NetREm: Network Regression Embeddings reveal cell-type transcription factor coordination for gene regulation

Saniya Khullar[1,2], Xiang Huang[1], Raghu Ramesh[1,4], John Svaren[1,5], Daifeng Wang[1,2,3,*]

[1]Waisman Center, University of Wisconsin-Madison, Madison, WI, 53705, USA
[2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, 53076, USA
[3]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, 53706, USA
[4]Comparative Biomedical Sciences Training Program, University of Wisconsin-Madison, Madison, WI, 53706, USA
[5]Department of Comparative Biosciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, WI, 53706, USA
[*]Corresponding author: daifeng.wang@wisc.edu

Abstract

Transcription factor (TF) coordination plays a key role in gene regulation such as protein-protein interactions (PPIs) and DNA co-bindings. Single-cell technologies facilitate gene expression measurement for individual cells and cell-type identification, yet the connection between TF coordination and gene regulation of various cell types remains unclear. To address this, we have developed a novel computational approach, Network Regression Embedding (NetREm), to reveal cell-type TF-TF coordination activities for gene regulation. NetREm leverages network-constrained regularization using prior interaction knowledge (e.g., protein, chromatin, TF binding) to analyze single-cell gene expression data. We test NetREm by simulation data and apply it to analyze various cell types in both central and peripheral nerve systems (PNS) such as neuronal, glial and Schwann cells as well as in Alzheimer's disease (AD). Our findings uncover cell-type coordinating TFs and identify new TF-target gene candidate links. We also validate our top predictions using Cut&Run and knockout loss-of-function expression data in rat and mouse models and compare our results with additional functional genomic data including expression quantitative trait loci (eQTL) and Genome-Wide Association Studies (GWAS) to link genetic variants to TF coordination. NetREm is open-source available at https://github.com/SaniyaKhullar/NetREm.

Introduction

Transcription Factors (TFs) are proteins that work together to regulate (activate or repress) target gene (TG) expression in a coordination fashion, especially at the cell-type level [1]. TFs bind to cognate DNA sequence-specific TF binding sites (TFBSs) on regulatory elements (e.g., enhancers and promoters) to mediate transcription of their respective TGs. Nonetheless, these gene regulatory mechanisms are multi-faceted; regulatory elements of TGs are formed by combinatorial interactions of multiple TFs within regulatory elements that form transcriptional regulatory modules (TRMs) [2] to ultimately govern transcription initiation [3]. Most TFs cooperate with other TFs (rather than operate in

isolation), working in concert to regulate gene expression utilizing mechanisms like co-binding or tethered-binding [3]. TFs can be part of stable complexes (e.g. heterodimers [4]) or they can enhance binding affinity of other TFs to nearby TFBSs (synergistic activation) to regulate TGs [4–6]. It is not uncommon for the regulation of one TG to necessitate interactions with 10-15 TFBSs [7]. Despite the prevalence of such TF-TF coordination, the underlying intricacies of this phenomena are not yet fully comprehended [4]. Further, a set of core TFs can determine cell-type-specific transcription profiles [8, 9]. Indeed, TF binding grammar is complex [7], context-specific and cell-type-specific, depending on other proteins and chromatin structure around TFBSs. Since the diversity of TF binding is core for cell-type specificity[8], models of TG regulation by TFs, e.g. gene regulatory networks (GRNs) must incorporate complex combinatorial analyses of TFs and their intricate interactions with one another.

Coordination among TFs regulates TG expression by modulating TF binding stability, localization, or post-translational modifications, affecting their regulatory function, activity, signaling. Contemporary research indicates a heightened adaptability among TFs in TG regulation, evidenced by phenomena like nucleosome-mediated cooperativity and the dynamic organization of TFBSs. These observations imply potential cooperativity even among disparate TFs that lack physical interaction [10]. Historically, classical models of TF cooperativity were predicated on direct physical protein-protein interactions (PPIs) among TFs; however, it is now understood that TFs occupying the same regulatory region need not interact directly to jointly coordinate the transcriptional activation of a TG. For example, the binding of one TF can alter local chromatin configurations or prompt DNA conformational transitions, such as nucleosomal DNA unwrapping or nucleosome eviction, thereby rendering the DNA more amenable to subsequent TF binding [10]. Additionally, through the tethered-binding mechanism involving coactivators and/or corepressors, TFs can exert regulatory influence over TGs via the direct or indirect cooperative recruitment of intermediary proteins, such as p300/CBP and the mediator complex [11]. Even in the absence of direct PPIs, the coordinate effort of many TFs is indispensable for co-factor recruitment and the establishment of nucleosome-free regions, hallmark features of enhancers and promoters, which characterize the distinct transcriptional profiles of specific cell-types.

Recent single-cell data analyses [12, 13] show cell-type-specific expression dynamic patterns, implying that gene regulation is cell-type-specific. However, the mechanistic role and extent of TF associations with other TFs (e.g. cooperativity) to drive TG regulation across cell-types remain unclear [3, 4, 14–16]. For instance, neuronal and glial cells are important in nervous system development, function, repair and processes like myelination, synaptogenesis, neuroplasticity. Thus, understanding TF coordination, at the cell-type level, is crucial for dissecting the GRNs that govern processes fundamental to cognition, movement, behavior and potential dysregulation in neurodegenerative diseases. State-of-the-art tools (e.g. BEELINE [17], SCENIC [18], GRNBoost [19], Signac [20], scGRNom [21]) build cell-type GRNs using co-expression, information-theoretic, machine learning-based, and/or multiomics integration techniques to infer potential cell-type-specific relationships among individual candidate TFs and their TGs that they help regulate (e.g. TF-TG links). Nonetheless, these tools mainly consider TFs in isolation, overlooking the interdependent TF-TF interactions critical for cell-type gene regulation. Tools like SCINET [22] reconstruct cell-type-specific interactomes, but they do not show gene regulation.

Previous studies leveraged network-regularized regression models to identify disease-associated genes and gene networks, incorporating existing biological information as prior knowledge [22, 23]. Nonetheless, these models have not been tailored to elucidate TF coordination in gene regulation. To

bridge this gap, we develop NetREm (Network Regression Embeddings), a novel computational framework designed to infer cell-type TF coordination activities for gene regulation. Building on established network-regularized regression techniques, NetREm integrates multimodal data (e.g., epigenomic markers, TF binding profiles, PPIs of TFs, derived TF-TF colocalizations, gene expression) to construct robust predictive models for TF and TG interactions. Our TF-TG regulatory networks complement existing tools that focus on predicting GRNs. One distinctive feature of NetREm is its ability to generate network regression embeddings for individual TGs. These embeddings aid in identifying core and novel cell-type TFs and in quantifying coordination among TFs (e.g., TF-TF coordination network) for cell-type gene regulation of TGs. NetREm can reveal TRMs comprising co-associated TFs and can prioritize TGs and coordination among TFs that significantly vary across different cell types or disease states. As demonstrations, we apply NetREm to simulation data and various cell types in both central and peripheral nerve systems (PNS) such as myelinating Schwann cells (mSCs) and non-myelinating Schwann cells (nmSCs), as well as in Alzheimer's disease (AD) and control states in various neuronal and glial cells. However, NetREm is an open-source tool for general-purpose use.

Results

Overview of Network Regression Embeddings (NetREm)

NetREm offers a robust, multiomics, computational approach to build and integrate networks of Transcription Factor (TF) to target gene (TG) regulatory links (complementary to a gene regulatory network (GRN)) and TF-TF coordination in a cell-type-specific manner (**Fig. 1**). Applicable to both single-cell and bulk data, NetREm is for situations where predictors are network-structured and collinearity is expected. NetREm leverages a comprehensive, ground truth protein-protein interaction (PPI) network to globally capture direct, verified interactions between TFs. Starting with multimodal data, including gene expression data and a PPI network, NetREm applies a 2-step optimization process for each TG. First, it formulates a PPI network-regularized regression problem to sift through candidate TFs, aiming to find those most likely to regulate a TG in the given cell type. Second, to solve this problem, NetREm employs Singular Value Decomposition (SVD) to create new latent space gene embeddings (i.e. network regression embeddings), which are then used in a Lasso regression model to predict TG expression. In the process, NetREm not only identifies key TFs but also learns TF-TF coordination, capturing potential relationships like cooperativity or competition (antagonism) among TFs. Coordination among TFs with TF-TF links in the input PPI are referred to as TF-TF interactions as such TF coordination events may be mediated by direct physical interactions (e.g. cell-type specific PPI subnetwork); the remaining novel TF-TF links may refer to indirect (e.g. nucleosome-mediated cooperativity) or direct interactions among TFs (potential links in future PPIs), which follow-up experiments may investigate. In fact, a strength of NetREm is this ability to go beyond the input PPI to discover novel TF-TF coordination links whose true nature remains unknown: either these are PPI links that have not yet been discovered or the TFs may interact indirectly with one another. This makes NetREm invaluable for constructing comprehensive cell-type TF-TG regulatory networks (linking TFs to their TGs) and TF-TF coordination networks (linking interacting TFs). As mentioned above, we apply NetREm to human data to predict TF-TG regulatory networks and TF-TF coordination networks in non-myelinating (nmSCs) and myelinating (mSCs) Schwann cells (SCs) and in 8 major neuronal and glial cell-types in control versus Alzheimer’s disease (AD) states.

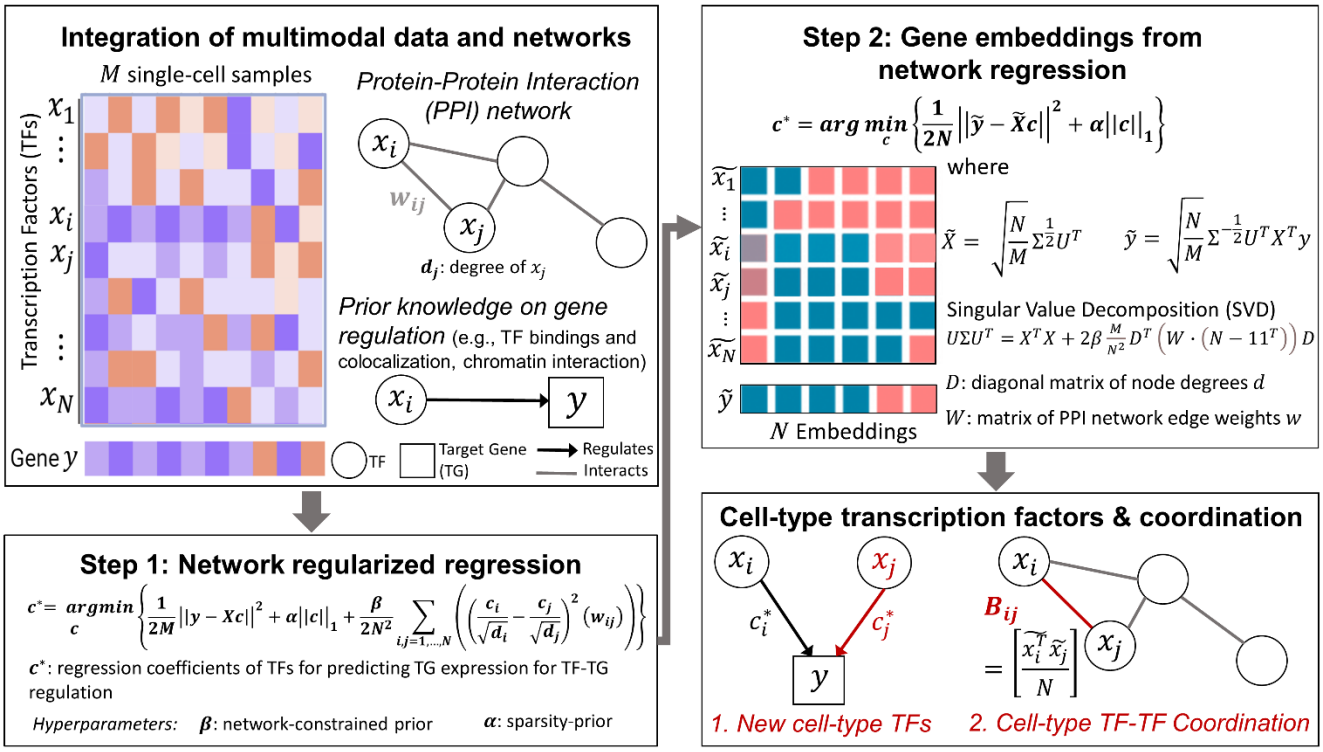


Figure 1 Overview of Network Regression Embeddings (NetREm), a computational framework for inferring cell-type TF-TF coordination for gene regulation. We apply NetREm for each target gene (TG) and aggregate the results to yield our final outputs at the cell-type level. The pipeline integrates multimodal data including gene expression ($X \in \mathbb{R}^{M \times N}$, $y \in \mathbb{R}^{M}$) for M samples, protein-protein interactions (PPI), and optional prior knowledge on gene regulation (i.e. reference gene regulatory network (GRN)) to select only relevant candidate Transcription Factor (TF) predictors for the given TG to reduce the dimensionality of X. **Step 1** involves setting up a PPI network-regularized regression problem to identify optimal TFs (out of the N candidate TFs) for a TG, guided by network prior hyperparameter β and sparsity prior hyperparameter α. **Step 2** transforms the original gene expression data into new latent space embeddings ($\tilde{X} \in \mathbb{R}^{N \times N}$, $\tilde{y} \in \mathbb{R}^{N}$) using Singular Value Decomposition (SVD). These embeddings are then subjected to Lasso regression to yield optimized and biologically relevant coefficients c^* for that given TG. NetREm's first output links optimal TFs to this TG by regression coefficients c^* (positive: potential activator, negative: potential repressor), in a TF-TG regulatory network that may be complementary to a GRN. These TF-TG regulatory links are likely to reflect the underlying biology of TF-TF coordination among predictors and may discover novel cell-type TFs for the TG. The second output is a TG-specific TF-TF coordination network predicting how TFs relate to each other (B_{ij} is a coordination score of TF_i and TF_j, cooperate: $B_{ij} > 0$, compete: $B_{ij} < 0$) in indirect or direct ways to regulate this given TG. NetREm's outputs thus identify novel TFs (in the complementary GRNs, like x_j) and infer potential coordination among TFs, like cooperativity or competition, within the cell-type.

