ChIP-seq data across biological contexts.

The Elevated collection relaxes the criteria to allow orthologous genes reaching the cut-off in three of the four rankings. This resulted in 3,106 Elevated TR-gene pairs, with 211 of the 216 available TRs having at least one gene in their set (median = 10). TRs with the largest Elevated collection closely overlapped with those having the largest Stringent sets, reinforcing the notion of preserved target genes among these TRs. The Species-specific level encompasses two groups of TRs: those that have ChIP-seq data in both species and those in only one. This is reflected in Fig. 5B, where we show the count of reproducible interactions for each group. The left panels display TRs with ChIP-seq in only one species and were thus ineligible for consideration in the Stringent or Elevated tiers. In human, this corresponded to 99 TRs with a median of 11 interactions. TFDP1 led with 93 genes supported by both aggregated single cell coexpression and binding evidence. In mouse, all 89 available TRs were associated with at least one gene (median = 18), with the interferon TR Irf8 having a maximum of 91 genes, including numerous immunity-associated genes such as *Mpeg1*, *Ctss*, *Cd180*, *Xcr1*, and *Trim30a*.

Highlighting ASCL1

We conclude by focusing on ASCL1, emphasizing that this exploration of ASCL1 regulatory targets is just one example made possible by the information we have summarized and made available for community use. In Fig. 6A we present the genes in each tier of evidence for ASCL1, along with their curation status from the 39 available ASCL1 targets in the literature corpus. Human ASCL1 was measured in 61 of 120 scRNA-seq datasets, and in mouse 65 of 103. Regarding ASCL1 binding data, there were 10 ChIP-seq datasets in human — largely in cancer cell lines — as well as 10 in mouse, mostly in neuronal and embryonic contexts.

Two genes fit the Stringent criteria used for this report: the literature-curated ASCL1 target and Notch signalling ligand DLL3 (Henke et al., 2009), and the cell cycling phosphatase CDC25B, which was not in the low-throughput literature collection but is nevertheless discussed elsewhere as a target of ASCL1 (Castro et al., 2006). The Elevated set consisted of 26 genes, with 6 narrowly missing the Stringent criteria (indicated by lighter shading in Figure 6A). Among them are well-described and literature-curated ASCL1 targets, such as the Notch effector HES6 (Nelson et al., 2009) and the neuroendocrine regulator INSM1 (Jacob et al., 2009; Jia et al., 2015). ASCL1 and INSM1 serve as markers for neuroendocrine tumours, such as for small cell lung carcinoma (SCLC; Zhong et al., 2022). Another Elevated ASCL1 gene, CKB, has upregulated expression in both SCLC (Borromeo et al., 2016; Qu et al., 2022) and ASCL1-high atypical teratoid/rhabdoid tumours (Tamrazi et al., 2019), suggesting an ASCL1 interaction with oncogenic potential across various contexts. We additionally draw attention to the BAF chromatin remodeler BCL7A, for which we found no ASCL1 connection in the literature, and which is also associated with diverse cancers (Baliñas-Gavira et al., 2020; Liu et al., 2021).

Other Elevated interactions help characterize ASCL1 as a regulator of both neuronal and oligodendrocyte lineages. This includes the cell cycle regulator GADD45G (Huang et al., 2010), the neuronal tubulin TUBB2B (Mazurier et al., 2014; Lin et al., 2017), and acetylcholine receptor subunit CHRNA4 (Ueno et al., 2012). PPP1R14B and ASCL1 expression was used to define a primitive oligodendrocyte progenitor population (Weng et al., 2019). We were unable to find (from a low-throughput study or otherwise) a direct connection between ASCL1 and the neuronal adhesion ADGRG1 (Simão et al., 2018), the cortical-marker and calcium-binding regulator KCNIP3 (Ragazzini et al., 2023), or the neuronal splicing factor CELF3 (Yu et al., 2017), although the latter is used as a neuroendocrine marker to characterize ASCL1-high SCLC subtypes (Zhang et al., 2018). Finally, we highlight REPIN1, an Elevated gene that lacked any ASCL1 connection in the literature that is also generally understudied.

The next tier, of Species-Specific sets, each comprised 19 genes. PRDX2, for example, is a neuronal-enriched mitochondrial gene that has been shown to enhance ASCL1-induced astrocyte-to-neuron reprogramming (Russo et al., 2021). HEPACAM2 is another gene implicated in cancer (Deprez et al., 2020; Yamada et al., 2022) that we could not find a direct ASCL1 association in the literature. TMEM61, lacking a 1:1 mouse ortholog, was only eligible for consideration in the Human-specific set, while the reciprocal applied to the mouse Nbl1. Of the 27 genes in the final tier, the Mixed-Species set, we highlight CXXC5. This zinc finger TR was initially characterized as a bone morphogenic-responsive regulator of Wnt signaling in neural stem cells (Andersson et al., 2009), and has been further described as a signal integrator in development and homeostasis with tumour suppressive qualities (Xiong et al., 2019). These examples collectively illustrate the diverse roles of essential TRs, such as ASCL1, in development and disease.

