Identifying Reproducible Transcription Regulator Coexpression Patterns with Single Cell Transcriptomics

Alexander Morin[1,2,3], Chingpan 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, rather than cell type compositional effects that emerge from bulk tissue samples. 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, 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 networks. Our workflow therefore explicitly prioritizes the patterns that are most reproducible across cell types, helping to characterize each TR's coexpression patterns. Towards this goal, we quantify the similarity of TR coexpression profiles within and across species, characterizing the consistency of each TR's coexpression across datasets. 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.

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. Techniques like scRNA-seq reveal heterogeneity within and across populations of cells, enabling the computational clustering and human annotation of cells into discrete biological types. The scale with which scRNA-seq data has been generated has created an unprecedented opportunity to not only study cell types across tissues and conditions, but also to understand the reproducibility of genomic patterns. This is important because, despite its power, scRNA-seq data is sparse owing to both biological and technical factors (Heumos et al., 2023).

Gene regulation is a field that stands to greatly benefit from the single cell era. A primary objective is to learn the repertoire of gene targets that are functionally influenced by proteins that control gene activity, such as transcription regulators (TRs). The ultimate goal is to create a temporal and context-specific map of these interactions across cell types, a task greatly assisted by single cell data. 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, involving a multitude of possible interactions and the diverse molecular modes through which TRs exert their activity (Lambert et al., 2018, Rothenberg 2019).

A traditional and widely-adopted approach for predicting TR-target relationships involves gene coexpression analysis, a method that calculates covariation between TR and gene expression using transcriptomic data, albeit most often from bulk samples. This analysis is often cast as generating a predicted gene regulatory network, where the strength of covariation (such as Pearson's correlation) between genes 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 co-variation 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.

As mentioned above, coexpression inference has relied on “bulk” transcriptomics datasets due to their historical availability. The abundance of such data has facilitated meta-analytic studies, demonstrating the existence of reproducible coexpression patterns across diverse studies (Lee et al., 2004; Ballouz et al., 2015). However, bulk coexpression analysis suffers from a critical confound that complicates the characterization of dynamic processes like gene regulation. Bulk samples typically comprise a mixture of cell types, each varying in proportions across samples. Consequently, variations in gene transcripts are often driven by differences in cell type proportions across samples (McCall et al., 2016; Farahbod and Pavlidis, 2019; Farahbod and Pavlidis, 2020). Coexpression in bulk may then primarily reflect differences in gene expression *between* cell types, rather than capturing a coordinated regulatory process *within* cells of the same type.

The emergence of scRNA-seq has made it possible to address this limitation. By treating cells of a defined type as a population, coordinated expression patterns can be more directly attributed to cellular processes, reducing the influence of varying mixtures of distinct cell types. Single cell coexpression has proliferated just like its bulk counterpart, and is often used conjointly with other data modalities to refine predictions of regulatory interactions (Aibar et al., 2017). However, caution in interpreting single cell coexpression is still warranted due to technical concerns, especially the sparsity of the data. This underscores the utility of meta-analytic approaches that seek to aggregate coexpression networks in search of reproducible interactions (Lee et al., 2004; Mistry et al., 2013; Ballouz et al., 2015).

Correspondingly, the benefits of this meta-analytic framework have been extended to single cell coexpression (Crow et al., 2016; Crow and Gillis, 2018) and further applied to answer biological questions. For example, Harris et al. (2021) aggregated scRNA-seq coexpression networks across multiple datasets to characterize replicable coexpression patterns in neurons. Their focus was on commonly expressed neuronal genes and the preservation of the global network structure, rather than any specific gene coexpression profile. Similarly, Suresh et al. (2023) built aggregate networks using middle temporal gyrus scRNA-seq data from five primate species, demonstrating the conservation of coexpression signals across primates and highlighting human-novel patterns. Lastly, Werner and Gillis (2023) explored the commonalities and differences of single cell coexpression in neural primary versus organoid tissues.

We drew inspiration from these works, as well as 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.

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. The median human dataset had 74,148 cells and 14 cell types; in mouse 36,755 cells and 12 cell types (Fig. 1B). 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.

Given the emphasis on correlation patterns in this work, we imposed the requirement that genes must exhibit non-zero read counts in a minimum of 20 cells within a cell type to be considered “measured.” This condition aims to reduce the calculation of correlations derived from excessively sparse data (Ballouz et al., 2015). Applying this definition of gene measurement, a median human dataset measured 15,341 protein coding genes, and for mouse 13,996 genes (Supplemental Fig. 1). Note that while a gene may be “measured” in a dataset — and thus viable for coexpression analysis — it might be only measured in a single cell type, while other genes are measured in multiple cell types.

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). The result can be conceptualized as a standard gene-by-gene coexpression matrix, save that the elements represent the aggregated strength of pairwise transcript covariation that was calculated separately for each cell type. Notably, unlike in Harris et al. (2021), where information was consolidated across 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 coexpressions in following analyses (Methods). We also show the most consistently negative coexpressed gene pair 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 (Discussion).

Similarity of TR-target profiles

Before prioritizing reproducible TR-gene interactions, we examined the concordance of the TR coexpression profiles across datasets. We expected that distinct profiles generated for the same TR and similar contexts would have elevated similarity relative to mismatched contexts or gene

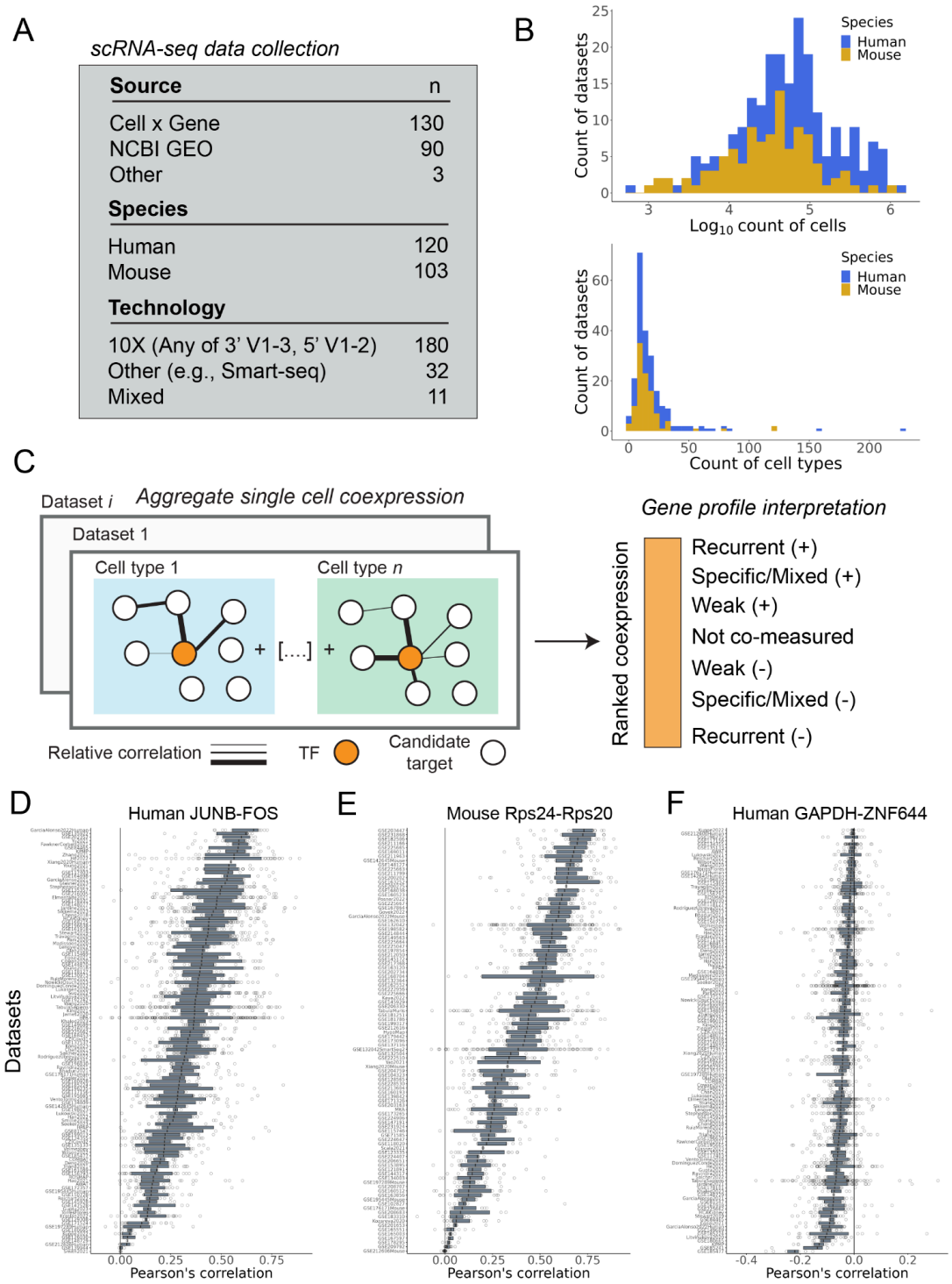


Figure 1. Overview of study design. (A) Counts of datasets by source, technology, and species. (B) Top panel: Counts of cells across the dataset corpus. Bottom panel: Counts of cell types. (C) Schematic of the single cell coexpression aggregation framework and the interpretation of an individual gene coexpression profile. (D, E) Examples of the most reproducible positively coexpresed 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.

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 considered multiple means of determining similarity (Methods), reporting here on the size of the overlap (*K)* of the top positively coexpressed (Top_K) or negatively coexpressed ($Bottom_K$) genes between each pair of gene profiles. 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. Our analysis was restricted to TRs measured in at least five networks.

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} = 177/200) was between *CENPA* profiles from two studies that, while not exactly sharing annotated cell types, both assayed the developing human intestine (Fawkner-Corbett et al., 2021; Elmentaite et al., 2021). The maximal human $Bottom_{200}$ (158/200) belonged to the repressive nuclear receptor *NCOR1* in a comparison between profiles generated from tissue atlas resources (Uhlén et al., 2015; The Tabula Sapiens Consortium et al., 2022). The highest Mouse Top_{200} (150/200) was associated with *E2f8*, derived from two studies of the blood-brain barrier; one of these datasets, in combination with a study on the aging brain, also contributed to *Elk3* having the highest $Bottom_{200}$ score of 149/200 (Kaya et al., 2022; 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.