Simulation study

First, we test NetREm on simulated gene expression data for 100,000 samples (training: 70,000, testing: 30,000), where X (TF predictors) and y (target gene TG) are drawn from a standard normal distribution (mean: 0, standard deviation (std): 1, **Fig. S1A**). We standardize X and center y. The 5

candidate TFs [TF_1, TF_2, TF_3, TF_4, TF_5] have Pearson (r) correlations $cor(TF, TG)$ = [0.9, 0.5, 0.4, -0.3, -0.8] with TG, respectively (**Fig. 2A, Fig. S1B-C**). We build a prior undirected ground truth network (our constraint) of direct TF-TF interactions among predictors with strong edge weights of 0.8 between TF_1 and TF_2 and 0.95 between TF_4 and TF_5. The default edge weight is 0.1 for any missing interactions. Our simulation shows how NetREm outputs TF-TG regulatory links (via coefficients c^*) and TF-TF coordination links (via B network of coordination scores) to predict the behavior and coordination among TF predictors to help regulate TG expression.

NetREm predicts gene expression of TG based on expression levels of 5 TFs, subject to the network constraint. We set β = 1 and use LassoCV (NetREm model type) to determine the optimal α^* of 9e-4, exclude the y-intercept, and retain our default degree threshold of 0.5 (so edges with w < 0.5 are not counted towards TF node degrees in degree matrix D). Our 2 outputs for this TG are coefficients c^* relating TFs to TG expression in a TF-TG regulatory network (**Fig. 2B**) and a TF-TF coordination network (**Fig. 2C**) of coordination scores B. We compare NetREm to 4 benchmark linear models: Linear Regression[24] and 3 regularization models (ElasticNetCV[25], LassoCV[26], RidgeCV[27]). All 4 models are fit with defaults and no y-intercept. Low ranks for coefficients imply high absolute values, i.e. more important predictors of TG expression have higher magnitudes for c^*.

Absolute values of coefficients for TF_2 and TF_4 are higher in NetREm than in benchmarks (**Fig. S1D**), highlighting NetREm's ability to prioritize these TFs due to their strong corresponding network links with TF_1 and TF_5. Unlike benchmarks, which favor TF_3 over TF_4 (since TF_3 has a stronger magnitude of correlation with y than TF_4 does), NetREm prioritizes TF_4 over TF_3 because TF_4 strongly interacts with TF_5 that highly negatively correlates with TG, while TF_3 has weak prior network connectivity with other TFs. Sensitivity analysis (**Fig. 2D-E**) for fixed α and β confirms that the magnitude of coefficients for TF_2 and TF_4 increase as their respective simulated prior network connections with TF_1 and TF_5 strengthen. Test MSE values are similar, with benchmarks at approximately 0.130, NetREm at 0.138.

One of the most salient features of NetREm is its ability to integrate prior network structure knowledge into the regression model. This extends beyond the realm of biological systems to any domain where relational structures among predictors exist. By doing so, NetREm enriches the model's interpretability and allows for the identification of key TFs that may not exhibit strong individual correlations with the dependent variable, TG. Our method has the unique advantage of providing a weighted network of coordination among TFs (i.e. direct and/or indirect TF-TF interactions) specific for the TG in this cell-type (**Fig. 2C**), which these 4 benchmark models cannot do. In our TG-specific TF-TF coordination network, TF_1 and TF_2 have positive interactions with each other, as do TF_4 and TF_5. NetREm predicts novel TF-TF coordination links that are not found in the input network, like negative and antagonistic relations between TF_1 and TF_5 (coordination score: -0.73) and between TF_3 and TF_5 (coordination score: -0.57); links like these correspond to potential direct or indirect TF-TF relations. TF_3 has a weaker positive relation with TF_1 than TF_2 does (0.11 vs. 0.38), which may explain its relatively small coefficient of 5e-3. Ultimately, NetREm suggests that TF_1 and TF_2 work together, directly, in a coordinated manner to increase TG expression, while TF_4 and TF_5 collaborate, directly, to decrease it.

We also test the robustness by running 1,000 simulations of NetREm and benchmark models (**Supp. Methods and Materials** in **Supp. File 5**), which show that NetREm has reduced standard deviations for all coefficients (**Table S1**), denoting less variability and increased stability. Pairwise 1-sided t-tests reveal NetREm's significantly heightened coefficient magnitudes for TF_2 and TF_4, and reduced values

for TF_3 against benchmarks (p < 2e-16, **Table S2**). Further, we run 1,000 simulations where we benchmark 7 values for β (**Table S3, Fig. S2B-C**), illustrating the impact of β on the bias-variance trade-off and how excessive network regularization penalization (high β) can over-constrain models, potentially causing NetREm to struggle and its predictions to thereby suffer.

Overall, NetREm's network constraint regularization enhances model robustness as it is less sensitive to noise in the data. This is especially beneficial in high-dimensional datasets where the risk of multicollinearity and overfitting are prevalent. Leveraging network pre-knowledge, NetREm accurately discerns and assigns each TF's influences, yielding dependable, detailed insights into individual roles, undeterred by intercorrelations. NetREm offers superior generalizability and consistency, potentially capturing intricate underlying structures and interactions in the data that might be oversimplified by benchmark models. This balanced integration of predictive accuracy and structural interpretability positions NetREm as a computationally efficient and methodologically sound choice. In summary, NetREm integrates TF-TG regulatory links and TF-TF coordination scores, holistically, to understand better the TF behaviors involved in regulating TG and deepen our grasp of TF regulatory dynamics.

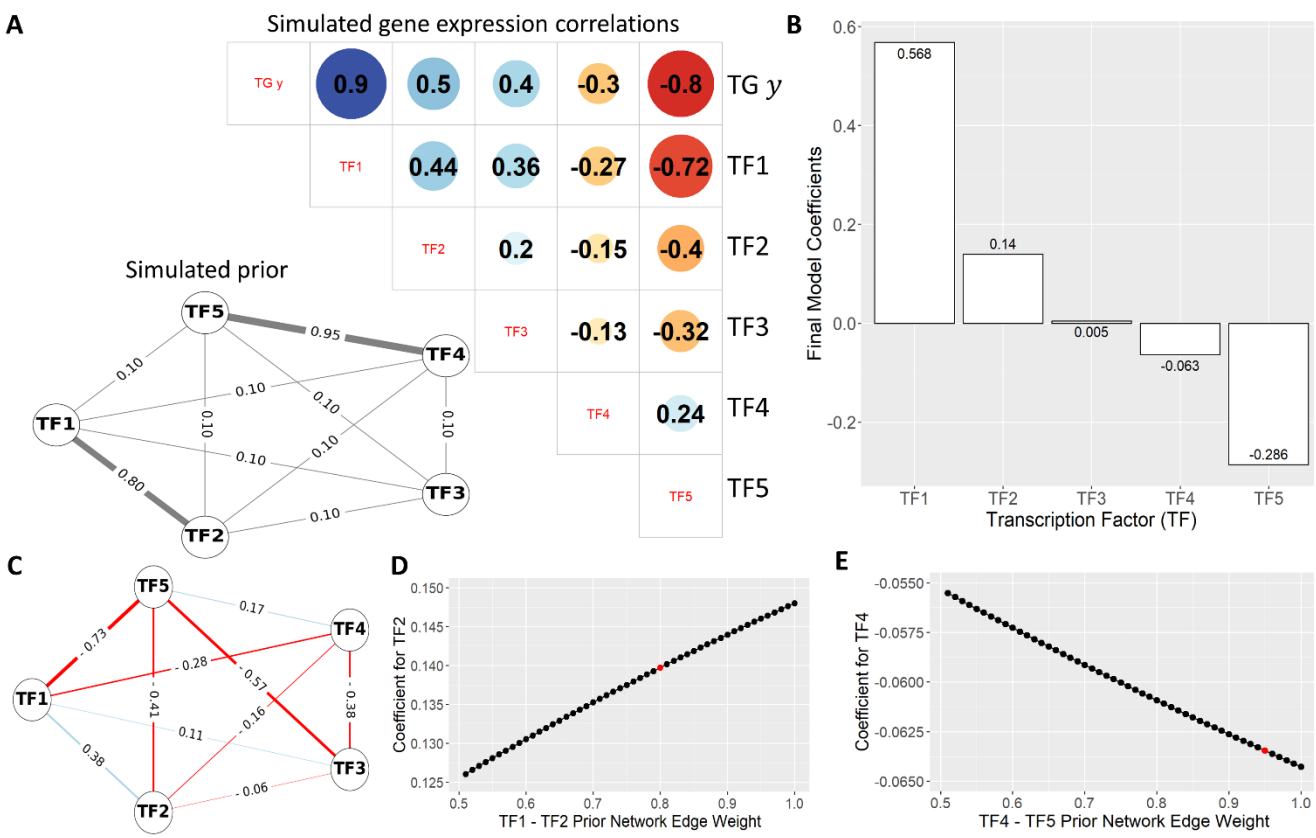


Figure 2 Simulation study for 5 transcription factors (TFs) and one target gene (TG). **A)** The bottom left shows the prior biological network with default edge weights (0.1), with stronger direct and verified connections for TF_1- TF_2 (0.8) and TF_4- TF_5 (0.95). The top right presents a Pearson correlation matrix among the TFs and TG in the training data where: $cor(TF, TG)$ = [0.9, 0.5, 0.4, -0.3, -0.8]. Dot sizes represent magnitude, and colors indicate positive (blue) or negative (red) correlations. **B)** Coefficients c^* for the TFs in the TF-TG regulatory network, based on a model fit using optimal $\alpha^* = \alpha$ = 9.0e-4 (based on LassoCV), β =1, and no y-intercept. **C)** The resulting TG-specific TF-TF coordination network (coordination scores: B), where red and blue links signify negative and positive interactions, respectively. Direct TF-TF interactions are between TF_1- TF_2 and TF_4- TF_5, while other TF-TF

coordination links with scores > 0.1 may be novel (and refer to direct or indirect TF relations). **D)** Effects of varying the TF_1-TF_2 edge weight in the original simulated biological prior network, from 0.51 (default minimum degree threshold is 0.5) to 1 in 0.01 increments, holding all other parameters fixed (e.g. utilizing Lasso model type for NetREm with α^*). Respective TF_2 coefficient increases monotonically from 0.126 to 0.148. The red dot signifies the selected edge weight used for the simulated prior network. **E)** A similar sensitivity analysis for TF_4 and TF_5 shows that TF_4 coefficient increases in magnitude, becoming more negative from -5.5e-2 to -6.4e-2, as the TF_4-TF_5 edge weight is perturbed.

Gene regulatory links between transcription factors and target genes in myelinating and non-myelinating human Schwann cells

With the emergence of new single-cell epigenomic data from many human tissues, it is now possible to model gene regulatory networks (GRNs) in novel contexts. The first application of our method is to analyze Schwann cells (SCs), which play a pivotal role in maintaining, regenerating, and supporting neurons in the peripheral nervous system (PNS). SCs, the major myelinating cells in the PNS, are derived from neural crest, and exhibit tremendous flexibility in not only myelination but also in several other tissues. In addition, they exhibit as terminally differentiated cells that can reverse their differentiation status after nerve injury and promote nerve regeneration[28]. There is considerable diversity in the differentiation status of SCs as characterized in single-cell studies in rodent nerve[29, 30]. The 2 main SC phenotypes are 1) myelinating SCs (mSCs) associated with larger diameter axons (>1 micron) and 2) non-myelinating SCs (nmSCs) that wrap a bundle of smaller diameter axons (typically sensory axons), which is referred to as a Remak bundle. While mSCs envelop axons in myelin sheaths to enhance conduction speed, nmSCs support sensory axon function and interactions without forming myelin, contributing to overall nerve integrity.

Uncovering TF-TG regulatory mechanisms modulating cellular processes in mSCs and nmSCs may be helpful for understanding and treating debilitating nerve injuries, demyelinating disorders, hereditary neuropathies. For instance, mutations (e.g. single nucleotide polymorphisms (SNPs)) affecting SC function are the most prevalent cause of a genetic neuropathy known as Charcot-Marie-Tooth disease (CMT) [31], one of the most common heritable diseases affecting the nervous system. Some of these mutations affect major transcriptional regulators of SC differentiation, such as EGR2 and SOX10[32], but transcriptional networks coordinating SC function remain largely unexplored. In fact, key TFs like EGR2, TEAD1, and SOX10 orchestrate regulation of myelination in the PNS, often cooperating to govern shared TGs [32, 33]. As several TFs within SCs lack cell-type specificity, uncovering TF-TF coordination networks crucial for SC lineage maturation and differentiation into mSCs or nmSCs also holds significance [34].

However, the distinct transcriptional regulatory networks and TF-TF coordination underlying these SC states have not been characterized. To this end, we apply NetREm to emerging single-cell data relating to human SCs in the Dorsal Root Ganglion (DRG) [35] (**Supp. File 5**). In particular, NetREm predicts TF-Regulatory Element-TG links for 9,901 TGs in mSCs and 7,075 TGs in nmSCs (**Supp. File 1**). Among these, 6,290 TGs are shared, 3,611 are mSC-specific, 785 are nmSC-specific. EGR2 is a TF in mSCs (not in nmSCs) [36, 37].

We define Loss-of-Function (LOF) TGs as those exhibiting altered expression due to the knockdown of a TF in studies of various mammals, implying their dependency on the presence of that TF (**Supp. File 5**). Validated direct TGs are a subset of overall LOF TGs that also have evidence of nearby TF binding, as identified by rodent ChIP-seq binding peaks for the TF. With this dual confirmation, we gain increased confidence in validated direct TGs. However, we acknowledge that LOF TGs may serve as potential valid candidate TGs for the TFs, being regulated possibly through indirect, if not direct, mechanisms [3]. While hundreds of TFs may be expressed in each cell-type, the identity of a

given cell-type is often attributed to the regulatory activity and coordination among just a few core TFs that shape its unique transcriptional landscape[8, 9].

For 6 core TFs (EGR2, NR2F2, SOX10, SREBF1, TEAD1, YY1) in nmSCs and mSCs (**Fig. 3A**), we provide the number of NetREm-predicted validated direct, LOF overall, and novel candidate TGs and report counts of TGs with strong Tibial Nerve eQTL (SNP-TG associations) support [38, 39]. Our evaluation of eQTL support is based on mapped SNP-TF-Regulatory Element-TG relations (from our TF-TG regulatory network output) (**Supp. File 5**). We identify instances where the eQTL SNP (eSNP) is predicted to strongly alter binding of the given TF to a regulatory element for this respective TG that it regulates; then, we use the eQTL data to determine if NetREm's prediction also has external support of this eSNP's association with changes in expression of this TG. For example, we predict 873 YY1 TGs in nmSCs (80 direct, 134 LOF, 659 potential novel candidates); of these predicted TGs, 15 direct (18.8%), 46 LOF (21.4%), 140 other YY1 nmSC TGs (21.2%) have eQTL validation.