Lastly, we summarize the compiled evidence for the Notch ligand encoding *DLL3*, a well-established and curated ASCL1 target (Henke et al., 2009) that was present in the Stringent collection. DLL3 ranked fourth in the ASCL1 coexpression rankings in both species, making it one of ASCL1's most reproducible coexpression partners. Figs. 6B,C illustrates the distribution of Pearson's correlations for the 238 annotated cell types from

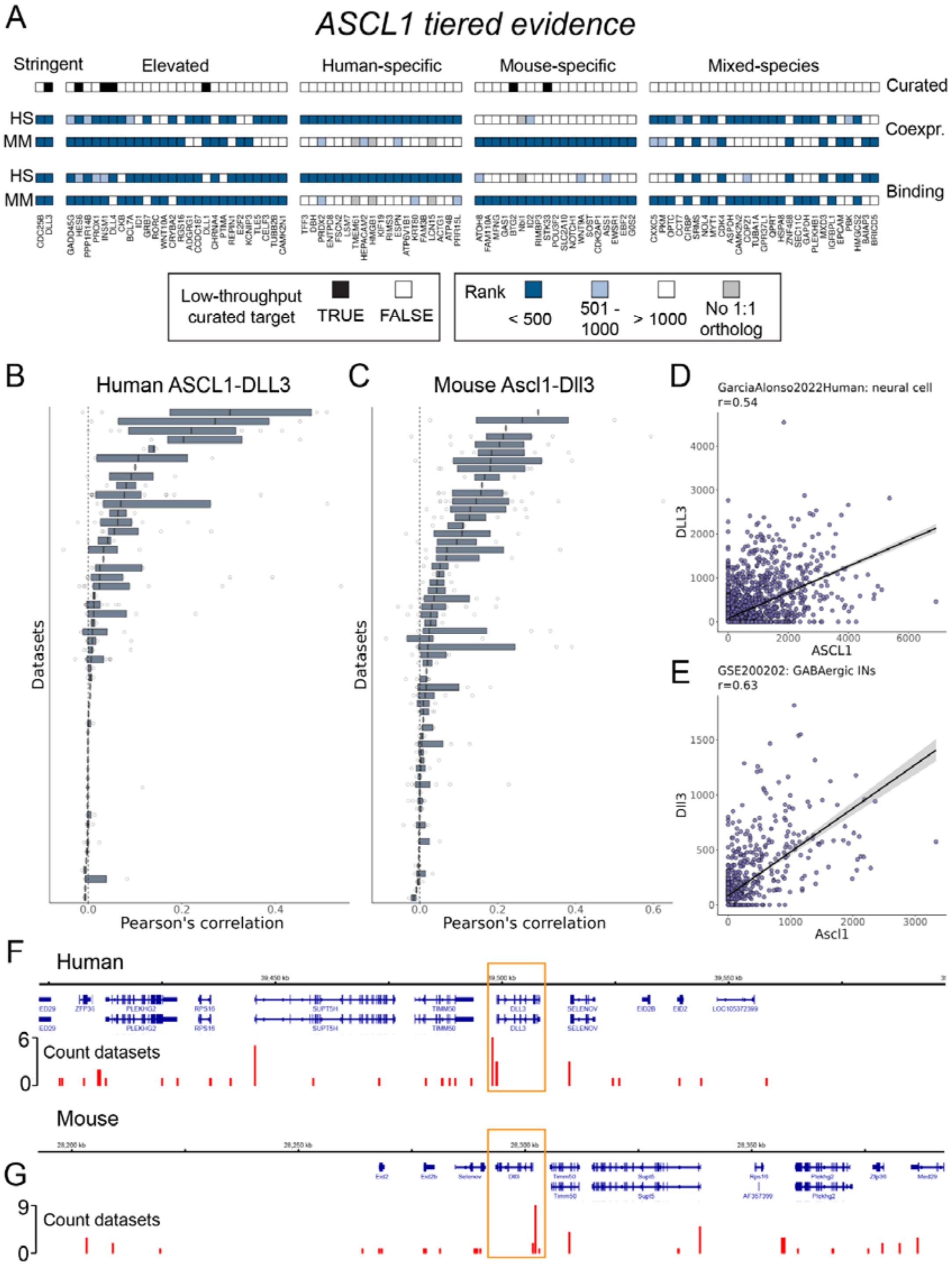


Figure 6. Reproducible ASCL1 interactions. (A) Heatmap representing the tiered evidence for ASCL1 candidate targets. (B, C) Distribution of Pearson's correlations for ASCL1-DLL3 in (B) human and (C) mouse, as in Fig. 1E-G. (D, E) Scatterplot of the CPM values for ASCL1 and DLL3 for the cells belonging to the cell type that had the highest correlation in the entire corpus for (D) human and (E) mouse. (F, G) Genome track plots centered on DLL3 (yellow boxes) in (F) human and (G) mouse, where the base of the red bars indicates ASCL1 binding regions, and the height indicates the count of ASCL1 ChIP-seq datasets with a peak in the region.