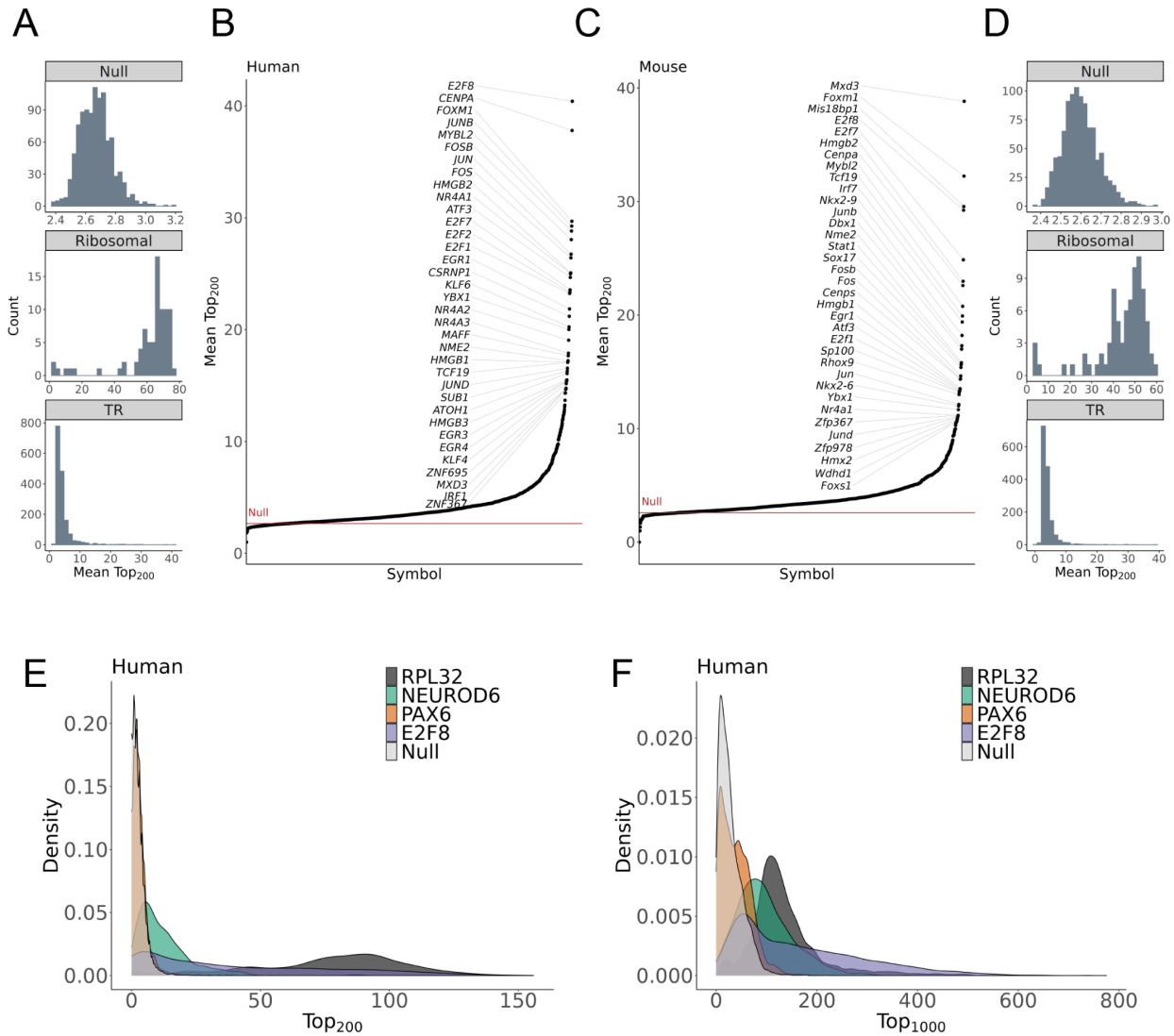


Figure 2. Similarity of TR profiles. (A) Histograms of Top200 overlap (note differences in X axis scale between plots) Top panel: Null generated from 1000 iterations of sampling one TR profile from each of 120 human datasets and calculating the average size of the Top_{200} overlap between every pair of sampled profiles, representing a null background setting. Middle panel: Histogram of the average Top_{200} overlap of all dataset pairs for each of 82 ribosomal genes representing a “best case” scenario. Bottom panel: Histogram of the average Top_{200} overlap of all dataset pairs for 1,606 human TRs. (B) The average Top_{200} overlap of all human TRs, with the red line indicating the average null overlap. (C,D) Same as in A,B, save for 103 mouse experiments and 1,484 TRs. (E,F) The distribution of (E) Top_{200} and (F) Top_{1000} overlaps between every pair of *PAX6* and *NEUROD6* profiles in human, with ribosomal *RPL32*, TR *E2F8*, and a representative null sample included for reference.

In Figs. 2A,D, 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 (red lines in Figs. 2B,C). The ribosomal genes, approximating an empirical "upper bound," averaged 61.1/200 in human and 44.2/200 in mouse. The distribution of average Top_{200} values was highly skewed for TRs, with 87.0% of human TRs and 88.3% of mouse TRs having an average Top_{200} value greater than the average of the null samples. However, for many TRs these differences were relatively small in magnitude. 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 Fig. 2.

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. 2B,C). The E2F family are well characterized regulators of the cell cycle, with E2F8 described as contributing to a negative feedback loop of other E2F members in the later stages of the mitotic cycle (Ly et al., 2017; Emanuele et al., 2020). Given its consistent positive profile, it is perhaps counterintuitive that *E2F8* traditionally has been considered a repressive E2F member (Christensen et al., 2005) — although the delineation between activating and repressive functionality within this family has been questioned (Lv et al., 2017). We re-emphasize that coexpression does not necessarily imply a direct regulatory relationship. However, the consistency and conservation of coexpression partners does suggest shared biological functionality (Lee et al., 2020; Harris et al., 2021).

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 the 1,228 orthologous pairs of TRs analyzed in the mean Top_{200} (*Spearman's correlation* = 0.63) and $Bottom_{200}$ (*Spearman's correlation* = 0.75) lists (Methods; Supplemental Fig. 3). Examples of regulators with consistent negative coexpression profiles in both species include the aforementioned *NCOR1*, the histone methyltransferase-encoding *KMT2C,* nuclear factor *NFAT5*, and the dual RNA and DNA-binding *SON*.

TRs with more context-restricted functions might be expected to exhibit relatively low cross-dataset similarity in our broad corpus, owing to the distinct gene sets that accompany these specialized roles. However, this turns out to not necessarily be the case. We contrast two TRs, PAX6 and NEUROD6, to illustrate. PAX6 is necessary for the development and function of several nervous and pancreatic tissues (Wen et al., 2009; Yeung et al., 2016). In this collection, the mean human PAX6 Top_{200} value was negligibly greater than the average null, marginally improving at K=1000 (Figs. 2E, F). While this does not exclude the existence of recurrent PAX6 targets within these comparisons, the low average overlap may be explained by the distinct expression profiles of PAX6-expressing tissues. In contrast, *NEUROD6,* also a neural TR, had one of the most consistent TR profiles in human (mean Top_{200} rank 46th out of 1,606 TRs), despite being measured in fewer datasets (n = 22/120) than PAX6 (n = 85/120). Thus, while both NEUROD6 and PAX6 can be characterized as specialized regulators, NEUROD6 may have a more consistent profile in this corpus due to information being drawn from a more restricted context.

Lastly, in an attempt to identify predominantly repressive factors, we examined if any TRs had more consistent negative profiles than positive profiles, relative to the respective nulls. Using *K*=200 did not reveal any convincing examples of TRs with such a divergence (Supplemental Fig. 3). However, using *K*=1000 did provide examples of TRs whose negative, but not positive, coexpression profiles exhibited greater overlap than shuffled profiles. A leading example in human was the Ikaros zinc finger *IKZF5,* a hematopoietic TR implicated in platelet deficiency and which is reported to be understudied relative to other Ikaros genes (Lentaigne et al., 2019).

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), and then, for each TR, aggregating their profiles *across* datasets.

We explored multiple summarization strategies, opting for a straightforward approach that reduced the influence of data sparsity. We ultimately averaged TR profiles across networks, applying a minor imputation step for non-measured gene pairs to standardize their “missing” status across profiles with variable gene measurement (Methods). 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. We also considered an alternative prioritization scheme that assigned each TR-gene pair the best rank achieved in any single network, to emphasize specific interactions (Morin et al., 2023). However, we found that this approach rewarded outlier behavior that was technical in nature. Nevertheless, we include these rankings in our final summarizations.

As before, we used the set of 82 L/S ribosomal genes as a “sanity check” of our workflow. We created an aggregate profile for each ribosomal gene, and then examined their top 82 ranked coexpressed gene partners (Supplemental Fig. 4). In both mouse and human, we found that, on median, 71 out of the 82 top ranked partners were other L/S ribosomal genes. Note that these values are slightly depressed by our choice to limit the analysis to one-to-one orthologues; in most cases, additional ribosomal proteins are in the top 82. This validates the prioritization of known biological coexpression. The notable exceptions fortuitously illustrate a point about tissue-specific patterns. One was human *RPL39L*, a paralog of *RPL39,* which has been described for its involvement in alternative ribosomal activity in the testes and spermatogenesis (Li et al., 2022). We observed that the top ranked *RPL39L* coexpressed partner was *PBK*, also implicated in testes function and spermatogenesis (Miki et al., 2020). Similarly, *RPL3L*, a heart and skeletal muscle-specific paralog of *RPL3* (Shiraishi et al., 2023) had only 2 out of 82 of its top partners belonging to the L/S ribosomal set. Notably, the cardiac myosin genes *MYH7* and *MYL2* ranked 5th and 6th, respectively, in *RPL3L’s* aggregate ranking of coexpressed partners. These results show that while robust context-independent coexpression patterns can be readily observed, context-specific patterns can also be discovered in our data.

Recovery of literature-curated TR-target interactions

Equipped with a unified single cell coexpression profile for each human and mouse TR, we wanted to assess the concordance of these rankings with an orthogonal line of regulation evidence. We reasoned that, 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 (Methods). 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).

There are two important considerations to this benchmark. First, there is an imbalance of curated targets (positive labels) for each TR, coupled with the absence of a definite set of negative interactions — thus all genes lacking curation are treated as negatives. Second, the precision-recall (AUPRC) and receiver operating characteristic (AUROC) values, while typically better than random, are also small in magnitude (Fig. 3B). This outcome is primarily due to the incomplete nature of the literature curation corpus (De Smet and Marchal, 2010), and the inherent complexity of benchmarking gene regulation, where no single line of evidence is exhaustively performant (Garcia-Alonso et al, 2019). Hence, our focus is on evaluating performance relative to a null. And so while not without caveats (Morin et al., 2023), this benchmark provides a relative sense of the regulation information contained in the aggregate profiles.