Next, we examine instances where, even for validated direct TGs exhibiting low correlations in training gene expression data with any of the 6 core SC TFs, NetREm can accurately identify the respective TFs as their top predictors, indicated by a low rank (**Fig. 3B**). For example, SOX10 is weakly correlated with its direct TG SIPA1L2 (a top candidate modifier gene for demyelinating CMT disease type 1A [31]), but is predicted to have rank 19. Similarly, despite its low correlation (r = 0.02) with TEX2, TEAD1 is assigned as its activator with the highest coefficient (rank 1); this is further confirmed by a LOF study where TEX2 has $\log_2(\text{Fold Change})$ of -0.72 with TEAD1 knockout [40]. Despite EGR2 having a higher magnitude of correlation with direct TG SOX5 than TEAD1 does (0.09 vs. 0.07), NetREm assigns a lower magnitude for EGR2's coefficient (rank 15 vs. rank 9).

In **Fig. 3C**, we focus on gene enrichment analysis for 3 groups of predicted TGs (EGR2 mSC, SOX10 mSC, SOX10 nmSC TGs). Studies show how SOX10 is co-expressed with other proteins like EGR2 at various stages in SC development[41, 42]. We find 1,368 SOX10 TGs (mSC-specific: 849, nmSC-specific: 367, direct: 295, LOFs: 884), with 47 and 109 common direct and LOF TGs, respectively, in both SCs. When compared with SOX10, 563 TGs are EGR2-specific. These 3 groups of TGs are strongly enriched for nervous system-related terms. For instance, mSC SOX10 TGs are enriched for PNS abnormalities in conduction and myelination. **Fig. 3D** presents epigenome tracks in rats for 3 novel candidate SOX10 mSC TGs (GPR157, CEP95, SCAMP5) with SOX10 as among their top 3 predictors. These TGs show strong SOX10 binding in open chromatin regions [43] in the rat PNS.

We predict SNP rs55927047 enhances TEAD1's ability to bind to its TF Binding Site (TFBS) on a FOXN2 promoter in open chromatin for adult SCs, thereby activating FOXN2 in mSCs (**Fig. 3E**). This is supported by overlapping orthologous rat PNS TEAD1 ChIP-seq peaks with this promoter. Linkage Disequilibrium (LD) analysis reveals a significant positive correlation between this SNP and another Tibial Nerve eSNP, rs79073127, which also increases FOXN2 expression. Further, rs79073127 associates with higher risk of inflammatory polyneuropathy based on the Pan UK Biobank Genome Wide Association Study (GWAS) [44]. Our GWAS-eQTL colocalization analysis flags FOXN2 as a risk gene for this condition with PP4 posterior probability of 74.6%. Results like these further corroborate NetREm's capability to potentially uncover biologically relevant signals and identify new TGs for TFs.

Moreover, we compare relative gene expression percentiles in Genotype-Tissue Expression (GTEx) SCs [38, 45] with corresponding values of 4 other central nervous system (CNS) cell types [46] for the predicted TGs for each of the respective TFs in nmSCs and mSCs. We use 1-sided 2-sample t-tests at the 5% p-adjusted level. RXRG TGs in nmSCs have significantly higher relative gene expression percentiles in GTEx SCs (p-adj < 2e-6) (**Fig. 3F**). This is particularly noteworthy since the same pattern is not observed for the remaining nmSC TGs (controls) not predicted to be strongly regulated

by RXRG, respectively (p-adj ≈ 1). RXRG mSC TGs do not have statistically higher expression percentiles, which may be attributed to RXRG's preferential expression in nmSCs over mSCs based on findings from single-cell data in both rodents and humans [30]. In addition, NR2F2, SREBF1, and YY1 show the same statistical significance results for their TGs in nmSCs (but not in mSCs), while EGR2 is significant in mSCs for all comparisons except for Microglia (p-adj = 5.1e-2). Overall, we find 24 TFs (including SOX10 and TEAD1) out of 240 (across mSCs and nmSCs) have these significantly higher results for both their respective nmSC TGs and their mSC TGs over the corresponding controls, for all 4 CNS cell-type comparisons with SCs. Thus, our predicted TGs for our core SC TFs indeed tend to typically be highly-expressed genes that play crucial cell-type-specific roles in SCs.

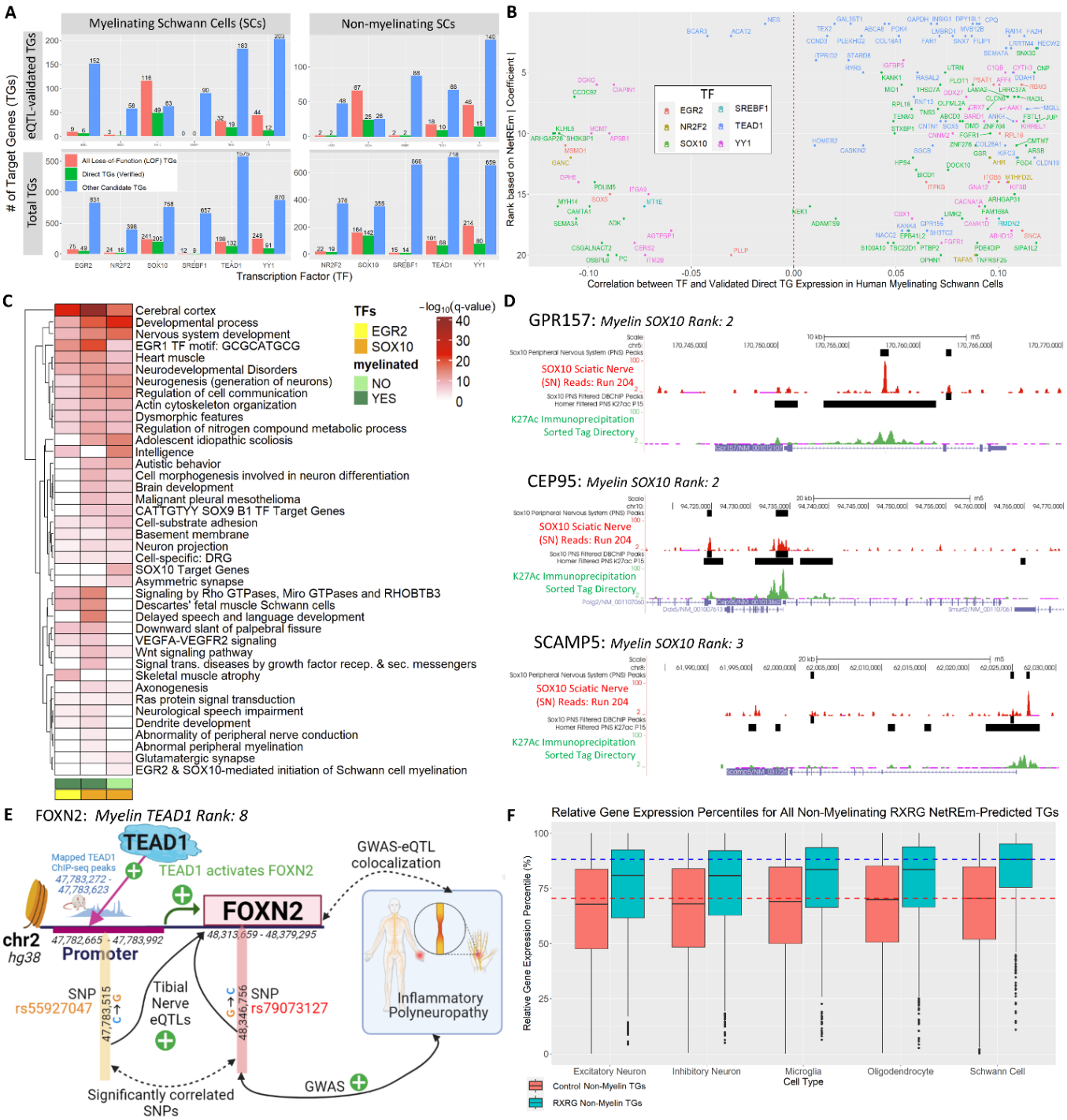


Figure 3 Gene regulatory links between transcription factors (TFs) and target genes (TGs) in myelinating (mSCs) and non-myelinating (nmSCs) human Schwann Cells (SCs) **A)** Bar plot categorizing TGs found for core TFs in mSCs (6 TFs) and nmSCs (missing TF EGR2). The top panel displays TGs with expression quantitative trait loci (eQTL) validation. The bottom panel reveals original counts. **B)** The x-axis: correlation between the core SC TF and validated direct TG (red: EGR2 TG, gold: NR2F2 TG, green: SOX10 TG, turquoise: SREBF1 TG, blue: TEAD1 TG, pink: YY1 TG) based on training gene expression data for mSCs. The y-axis: rank of absolute value of NetREm's regression coefficient c^* for that TF for that given TG, where smaller rank values imply a greater magnitude for the coefficient (and stronger relationship). We show direct TGs where the respective TF is among the top 20. **C)** Select gene enrichments for all EGR2 mSC, SOX10 mSC, and SOX10 nmSC TGs. Enrichments are hierarchically clustered and $-\log_{10}(q)$ values are reported. **D)** Epigenome tracks in rats (rn5 reference genome) for 3 potential novel candidate SOX10 mSC TGs with SOX10 as a top predictor TF. These tracks correspond to rat sciatic nerve ChIP-seq peaks, SOX10 peaks in the peripheral nervous system (PNS), histone modifications associated with enhancers (H3K27ac immunoprecipitation sorted tags). **E)** Tibial Nerve eQTL SNP (eSNP) rs55927047 (chromosome 2: 47,783,515 change in DNA base from C to G) located in the FOXN2 promoter (overlaps with orthologous TEAD1 ChIP-seq binding regions in rats) may strongly boost TEAD1 affinity for binding to the FOXN2 promoter to activate expression (given $c^* > 0$) of FOXN2, a candidate GWAS-eQTL colocalized TG biomarker for inflammatory polyneuropathy. FOXN2 is a novel TG for TEAD1 that has experimental support as a proximal TEAD1 TG based on ChIP-seq binding within 100 kilobases of its Transcription Start Site (TSS). Genome coordinates are based on the hg38 human reference genome. **F)** Boxplot that compares relative gene expression percentiles for all RXRG nmSC TGs in pooled SCs (GTEx Schwann cell [38, 45]) with those in the other 4 cell types (Microglia, Oligodendrocytes, Excitatory Neurons, Inhibitory Neurons [46]). The median relative gene expression percentile for RXRG nmSC TGs in SCs overall is 88.1 relative to 70.4 in controls (non-TGs for RXRG in nmSCs).

Coordination among transcription factors for gene regulation in myelinating and non-myelinating human Schwann cells

We benchmark test Mean Square Error (MSE) values across 5 regression models (prior to filtering for absolute value coefficients) for mSCs and nmSCs (**Fig. 4A**). One-sided 2-sample t-tests confirm that NetREm's test MSE values are significantly lower than those of the 4 benchmarks (mSCs: p < 1.4e-72, nmSCs: p < 3.0e-11, **Table S4**). NetREm predicts 122,556 mSC TF-TG links (eQTL validation: 10,203) and 142,168 nmSC links (eQTL validation: 10,904). **Fig. S3** depicts principal hub TFs for SCs, highlighting top TFs regulating the most eQTL-validated TGs across SCs. RXRA governs the most eQTL-validated TGs (484), predominantly in mSCs (308). EGR1, with 457 nmSC eQTL-validated TGs, aligns with studies emphasizing its significance in nmSCs, SC precursors, immature SCs[47]. Additionally, we pinpoint other central SC TFs like YY1 (400 eQTL-valid TGs out of 1,794).

NetREm also predicts a TF-TF coordination network for mSCs and nmSCs (**Supp. File 2**). In **Fig. 4C**, we visualize NetREm's predicted mSC TF-TF network for the 25 mSC-specific TFs and their 74 TF-TF links found in the input PPI network (excluding novel links for diagram simplicity). That is, here, we subset the TF-TF coordination network for TF-TF links with physical interactions found in the input PPI (i.e. direct TF-TF interactions). Notably, the average TF-TF relationship between mSC-specific TFs EGR2 and PPARD is the strongest with $\bar{B}$ = 7.9e-2. To provide further context, we annotate TF-TF relationships via the Contextual PPI [18]. ESR2 and PPARD interact in PNS neoplasms (abnormal

cancerous growth impacting SCs and related cells, e.g. Schwannomas[48]) with $\bar{B}$ = 3.4e-2. **Fig. S4A** shows TF-TF interactions for 30 nmSC-specific TFs; for instance, SOX4 regulatory activity in nmSCs may be attributed to the absence of its verified repressor EGR2 (NetREm coefficient c^*= -3.81e-2; verified as repressor based on LOF data, **Supp. File 5**). We uncover 65 common TGs for RXRG, TEAD1, and YY1 in nmSCs (**Fig. S4B**, TEAD1: 7 valid direct, 5 LOF only; YY1: 7 valid direct, 11 LOF only; RXRG: 37 ChIP-seq annotated) and 15 in mSCs (TEAD1: 2 valid direct, 2 LOF only; YY1: 3 valid direct, 3 LOF only; RXRG: 11 ChIP-seq annotated), suggesting greater cooperativity among these TFs in nmSCs. ANKRD11 and SORCS1 are regulated by these 3 TFs in both SC types.

Next, we highlight NetREm's ability to uncover strong novel coordination among TFs that were previously unknown (**Fig. S5**). We focus on 7 SC TFs that we have additional experimental validation data for: EGR2, NR2F2, RXRG, SOX10, SREBF1, TEAD1, YY1 (**Supp. File 5**). Our comprehensive input PPI network does not have RXRG-TEAD1 and RXRG-YY1 links (**Fig. S5A**). RXRG strongly positively correlates with TEAD1 (r = 0.5) and YY1 (r = 0.7) in nmSCs but negatively in mSCs (-0.89 and -0.72, respectively) in mouse Sciatic Nerves. We find that RXRG and YY1 have 239 (Jaccard Similarity (JS) = 0.16) and RXRG and TEAD1 have 192 common TGs in nmSCs. After mapping Tibial Nerve eQTLs to TF-TG regulatory networks, we find RXRG and YY1 have 26 common eQTLs (SNP-TGs) and eSNPs, and 29 common eTGs. Next, we examine various statistical aggregations for our raw B matrix of TF-TF coordination scores: across mSCs (46,010 non-zero TF-TF coordination links, 9,901 TGs) and nmSCs (48,180 non-zero TF-TF coordination links, 7,075 TGs). RXRG has positive B coordination scores for 5,572 nmSC TGs with YY1 and 5,179 with TEAD1. In nmSCs, RXRG interactions are particularly strong with TEAD1 ($\bar{B}$: 0.16, top 93.3%; median B: 0.23, top 94.9%; standard deviation (std): 0.1) and YY1 ($\bar{B}$: 0.04, top 73.8%; median B: 0.05, top 81.3%; std: 0.02) (**Fig. S5B-C**). Out of 15,864 ChIP-Seq H3K27 enhancer peaks shared between PNS and S16 rat (rn5) SC lines, RXRG shares 3,450 and 2,017 peaks with YY1 and TEAD1, respectively (**Fig. S5D**). In fact, 43.9% of RXRG peaks have YY1 binding (most colocalization/cobinding among the other 6 TFs) and 25.7% of RXRG peaks have TEAD1 binding (**Fig. S5E**). Conversely, 24% of TEAD1 peaks and 28% of YY1 peaks have RXRG binding. Further, 119 common TGs have ChIP-seq and H3K27Ac support in rats for RXRG and TEAD1 (and 148 for RXRG-YY1) in nmSCs. Thus, NetREm discovers very strong novel RXRG-YY1 and RXRG-TEAD1 TF-TF coordination links.