54 human datasets in which ASCL1 and DLL3 were co-measured (275 cell types in 61 datasets for mouse). Notably, despite being one of the most reproducible ASCL1 coexpressions, this association is not universal across all cell types. Figs. 6D,E shows the scatter plots of the individual cell types in which the greatest correlation was found: in human, annotated as “neural cells” ($r = 0.54$; Garcia-Alonso et al., 2022), and in mouse, “GABAergic INs” (interneurons) ($r = 0.63$, Hamed et al., 2022). Given the importance of ASCL1 regulation of Notch signalling in neuronal cells (Castro et al., 2006; Castro et al., 2011; Lampada and Taylor, 2023), these collective observations support that our resource can still prioritize specific interactions.
In Figs. 6F,G, we demonstrate the ASCL1-DLL3 binding evidence; DLL3 was ranked 493rd in the human aggregate binding profile and 81st in mouse. In human, this corresponded to 83 discrete bound regions (Methods) within 500Kb of either direction of the DLL3 TSS, and 25 within 100Kb; in mouse 73 regions within 500Kb and also 25 within 100Kb. We calculated which regions were most frequently bound by ASCL1 across datasets, reasoning that this may help prioritize functional ASCL1-DLL3 enhancers (while being cognizant of biasing factors like open promoters). Using the 500Kb cut-off in human, we found that 20 sites were bound in more than one dataset, and that a region approximately 775 base pairs upstream of the DLL3 TSS had a maximum count of 6. In mouse, 28 regions were bound across multiple datasets, with the most frequently bound region (nine of ten datasets) occurring approximately 400 base pairs upstream of the DLL3 TSS.

Discussion

In this study we pursued two main objectives. First, we aimed to understand the behavior of the meta-analytic strategy of aggregating single cell coexpression networks (Crow et al., 2016), applying this methodology across a large and broad corpus of scRNA-seq studies. We believe this technique holds great potential in uncovering robust gene coexpression patterns free from the confounding effect of cellular composition. However, before considering specific cell types or conditions, we sought to calibrate expectations using a large collection of heterogeneous data. This objective aligned with our second aim of identifying reproducible transcription regulator coexpression patterns. We wished to assess how well this information aligns with other lines of regulation evidence, and to provide an organized summary of this information as a community resource (https://doi.org/10.5683/SP3/HJ1B24).

While prior work has nominated TR-target interactions across a large and context-independent corpus of data (Garcia-Alonso et al., 2019; Keenan et al., 2019; Müller-Dott et al., 2023), to our knowledge ours is the first to do so using a broad range of single cell transcriptomics. Our literature curation benchmark strongly supports the ability of this resource to prioritize curated targets, and we further find numerous examples of reproducible and conserved coexpressed TR-gene partners also supported by ChIP-seq evidence. Collectively, this suggests that this information can help prioritize interactions when direct experimental evidence is lacking. Our benchmarks additionally provide insight into the TRs whose activity is more challenging to uncover, given the considered genomics data (Supplemental Tables 2-3).

Our workflow prioritizes interactions that are most common across contexts, akin to our prior study (Morin et al., 2023). Overall, it is not surprising that the most reproducible relationships tend to relate to processes shared by many cell types. This may be partly a function of expression levels (Crow et al., 2016), but it is logical that the dynamics of processes like the cell cycle are more readily captured by changing transcript levels. We still find evidence for highly context-specific interactions: as long as there is enough supporting data such patterns can emerge. Conversely, if a TR’s activity is highly pleiotropic, our framework will tend to only prioritize the partners shared across data. That we are able to observe reproducible patterns in this heterogenous collection raises our confidence in applying this framework to specific contexts in future work, such as identifying tissue-specific versus global partners for TRs like PAX6.

Repression is more difficult to infer from coexpression than activation, for reasons we discuss in the Supplemental Material. Similarly, differential interactions are more difficult to characterize than those that are reproducible, requiring evidence of absence. While these considerations motivated our focus on the top reproducible coexpression patterns, the data we have organized can help potentiate the discovery of divergent regulatory interactions. Suresh and colleagues (2023), for example, used single cell coexpression of human and primate data to nominate both conserved and human-novel coexpression patterns. Given that "TR-rewiring" (differential TR activity) is hypothesized to be a primary driver of phenotypic variation, it would be valuable to assess the degree to which differential coexpression between species in matched contexts can reveal distinct regulatory activity.