We first examined the effectiveness of the aggregate profiles in recovering curated targets relative to the individual TR profiles that compose the aggregate. We treated the AUCs obtained from the individual profiles as a distribution, and computed the observed AUC quantile of the aggregate profile (*Quant_aggregate)*. A *Quant_aggregate* value of 1 indicates that the aggregate ranking outperformed (better prioritized curated targets) every individual TR profile. On average, the aggregate profiles moderately outperformed the expected AUC value from an individual profile (Supplemental Fig. S5). The median AUROC *Quant_aggregate* of human TRs was 0.76, with 116/451 (25.7%) having values above 0.9. In mouse, these respective values were 0.78 and 150/434 (34.6%). 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)*. 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. 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,

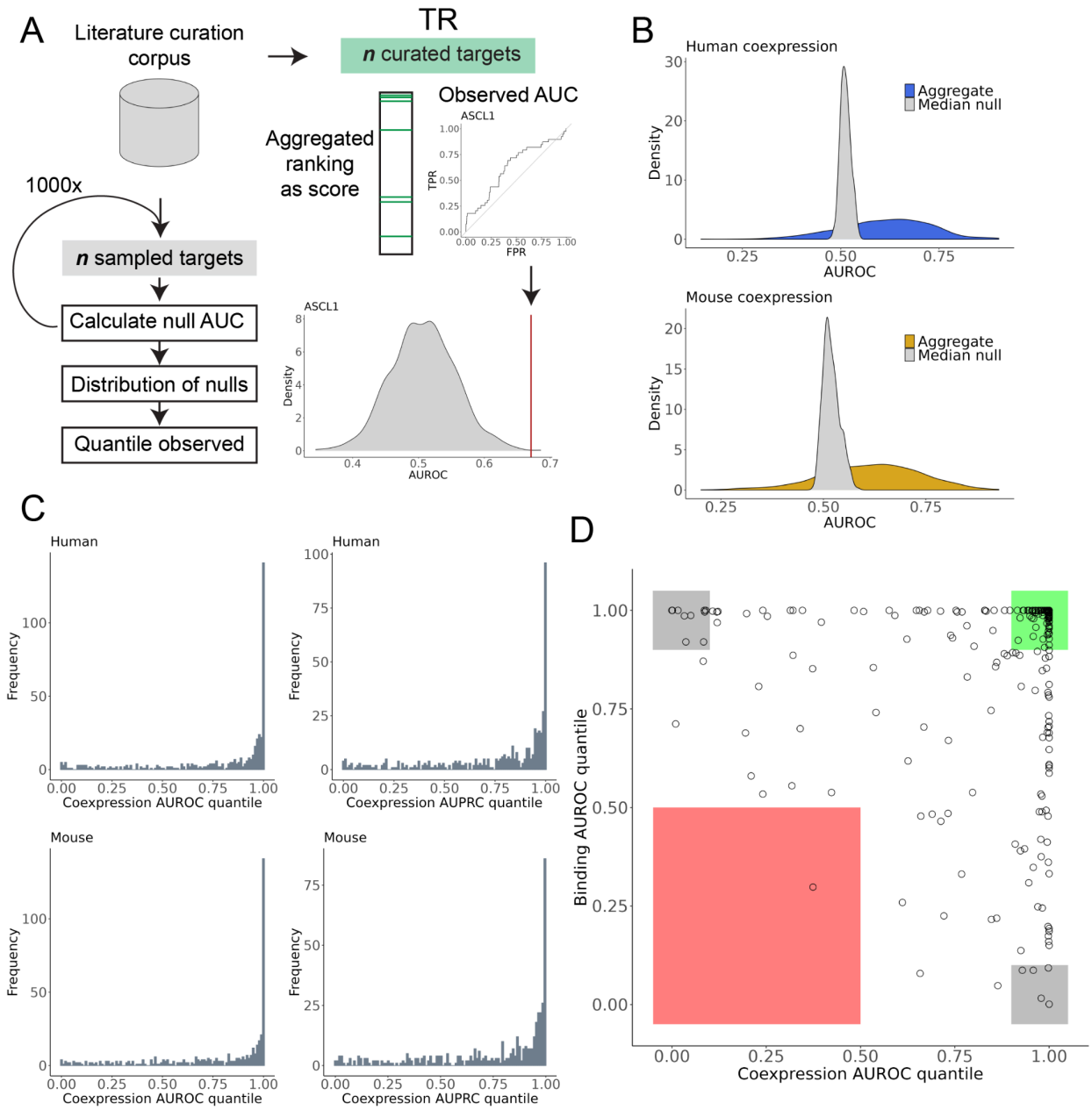


Figure 3. Recovery of literature curated targets by aggregate rankings. (A) Schematic of literature curation evaluation. (B) Distributions of the observed AUROCs for 451 human and 434 mouse aggregate TR coexpression profiles, along with the distribution of the median null AUROCs generated for each profile. (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.

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 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, would outperform coexpression. However, in our hands the aggregate binding evidence was on par with coexpression in its ability to predict known targets (Supplemental Fig. 5).

Among TRs with both binding and coexpression data, a notable fraction were effective in the benchmark for both data types, as demonstrated for human TRs in Fig. 3D. In human, 131 of 253 (51.8%) TRs had AUCs (AUPRC or AUROC) *Quant_binding* > 0.9 and *Quant_coexpression* > 0.9; in mouse 129 of 241 (53.5%). 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 85 of 131, mouse 85 of 129), 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, 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 0.17 in human and mouse, respectively. Focusing on human, this corresponded to the coexpression aggregate having 7 of 91 LEF1 curated targets within the top 500 of its ranking, while the binding aggregate had only two curated targets at this same cut-off. Conversely, examples of TRs whose binding, but not coexpression, profiles were performant in both species include ARNTL, CLOCK, CREB1, MEF2D, MYC, and MYCN, and YY1.

Lastly, we re-evaluated the coexpression benchmark, reversing the ranks to prioritize negative coexpression for scoring. Our goal was to identify any TRs that performed poorly using positive coexpression (AUCs *Quant_coexpression* < 0.5) but displayed improved performance using negative coexpression (*Quant_reverse* > 0.9), with the aim of uncovering repressive interactions. These reversed rankings were much less performant as a group (Supplemental Fig. 5), consistent with the observation that the literature corpus is enriched for activating interactions (Chu et al., 2021). We did, however, identify 26 TRs in human and 22 in mouse that exhibited enhanced performance upon reversing the rankings, with 9 of these factors occurring in both species. Moreover, 13 of these 26 human TRs and 8 of the 22 mouse TRs also had performant binding aggregations (*Quant_binding* > 0.9), and three were common to both species: MAX, NRF1, and YY1. Collectively, these observations may point to TRs with more readily identifiable repressive activity: their negative, but not positive, coexpression profiles are able to retrieve their curated targets relative to sampled targets, and these curated targets are also associated with elevated binding signal in the respective TR ChIP-seq experiments.

Identification of orthologous 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). Coexpression analysis provides an attractive means to uncover common and divergent interactions. Prior studies have explored the preservation of coexpression network structure between mouse and human (Monaco et al., 2015) and across 14 species using bulk data (Lee et al., 2020), as well as investigated common and human-novel single cell coexpression signals across primate brains (Suresh et al., 2023). Here, our specific aim was to identify the extent to which individual TR aggregate coexpression profiles were preserved between mouse and human. In the following analyses, we subset each species' set of rankings to include only 16,698 one-to-one

orthologous protein coding genes between mouse and human (Methods), encompassing 1,241 orthologous TRs.

The distribution of Spearman correlations between the orthologous rankings indeed suggests a degree of preservation (Fig. 4A). The median correlation was 0.71, albeit with appreciable spread. While there are TRs with low correlation 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. 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 (Patel et al., 2012; Suresh et al., 2023) and illustrated in Fig. 4B. For each TR and species, we selected the top 200 coexpressed partners (Top_{200}). We next calculated the overlap of this gene set with the Top_{200} gene set of every TR in the other species, treated the mismatched TR overlaps as a distribution, then determined the quantile of the observed Top_{200} for the matched ortholog TR. This procedure was then repeated for the reciprocal species. The result is a pair of relative rankings for each orthologous TR: how well a human TR's ranking recovered its mouse ortholog relative to all other mouse TRs (*Quant_human*), and the recovery of the mouse ranking across human TRs (*Quant_mouse*).

As demonstrated in Fig. 4C, there was considerable preservation of single cell aggregate TR coexpression profiles between mouse and human. The median Top_{200} *Quant_human* was 0.971, with 174/1,241 (14.0%) TRs having a perfect value of 1; in mouse these values were 0.976 and 171/1,241 (13.8%), respectively. These relative values correspond to a median Top_{200} overlap of15 genes, with CENPA and HMGB2 each having a maximal Top_{200} of 154 genes (Fig. 4D). Logically, many of these highly preserved TRs also had similar profiles within species, and those that were weakly preserved generally lacked consistency within species (Fig. 2; Supplemental Fig. 6). These findings collectively contribute to characterizing the extent to which each TR can be defined by a set of coexpressed gene partners, facilitating inferences into their biological roles. While the most preserved TRs were led by regulators of housekeeping processes such as cell division, we also observe this preservation among more specific TRs, such as NEUROD6 (Quant_human and Quant_mouse = 1, Top_{200} = 51).

In Fig. 4E we illustrate this overlap procedure for ASCL1, another context-restricted TR that showed agreement between species. Of the 200 genes that were most consistently

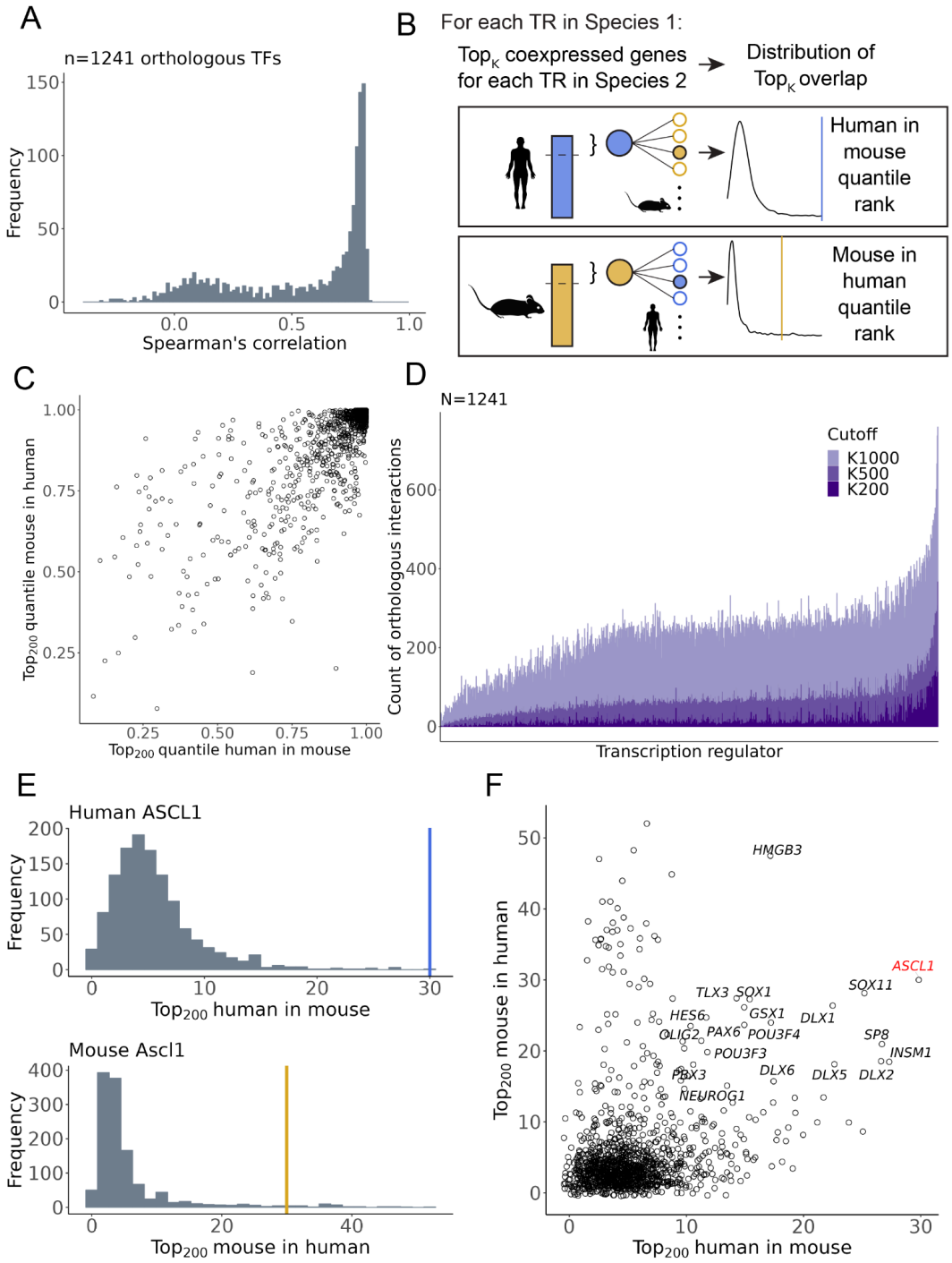


Figure 4. Preservation of mouse and human single cell coexpression profiles. (A) Distribution of Spearman's correlations between the aggregate single cell coexpression profiles of 1,241 orthologous TRs. (B) Schematic of the *Top$_K$* workflow, adapted from Suresh et al., 2023. (C) Scatterplot of the quantiles of each TR's observed *Top$_{200}$* overlap with its ortholog, relative to all other TRs. (D) Count of the *Top$_K$* overlaps across three cut-offs. (E) Distributions of *Top$_{200}$* overlaps of human ASCL1 across mouse TRs (top panel) and mouse Ascl1 across human TRs (bottom panel), with bars indicating ASCL1's position in the cross-species comparison. (F) Scatter plot of the ASCL1 *Top$_{200}$* overlaps.

coexpressed with human ASCL1, 30 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, resulting in Quant_human = 1. In the reciprocal comparison, where mouse Ascl1 was queried against all human TRs, human ASCL1 ranked 35st (Quant_mouse = 0.97). The 34 human TRs with a greater overlap with mouse Ascl1 did not have a sizable overlap in the reciprocal comparison, save for HMGB3 (Fig. 4F). 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).