Analysis of TEAD1 Cut&Run read density centered across EGR2 peaks reveals colocalization of TEAD1 at approximately 40% of EGR2 TFBSs (**Fig. 4C**), supporting the predicted model for coordination among EGR2 and TEAD1 in mSCs. In addition, we find that YY1 Cut&Run reads overlap roughly 70% when centered on RXRG peaks and SOX10 peaks. CDH19 is a SC marker which peaks in expression at the promyelinating SC stage[49]. NetREm predicts that Tibial Nerve eSNP rs17799413 associates with decreased CDH19 expression (slope: 2.44e-1), by strongly altering binding of TEAD1 in both SCs and RXRG and YY1 in nmSCs. Cut&Run sequencing reveals binding peaks for TEAD1 and RXRG that are 40 kilobases (kb) upstream of CDH19's Transcription Start Site (TSS) (**Fig. 4D**) on the DNA. NCAM1 codes for an adhesion molecule preferentially expressed in nmSCs[50] and is a YY1 LOF TG and SOX10 direct TG. NetREm predicts EGR2 as one of the TFs regulating NCAM1 in mSCs and RXRG, SREBF1, and TEAD1 as some of the TFs for NCAM1 in nmSCs; SOX10 and YY1 regulate NCAM1 in both SCs. Tibial Nerve eSNP rs10749999 associates with increases in NCAM1 expression (slope: 9e-2) and may boost YY1 binding in both SCs and TEAD1 binding in nmSCs. Active S16 rat cell line SC enhancers roughly 130 kb and 200 kb upstream of the NCAM1 TSS have TEAD1 and YY1 binding. NCAM1 has binding of TEAD1, RXRG, and YY1 at the promoter region surrounding its gene locus. This further supports our prediction of joint regulation and coordination by these 3 TFs (e.g. our novel RXRG-TEAD1 and RXRG-YY1 links).

Fig. 4E illustrates how we may integrate NetREm's TF-TG regulatory networks (and corresponding coefficients c^*) and TF-TF interaction networks (and respective raw B coordination scores) with GWAS SNPs to yield stronger insights on gene regulatory mechanisms that may associate with disease phenotypes. Studies have observed the role of CEP192 in the formation and function of the mitotic spindle[51] and suggest that dysregulation in mitotic checkpoint regulators may lead to abnormal insulin signaling that may cause diabetes[52]. Here, Tibial Nerve eSNP rs11663049 (associated with decreased risk of polyneuropathy in diabetes based on the Pan UK Biobank GWAS[44]) is connected to decreased CEP192 expression in nmSCs. NetREm predicts this SNP strongly decreases binding of 3 activators (positive coefficients: RARB, RXRG, NR4A2) to a CEP192 promoter. We predict a novel TF-TF link between RXRG and NFKB1. Indeed, the Contextual PPI predicts that RXRG-NR4A2 and NR4A2-THRB links are associated with PNS neoplasms. GWAS-eQTL colocalization is significant for CEP192 (PP4 of 69%). Hence, NetREm showcases how non-coding disease-associated eSNPs may be associated with aberrations in TF coordination and TF bindings that may help explain changes in TG expression that correlate with the disease phenotype.

Tibial Nerve eSNP rs1544474 located within a LAMB1 enhancer region in open chromatin in adult SCs may impact regulation of LAMB1 (a validated direct SOX10 TG required for Schwann cell development), by altering the binding of a potential transcriptional regulatory module (TRM) of TFs (**Fig. 4F**). We predict this SNP disrupts bindings of activators (both SCs: SOX10; mSCs: FOXO1, FOXJ2; nmSCs: EP300, HDAC), ultimately cascading to significantly decreased LAMB1 expression (eQTL slope: -0.28, p: 1.4e-21). NetREm identifies SOX10 as a TF in both TF-TG complementary GRNs for SCs. On average, our TRM shows strong cooperativity between SOX10 and EP300 in nmSCs ($\bar{B}$ = 0.412), followed by SOX10 with FOX1 in mSCs ($\bar{B}$ = 0.351). Moreover, a study [53] performing BioID (proximity-dependent biotinylation) analysis finds that SOX10 and EP300 have strong TF-TF physical protein interactions in humans (Significance Analysis of INTeractome (SAINT) score = 1, the highest). Studies have observed that SOX family TFs like SOX10 exhibit cell-type-specific roles and regulate diverse intracellular events, achieving specificity through binding partner TFs that facilitate TG regulation by binding to nearby SOX sites on the DNA; this cooperation with partner factors (typically various types of TFs) is dynamic, allowing SOX family TFs to manage various events within and across cell-types [54]. That is, partner TFs for SOX TFs may differ across cell-types and/or cellular contexts [54]. NetREm discovers 3 novel TF-TF nmSC coordination links (SOX10 with RFX1 and RXRB, RXRB-ETS1) and a TBX2-FOXJ2 mSC coordination link. This example highlights how NetREm unveils TF-TF coordination networks, such as cooperativity, within the regulation of a shared TG. Indeed, multiple TFs coordinate with each other, working together at enhancers, sometimes directly by PPIs or via indirect forms of cooperativity[11].

Further, we compare NetREm with RTNduals[55] for predicting TF-TF coordination links that are annotated as highly-confident direct physical interactions in humans based on strong SAINT scores in experiments [53] (**Supp. File 5**). We fit NetREm and RTNduals for the same TFs and gene expression data for SCs (GTEx SCs[38, 45]: 83 TFs) as RTNduals returns no results for mSCs or nmSCs from the respective DRG gene expression data [35] (potential drawback of RTNduals). For GTEx SC gene expression data, RTNduals returns 267 TF-TF links. Thus, we then select the top 267 respective TF-TF links from NetREm based on $|B|$. NetREm has superior performance over RTNduals as it predicts 29 TF-TF coordination links in SCs with strong BioID and/or affinity purification mass-spectrometry (AP-MS) experimental validation, while RTNduals only predicts 7 (**Fig. S6**). This suggests that integrating the PPI as ground truth input data in NetREm aids in revealing more validated and meaningful TF-TF relations.

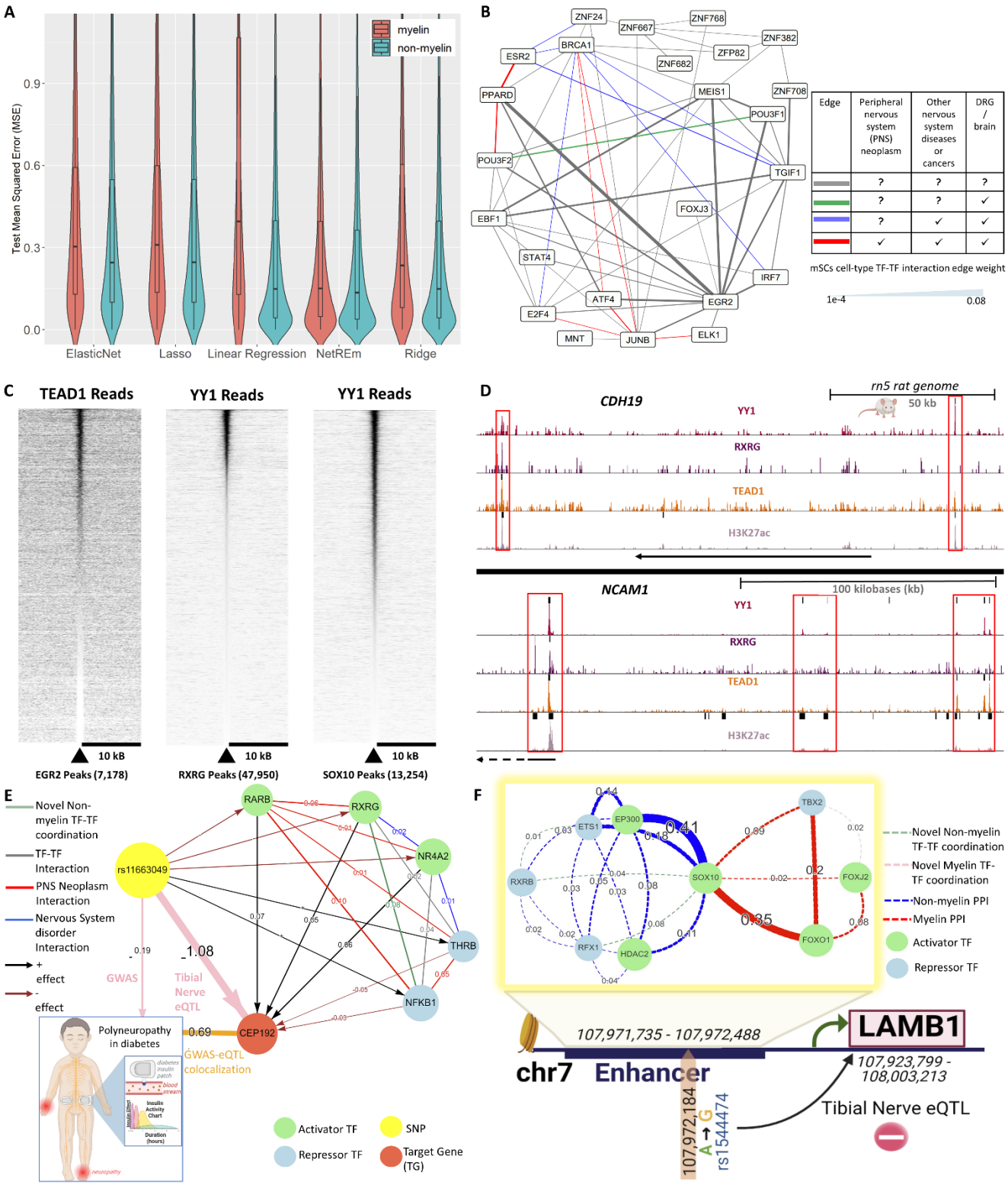


Figure 4 Coordination among transcription factors (TFs) for gene regulation in myelinating (mSCs) and non-myelinating (nmSCs) human Schwann cells (SCs) **A)** Density boxplot comparing test Mean Squared Error (MSE) values for 5 models—NetREm, ElasticNet, Ridge, Lasso, Linear Regression—

in mSCs and nmSCs. NetREm outperforms the 4 benchmarks in both cell types, with the lowest median and average test MSEs. For mSCs, NetREm achieves a median MSE of 0.15 and average of 0.3, while for nmSCs, the corresponding values are 0.14 and 0.28. **B)** Network of 25 mSC-specific TFs, with edge weights representing average B coordination scores ($\bar{B}$) across 9,901 TGs in the mSC NetREm TF-TG Regulatory Network. Higher $\bar{B}$ indicates stronger predicted TF-TF coordination. Edges are annotated based on the Contextual PPI [56]. **C)** Heatmaps show read density overlaps of the TEAD1 Cut&Run assay centered on EGR2 peaks, YY1 Cut&Run read density centered on RXRG peaks, and YY1 Cut&Run read density centered on SOX10 peaks. **D)** Cut&Run tracks of YY1, RXRG, and TEAD1, along with ChIP-seq tracks of H3K27ac, are shown from the S16 Schwann cell line in rats. Boxes highlight enhancer regions where TFs colocalize across genes CDH19 (top tracks) and NCAM1 (bottom tracks). These 2 TGs are impacted by Tibial Nerve eQTL SNPs (eSNPs) that alter regulatory TF binding at their loci. **E)** While **Fig. 3D** focuses on the impact of a single eSNP on 1 TF, this panel provides a more extensive example for TG CEP192 in nmSCs, which GWAS-eQTL colocalizes with decreased risk of polyneuropathy in diabetes. Tibial Nerve eSNP (yellow node) is associated with decreased risk of this condition and decreased CEP192 expression by disrupting bindings (brown arrows) of 3 activator TFs (positive coefficients $c^* > 0$, green nodes) and boosting bindings (black arrows) of potential repressor TFs ($c^* < 0$, blue nodes) to a CEP192 regulatory region. CEP192-specific raw TF-TF B coordination scores are undirected links, which are colored if they are associated with PNS Neoplasms (red) or Nervous System-related disorders (blue) based on the Contextual PPI [56] or if they are novel (not in input PPI). **F)** Tibial Nerve eSNP potentially influencing regulation of LAMB1, a confirmed SOX10 direct TG, through a specific set of activator and potential repressor TFs that may form distinct transcriptional regulatory modules (TRMs) in mSCs and nmSCs. Average coordination scores ($\bar{B}$) for that Schwann cell type are shown. Blue and red links refer to TF-TF interactions in nmSCs and mSCs, respectively, with support from the input PPI; the corresponding pink and green links are for the remaining TF-TF coordination links that are novel (and may constitute direct and/or indirect TF-TF relations, which follow-up studies may investigate).

Prediction and comparative analysis of cell-type coordination among transcription factors for gene regulation across neuronal and glial cell types in Alzheimer's disease

Cell-type-level TF coordination in the brain is key for proper functioning of neurons (e.g. synapse plasticity, neurotransmission, communication, neurogenesis, synaptogenesis). Disruptions to these networks have been found in brain-related disorders like Alzheimer's disease (AD) [57], a progressive neurodegenerative disorder characterized by memory loss, cognitive decline, neuroinflammation, behavioral changes. Studies have observed the role of PPIs and proteomic links in the behavioral and psychological symptoms of dementia and AD [58]. There is a strong need to identify master regulators in AD and develop therapies and pharmacological interventions to target their signaling pathways, to help with on-going research into treating and/or delaying AD[59]. We use NetREm to construct 16 TF-TG regulatory networks (**Supp. File 3)** and 16 corresponding TF-TF coordination networks (**Supp. File 4**) for 8 cell types related to AD and control stages. Reverse engineering changes in these networks across cell-types and among disease versus controls may help illuminate the path towards a more holistic understanding of the core molecular drivers of AD.