Numerous methods have been developed for gene regulatory network reconstruction using single cell coexpression, with multiple benchmarks concluding that no algorithm dominates (Chen and Mar, 2018; Pratapa et al., 2020; Nguyen et al., 2021; McCalla et al., 2023). In particular, McCalla and colleagues (2023) emphasized the favorable performance of Pearson’s correlation (as used in this study) relative to more complex models. This aligns with observations by Harris et al., 2021, who found that aggregating single cell coexpression using the computationally efficient Pearson’s correlation provided results that were consistent with alternative similarity metrics (Skinnider et al., 2019). Indeed, we feel that the most important ingredient in the analysis is the aggregation of data because the sparsity of the data is difficult to address otherwise. Our focus on simplistic approaches supports that our conclusions are generalizable to more complex forms of coexpression analysis (Crow et al., 2016).

We believe that the organized information we provide will be a valuable community resource. Beyond lists of genes plausibly regulated by each TR, the interactions identified in this study can assist studies examining the conservation of regulatory interactions, or the chromatin factors commonly coexpressed with each TR. Highly ranked interactions could be used for benchmarking predictive methods, or further dissected towards our understanding of the chromatin and sequence features that are characteristic of reproducible interactions. Future work may find it fruitful to construct context-specific aggregations to contrast against this heterogeneous collection, or to further integrate this resource with other lines of regulation evidence, as we did with the ChIP-seq data.

Data Availability

All summarized rankings, scored ChIP-seq experiments, and GO analysis results are made available as R objects in the Borealis data repository (https://doi.org/10.5683/SP3/HJ1B24). The identifiers and associated data links of the analyzed scRNA-seq experiments are found in Supplemental Table 1 and summaries of the curation benchmark are found in Supplemental Tables 2-3. The code to reproduce the analysis is located at https://github.com/PavlidisLab/TR_singlecell.

Funding

This work was supported by National Institutes of Health grant MH111099 (https://www.nih.gov/) and Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2016-05991 (https://www.nserc-crsng.gc.ca/) and the Canadian Foundation for Innovation Leaders Opportunity Fund held by P.P. A.M. had funding support from the Canadian Institutes of Health Research Canada Graduate Scholarship (CIHR-CGS), Natural Sciences and Engineering Research Council of Canada - Collaborative Research and Training Experience (NSERC-CREATE), and UBC Institute of Mental Health (IMH) Marshall Scholars programs. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interest

The authors have no competing interests to declare.

Acknowledgments

We thank Dr. Marine Louarn, Taylor Kim, Joey He, Wilson Tu, and Alice Ma for their work on expanding the literature curation collection of Chu et al., 2021. We are additionally thankful to Mahan Rafieenaini for his assistance in identifying scRNA-seq datasets. The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org), corresponding to Mathys et al., 2019 (https://www.synapse.org/Synapse:syn18485175; coded ROSMAP in the Metadata).

Author contributions

A.M. conceived and designed the analysis, conducted data collection and analysis, and wrote the manuscript. C.C. contributed to methodological development and data analysis, as well as input on the manuscript. P.P. provided oversight, contributed to study conception and design, and co-wrote the manuscript.