Examination of the $Bottom_{200}$ also suggested a degree of preservation, albeit attenuated compared to the reproducible positive interactions (median $Bottom_{200}$ *Quant_human* = 0.88; *Quant_mouse* = 0.93; Supplemental Fig. 6). Consistent with our prior comparisons (Supplemental Fig. 2D), we observed a higher magnitude of baseline overlap in the bottom of the lists (median $Bottom_{200}$ = 30). While this may in part be a byproduct of our analysis approach, 21 TRs still achieved a perfect $Bottom_{200}$ quantile for both species, with ZNF532 having the largest overlap of this set ($Bottom_{200}$ = 108). A further five TRs — SOX17, SOX18, SUB1, THRA, and GLMP — also achieved a perfect Top_{200} quantile between species.

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 co-variation. 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; 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.

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:

> **1. Stringent**: Required a gene's presence in the Top 500 of both coexpression and binding in both species.
>
> **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.
>
> **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.

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 541 TR-gene pairs corresponding to 102 TRs and 355 unique genes. These genes were unevenly distributed among TRs, with 82 of 355 occurring in more than one TR's Stringent collection. For example, three genes — TRIB1, TNFAIP3, and MCL1 — each appeared across nine TR's Stringent sets: ATF3, CEBPB, CEBPD, FOS, FOSL1, FOSL2, JUN, JUNB, JUND, NFKB1, BHLHE40, IRF4, and REL. This suggests that these three genes are common terminal end points of AP-1 functionality.

Consistent with these observations, the TRs with the largest Stringent collection featured multiple AP-1 members, led by FOSL1 with 30 genes, along with immunity TRs such as STAT1, STAT2, and IRF1. More specialized TRs also had among the largest Stringent sets, such as the

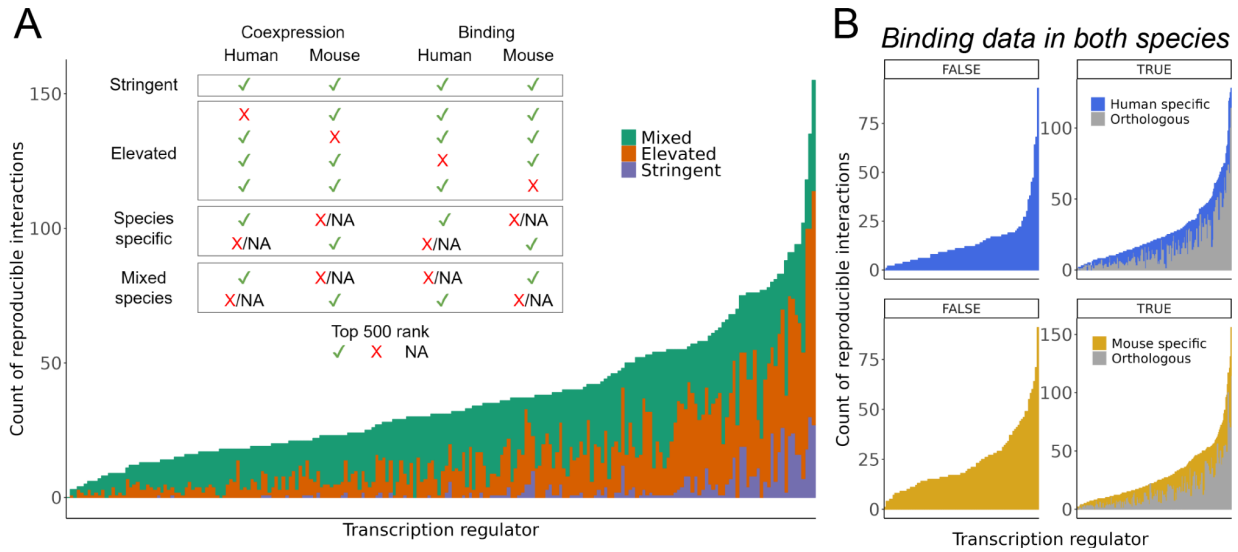


Figure 5. Count of interactions supported across methods and species. (A) Inset: criteria used to group interactions into tiers. Bar chart: Count of unique interactions gained in each orthologous tier (Stringent, Elevated, and Mixed-Species) for the 216 TRs with binding data in both species. (B) Count 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.

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 highly influenced by the limited coverage of ChIP-seq data across biological contexts (see Discussion).

The Elevated collection relaxes the criteria to allow orthologous genes reaching the cut-off in three of the four rankings. This resulted in 3,165 Elevated TR-gene pairs, with 211 of the 216 available TRs having at least one gene in its 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. MAFB, another TR implicated in hematopoiesis, had the largest Elevated set (n = 29) of TRs lacking any genes in their Stringent collection.

Of these 29 Elevated MAFB genes, 24 had a Top 500 coexpression ranking in both species and binding evidence in mouse, but not human (Supplemental Fig. 7). Further, none of these 24 genes had a human binding ranking that was particularly close to the cut-off. Determining if this signifies true species-specific differential binding of MAFB among these genes — while still preserving their correlation of transcripts — or reflects imbalanced ChIP-seq coverage between species is beyond the scope of this study. Nevertheless, this led us to examine if a single data type or species was more frequently the “absent” ranking across the Elevated collection (Supplemental Fig. 7). We generally see a split between species, while the binding rankings tended to be more frequently absent.

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 101 TRs with a median of 11 interactions. TFDP1 led with 93 genes being 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. Correspondingly, this set contained numerous immunity-associated genes, with the highest rank products belonging to Mpeg1, Ctss, Cd180, Xcr1, and Trim30a.

For TRs with ChIP-seq data in both species, genes that are gained beyond the Stringent and Elevated levels (right panels Fig. 5B) could either make the reporting cut-offs exclusively in one species, or lack a one-to-one orthologous match between species. Regarding the latter, our

findings indicate that genes added in the Species-specific collections were seldom due to their lack of an orthologous match, save for a few immune TRs (Supplemental Fig. 7). It is therefore possible that the interactions with highly differential rankings between species may be enriched for candidates of evolutionary-divergent gene regulation. However, this comparison is greatly complicated by the uneven biological coverage of the binding data between species, and in this study we prioritized reporting on the reproducible interactions.

Highlighting ASCL1

We conclude by focusing on ASCL1, an essential pioneer nervous system regulator that is also studied for its involvement in cancer. We emphasize 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. We further note that ASCL1 will be variably co-measured (and thus eligible for calculating coexpression) with other genes among these datasets. For example, there were 2,698 genes that were co-measured in 30 or fewer datasets with human ASCL1, and 7,362 genes that were co-measured in all 61 datasets. 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 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,

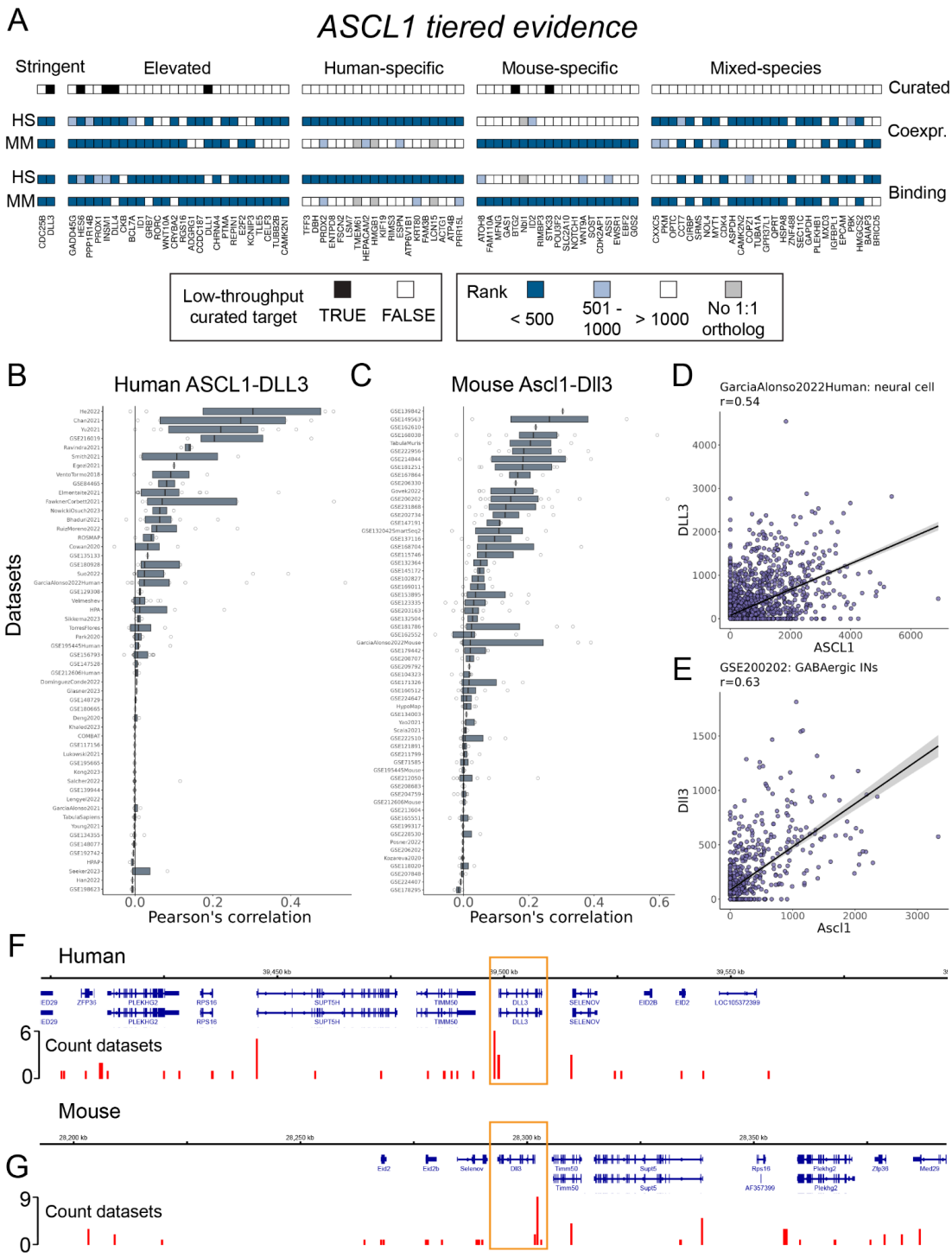


Figure 6. Reproducible ASCL1 interactions. (A) Heatmap representing the tiered evidence for ASCL1 candidate targets. (B, C) Distribution of Pearson's correlations for ASCL1-DLL3 in (B) human and (C) mouse, as in Fig. 1E-G. (D, E) Scatterplot of the CPM values for ASCL1 and DLL3 for the cells belonging to the cell type that had the highest correlation in the entire corpus for (D) human and (E) mouse. (F, G) IGV track plots centered on DLL3 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.