Fig. 5A illustrates raw B coordination scores for TF-TF links for target gene OLFM3 in Microglia (Mic) and APOA2 in Inhibitory Neurons (InNs). APOA2 shows distinct differences in TF coordination in AD InNs compared to baseline controls (2-sided t: 55.76, p-adj < 1e-100). Similarly, OLFM3, an AMP-AD nominated AD candidate TG [19], displays notable differences in Mic (2-sided t: 2.33, p-adj < 2e-2). Some TF-TF links exhibit strong coordination exclusively in AD, such as MEF2C-RUNX1 and MEF2A-RUNX1 for OLFM3, MEF2C-RORA and MEF2C-THRB for APOA2. While HIF1A-RUNX1 is an AD TF-TF link highly specific for OLFM3 in AD Mic (and not controls), this link is strong on average (across TGs in Mic) in both AD and control Mic ($\bar{B} \geq 0.5$) (**Fig. 5B**); we also observe this for MEF2C-THRB for APOA2 in InNs. RUNX1 binds to numerous Mic-specific genes [20]. Coordination between SPI1 and MEF2C plays pivotal roles in Mic enhancer landscapes [21]. Several TF-TF links, like SPI1-RUNX1 in AD Mic and FOXO3-MEF2C and MEF2A-MEF2C in all 4 TF-TF coordination networks, are annotated AD PPI links by the Contextual PPI [18]. AD Mic exhibit strong links (e.g. FOXO3-RUNX1, MEF2C-THRB, STAT3-MEF2A, NR3C1-RUNX1) that are not present in InNs. In contrast, InNs feature coordination among TFs like YY1 and MEF2C, and AD InNs display robust MEF2C-ZBTB7A links.

In **Fig. 5C**, we present a multifaceted network weaving together TF-TG regulatory links, TF-TF coordination networks, phenotypic outcomes, SNPs. Our attention is drawn, particularly, to the regulation of ANXA11, a critical player in diverse cellular functions (e.g. cell division, apoptosis, neutrophil function) and in signaling pathways like MAPK and P53[60]. Mutations in ANXA11 have been correlated with nervous system diseases like Amyotrophic Lateral Sclerosis (ALS)[61] and high risk of sarcoidosis[62], an inflammatory condition frequently affecting lungs and lymph nodes, leading to poorer health overall. In a previous study[63], we highlighted that ANXA11 positively correlates with Mild Cognitive Impairment and belongs to a gene co-expression module positively correlated with Controls in the Hippocampus (brain region crucial for learning and memory). Here, NetREm offers nuanced insights on how non-coding SNPs associated with ANXA11 expression may relate to various medical conditions. We predict SNP rs11202929 may enhance binding affinity of 9 activators (BACH1, BATF, ETV5, IRF7, KLF12, KLF7, NR3C1, SPI1) to TFBSs on ANXA11's enhancer, specifically in open chromatin in adult Mic. SPI1 is a core TF regulating AD-associated genes in Mic[64]. This SNP is protective against AD (GWAS effect size < 0) and significantly positively associates with increased ANXA11 expression (eQTL effect > 0). Our statistical analyses reveal that ANXA11 has significantly higher gene expression levels in control Mic than in AD Mic. In addition, coordination among TFs in control Mic are discernibly stronger (1-sided t test) than they are in AD Mic, pointing to a potentially disrupted network of TF cooperativity during AD progression. We uncover other risk SNPs for AD and cognitive functioning; SNP rs12412257 associates with lower ANXA11 expression and reduced ability for word interpolation, a key measure of fluid intelligence and reasoning in the Pan UK Biobank GWAS[44]. Mini-Mental State Exams and psycholinguistic analyses like word interpolation are key predictive markers of AD onset [63, 65]. GWAS-eQTL colocalization analyses for both SNPs are significant, revealing that ANXA11 expression positively associates with reduced risk of AD and with improved word interpolation ability. Overall, NetREm provides a powerful framework for deepening our understanding of complex GRNs and their implications across a spectrum of health conditions.

Next, we compare the predicted TF coordination networks of NetREm and RTNduals[55], using identical input training gene expression data and the same list of TFs across 8 cell types (**Fig. S6**). Our focus is on Controls and 4 CNS cell types from processed Lake gene expression data [46] due to the potential perturbations of TF-TF coordination in AD. NetREm predominantly excels over RTNduals in predicting experimentally validated TF-TF interactions, except in Control Mic. For instance, in Control Pericytes, 9.4% of NetREm's predicted TF-TF coordination network links, compared to 5.9% by RTNduals, have experimental validation [42]. This highlights how NetREm can potentially discover novel TF-TF interactions not currently represented in existing PPI databases, thereby contributing to ongoing efforts to broaden our understanding of PPIs[66].

For each of the 8 cell-types, we build an XGBoost model [67] to detect TGs with altered TF-TF B coordination scores from Control to AD stages that may predict TF-TF links annotated with neurodegenerative diseases in the Contextual PPI [18]. Input data consists of absolute value differences in maximum absolute value scaled B-matrix coordination scores across TGs for Control versus AD states, for the given cell type. Final data for each cell-type contains an equal number of TF-TF links in both classes for balanced classification (**Supp. File 5**). We evaluate performance via stratified 5-fold Cross Validation (CV) and provide metrics and average Receiver-Operator Characteristic (ROC) (**Fig. 5D)**. Besides ExNs, all cell types have strong average performance (area under ROC curve (AUC) $\geq$ 0.87, area under Precision-Recall Curve (AUPR) $\geq$ 0.89, accuracy $\geq$ 88.7%). We filter our final classifier models for the top 500 genes (based on high feature importance gains). **Fig. 5E** visualizes select gene enrichments for these TGs for the respective cell-types. We note the presence of several terms related to neurodegeneration, gliomas, neuropathies, cell death, autophagy (typically impaired in many neurodegenerative disorders and associated with the buildup of toxic aggregated proteins[68]), immune responses, cancers, learning/memory, cell-type-specific functions/disorders. Out of these top 500 genes, Astrocytes have 26 and 28 common genes with Oligodendrocytes (Oligo) and Oligo progenitor cells (OPCs), respectively; this supports recent findings on cross-talk and communication between Astrocytes and Oligo (often via cell-cell contact, cytokines, other signaling molecules) being core for glial development, onset and progression of neuroinflammatory diseases (e.g. AD, Schizophrenia), regeneration and repair of neurons [69].

In our exploration of the 16 TF-TG Regulatory Networks, we delve deep into commonalities and distinctiveness of TF-TG links and perform hierarchical clustering to unravel further intricate dynamics of regulatory behavior across the cell types and AD versus Controls. Typically, networks corresponding to a specific cell-type manifest remarkable coherence between conditions, delineating into 2 principal clusters for TF-TG Overlaps. The first cluster resonates with a convergence of AD and Control links, predominantly for InNs, Microglia, and ExNs, accentuating a pronounced overlap between Mic and ExNs. The second cluster encapsulates remaining cell types, revealing divergent behavior of AD and Control Oligo and distinct separation between AD and Control Pericytes. AD Pericytes share similar regulatory behavior with AD OPCs (23,157 shared TF-TG links) while Control Pericytes tend to cluster with Control Endothelial Blood-Brain Barrier (BBB) cells (18,871 shared TF-TG links, Jaccard Similarity (JS) = 0.09). Pericytes help maintain the BBB and facilitate neuroinflammatory processes[58], which may be central to AD progression[71].

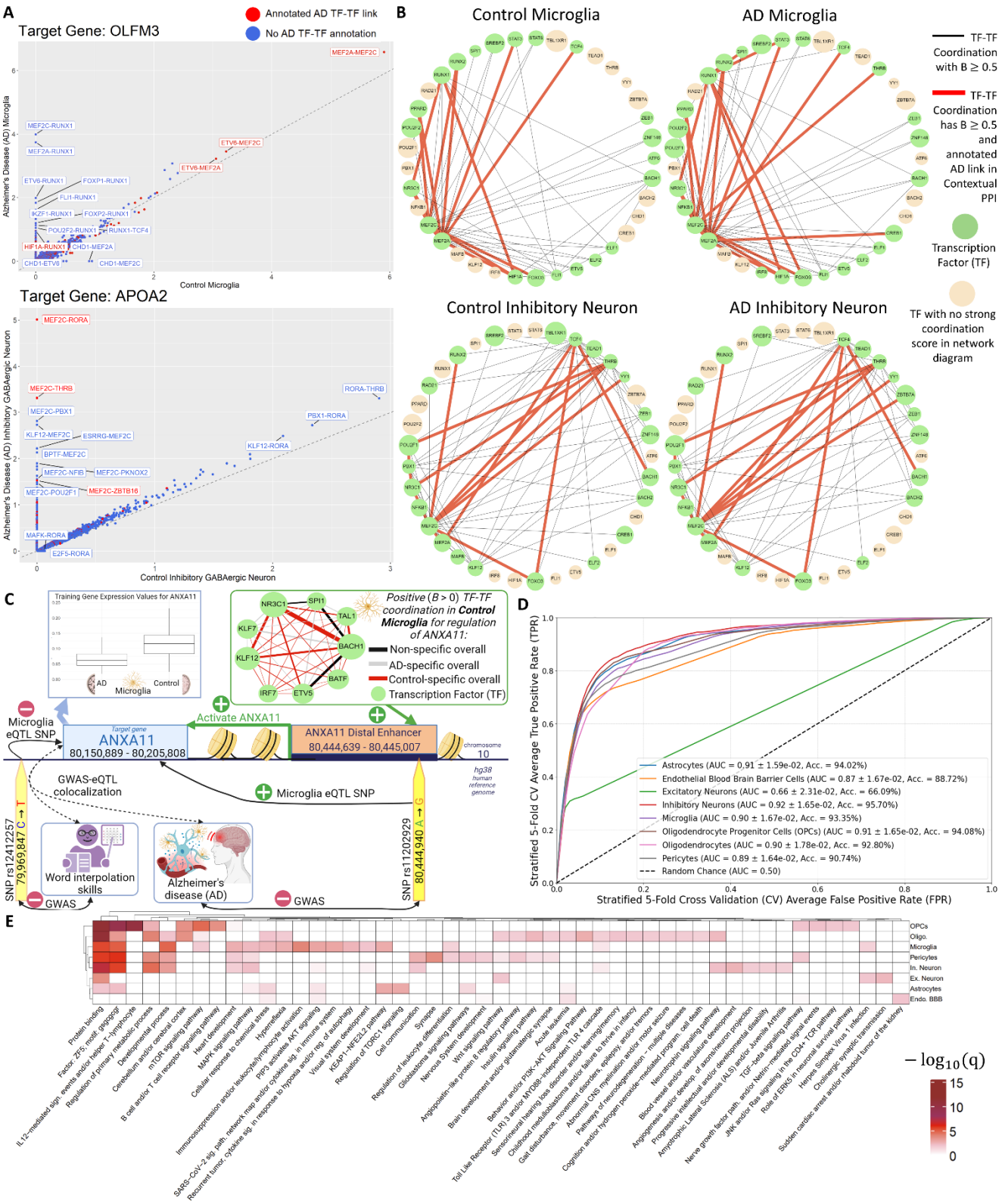


Figure 5 Prediction and comparative analysis of cell-type coordination among transcription factors (TFs) for gene regulation across neuronal and glial cell types in Alzheimer's disease (AD) **A)** Scatterplots comparing raw TF-TF links and corresponding B coordination scores for 2 target genes

(TGs), OLFM3 and APOA2, in Control and Alzheimer's Disease (AD) states for Microglia (Mic) and Inhibitory Neurons (InNs), respectively. Red points indicate TF-TF links annotated in the Contextual PPI [56] for AD, while blue points represent links not currently annotated for AD. The gray dashed line marks the 45-degree line. **B)** Circular network diagrams visualizing average TF-TF coordination (direct and/or indirect interactions) in AD/control Mic and InNs, focusing on select TFs denoted by light green nodes. These TFs exhibit strong and potentially cooperative coordination values ($\bar{B} \geq 0.5$). Red links are AD-related in the Contextual PPI[56]. TFs not expressed or lacking strong links in this filtered network are in peach. **C)** TF-TF coordination for 9 TFs that may cooperate to activate ANXA11 in NetREm's TF-TG regulatory network for in Control Mic (typically higher expression for ANXA11 than in AD). Edge widths for TF-TF coordination links are from original, raw, B coordination scores for ANXA11 regulation in Controls and colors are based on statistical significance of TF-TF links across all TGs in both Mic networks. Light gray links are significantly higher in AD; red links are significantly higher in Controls. Black links are not significant. One SNP correlates with decreased AD susceptibility, increases binding of this TF-TF network, and positively correlates with increased ANXA11 expression. A risk SNP for reduced cognitive ability (e.g. poor word interpolation skills) is linked to decreased ANXA11 expression. **D)** Average Stratified 5-fold Cross Validation (CV) Receiver Operator Characteristic (ROC) curves for XGBoost models designed to predict TF-TF links annotated in neurodegenerative diseases (class 1) in the Contextual PPI[56] based on balanced class data (equal # of links in class 0 and class 1). **E)** Heatmap of gene enrichment analysis terms for optimal genes identified by XGBoost models across 8 cell types, categorized into 4 groups. Hierarchical clustering is performed on rows (cell types) and columns (terms), and the heatmap shows $-\log_{10}(q)$ values.

Discussion

In this paper, we present NetREm, a computational approach to identify embeddings from network regularized regression of single-cell gene expression data to predict cell-type interactions of TFs for gene regulation. NetREm particularly enables a deeper understanding of how coordination among TFs contribute to target gene (TG) regulation at the cell-type level, a fundamental biological question that remains underexplored. Gene expression data is often nascent and limited for cell types and may not capture the entirety of the gene regulatory network (GRN) relationships between TFs and TGs. Hence, it is important to integrate additional omics data, such as protein-protein interaction (PPI) networks, into the regression problem of GRN prediction, which is precisely what NetREm accomplishes. NetREm uncovers instances of coordination among TFs, some which may be mediated by direct, physical PPI interactions and others that may be indirect. NetREm's TF-TF interaction networks (subsets of TF-TF coordination networks for TF-TF links with physical validation data) mostly tend to be cell-type-specific subnetworks of the original PPI, which may help characterize the original global PPI at the cell-type level, an on-going research problem [72–74]. Nonetheless, NetREm predicts and flags novel TF-TF links (i.e. remaining TF-TF coordination links) for follow-up studies to determine their true nature (i.e. whether direct or indirect interactions), which may assist computational efforts for finding undiscovered PPI links (i.e. link prediction problem) [75] and improve our understanding of indirect interactions among TFs that are involved in gene regulation.