Citations

1. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nature Methods. 2017 Nov;14(11):1083–6.
2. Al-Jaberi N, Lindsay S, Sarma S, Bayatti N, Clowry GJ. The Early Fetal Development of Human Neocortical GABAergic Interneurons. Cerebral Cortex. 2015 Mar;25(3):631–45.
3. Andersson T, Södersten E, Duckworth JK, Cascante A, Fritz N, Sacchetti P, et al. CXXC5 is a novel BMP4-regulated modulator of Wnt signaling in neural stem cells. J Biol Chem. 2009 Feb 6;284(6):3672–81.
4. Aslanpour S, Rosin JM, Balakrishnan A, Klenin N, Blot F, Gradwohl G, et al. Ascl1 is required to specify a subset of ventromedial hypothalamic neurons. Development. 2020 Jan 1;dev.180067.
5. Baliñas-Gavira C, Rodríguez MI, Andrades A, Cuadros M, Álvarez-Pérez JC, Álvarez-Prado ÁF, et al. Frequent mutations in the amino-terminal domain of BCL7A impair its tumor suppressor role in DLBCL. Leukemia. 2020 Oct;34(10):2722–35.
6. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics. 2015 Feb 24;btv118.
7. Ballouz S, Pavlidis P, Gillis J. Using predictive specificity to determine when gene set analysis is biologically meaningful. Nucleic Acids Res. 2017 Feb 28;45(4):e20–e20.
8. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-995.
9. Borromeo MD, Savage TK, Kollipara RK, He M, Augustyn A, Osborne JK, et al. ASCL1 and NEUROD1 Reveal Heterogeneity in Pulmonary Neuroendocrine Tumors and Regulate Distinct Genetic Programs. Cell Rep. 2016 Aug 2;16(5):1259–72.
10. Bravo González-Blas C, De Winter S, Hulselmans G, Hecker N, Matetovici I, Christiaens V, et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. Nat Methods. 2023 Sep;20(9):1355–67.
11. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett. 2004 Aug 27;573(1–3):83–92.
12. Castro DS, Martynoga B, Parras C, Ramesh V, Pacary E, Johnston C, et al. A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets. Genes Dev. 2011 May 1;25(9):930–45.
13. Castro DS, Skowronska-Krawczyk D, Armant O, Donaldson IJ, Parras C, Hunt C, et al. Proneural bHLH and Brn Proteins Coregulate a Neurogenic Program through Cooperative Binding to a Conserved DNA Motif. Developmental Cell. 2006 Dec 1;11(6):831–44.
14. Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinformatics. 2018 19;19(1):232.
15. Chu ECP, Morin A, Chang THC, Nguyen T, Tsai YC, Sharma A, et al. Experiment level curation of transcriptional regulatory interactions in neurodevelopment. PLOS Computational Biology. 2021 Oct 19;17(10):e1009484.
16. Crow M, Gillis J. Co-expression in Single-Cell Analysis: Saving Grace or Original Sin? Trends in Genetics [Internet]. 2018 Aug 23 [cited 2018 Aug 28]; Available from: http://www.sciencedirect.com/science/article/pii/S0168952518301288
17. Crow M, Paul A, Ballouz S, Huang ZJ, Gillis J. Exploiting single-cell expression to characterize co-expression replicability. Genome Biology. 2016;17:101.
18. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. The functional consequences of variation in transcription factor binding. PLoS Genet. 2014 Mar;10(3):e1004226.

19. Deprez M, Zaragosi LE, Truchi M, Becavin C, Ruiz García S, Arguel MJ, et al. A Single-Cell Atlas of the Human Healthy Airways. Am J Respir Crit Care Med. 2020 Dec 15;202(12):1636–45.

20. Domínguez Conde C, Xu C, Jarvis LB, Rainbow DB, Wells SB, Gomes T, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. 2022 May 13;376(6594):eabl5197.

21. Elmentaite R, Kumasaka N, Roberts K, Fleming A, Dann E, King HW, et al. Cells of the human intestinal tract mapped across space and time. Nature. 2021 Sep 9;597(7875):250–5.

22. Emanuele MJ, Enrico TP, Mouery RD, Wasserman D, Nachum S, Tzur A. Complex Cartography: Regulation of E2F Transcription Factors by Cyclin F and Ubiquitin. Trends Cell Biol. 2020 Aug;30(8):640–52.

23. Farahbod M, Pavlidis P. Differential coexpression in human tissues and the confounding effect of mean expression levels. Bioinformatics. 2019 Jan 1;35(1):55–61.

24. Farahbod M, Pavlidis P. Untangling the effects of cellular composition on coexpression analysis. Genome Res. 2020 Jun 24;30(6):gr.256735.119.

25. Fawkner-Corbett D, Antanaviciute A, Parikh K, Jagielowicz M, Gerós AS, Gupta T, et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. Cell. 2021 Feb;184(3):810-826.e23.

26. Ferreira R, Ohneda K, Yamamoto M, Philipsen S. GATA1 function, a paradigm for transcription factors in hematopoiesis. Mol Cell Biol. 2005 Feb;25(4):1215–27.

27. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019 Aug;29(8):1363–75.

28. Garcia-Alonso L, Lorenzi V, Mazzeo CI, Alves-Lopes JP, Roberts K, Sancho-Serra C, et al. Single-cell roadmap of human gonadal development. Nature. 2022 Jul 21;607(7919):540–7.

29. Gitter A, Siegfried Z, Klutstein M, Fornes O, Oliva B, Simon I, et al. Backup in gene regulatory networks explains differences between binding and knockout results. Mol Syst Biol. 2009;5:276.

30. Hamed AA, Kunz DJ, El-Hamamy I, Trinh QM, Subedar OD, Richards LM, et al. A brain precursor atlas reveals the acquisition of developmental-like states in adult cerebral tumours. Nat Commun. 2022 Jul 19;13(1):4178.

31. Harris BD, Crow M, Fischer S, Gillis J. Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain. Cell Systems [Internet]. 2021 May 19 [cited 2021 May 25]; Available from: https://www.sciencedirect.com/science/article/pii/S2405471221001538

32. Henke RM, Meredith DM, Borromeo MD, Savage TK, Johnson JE. Ascl1 and Neurog2 form novel complexes and regulate Delta-like3 (Dll3) expression in the neural tube. Dev Biol. 2009 Apr 15;328(2):529–40.

33. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet. 2023 Aug;24(8):550–72.

34. Hu Y, Comjean A, Rodiger J, Chen W, Kim AR, Qadiri M, et al. FlyRNAi.org 2025 update-expanded resources for new technologies and species. Nucleic Acids Res. 2025 Jan 6;53(D1):D958–65.

35. Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, et al. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics. 2011 Aug 31;12:357.

36. Hu Z, Killion PJ, Iyer VR. Genetic reconstruction of a functional transcriptional regulatory network. Nat Genet. 2007 May;39(5):683–7.

37. Huang HS, Kubish GM, Redmond TM, Turner DL, Thompson RC, Murphy GG, et al. Direct transcriptional induction of Gadd45gamma by Ascl1 during neuronal differentiation. Mol Cell Neurosci. 2010 Jul;44(3):282–96.

38. Jacob J, Storm R, Castro DS, Milton C, Pla P, Guillemot F, et al. Insm1 (IA-1) is an essential component of the regulatory network that specifies monoaminergic neuronal phenotypes in the vertebrate hindbrain. Development. 2009 Jul;136(14):2477–85.

39. Jia S, Wildner H, Birchmeier C. Insm1 controls the differentiation of pulmonary neuroendocrine cells by repressing Hes1. Dev Biol. 2015 Dec 1;408(1):90–8.

40. Kang Y, Patel NR, Shively C, Recio PS, Chen X, Wranik BJ, et al. Dual threshold optimization and network inference reveal convergent evidence from TF binding locations and TF perturbation responses. Genome Res. 2020 Mar;30(3):459–71.

41. Keenan AB, Torre D, Lachmann A, Leong AK, Wojciechowicz ML, Utti V, et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res. 2019 Jul 2;47(W1):W212–24.

42. Kito-Shingaki A, Seta Y, Toyono T, Kataoka S, Kakinoki Y, Yanagawa Y, et al. Expression of GAD67 and Dlx5 in the Taste Buds of Mice Genetically Lacking Mash1. Chemical Senses. 2014 Jun 1;39(5):403–14.

43. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018 Feb 8;172(4):650–65.

44. Lampada A, Taylor V. Notch signaling as a master regulator of adult neurogenesis. Front Neurosci. 2023 Jun 29;17:1179011.

45. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res. 2004;14:1085–94.

46. Lee J, Shah M, Ballouz S, Crow M, Gillis J. CoCoCoNet: conserved and comparative co-expression across a diverse set of species. Nucleic Acids Res. 2020 Jul 2;48(W1):W566–71.

47. Li X, Zheng Y, Hu H, Li X. Integrative analyses shed new light on human ribosomal protein gene regulation. Sci Rep [Internet]. 2016 Jun 27 [cited 2018 Dec 25];6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4921865/

48. Lin H, Zhu X, Chen G, Song L, Gao L, Khand AA, et al. KDM3A-mediated demethylation of histone H3 lysine 9 facilitates the chromatin binding of Neurog2 during neurogenesis. Development. 2017 Oct 15;144(20):3674–85.

49. Liu C, Martins AJ, Lau WW, Rachmaninoff N, Chen J, Imberti L, et al. Time-resolved systems immunology reveals a late juncture linked to fatal COVID-19. Cell. 2021 Apr 1;184(7):1836-1857.e22.

50. Liu J, Gao L, Ji B, Geng R, Chen J, Tao X, et al. BCL7A as a novel prognostic biomarker for glioma patients. J Transl Med. 2021 Aug 6;19(1):335.

51. Liu YH, Tsai JW, Chen JL, Yang WS, Chang PC, Cheng PL, et al. Ascl1 promotes tangential migration and confines migratory routes by induction of Ephb2 in the telencephalon. Sci Rep. 2017 Mar 9;7(1):42895.

52. Lunden JW, Durens M, Phillips AW, Nestor MW. Cortical interneuron function in autism spectrum condition. Pediatr Res. 2019 Jan;85(2):146–54.

53. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nature Methods [Internet]. 2012 [cited 2012 Jul 18]; Available from: http://www.nature.com/nmeth/journal/vaop/ncurrent/abs/nmeth.2016.html

54. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer’s disease. Nature. 2019 Jun 20;570(7761):332–7.

55. Mazurier N, Parain K, Parlier D, Pretto S, Hamdache J, Vernier P, et al. Ascl1 as a novel player in the Ptf1a transcriptional network for GABAergic cell specification in the retina. PLoS One. 2014;9(3):e92113.

56. McCall MN, Illei PB, Halushka MK. Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome. The American Journal of Human Genetics. 2016 Sep 1;99(3):624–35.