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 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; Casto et al., 2011; Lampada and Taylor, 2023), these collective observations support the ability of our resource to 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 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, or at least help identify other regulatory factors that commonly participate in the same regulatory networks. Our findings additionally provide insight into the TRs whose activity is more challenging to uncover, given the considered genomics data.

Our workflow explicitly prioritizes the interactions that are most common across contexts, akin to our prior study (Morin et al., 2023). This comes with notable caveats: the data coverage of biological contexts, the variability of a TR's functional gene targets across these contexts, and the ability of transcriptomics to identify the dynamics of these interactions. We also do not equate reproducibility as an absolute metric of biological importance. It is possible that a developmentally critical interaction has a positive association in one cell type and a negative association in another, which would tend to be de-prioritized by aggregation. Furthermore, we emphasize that reproducibility does not imply universality, as exemplified by the ASCL1-DLL3 interaction (Figs. 6B,C).

It is not surprising that the most reproducible interactions tend to correspond to fundamental cellular processes. This may not solely be a result of the breadth of the contributing data: prior work has shown that highly expressed genes tend to have stronger coexpression (Crow et al., 2016). Additionally, it is logical that the dynamics of processes like the cell cycle are more readily captured by changing transcript levels. What is perhaps surprising is that we still find evidence for highly context-specific interactions. This is still, however, sensible: if a TR's expression is highly restricted to one context, then our aggregation framework will only draw information from that context, and as long as there is enough supporting data such patterns can emerge. Conversely, if a TR's activity is highly pleiotropic, our framework will only prioritize the targets shared across data.

Repression is difficult to infer with coexpression for multiple reasons. Consistent with our prior observations in bulk data (Lee et al., 2004), negative correlations tend to be smaller in magnitude and less consistent than positive values (Fig. 1F). A predicted positive interaction means that a TR and gene are both detected across samples, but lack of coexpression can occur when a gene is present and the TR absent, necessitating a differential expression framework to ensure the TR's presence. Additionally, it is possible that a TR strongly represses

a target, resulting in the target having zero expression and thus no variation to calculate coexpression. 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 evolutionary divergent regulatory interactions.

Single cell coexpression analysis has received comparatively less attention than differential expression analysis in the literature (Zhang et al., 2021), so there is less consensus on how best to conduct it. SCENIC (Aibar et al., 2017) has become a popular method for single cell coexpression. However, its objective — partitioning cells into groups based on shared TR coexpression and motif enrichment — differs from our goal of identifying reproducible TR coexpression partners across predefined cell types ("TR-centric"). Skinnider and coauthors (2019) benchmarked similarity metrics in single cell coexpression analysis, highlighting the performance of metrics of proportionality. However, as noted by Harris et al., 2021, these performances were often similar across the board, and in their hands the computationally efficient Pearson's correlation (as used in this study) resulted in aggregate networks consistent with those constructed using measures of proportionality. 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. Nevertheless, a comprehensive assessment of the stability of reproducible interactions across various metrics and data normalization procedures, although computationally demanding, could be worthwhile.

While this sparsity affects every level of our handling of the data, we highlight its interaction with the aggregation strategy. A prior application of this approach focused on networks comprising well-measured genes in a specific cell type (Harris et al., 2021). In contrast, our goal was to compare global patterns, prompting us to retain the full complement of protein coding genes. This necessitated two imputation steps, which we emphasize were done not to interpolate missing observations, but to better reflect their "missingness" during the ranking procedures (Methods). While these steps alleviated the positive relationship between a gene pair's sparsity and its aggregate ranking, further improvements may involve selecting a prior that more effectively mitigates the impact of uneven gene coverage across datasets.

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 features that are characteristic of reproducible interactions. Following 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.

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 (V8; Hu et al. 2011), keeping only genes with a score of at least five that were also reciprocally the best score between mouse and human and excluding genes with more than one match. 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, version 3

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. When distinct collections contained an overlapping set of cells (i.e., an integration of experiments) we excluded the identical cell IDs found in the later collection. 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 note that the Cell x Gene datasets typically had already undergone filtering that was at least as stringent as these requirements. All gene count matrices were fixed to the RefSeq Select protein coding genes. 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 normalized gene by cell count 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 and rank-standardized into the range [0, 1]. 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 coexpression is not calculated from overly-sparse populations. WGCNA (Langfelder and Horvath, 2012; V1.72-1) was used in Step 3 for its efficient correlation implementation. Exploratory analyses using Spearman's correlation instead of Pearson's gave similar results (not shown). The zero imputation in Step 4 is to ensure the ranking procedure includes non-measured genes, placing them in between positive and negative correlations.

Gene profile similarity

Coexpression profiles may not have a full complement of measured genes, and thus contain tied ranks corresponding to missing values. Consequently, metrics of similarity that compare all of two lists, such as Spearman's correlation, are inappropriate. We calculated set overlap metrics between lists, Jaccard values after binarizing gene lists by both the Top_K and $Bottom_K$ status, as well as AUC metrics, where one list is treated as a score and the Top_K genes of the other list as labels. While there was agreement between these metrics of similarity, we favoured the interpretability of reporting the size of the Top_K and $Bottom_K$ overlaps.

Aggregating TR profiles and the effect of gene measurement sparsity

We also explored various methods of aggregating TR profiles from distinct networks into a single ordered list, and again found that they largely agreed. Opting for simplicity, we averaged the rank-standardized values and used the resulting ordering as a TR's aggregate profile. Notably, we observed a correlation between a gene's position in the aggregate and the count of times that gene was measured, regardless of the aggregation method used. Because each profile had variable measurement, there was variable delineation between the positive coexpression values, the non-measured gene pair ties, and negative coexpression values. Therefore, for a given TR's set of profiles, we imputed all tied values to the median rank-standardized value across the profiles before averaging. This imputation standardized the values of missing values and alleviated the relationship between a gene's aggregate position and its measurement count.

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 (TRRUST: Han et al., 2018; InnateDB: Lynn et al., 2008; TFactS: Essaghir et al., 2010; TFe: Yusuf et al., 2012; HTRIdb: Bovolenta et al., 2012; CytReg: Carrasco Pro et al., 2018; ORegAnno: Lesurf et al., 2016; ENdb: Bai et al., 2019), and then significantly expanded upon neurologically-relevant TRs. Since this publication, we have further expanded this collection, to a total of 27,627 interactions encompassing 773 TRs and 5,899 gene targets. For all analyses, we considered the presence of any experiment (e.g., EMSAs, reporter assays) in this collection supporting an interaction as a positive label. If a target gene was in the orthologous gene set, we allowed it to be counted as a label for either species. We then used each TR's aggregate profile's ranking as a score and its curated targets as labels, calculating AUC metrics using the ROCR package (Sing et al., 2005; V1.0-11). We generated a null distribution of AUC values for each aggregate TR profile, randomly sampling from the entire literature curation corpus a number of targets equal to the count of curated targets for the given TR, repeating this process 1000 times.

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), based on the exponential decay function introduced by Ouyang et al. (2009):

$$S_g = \sum_{k=1}^{K} e^{-\frac{d_k}{d_0}}$$

Where S is the binding score for a gene (g) in one TR experiment, K is the number of peak summits within 1Mbp of the gene TSS, d_k represents the absolute distance in bps between the TSS and the peak summit, and d_0 is the decay constant, set to 5,000 as in the original publication. Unibind ChIP-seq experiments may be "duplicated," corresponding to TRs matched to multiple binding motifs. In such cases, we averaged the duplicated binding scores. To alleviate batch/technical considerations, we bound all de-duplicated gene binding vectors into a gene by experiment matrix, added 1, and applied a $\log_{10}$ transformation followed by quantile normalization (preprocessCore R package version 1.48). This process was done separately for each species' experiments. To generate an aggregate binding profile, we averaged the gene binding vectors specific to each TR.

ASCL1 binding region analysis

GenomicRanges (Lawrence et al., 2013; V1.50.2) was used for all analyses. ASCL1 had duplicated ChIP-seq datasets in Unibind (see above), and so we took the union of peaks for datasets with the same experiment ID. All peaks were resized to ~300 base pairs, and a "consensus" list of bound regions was generated for all discrete bound regions across datasets. We calculated the count of individual datasets that overlapped this consensus set, and plotted these counts using igvR (Shannon 2023; V.1.22).

Competing interest

The authors have no competing interests to declare.

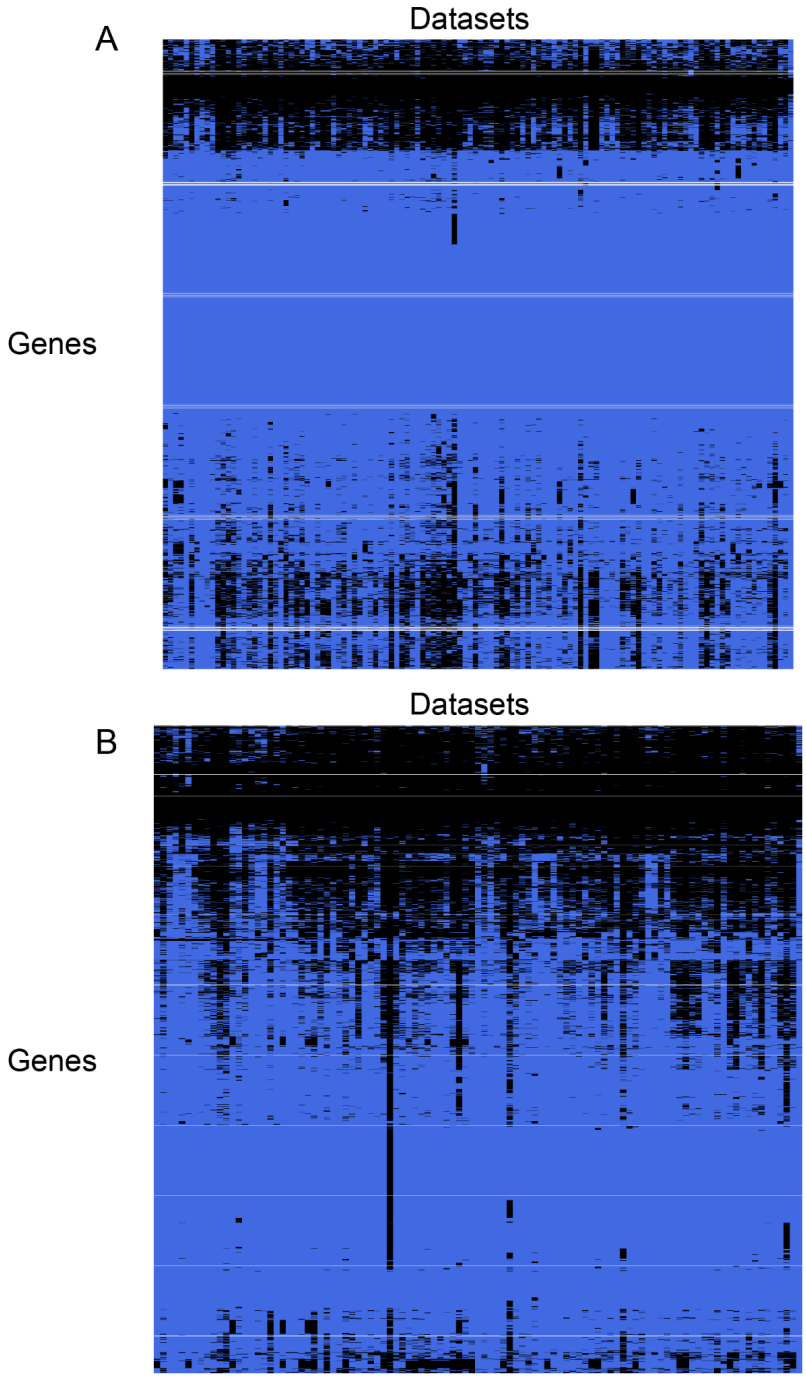
Acknowledgments

We would like to thank Dr. Marine Louarn, Taylor Kim, Joey He, Wilson Tu, and Alice Ma for their work on expanding the literature curation collection (Chu et al., 2021). We are additionally thankful to Mahan Rafieenaini for his assistance in identifying scRNA-seq datasets. We are also grateful to the Religious Orders Study and Memory and Aging Project (ROSMAP) for the use of the following dataset: https://doi.org/10.7303/syn2580853 (coded ROSMAP in the metadata).