As demonstrations, we apply NetREm to investigate the TF crosstalk involved in regulating TGs in myelinating (mSCs) and non-myelinating (nmSCs) Schwann cells (SCs), as well as in Alzheimer's

disease (AD) and control states in various neuronal and glial cells. We identify a variety of potential new TF-TF coordination links, especially in AD. In fact, disruptions in PPIs in neurons and glial cells may be pivotal in many neurobiological disorders as PPIs help mediate or regulate most functions of neurons [57]. Hopefully, the insights gained from approaches like NetREm may help pave the way for targeted therapies in neurodegenerative diseases and regenerative medicine. Overall, NetREm can be used in general to predict context-specific TF interactions across broader conditions such as various tissue types and disease types. Importantly, the applicability of NetREm extends beyond these specific cases to any discipline where an inherent network structure exists among predictors.

Currently, NetREm learns embeddings using singular value decomposition (SVD). In the future, we can integrate nonlinear dimensionality reduction into NetREm to reveal nonlinear and prior knowledge regularized latent representations. In addition to gene expression regression, NetREm can be extended to other emerging single-cell omics such as scATAC-seq to identify TF and chromatin interactions on open accessible regions at the cell-type level. Moreover, the networks for regularization can incorporate other important biological knowledge such as signaling pathways, which may shed light on more systematic insights of direct and indirect TF interactions (i.e., coordination among TFs) on gene regulation.

Methods and Materials

NetREm Methodology

Integrating multimodal data and networks in NetREm workflow

Preliminary Definitions:
We start with single-cell gene expression data for M cell samples and h genes, including g TGs $\{y_k\}_{k=1}^{g}$ and f potential Transcription Factors (TFs), where $g \leq h$ and $f < h$. Some TGs may also function as TFs. NetREm identifies $\{N_k\}_{k=1}^{g}$ candidate TFs for each TG, with $N_k \leq f$. We note that the value of N_k may vary for different TGs. For simplicity, we focus on explaining the pipeline for a single TG, denoted as y, and use N to represent its candidate TFs. These steps are applicable to all TGs in the cell-type and are run separately for each TG.

Optional Prior Knowledge on cell-type gene regulation:
Although optional, constructing a prior candidate Gene Regulatory Network (GRN) or link from TFs to TGs, enhances NetREm's performance via initial feature selection of biologically relevant TFs for the TG, reducing the number of TFs from f to N. In our applications, we use multi-omics data, specifically epigenetic data of open chromatin regions mapped to the TG, to identify potential regulatory elements that interact in various configurations (e.g., promoter-enhancer, promoter-promoter) or separately. Using databases of Position Weight Matrices (PWMs), we map TF motifs to these elements to predict TF Binding Sites (TFBSs). We thus link TFs to regulatory elements to the TG. Then, we prune this initial list of TFs via methods like relative gene expression values for TFs and motif matching scores.

To identify potentially novel or overlooked TFs, we augment this pruned TF list based on predicted TF-TF colocalization and Gene Ontology Molecular Function (GOMF) similarities[76]. Only TFs with

known links in our input Protein-Protein Interaction (PPI) network are added. This results in N candidate TFs for our TG as our prior knowledge base.

NetREm is then applied to identify which of these N TFs are predictive of the TG y expression, considering the network structure among the TFs. For these M cell samples, our X matrix contains single-cell expression data for these N TFs, while y contains expression data for our given TG. Each cell sample represents the gene expression profile of an individual cell. In bulk gene expression datasets for humans, for example, M can represent individual patients, and each sample can correspond to average gene expression values across cells of a specific cell-line or tissue. By following this comprehensive approach, NetREm effectively integrates multi-omics data and PPI networks to discover key regulatory TFs and TF-TF coordination events for each TG in a cell-type-specific manner.

Step 1: Network regularized regression

Given a gene expression dataset with M samples and N features represented as a matrix $X \in \mathbb{R}^{M \times N}$ (e.g. N candidate TF gene expression values for each sample) and a corresponding response vector $y \in \mathbb{R}^{M}$ (e.g. TG expression value for each sample), we want to learn a linear predictor $y \approx Xc^*$ for a given TG. The optimal coefficients $c^* \in \mathbb{R}^N$ are found by solving the following optimization problem (based on the formulation from [22]):

$$c^* = \underset{c}{\operatorname{argmin}}\, f(c) = \frac{1}{2M}\left|\left|y - Xc\right|\right|^2 + \alpha\left|\left|c\right|\right|_1 + \frac{\beta}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(\frac{c_i}{\sqrt{d_i}} - \frac{c_j}{\sqrt{d_j}}\right)^2 \qquad Eq\ (1)$$

Three terms are added together to form the objective function in equation $Eq\ (1)$. The first data-fitting term $\frac{1}{2M}\left|\left|y - Xc\right|\right|^2$ ensures that Xc is close to y, and $\frac{1}{2M}$ is a normalization factor to make it invariant to the sample size M. The second sparsity-prior term $\alpha\left|\left|c\right|\right|_1$ favors a sparse solution (a small number of non-zero entities) for c^*, which helps to simplify the model and boost model reliability. The third network-prior term $\frac{\beta}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(\frac{c_i}{\sqrt{d_i}} - \frac{c_j}{\sqrt{d_j}}\right)^2$ ensures that two TFs interact with each other to have similar weights in an undirected graph (our network prior). Each node i in this network prior graph has degree d_i and each edge has weight w_{ij} that indicates how much interaction nodes i and j have (e.g. strength of direct TF-TF interactions). In protein-protein interaction (PPI) networks, edges are given positive weights; this is our input PPI. We connect edges that are missing or unknown with an artificial default edge weight, so that the PPI network is fully connected. In $Eq\ (1)$, the network-constrained prior, $\beta \geq 0$, decides the strength of the network regularization penalty, while $\alpha \geq 0$ impacts the Lasso L1 regularization penalty (affecting sparsity of the optimal solution c^*).

Step 2: Gene embeddings from network regression

We have developed a novel method that transforms the original problem into a Lasso problem in a new space with cell-type-specific TF-TF interactions. First, we represent the network-prior term in a more compact matrix-vector form as:

$$\frac{\beta}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(\frac{c_i}{\sqrt{d_i}} - \frac{c_j}{\sqrt{d_j}}\right)^2 = \frac{\beta}{N^2}c^T A c$$

where $A \in \mathbb{R}^{N \times N}$ is a symmetric matrix that can be computed as:

$$A = D^T(W \cdot V)D,$$

where $D = diag\left(1/\sqrt{d}\right) \in \mathbb{R}^{N \times N}$ is a diagonal matrix with diagonal elements $\sqrt{d_i}$, $W \in \mathbb{R}^{N \times N}$ is the edge weight matrix with entities w_{ij}, $V = N - 11^T \in \mathbb{R}^{N \times N}$is a constant matrix where $1 \in \mathbb{R}^{N \times 1}$ a N dimensional all 1 column vector, and $W \cdot V$ is the element wise multiplication of W and V.

Using this matrix-vector representation, we can reformulate $f(c)$ in $Eq\ (1)$ as

$$\begin{aligned} f(c) &= \frac{1}{2M}\left|\left|y - Xc\right|\right|^2 + \alpha\left|\left|c\right|\right|_1 + \frac{\beta}{N^2}c^T Ac \\ &= \frac{1}{2M}c^T\left(X^T X + 2\beta\frac{M}{N^2}A\right)c - \frac{1}{M}y^T Xc + \alpha\left|\left|c\right|\right|_1 + \frac{1}{2M}y^T y \\ &= \frac{1}{2N}\left|\left|\tilde{y} - \tilde{X}c\right|\right|^2 + \alpha\left|\left|c\right|\right|_1 + \frac{1}{2M}y^T y - \frac{1}{2N}\tilde{y}^T\tilde{y}\,X, \qquad Eq\ (2) \end{aligned}$$

where $\tilde{X} \in \mathbb{R}^{N \times N}$ and is symmetric (i.e. $\tilde{X} = \tilde{X}^T$) and $\tilde{y} \in \mathbb{R}^N$ satisfies:

$$\frac{1}{N}\tilde{X}^T\tilde{X} = \frac{1}{M}X^T X + \frac{2\beta}{N^2}A \qquad Eq\ (3a)$$

$$\frac{1}{N}\tilde{y}^T\tilde{X} = \frac{1}{M}y^T X \qquad Eq\ (3b)$$

Finally, we can reformulate $Eq\ (1)$ as a conventional Lasso problem by omitting the constant term $\frac{1}{2M}y^T y - \frac{1}{2N}\tilde{y}^T\tilde{y}$, which we can solve using existing standard Lasso regression solvers:

$$c^* = \underset{c}{\operatorname{argmin}}\,\tilde{f}(c) = \frac{1}{2N}\left|\left|\tilde{y} - \tilde{X}c\right|\right|^2 + \alpha\left|\left|c\right|\right|_1 \quad Eq\ (4)$$

To compute $\tilde{X}$ and $\tilde{y}$ we perform a Singular Value Decomposition (SVD) on $X^T X + 2\beta\frac{M}{N^2}A \in \mathbb{R}^{N \times N}$:

$$X^T X + 2\beta\frac{M}{N^2}A = U\Sigma U^T,$$

And then compute

$$\tilde{X} = \sqrt{\frac{N}{M}}\Sigma^{\frac{1}{2}}U^T$$

$$\tilde{y} = \sqrt{\frac{N}{M}}\Sigma^{-\frac{1}{2}}U^T X^T y$$

Thus, X and y are transformed to a new latent space of gene embeddings that incorporate the PPI network information: $\tilde{X} \in \mathbb{R}^{N \times N}$ and $\tilde{y} \in \mathbb{R}^{N \times 1}$, respectively, in $Eq\ (4)$. We perform Lasso regression to solve $Eq\ (4)$ for $\tilde{X}$ and $\tilde{y}$, determining the optimal coefficients c^*.

Output 1: Identification of potential novel cell-type TFs in TF-TG regulatory network

All in all, NetREm constructs more comprehensive TF-TG regulatory networks (complementary to a gene regulatory network (GRN)) that identify the N^* optimal TFs (out of N candidates) that regulate

this TG, where $N^* \leq N$. This TF-TG regulatory network is weighted by Lasso regression coefficients c^*. The significance of each TF in the regulation of the TG is gauged by the absolute value of its coefficient. TF_i with coefficient $c_i^* > 0$ is predicted to activate the TG, functioning as an activator. Conversely, TFs with negative coefficients are anticipated to repress the TG, acting as repressors. NetREm has the potential to unearth novel, cell-type-specific TFs involved in TG regulation. These TF-TG regulatory links likely capture truer biological interactions among TFs. We establish these relationships through TF Binding Sites on regulatory elements where TFs are probable to bind.

To construct a cell-type-specific TF-TG regulatory network, we weave together individual TF-TG regulatory network links across all TGs in the cell type. Our comprehensive cell-type-specific TF-TG regulatory network relates TFs to regulatory elements to the TGs they regulate (Output 1). NetREm evaluates its predictive performance for each TG through various metrics, including Mean Square Error (MSE), normalized MSE, signal-to-noise ratio (SNR), and peak SNR (PSNR), in both training and testing data. We can further refine our predicted TF-TG regulatory network links by applying certain filters. For example, we may include only those TFs with absolute value coefficients above a minimum threshold (c_{min}, default = 0) and TGs that meet specific model performance criteria.

Output 2: Cell-type-specific TF-TF interactions

The 2nd output is a weighted TG-specific TF-TF coordination network given by an adjacency matrix of coordination scores B. When applying the aforementioned procedure to transform the original single-cell gene expression dataset X, y to the new dataset $\tilde{X}, \tilde{y}$, we embed the network-prior (e.g. TF-TF interaction) in the new space by using $Eq\ (3a)$: $\frac{1}{N}\tilde{X}^T\tilde{X} = \frac{1}{M}X^TX + \frac{2\beta}{N^2}A$ and generate a new inner-product space that depicts TF-TF coordination particular to a certain cell type. To be more precise, two terms are combined to generate the new normalized inner product space denoted by $\frac{1}{N}\tilde{X}^T\tilde{X}$. The first term $\frac{1}{M}X^TX$ represents the original normalized inner product space from column vectors $x_1, x_2, \ldots, x_N$ of X, where $x_i \in R^M$ represents the ith TF representation (i.e. TF_i) in the M samples for a specific cell type. The second term $A = D^T(W \cdot V)D$ purely depends on the TF-TF interaction weights and is cell-type independent, which can be retrieved from a conventional database of protein-protein interaction (PPI) networks. By combining the cell-type-specific data term $\frac{1}{M}X^TX$ with the cell-type-independent TF-TF interaction term $\frac{2\beta}{N^2}A$, the new normalized inner product space $B = \frac{1}{N}\tilde{X}^T\tilde{X}$ encodes the TG-specific TF-TF coordination scores for that given target gene (TG) in the cell-type. Thus, the TG-specific coordination score between any two TFs, TF_i and TF_j, for a TG can be calculated as $B_{ij} = \frac{\tilde{x_i}^T\tilde{x_j}}{N}$. Hence, our gene embeddings learned by NetREm for TFs, our $\tilde{X}$, can be utilized to build a TF-TF coordination network showing how TFs relate to each other in a network, in terms of a B matrix, to regulate that given TG (i.e. weighted and signed TG-specific TF-TF coordination).

Our input protein-protein interaction (PPI) network generally attributes positive weights to all protein pairs, even those with antagonistic, competing roles in biological pathways [77]. Additionally, the PPI network encompasses direct global interactions among TFs, not necessarily cell-type-specific ones. To address this, we predict the nature of TF interactions as either cooperative or antagonistic at the cell-type-level. In our framework, $B_{ij} > 0$ suggests potential cooperativity, such as co-binding, between TF_i and TF_j. Conversely, $B_{ij} < 0$ indicates potential antagonism or competition. $B_{ij} = 0$ means neither TF_i nor TF_j is among the N candidate TFs for the given TG. For each TG, we identify up to $N^2 - N$

non-zero TF-TF links based on initial feature selection. While some TFs with non-zero B values may not appear in the final TF-TG regulatory network for a TG (i.e. have a regression coefficient c^* with absolute value $\leq c_{min}$), we still include them in our analysis; all N candidates contribute to network regularization and gene embeddings, and may indirectly influence TG regulation. We create a weighted and signed cell-type-specific network of coordination among TFs by averaging these B values across all TGs in the cell type to obtain $\bar{B}$.