57. McCalla SG, Fotuhi Siahpirani A, Li J, Pyne S, Stone M, Periyasamy V, et al. Identifying strengths and weaknesses of methods for computational network inference from single-cell

RNA-seq data. Steinmetz L, editor. G3: Genes, Genomes, Genetics. 2023 Mar 9;13(3):jkad004.

58. Mistry M, Gillis J, Pavlidis P. Meta-analysis of gene coexpression networks in the post-mortem prefrontal cortex of patients with schizophrenia and unaffected controls. BMC Neurosci. 2013 Dec;14(1):105.

59. Monaco G, Van Dam S, Casal Novo Ribeiro JL, Larbi A, De Magalhães JP. A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. BMC Evol Biol. 2015 Dec;15(1):259.

60. Morin A, Chu ECP, Sharma A, Adrian-Hamazaki A, Pavlidis P. Characterizing the targets of transcription regulators by aggregating ChIP-seq and perturbation expression data sets. Genome Res [Internet]. 2023 Jun 12 [cited 2023 Jun 12]; Available from: https://genome.cshlp.org/content/early/2023/06/12/gr.277273.122

61. Müller-Dott S, Tsirvouli E, Vazquez M, Ramirez Flores RO, Badia-I-Mompel P, Fallegger R, et al. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. Nucleic Acids Res. 2023 Nov 10;51(20):10934–49.

62. Nelson BR, Hartman BH, Ray CA, Hayashi T, Bermingham-McDonogh O, Reh TA. Acheate-scute like 1 (Ascl1) is required for normal delta-like (Dll) gene expression and notch signaling during retinal development. Dev Dyn. 2009 Sep;238(9):2163–78.

63. Nguyen H, Tran D, Tran B, Pehlivan B, Nguyen T. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. Briefings in Bioinformatics. 2021 May 20;22(3):bbaa190.

64. Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc Natl Acad Sci USA. 2009 Dec 22;106(51):21521–6.

65. Patel RV, Nahal HK, Breit R, Provart NJ. BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species. Plant J. 2012 Sep;71(6):1038–50.

66. Posner DA, Lee CY, Portet A, Clatworthy MR. Humoral immunity at the brain borders in homeostasis. Current Opinion in Immunology. 2022 Jun;76:102188.

67. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17(2):147–54.

68. Puig RR, Boddie P, Khan A, Castro-Mondragon JA, Mathelier A. UniBind: maps of high-confidence direct TF-DNA interactions across nine species. BMC Genomics. 2021 Dec;22(1):482.

69. Qu S, Fetsch P, Thomas A, Pommier Y, Schrump DS, Miettinen MM, et al. Molecular Subtypes of Primary SCLC Tumors and Their Associations With Neuroendocrine and Therapeutic Markers. J Thorac Oncol. 2022 Jan;17(1):141–53.

70. Ragazzini R, Boeing S, Zanieri L, Green M, D'Agostino G, Bartolovic K, et al. Defining the identity and the niches of epithelial stem cells with highly pleiotropic multilineage potency in the human thymus. Developmental Cell. 2023 Nov;58(22):2428-2446.e9.

71. Rothenberg EV. Causal Gene Regulatory Network Modeling and Genomics: Second-Generation Challenges. Journal of Computational Biology. 2019 Jul 1;26(7):703–18.

72. Russo GL, Sonsalla G, Natarajan P, Breunig CT, Bulli G, Merl-Pham J, et al. CRISPR-Mediated Induction of Neuron-Enriched Mitochondrial Proteins Boosts Direct Glia-to-Neuron Conversion. Cell Stem Cell. 2021 Mar 4;28(3):524-534.e7.

73. Salem S, Salem D, Gros P. Role of IRF8 in immune cells functions, protection against infections, and susceptibility to inflammatory diseases. Hum Genet. 2020 Jun;139(6–7):707–21.

74. Shen WK, Chen SY, Gan ZQ, Zhang YZ, Yue T, Chen MM, et al. AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. Nucleic Acids Research. 2023 Jan 6;51(D1):D39–45.

75. Simão D, Silva MM, Terrasso AP, Arez F, Sousa MFQ, Mehrjardi NZ, et al. Recapitulation of Human Neural Microenvironment Signatures in iPSC-Derived NPC 3D Differentiation. Stem Cell Reports. 2018 Aug;11(2):552–64.

76. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005 Oct 15;21(20):3940–1.

77. Skinnider MA, Squair JW, Foster LJ. Evaluating measures of association for single-cell transcriptomics. Nature Methods. 2019 Apr 8;1.

78. Sonawane AR, Weiss ST, Glass K, Sharma A. Network Medicine in the Age of Biomedical Big Data. Front Genet. 2019;10:294.