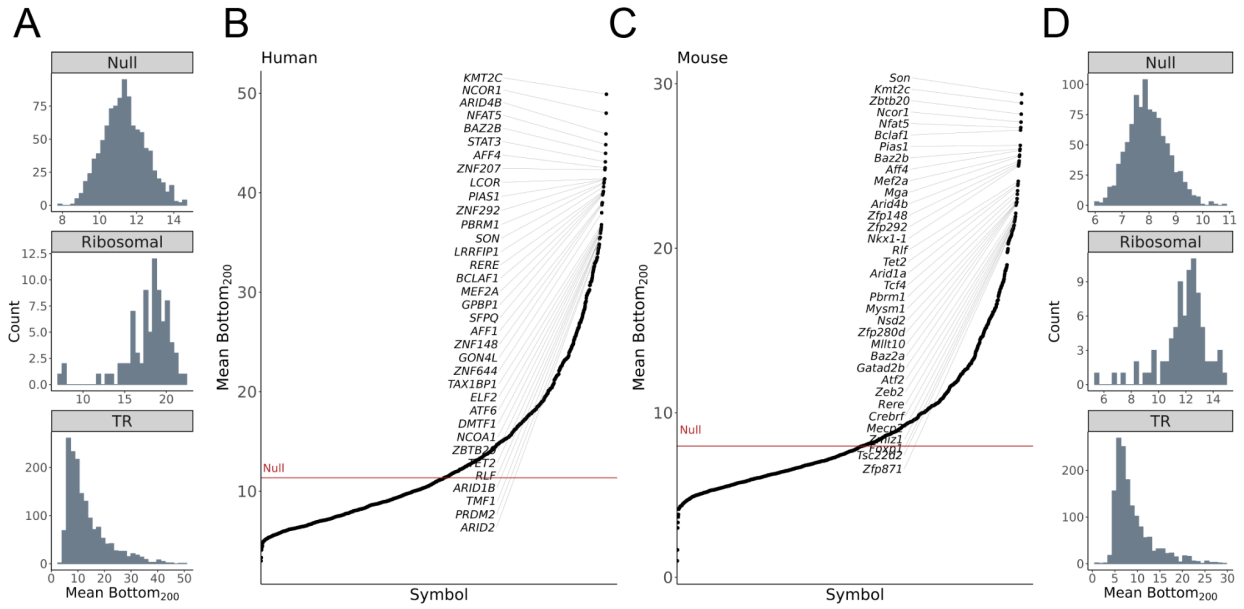
Author contributions

A.M. conceived and designed the analysis, conducted data collection and analysis, and wrote the paper. C.C. contributed to methodological development and data analysis, as well as provided editorial support. P.P. conceived and designed the analysis, provided data analysis support, and made substantial contributions to editing the paper.

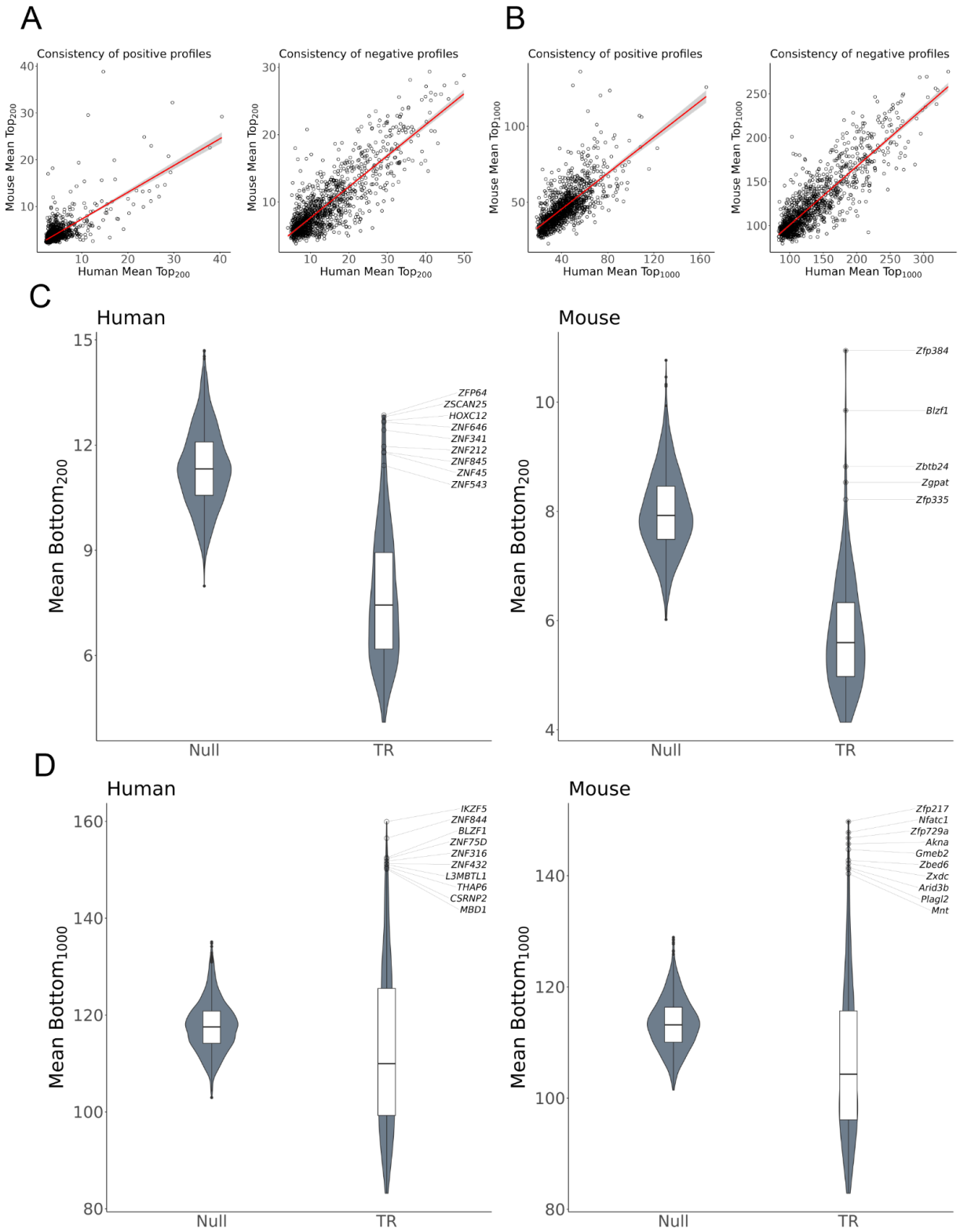
Supplemental Figures



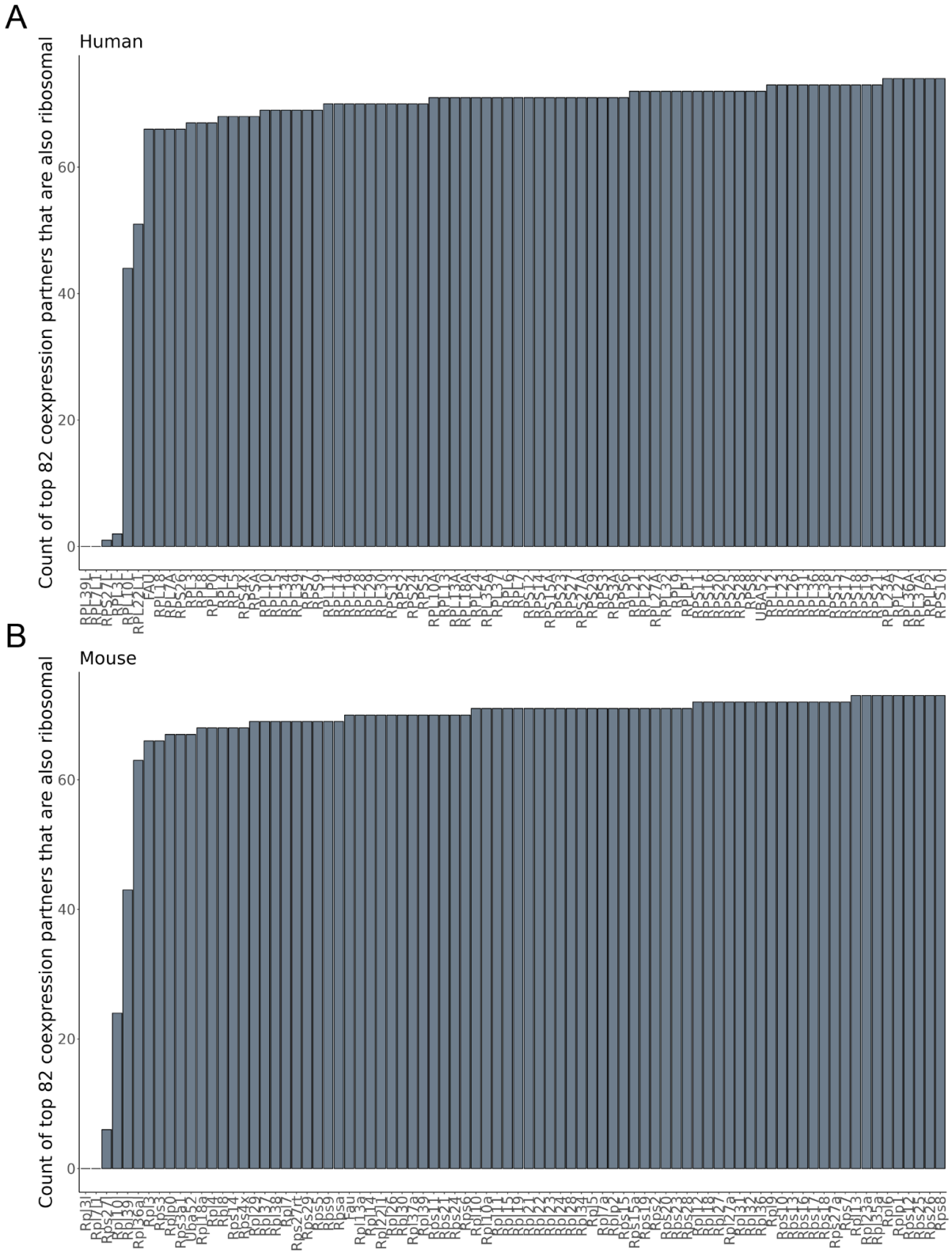
Supplemental Figure 1. Gene measurement coverage. (A) Binary heatmap indicating whether (blue) or not (black) a gene had non-zero counts in at least 20 cells in at least one cell type in a dataset, for 19,213 human protein coding genes and 120 datasets. (B) Mouse: 20,971 protein coding genes and 103 experiments.