In general, coordination among TFs for gene regulation can be direct (physical interactions) or indirect (i.e. TFs may not necessarily interact physically). Novel TF-TF links are not found in our input PPI; the TF pairs in these novel TF-TF links may interact directly or indirectly and follow-up studies and further exploration are needed to determine the true nature of their coordination. In cases where the TF-TF links are found in the input global PPI (i.e. there is documented ground truth data of physical interactions among these 2 TFs) with weight > default edge weight, we refer to those as TF-TF interactions (i.e. direct coordination among the TFs); these TF-TF interactions may serve as cell-type-specific and/or disease-specific subnetworks of our PPI. We only consider the TF-TF links in the PPI with weight w > default edge weight as being indicative of true direct relations among predictors.

To clarify further: When we mention the input PPI, we refer to the original PPI of direct, experimentally validated TF-TF interactions with weight w > default edge weight (that we assign to missing links in input PPI); this is our input into NetREm and what we consider as biologically meaningful. Ultimately, NetREm will add artificial, default edge weights to any missing links in this input PPI to make this a fully-connected network for the network regularized regression and to enable the discovery of novel TF-TF coordination links. Thus, novel TF-TF coordination links identified by NetREm are a subset of these TF-TF edges that are missing in the input PPI, but added by NetREm with weight w = default.

We also employ an integrated Contextual PPI database [56] of dummy (0 = absent, 1 = present) values, annotated for various diseases and tissues, as a key validation tool for our TF-TF links. For our machine learning analysis tasks, we utilize raw B coordination scores of TF-TF links for each TG in a matrix and then perform maximum absolute value scaling (since B values may also be negative), where we divide B coordination scores for each TG by the maximum absolute value for that TG. We thus have a matrix of scaled TF-TF link coordination scores (i.e. embeddings) as rows, for the respective TG as columns.

Overall, NetREm provides a holistic approach to understanding cell-type-specific gene regulatory mechanisms and enriches our understanding of core cell-type interactions (direct and indirect) among TFs that are involved in these processes. Please see **Supplementary Methods and Materials** (**Supp. File 5**) for details on the software implementation of NetREm.

Datasets and Pre-processing

We test NetREm on 2 different applications using multiomics datasets for nervous system-related cell-types for humans:

(1) Myelinating (mSCs) and non-myelinating (nmSCs) Schwann cells (SCs) in the peripheral nervous system (PNS).
(2) Alzheimer's disease and Control stages for 8 cell-types in the central nervous system (CNS). 4 Glial cells: Astrocytes, Oligodendrocytes (Oligo), Oligodendrocyte Progenitor cells (OPCs);

Microglia (Mic); 2 Neuronal cells: GABA-ergic Inhibitory neurons (InNs), Glutamatergic Excitatory neurons (ExNs); 2 Vascular and Blood-Brain Barrier (BBB) cells: Pericytes, Endothelial BBB cells.

Supplementary Methods (**Supp. File 5**) has more details on the methods used for both applications.

Single-cell gene expression data

Schwann Cells: We utilize Seurat [78] to normalize gene expression datasets for 13,886 genes in mSCs and nmSCs in the Dorsal Root Ganglion (DRG) L4,5 regions for 5 human donors [35]. We include only protein-coding genes and remove genes not expressed in at least 3 cells in the total data (comprised of other cell types). Our final data has 17,049 genes for 319 cells in mSCs (training: 223 cells, testing: 96 cells) and 2,468 cells in nmSCs (training: 1,727 cells, testing: 741 cells).

Alzheimer's disease (AD): We use the processed data for 24 individuals with AD pathology and 24 healthy controls provided by Gupta et. al [79] for all 8 neuronal/glial cell types. This data is from the original AD study [12] of 80,660 droplet-based single-nucleus transcriptomes (i.e. the cells) for 17,926 genes from the prefrontal cortex for 8 neuronal/glial cell types.

Validation data: We employ these 2 datasets to evaluate our NetREm models in our applications.

1. Lake et al. [46] provides data for 4 CNS cell types: ExNs, InNs, Mic, Oligo. This data undergoes preprocessing according to methods outlined in Jin et al. [21], which includes removing genes expressed in <100 cells, normalization and scaling via Seurat [78], imputing single-cell expression using the MAGIC algorithm. The final data comprises 15,694 genes across 13,709 ExNs, 6,045 InNs, 317 Microglia, and 2,657 Oligodendrocyte cells.
2. GTEx data from Eraslan et al. [38, 45] for SCs pooled from 5 tissues (esophagus mucosa/muscularis, heart, prostate, skeletal muscle). After applying Seurat [78] and MAGIC normalization, this dataset contains 14,145 genes across 1,430 SCs.

Position Weighted Matrices (PWMs) of TF Binding Motifs

We obtain PWMs for humans from R [80] Bioconductor [81] databases. Each motif is specific to 1 TF and a TF may have many motifs. When running motifmatchr [82] to search for motif matches for TFs, we use JASPAR2022 [83] for humans (species = 9606) and convert Position Frequency Matrices (PFMs) to PWMs using TFBSTools (toPWM() function, type = "prob"[84]). We have 949 PWMs corresponding to 680 TFs. To run monaLisa [85] for motif analysis and TF binding prediction, we use a comprehensive database of 8,903 PWMs (last update: March 4, 2022). We use MotifDb [86] for "hsapiens" to obtain 6,086 PWMs and motifbreakR's motifbreakR_motif data [87] to add 2,817 PWMs. We remove redundant motifs (found in both sources) for TFs when running monaLisa for our analysis.

Prior knowledge on gene regulation

Prior gene regulatory networks (GRNs)

To construct prior candidate TF-TG regulatory links (prior reference GRNs), we use a multi-step approach based on the available multiomics data. We explain our steps below and in **Supp. Methods and Materials (Supp. File 5)** as a potential guide. We start by identifying raw scATAC-seq peaks (regions of open and accessible DNA chromatin) in adult cell-types [88] relevant to our study. We map these peaks to TGs using various methods:

1. Overlapping scATAC-seq peaks with cis-Candidate Regulatory Elements (cCRE) [88] that are already mapped to TGs. We utilize all interacting cCRE regions for that cell-type (e.g. promoter-promoter, promoter proximal – promoter, distal (enhancer) – promoter) with Activity by Contact (ABC) scores.
2. Utilizing TxDb.Hsapiens.UCSC.hg38.knownGene [89, 90] to define default promoter-only regions. For Oligo, Mic, and Neurons (ExNs and/or InNs), we also utilize Plaq-seq data[91] on promoter-enhancer interactomes and enhancer regions in hg19. We run scGRNom[21] Step 1 (scGRNom_interaction) on this epigenomic data to predict all possible interactions between enhancers and promoters and annotate them with the associated TGs. We use rtracklayer [92] and the hg19tohg38.over.chain file from UCSC [95] to perform liftover [96] to map the associated enhancer and promoter regions to hg38 human genomic coordinates. We utilize these annotated interaction enhancer and promoter regions as Plaq-seq-based regulatory regions for the respective cell-types.
3. Characterizing the remaining unmapped peaks via expression quantitative trait loci (eQTL) data, which links single nucleotide polymorphisms (SNPs) to TGs whose expression levels they impact. We use GTEx v8 Tibial Nerve eQTL data for SCs [38, 39] and eQTL data from Byrois et al. for brain cells for our AD application [97].

With these mapped scATAC-seq regions, we predict TF binding using motifmatchr [82] and monaLisa tools [85], as well as our respective PWM databases. These TFs are predicted to recognize their sequence-specific motifs on TFBSs of various regulatory elements like promoters, enhancers, or eQTL-based regions, thereby regulating the expression of their respective TGs. Lastly, we refine the list of TFs for each TG based on pruning by motif-matching scores and TF relative expression levels.

Colocalization among Transcription Factors (TFs)
Next, for this pruned TF list, we add highly-similar TFs, combining the below 4 approaches, checking that these TF pairs have known PPI links (above our minimum weight of 0.01). Ultimately, this results in N candidate TFs for the TG. In general, we assume that a given motif binds to a single TF motif at a time. The only exception to this is when we have dimeric transcription factors (e.g. AP-1, JUN/FOS), which can be heterodimers of 2 different TFs.

<u>Approach 1:</u> We identify scATAC-seq peaks unique for that cell-type, ensuring no significant overlap with peaks from other cell-types [88]. See **Supp. File 5** for details. Using monaLisa [85], we predict TF binding within these specific peaks, noting the TFs and their binding locations. For each TF pair, a score of 1 is assigned per peak if they have non-overlapping potential TFBSs, suggesting cooperativity. Aggregate scores for all TF pairs are calculated across the peaks. Finally, we scale these pairs using a Jaccard Similarity (JS) index, with higher scores indicating greater colocalization. Here, colocalization is 2 TFs that bind within the same enhancer but not to the same motif.

$$JS(TF_i, TF_j) = \frac{\#\text{ of peaks with potential non} - \text{overlap in binding for } TF_i \text{ and } TF_j}{\#\text{ of peaks bound by } TF_i \frac{\text{and}}{\text{or}} TF_j}$$

<u>Approach 2:</u> We run motifmatchr [82] on all raw scATAC-seq peaks for the cell types and identify TFs with strong binding for each peak. Next, we calculate JS between any 2 TFs as:

$$JS(TF_i, TF_j) = \frac{\#\text{ of peaks bound by } TF_i \text{ and } TF_j}{\#\text{ of peaks bound by } TF_i \frac{\text{and}}{\text{or}} TF_j}$$

Then, we consider TFs with a high JS score as being very similar to each other (e.g. potentially sharing the same TF family of motifs).

Approach 3: We use data on known TF complexes (i.e. TF pairs with weight 1 in comprehensive PPI).
Approach 4: We employ clusterProfiler [76] to determine TF similarity based on Molecular Function (MF). Using the godata() function on the org.Hs.eg.db database[98], we generate semantic similarity data for Gene Ontology (GO) Molecular Function (GOMF). We then apply the mgeneSim() function with default settings, utilizing the Wang measure and BMA approach. This method helps infer genetic interactions and PPIs, as closely-related TFs often exhibit similar behavior [99].

In **Supp. Methods and Materials (Supp. File 5)**, we detail, step-by-step, how we integrate these data sources (along with others) to help determine the list of N candidate TFs for each target gene (TG, our y), which will be part of our input X when running NetREm for each TG in the given cell-type.

Protein-Protein Interaction (PPI) Networks

We construct a comprehensive human PPI network of direct interactions, with weights $w > 0$, from multiple sources. STRING v12 [77] provides 13.2 million links, whose average combined scores are scaled between 0 and 1. Hu.MAP 2.0 [100] contributes 17.5 million pairs with probability weights, which we used as weights. Other data, including protein complexes and PPIs, comes from MaayanLab [101], Human Reference Interactions (HuRI) [102], TF complexes in: PWMs, Contextual PPI [56], KEGG Pathway Networks [103] we web-scraped. These sources are integrated into our final PPI network, which has 31.7 million links. TFs found in complexes with each other in any source are assigned a weight of 1. For TF-TF links found in multiple sources, we use the highest weight for that link. In our applications, the PPI is filtered to include relevant TFs and links with edge weight w > 0.01. This is our final input PPI, our ground truth network, where links may correspond to direct, physical TF-TF interactions.

When we run NetREm, we assign a default edge weight of 0.01 to any TF links with missing edges to ensure a fully-connected network for NetREm's regularized regression and to help guide the discovery of novel TF-TF coordination links; nonetheless, the input PPI refers to the TF-TF links in the PPI with edge weight > 0.01. More details are in **Supp. Methods and Materials in Supp. File 5**.

NetREm Parameters and Evaluation

See **Supp. File 5** for more details on parameters we utilize for our applications (and hyperparameter tuning) as well as our evaluation approaches. We measure the performance of our NetREm models in various ways. These include running benchmark models (e.g. LassoCV[26], Linear Regression[24], RidgeCV[27], ElasticNetCV[25]) on the same input data. We utilize single nucleotide polymorphisms (SNPs) related to disease/traits in Genome-Wide Association Studies (GWAS) and/or to changes in gene expression via expression quantitative trait loci (eQTL) analysis. These SNPs are mapped onto our TF-TG regulatory networks using impacted TF Binding Sites and regulatory elements for TGs. Further, we perform gene enrichment analysis of TGs and relative gene expression percentile analysis. We have obtained experimental data for 7 core TFs (EGR2, NR2F2, RXRG, SOX10, SREBF1, TEAD1, YY1) in Schwann cells. In addition, we utilize a Contextual PPI database with annotated TF-TF links[56] associated with various diseases/traits. Finally, we utilize a recent study [53] on experimentally-validated physical interactions among TFs in humans as a gold standard to

evaluate the performance of NetREm to that of a comparable tool, RTNduals[55], on input gene expression data and a list of TFs for a given cell-type.

Availability of data and materials

We implement NetREm as an open-source software package in Python, available on Github: https://github.com/SaniyaKhullar/NetREm. Our analysis is done for human reference genome hg38 (GRCh38). The mapping of cis-Candidate Regulatory Elements (cCREs) for hg38 human reference genome coordinates is available here: http://catlas.org/catlas_downloads/humantissues/cCRE_hg38.tsv.gz. Activity by Contact (ABC) scores for interacting cCRE regulatory elements is found here: http://catlas.org/catlas_downloads/humantissues/ABC_scores/. Raw single-cell scATAC-seq peaks are available here: http://catlas.org/catlas_downloads/humantissues/Peaks/RAW/. Single-cell gene expression data for the Dorsal Root Ganglion (DRG) is available in Gene Expression Omnibus (GEO) with the accession number GSE169301. Tracks for SOX10 binding in Peripheral Nervous System (PNS) in rats is available in GEO with the accession number GSE64703 [43]. We have created a corresponding UCSC Genome Browser session for visualizing these SOX10 rat tracks here: https://genome.ucsc.edu/s/saniyaKhullar/updated_rn5. Some data for EGR2 have been deposited in NCBI GEO: GSE35132. We have provided our final imputed gene expression data for AD versus controls here: http://tiny.cc/ad_control_netrem. Raw depositions for ChIP-seq data for H3K27ac and Cut&Run data for YY1, RXRG, and TEAD1 are in progress at Gene Expression Omnibus (GEO).

Supplementary Data

- Supplementary File 1: Schwann Cell Transcription Factor – Target Gene interaction (|coef| >=0.05)
- Supplementary File 2: Schwann Cell Transcription Factor (TF) – TF coordination (|average B| >= 0.01, |coef| >= 0.025)
- Supplementary File 3: Cell-type TF-TG networks across AD and Control (|coef| >=0.05, test mean square error (MSE) <= 1)
- Supplementary File 4: Cell-Type TF-TF coordination across AD and Control (|average B| >= 0.01, test MSE <= 1)
- Supplementary File 5: Supplementary Figures, Tables, Methods/Materials

Acknowledgements

The authors are grateful for feedback and advice from Kalpana Hanthanan Arachchilage, Ph.D., Jerome Choi, Noah Cohen Kalafut, Sean Chang, and Harrison Pantera for this project. In addition, the authors appreciate feedback from Renu Poochie Khullar and Armaan Khullar for the manuscript.