79. Suresh H, Crow M, Jorstad N, Hodge R, Lein E, Dobin A, et al. Comparative single-cell transcriptomic analysis of primate brains highlights human-specific regulatory evolution. Nat Ecol Evol. 2023 Sep 4;1–14.

80. Szu J, Wojcinski A, Jiang P, Kesari S. Impact of the Olig Family on Neurodevelopmental Disorders. Front Neurosci. 2021 Mar 30;15:659601.

81. Tamrazi B, Venneti S, Margol A, Hawes D, Cen SY, Nelson M, et al. Pediatric Atypical Teratoid/Rhabdoid Tumors of the Brain: Identification of Metabolic Subgroups Using In Vivo 1H-MR Spectroscopy. AJNR Am J Neuroradiol. 2019 May;40(5):872–7.

82. Tutukova S, Tarabykin V, Hernandez-Miranda LR. The Role of Neurod Genes in Brain Development, Function, and Disease. Front Mol Neurosci. 2021 Jun 9;14:662774.

83. Ueno T, Ito J, Hoshikawa S, Ohori Y, Fujiwara S, Yamamoto S, et al. The identification of transcriptional targets of Ascl1 in oligodendrocyte development. Glia. 2012 Oct;60(10):1495–505.

84. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419.

85. van Lengerich B, Zhan L, Xia D, Chan D, Joy D, Park JI, et al. A TREM2-activating antibody with a blood-brain barrier transport vehicle enhances microglial metabolism in Alzheimer’s disease models. Nat Neurosci. 2023 Mar;26(3):416–29.

86. Wang B, Long JE, Flandin P, Pla R, Waclaw RR, Campbell K, et al. Loss of Gsx1 and Gsx2 function rescues distinct phenotypes in Dlx1/2 mutants. J of Comparative Neurology. 2013 May;521(7):1561–84.

87. Wang S, Sun H, Ma J, Zang C, Wang C, Wang J, et al. Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat Protoc. 2013 Dec;8(12):2502–15.

88. Wang X, He Y, Zhang Q, Ren X, Zhang Z. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. Genomics, Proteomics & Bioinformatics. 2021 Apr;19(2):253–66.

89. Wen JH, Chen YY, Song SJ, Ding J, Gao Y, Hu QK, et al. Paired box 6 (PAX6) regulates glucose metabolism via proinsulin processing mediated by prohormone convertase 1/3 (PC1/3). Diabetologia. 2009 Mar 1;52(3):504–13.

90. Weng Q, Wang J, Wang J, He D, Cheng Z, Zhang F, et al. Single-Cell Transcriptomics Uncovers Glial Progenitor Diversity and Cell Fate Determinants during Development and Gliomagenesis. Cell Stem Cell. 2019 May;24(5):707-723.e8.

91. Werner JM, Gillis J. Preservation of co-expression defines the primary tissue fidelity of human neural organoids. bioRxiv. 2023 Oct 17;2023.03.31.535112.

92. Xiong X, Tu S, Wang J, Luo S, Yan X. CXXC5: A novel regulator and coordinator of TGF-β, BMP and Wnt signaling. J Cell Mol Med. 2019 Feb;23(2):740–9.

93. Yamada Y, Bohnenberger H, Kriegsmann M, Kriegsmann K, Sinn P, Goto N, et al. Tuft cell-like carcinomas: novel cancer subsets present in multiple organs sharing a unique gene expression signature. Br J Cancer. 2022 Nov;127(10):1876–85.

94. Yeung J, Ha TJ, Swanson DJ, Goldowitz D. A Novel and Multivalent Role of Pax6 in Cerebellar Development. J Neurosci. 2016 Aug 31;36(35):9057–69.

95. Yu J, Mu J, Guo Q, Yang L, Zhang J, Liu Z, et al. Transcriptomic profile analysis of mouse neural tube development by RNA-Seq. IUBMB Life. 2017 Jul 10;

96. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014 Nov 20;515(7527):355–64.

97. Zhang W, Girard L, Zhang YA, Haruki T, Papari-Zareei M, Stastny V, et al. Small cell lung cancer tumors and preclinical models display heterogeneity of neuroendocrine phenotypes. Transl Lung Cancer Res. 2018 Feb;7(1):32–49.

98. Zhang Y, Cuerdo J, Halushka MK, McCall MN. The effect of tissue composition on gene co-expression. Brief Bioinform [Internet]. [cited 2019 Dec 9]; Available from: https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbz135/5669861

99. Zhong E, Pareja F, Hanna MG, Jungbluth AA, Rekhtman N, Brogi E. Expression of novel neuroendocrine markers in breast carcinomas: a study of INSM1, ASCL1, and POU2F3. Hum Pathol. 2022 Sep;127:102–11.

100. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018 Oct 3;1.