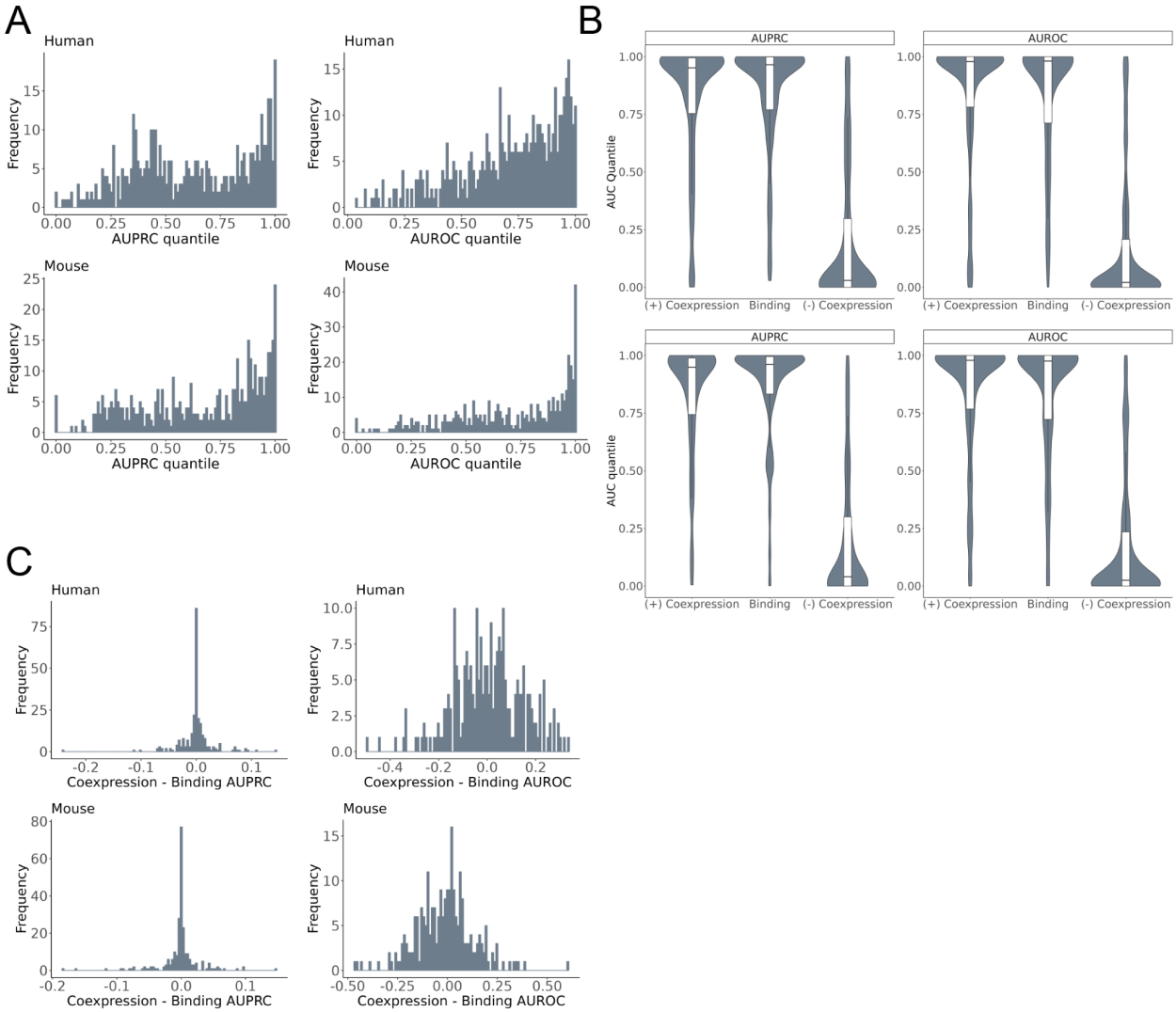
Supplemental Figure 2. Similarity of TR profiles. (A) Top panel: Histogram of 1000 iterations of sampling one TR profile from each of 120 human datasets and calculating the average size of the Bottom_{200} overlap between every pair of sampled profiles, representing a null background setting. Middle panel: Histogram of the average Bottom_{200} overlap of all dataset pairs for each of 82 ribosomal genes representing a “best case” scenario. Bottom panel: Histogram of the average Bottom_{200} overlap of all dataset pairs for 1,606 human TRs. Note that all three panels have different axis scales. (B) The average Bottom_{200} overlap of all human TRs, with the red line indicating the average null overlap. (C,D) Same as in A,B, save for 103 mouse experiments and 1,484 TRs.



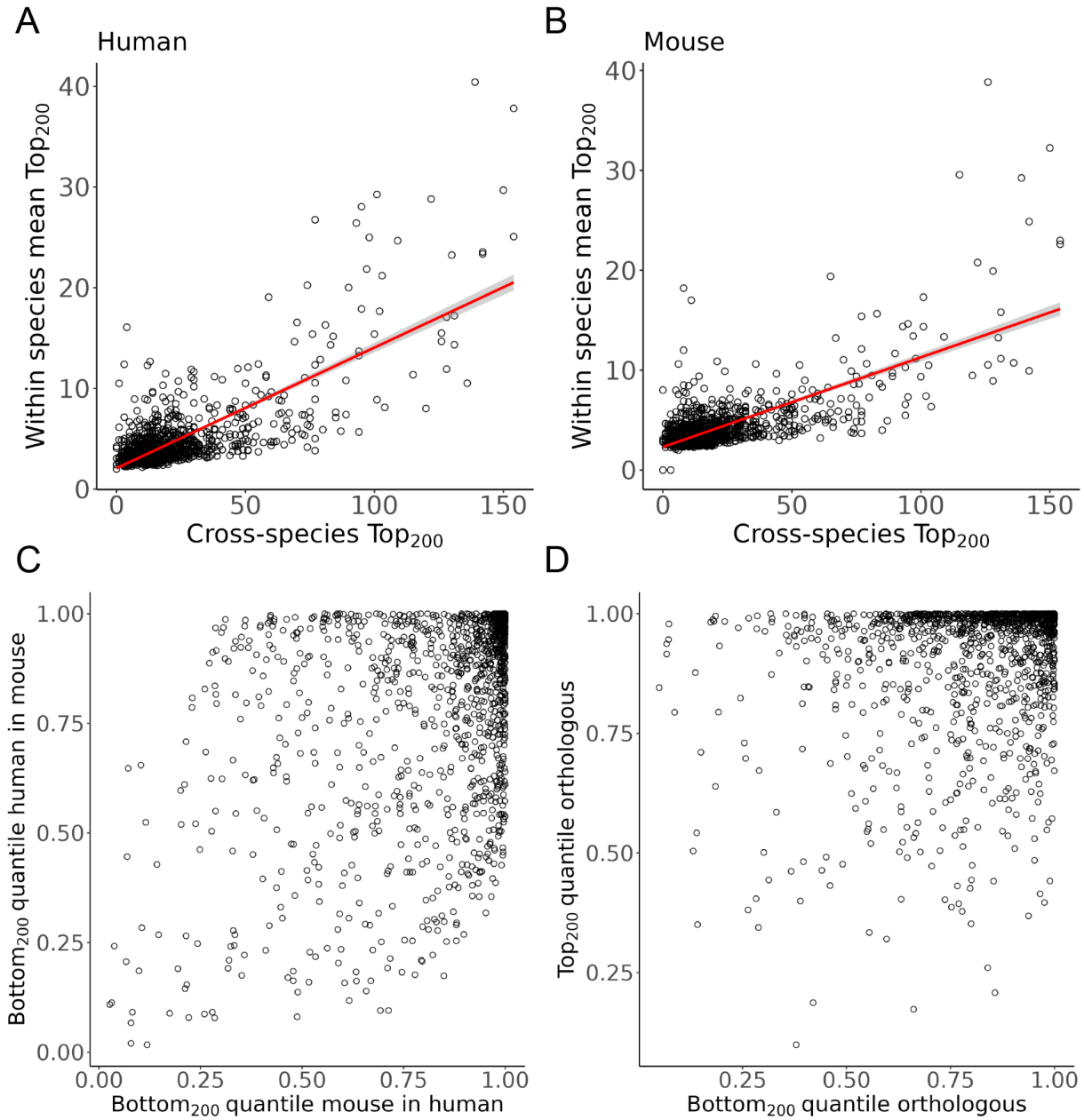
Supplemental Figure 3b. (A, B) Preservation of the consistency of positive and negative profiles between mouse and human for orthologous TR at (A) K=200 and (B) K=1000. (C, D) Examples of TR profiles with consistent negative but not positive profiles.



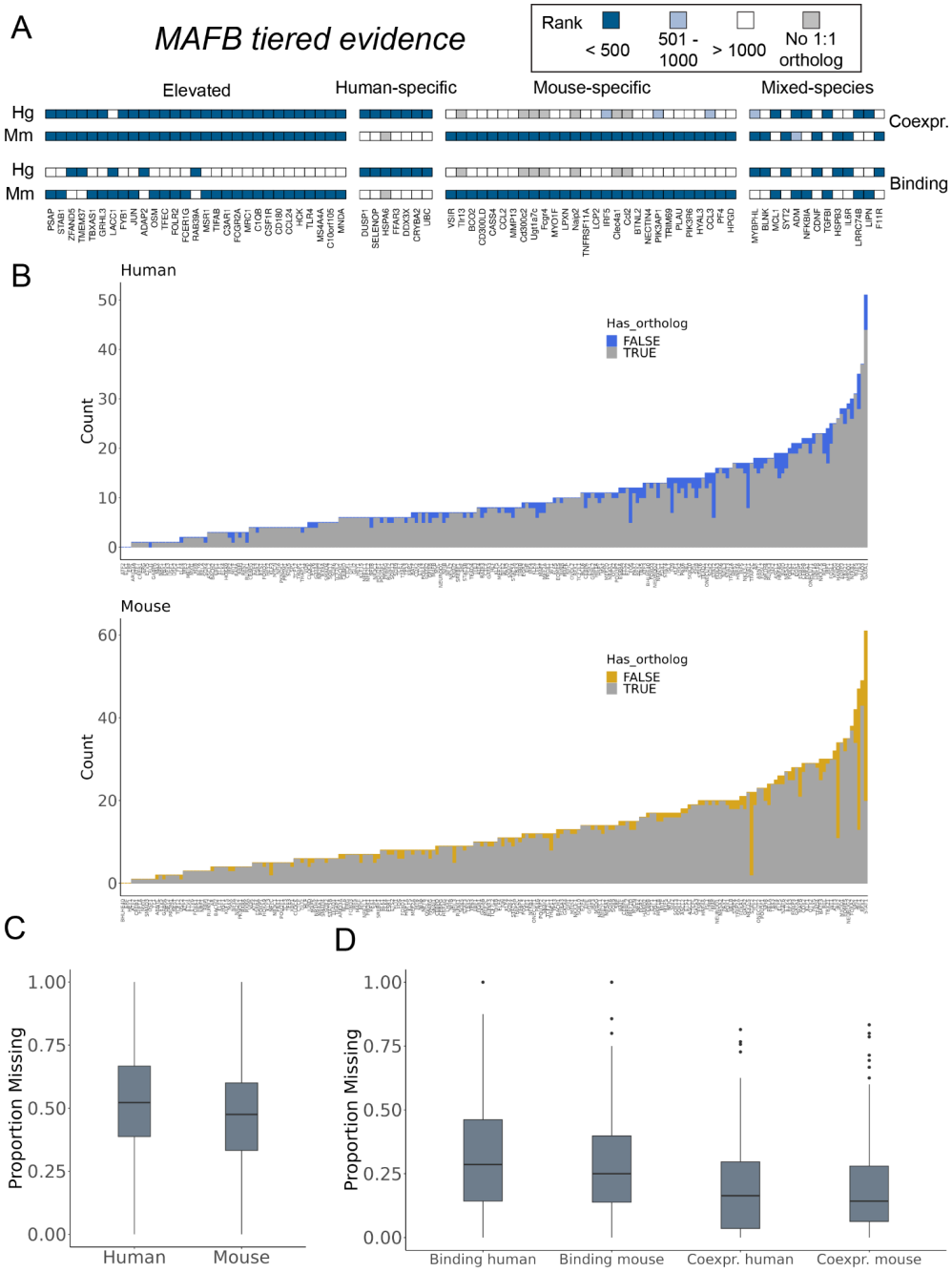
Supplemental Figure 4. L/S ribosomal gene aggregate profiles prioritize other ribosomal genes in (A) human and (B) mouse.



Supplemental Figure 5. Literature curated target recovery. (A) Histograms of the observed AUC quantiles for 451 human and 434 mouse aggregate TR coexpression profiles, relative to the corresponding individual TR profiles. A value of 1 indicates that an aggregate profile had an AUC greater than all individual profiles. (B) Distributions of the observed AUC quantiles for the 253 human and 241 mouse TRs that had binding and coexpression data. (C) Histograms of the difference between the raw AUC values between coexpression and binding aggregates. Positive values indicate that coexpression was better able to recover curated targets, negative values for binding.



Supplemental Figure 6. Comparison of single cell coexpression across species. (A, B) Scatterplots of the *Top*$_{200}$ overlap between orthologous TRs versus the average *Top*$_{200}$ between every unique pair of individual TR profiles in (A) human and (B) mouse. (C) Scatterplot of the quantiles of each TRs observed *Bottom*$_{200}$ overlap with its ortholog, relative to all other TRs. (D) Scatterplot of the *Top*$_{200}$ and *Bottom*$_{200}$ quantiles, where each point represents the average of *Quant_human* and *Quant_ortho.*



Supplemental Figure 7. (A) MAFB tiered evidence. (B) Related to Fig. 5B: Genes gained in the Species-Specific set for TRs with ChIP-seq data in both species may have made the cut-off in one species only, or lack an orthologous match. Barcharts show the count of interactions gained in the former scenario (coloured) or the latter (grey) for 216 TRs in human and mouse. (C, D) The Elevated collection required a gene to make the cut-off in three of four rankings. For each of 216 TRs, we tallied which rankings failed to make the cut-off among the Elevated genes, and represented this as a proportion. MAFB, for example, had 29 Elevated genes, 25 of which that did not make the cut-off in the human rankings (binding or coexpression), thus its Proportion Missing in human was 0.86. In (C) we show the distributions of these values by species. In (D) we show the breakdown by the individual rankings.

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. Nat 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. Bai X, Shi S, Ai B, Jiang Y, Liu Y, Han X, et al. ENdb: a manually curated database of experimentally supported enhancers for human and mouse. Nucleic Acids Research. 2019 Oct 29;gkz973.
6. 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.
7. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics. 2015 Jul 1;31(13):2123–30.
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 Research. 2012 Nov 26;41(D1):D991–5.
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. Bovolenta LA, Acencio ML, Lemke N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. BMC Genomics. 2012;13(1):405.
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. Carrasco Pro S, Dafonte Imedio A, Santoso CS, Gan KA, Sewell JA, Martinez M, et al. Global landscape of mouse and human cytokine transcriptional regulation. Nucleic Acids Research. 2018 Oct 12;46(18):9321–37.
13. 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.
14. 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.
15. Christensen J, Cloos P, Toftegaard U, Klinkenberg D, Bracken AP, Trinh E, et al. Characterization of E2F8, a novel E2F-like cell-cycle regulated repressor of E2F-activated transcription. Nucleic Acids Res. 2005;33(17):5458–70.
16. Crow M, Gillis J. Co-expression in Single-Cell Analysis: Saving Grace or Original Sin? Trends Genet. 2018 Nov;34(11):823–31.

17. Crow M, Paul A, Ballouz S, Huang ZJ, Gillis J. Exploiting single-cell expression to characterize co-expression replicability. Genome Biol. 2016 May 6;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. De Smet R, Marchal K. Advantages and limitations of current network inference methods. Nat Rev Microbiol. 2010 Oct;8(10):717–29.
20. 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.
21. 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.
22. 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.
23. 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.
24. Essaghir A, Toffalini F, Knoops L, Kallin A, Van Helden J, Demoulin JB. Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. Nucleic Acids Research. 2010 Jun;38(11):e120–e120.
25. 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.
26. Farahbod M, Pavlidis P. Untangling the effects of cellular composition on coexpression analysis. Genome Res. 2020 Jun;30(6):849–59.
27. 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.
28. 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.
29. 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.
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. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Research. 2018 Jan 4;46(D1):D380–6.
32. 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. 2021 Jul;12(7):748-756.e3.
33. He P, Lim K, Sun D, Pett JP, Jeng Q, Polanski K, et al. A human fetal lung cell atlas uncovers proximal-distal gradients of differentiation and key regulators of epithelial fates. Cell. 2022 Dec;185(25):4841-4860.e25.
34. 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.

35. 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.
36. 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.
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. Kaya T, Mattugini N, Liu L, Ji H, Cantuti-Castelvetri L, Wu J, et al. CD8+ T cells induce interferon-responsive oligodendrocytes and microglia in white matter aging. Nat Neurosci. 2022 Nov;25(11):1446–57.
42. 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.
43. 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.
44. 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.
45. Lampada A, Taylor V. Notch signaling as a master regulator of adult neurogenesis. Front Neurosci. 2023 Jun 29;17:1179011.
46. Langfelder P, Horvath S. Fast R Functions for Robust Correlations and Hierarchical Clustering. J Stat Soft [Internet]. 2012 [cited 2024 Feb 15];46(11). Available from: http://www.jstatsoft.org/v46/i11/
47. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. Prlic A, editor. PLoS Comput Biol. 2013 Aug 8;9(8):e1003118.
48. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res. 2004 Jun;14(6):1085–94.
49. 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.
50. Lentaigne C, Greene D, Sivapalaratnam S, Favier R, Seyres D, Thys C, et al. Germline mutations in the transcription factor IKZF5 cause thrombocytopenia. Blood. 2019 Dec 5;134(23):2070–81.
51. Lesurf R, Cotto KC, Wang G, Griffith M, Kasaian K, Jones SJM, et al. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D126–32.
52. Li H, Huo Y, He X, Yao L, Zhang H, Cui Y, et al. A male germ-cell-specific ribosome controls male fertility. Nature. 2022 Dec 22;612(7941):725–31.
53. Li X, Zheng Y, Hu H, Li X. Integrative analyses shed new light on human ribosomal protein gene regulation. Sci Rep. 2016 Jun 27;6(1):28619.

54. 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.
55. 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.
56. 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.
57. 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.
58. Lunden JW, Durens M, Phillips AW, Nestor MW. Cortical interneuron function in autism spectrum condition. Pediatr Res. 2019 Jan;85(2):146–54.
59. Lv Y, Xiao J, Liu J, Xing F. E2F8 is a Potential Therapeutic Target for Hepatocellular Carcinoma. J Cancer. 2017;8(7):1205–13.
60. Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, Barsky A, et al. InnateDB: facilitating systems‐level analyses of the mammalian innate immune response. Molecular Systems Biology. 2008 Jan;4(1):218.
61. 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
62. 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.
63. 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;99(3):624–35.
64. Miki Y, Devi L, Imai Y, Minami N, Koide T, Goel S. Deletion of the PDZ-binding kinase (Pbk) gene does not affect male fertility in mice. Reprod Fertil Dev. 2020;32(10):893.
65. 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.
66. 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.
67. 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.
68. 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.
69. Nord AS, West AE. Neurobiological functions of transcriptional enhancers. Nat Neurosci. 2020 Jan;23(1):5–14.
70. 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.

71. 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.
72. Posner DA, Lee CY, Portet A, Clatworthy MR. Humoral immunity at the brain borders in homeostasis. Current Opinion in Immunology. 2022 Jun;76:102188.
73. 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.
74. 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.
75. 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.
76. Rothenberg EV. Causal Gene Regulatory Network Modeling and Genomics: Second-Generation Challenges. Journal of Computational Biology. 2019 Jul 1;26(7):703–18.
77. 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.
78. Shannon P. igvR [Internet]. Bioconductor; 2018 [cited 2024 Jan 15]. Available from: https://bioconductor.org/packages/igvR
79. Shiraishi C, Matsumoto A, Ichihara K, Yamamoto T, Yokoyama T, Mizoo T, et al. RPL3L-containing ribosomes determine translation elongation dynamics required for cardiac function. Nat Commun. 2023 Apr 20;14(1):2131.
80. 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.
81. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005 Oct 15;21(20):3940–1.
82. Skinnider MA, Squair JW, Foster LJ. Evaluating measures of association for single-cell transcriptomics. Nat Methods. 2019 May;16(5):381–6.
83. Sonawane AR, Weiss ST, Glass K, Sharma A. Network Medicine in the Age of Biomedical Big Data. Front Genet. 2019 Apr 11;10:294.
84. 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;7(11):1930–43.
85. 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.
86. The Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018 Oct;562(7727):367–72.
87. The Tabula Sapiens Consortium*, Jones RC, Karkanias J, Krasnow MA, Pisco AO, Quake SR, et al. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science. 2022 May 13;376(6594):eabl4896.
88. 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.

89. 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.
90. 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.
91. 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.
92. 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.
93. 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;52(3):504–13.
94. 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.
95. 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.
96. 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.
97. 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.
98. Yeung J, Ha TJ, Swanson DJ, Goldowitz D. A Novel and Multivalent Role of Pax6 in Cerebellar Development. Journal of Neuroscience. 2016 Aug 31;36(35):9057–69.
99. 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 Sep;69(9):706–19.
100. 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.
101. Yusuf D, Butland SL, Swanson MI, Bolotin E, Ticoll A, Cheung WA, et al. The Transcription Factor Encyclopedia. Genome Biol. 2012;13(3):R24.
102. Zamboni M, Llorens-Bobadilla E, Magnusson JP, Frisén J. A Widespread Neurogenic Potential of Neocortical Astrocytes Is Induced by Injury. Cell Stem Cell. 2020 Oct 1;27(4):605-617.e5.
103. 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.
104. Zhang Y, Cuerdo J, Halushka MK, McCall MN. The effect of tissue composition on gene co-expression. Brief Bioinform. 2021 Jan 18;22(1):127–39.
105. 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.