Conflict of Interest

None declared.

Funding

National Institutes of Health, R01AG067025, RF1MH128695, R03NS123969, R21NS127432, R21NS128761 to D.W. and P50HD105353 to Waisman Center. S.K. was supported by an NLM training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM 5T15LM007359).

Contributions

D.W. conceived and designed the study. S.K., X.H., and D.W. built the methodology for NetREm. S.K., X.H., R.R., J.S., and D.W. contributed to writing the manuscript. S.K. and X.H. wrote the open-source code for NetREm. R.R. and J.S. ran experiments in rats and provided data for validation for Schwann cells. S.K., J.S., and D.W. analyzed the data. All authors read and approved the manuscript.

References

1. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018;172:650–65.
2. Guo Y, Gifford DK. Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. BMC Genomics. 2017;18:45.
3. Nie Y, Shu C, Sun X. Cooperative binding of transcription factors in the human genome. Genomics. 2020;112:3427–34.
4. Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions | Nature Communications. https://www.nature.com/articles/s41467-019-13888-7. Accessed 27 Mar 2023.
5. Detection of cooperatively bound transcription factor pairs using ChIP-seq peak intensities and expectation maximization - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6049898/. Accessed 27 Mar 2023.
6. Zhao Y. TFSyntax: a database of transcription factors binding syntax in mammalian genomes. Nucleic Acids Res. 2023;51:D306–14.
7. TF-COMB – Discovering grammar of transcription factor binding sites - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2001037022003051. Accessed 27 Mar 2023.
8. Lee B-K, Bhinge AA, Battenhouse A, McDaniell RM, Liu Z, Song L, et al. Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. Genome Res. 2012;22:9–24.
9. D'Alessio AC, Fan ZP, Wert KJ, Baranov P, Cohen MA, Saini JS, et al. A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity. Stem Cell Rep. 2015;5:763–75.
10. Mirny LA. Nucleosome-mediated cooperativity between transcription factors. Proc Natl Acad Sci. 2010;107:22534–9.
11. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13:613–26.
12. Single-cell transcriptomic analysis of Alzheimer's disease | Nature. https://www.nature.com/articles/s41586-019-1195-2. Accessed 27 Mar 2023.

13. Comprehensive functional genomic resource and integrative model for the human brain | Science. https://www.science.org/doi/10.1126/science.aat8464?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed. Accessed 27 Mar 2023.
14. Karczewski KJ, Tatonetti NP, Landt SG, Yang X, Slifer T, Altman RB, et al. Cooperative transcription factor associations discovered using regulatory variation. Proc Natl Acad Sci. 2011;108:13353–8.
15. Know Thy Friends, know ThyselF – Transcription Factor (TF) regulation resolved | Springer Nature Protocols and Methods Community. https://protocolsmethods.springernature.com/posts/know-thy-friends-know-thyself-transcription-factor-tf-regulation-resolved. Accessed 27 Mar 2023.
16. Hannenhalli S, Levy S. Predicting transcription factor synergism. Nucleic Acids Res. 2002;30:4278–84.
17. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data | Nature Methods. https://www.nature.com/articles/s41592-019-0690-6. Accessed 27 Mar 2023.
18. SCENIC: single-cell regulatory network inference and clustering | Nature Methods. https://www.nature.com/articles/nmeth.4463. Accessed 27 Mar 2023.
19. Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics. 2019;35:2159–61.
20. Single-cell chromatin state analysis with Signac | Nature Methods. https://www.nature.com/articles/s41592-021-01282-5. Accessed 27 Mar 2023.
21. Jin T, Rehani P, Ying M, Huang J, Liu S, Roussos P, et al. scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. Genome Med. 2021;13:95.
22. Network-constrained regularization and variable selection for analysis of genomic data | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/24/9/1175/206444. Accessed 27 Mar 2023.
23. Wang Y, Hu X, Jiang X, He T, Yuan J. Predicting microbial interactions by using network-constrained regularization incorporating covariate coefficients and connection signs. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2015. p. 635–8.
24. sklearn.linear_model.LinearRegression. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.LinearRegression.html. Accessed 22 Aug 2023.
25. sklearn.linear_model.ElasticNetCV. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.ElasticNetCV.html. Accessed 22 Aug 2023.
26. sklearn.linear_model.LassoCV. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.LassoCV.html. Accessed 22 Aug 2023.
27. sklearn.linear_model.RidgeCV. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.RidgeCV.html. Accessed 22 Aug 2023.
28. Ma KH, Svaren J. Epigenetic Control of Schwann Cells. The Neuroscientist. 2018;24:627–38.
29. Disentangling glial diversity in peripheral nerves at single-nuclei resolution | Nature Neuroscience. https://www.nature.com/articles/s41593-021-01005-1. Accessed 24 Oct 2023.
30. Gerber D, Pereira JA, Gerber J, Tan G, Dimitrieva S, Yángüez E, et al. Transcriptional profiling of mouse peripheral nerves to the single-cell level to build a sciatic nerve ATlas (SNAT). eLife. 2021;10:e58591.
31. Variation in SIPA1L2 is correlated with phenotype modification in Charcot– Marie– Tooth disease type 1A - Tao - 2019 - Annals of Neurology - Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/ana.25426. Accessed 14 Jul 2023.

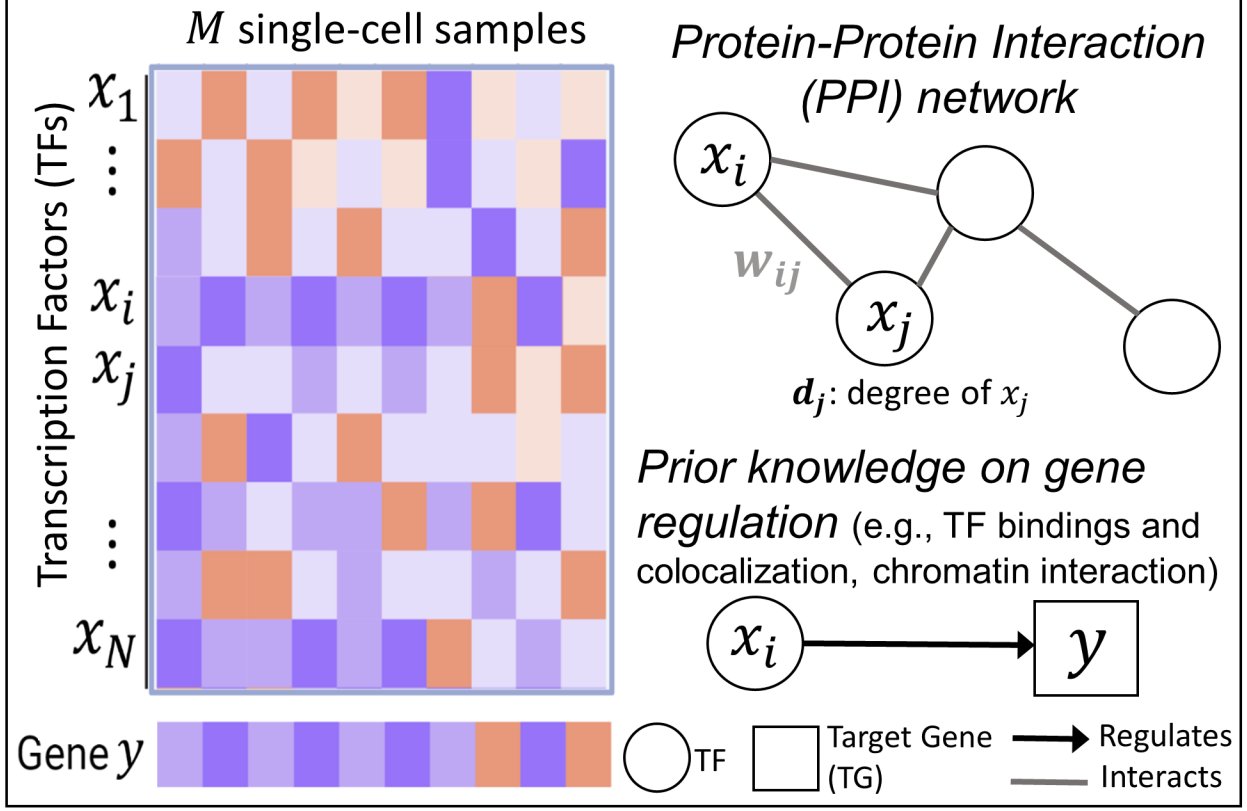
32. Srinivasan R, Sun G, Keles S, Jones EA, Jang S-W, Krueger C, et al. Genome-wide analysis of EGR2/SOX10 binding in myelinating peripheral nerve. Nucleic Acids Res. 2012;40:6449–60.
33. Lopez-Anido C, Poitelon Y, Gopinath C, Moran JJ, Ma KH, Law WD, et al. Tead1 regulates the expression of Peripheral Myelin Protein 22 during Schwann cell development. Hum Mol Genet. 2016;25:3055–69.
34. Denarier E, Forghani R, Farhadi HF, Dib S, Dionne N, Friedman HC, et al. Functional Organization of a Schwann Cell Enhancer. J Neurosci. 2005;25:11210–7.
35. Avraham O, Chamessian A, Feng R, Yang L, Halevi AE, Moore AM, et al. Profiling the molecular signature of satellite glial cells at the single cell level reveals high similarities between rodents and humans. PAIN. 2022;163:2348.
36. Balakrishnan A, Belfiore L, Chu T-H, Fleming T, Midha R, Biernaskie J, et al. Insights Into the Role and Potential of Schwann Cells for Peripheral Nerve Repair From Studies of Development and Injury. Front Mol Neurosci. 2021;13:608442.
37. Lopez-Anido C, Sun G, Koenning M, Srinivasan R, Hung HA, Emery B, et al. Differential Sox10 genomic occupancy in myelinating glia. Glia. 2015;63:1897–914.
38. GTEx Portal. https://www.gtexportal.org/home/datasets. Accessed 22 Aug 2023.
39. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
40. YAP and TAZ control peripheral myelination and the expression of laminin receptors in Schwann cells | Nature Neuroscience. https://www.nature.com/articles/nn.4316. Accessed 7 Sep 2023.
41. Qi J, Ma L, Guo W. Recent advances in the regulation mechanism of SOX10. J Otol. 2022;17:247–52.
42. Activation of Krox20 gene expression by Sox10 in myelinating Schwann cells.
43. GEO Accession viewer. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64703. Accessed 14 Jul 2023.
44. Pan UKBB | Pan UKBB. https://pan.ukbb.broadinstitute.org/. Accessed 25 Aug 2023.
45. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function | Science. https://www.science.org/doi/10.1126/science.abl4290. Accessed 1 Jun 2023.
46. Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol. 2018;36:70–80.
47. Svaren J, Meijer D. The molecular machinery of myelin gene transcription in Schwann cells. Glia. 2008;56:1541–51.
48. Potter GK, Feldman JS. Neoplasms of the peripheral nervous system. Clin Podiatr Med Surg. 1990;7:141–9.
49. Stratton JA, Kumar R, Sinha S, Shah P, Stykel M, Shapira Y, et al. Purification and Characterization of Schwann Cells from Adult Human Skin and Nerve. eNeuro. 2017;4.
50. Martini R, Schachner M. Immunoelectron microscopic localization of neural cell adhesion molecules (L1, N-CAM, and MAG) and their shared carbohydrate epitope and myelin basic protein in developing sciatic nerve. J Cell Biol. 1986;103 6 Pt 1:2439–48.
51. Joukov V, De Nicolo A, Rodriguez A, Walter JC, Livingston DM. Centrosomal protein of 192 kDa (Cep192) promotes centrosome-driven spindle assembly by engaging in organelle-specific Aurora A activation. Proc Natl Acad Sci. 2010;107:21022–7.
52. Choi E, Zhang X, Xing C, Yu H. Mitotic Checkpoint Regulators Control Insulin Signaling and Metabolic Homeostasis. Cell. 2016;166:567.
53. Human transcription factor protein interaction networks | Nature Communications. https://www.nature.com/articles/s41467-022-28341-5. Accessed 3 Sep 2023.

54. Stevanovic M, Drakulic D, Lazic A, Ninkovic DS, Schwirtlich M, Mojsin M. SOX Transcription Factors as Important Regulators of Neuronal and Glial Differentiation During Nervous System Development and Adult Neurogenesis. Front Mol Neurosci. 2021;14:654031.
55. Chagas VS, Groeneveld CS, Oliveira KG, Trefflich S, de Almeida RC, Ponder BAJ, et al. RTNduals: an R/Bioconductor package for analysis of co-regulation and inference of dual regulons. Bioinformatics. 2019;35:5357–8.
56. IID 2021: towards context-specific protein interaction analyses by increased coverage, enhanced annotation and enrichment analysis | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/50/D1/D640/6424757. Accessed 14 Jul 2023.
57. Mathew B, Bathla S, Williams KR, Nairn AC. Deciphering Spatial Protein–Protein Interactions in Brain Using Proximity Labeling. Mol Cell Proteomics. 2022;21.
58. Mao Y, Fisher DW, Yang S, Keszycki RM, Dong H. Protein-protein interactions underlying the behavioral and psychological symptoms of dementia (BPSD) and Alzheimer's disease. PLOS ONE. 2020;15:e0226021.
59. Vargas DM, De Bastiani MA, Zimmer ER, Klamt F. Alzheimer's disease master regulators analysis: search for potential molecular targets and drug repositioning candidates. Alzheimers Res Ther. 2018;10:59.
60. Mirsaeidi M, Gidfar S, Vu A, Schraufnagel D. Annexins family: insights into their functions and potential role in pathogenesis of sarcoidosis. J Transl Med. 2016;14:89.
61. Wang Y, Duan X, Zhou X, Wang R, Zhang X, Cao Z, et al. ANXA11 mutations are associated with amyotrophic lateral sclerosis–frontotemporal dementia. Front Neurol. 2022;13:886887.
62. Smith BN, Topp SD, Fallini C, Shibata H, Chen H-J, Troakes C, et al. Mutations in the vesicular trafficking protein annexin A11 are associated with amyotrophic lateral sclerosis. Sci Transl Med. 2017;9:eaad9157.
63. Khullar S, Wang D. Predicting brain-regional gene regulatory networks from multi-omics for Alzheimer's disease phenotypes and Covid-19 severity. Hum Mol Genet. 2023;32:1797–813.
64. Rustenhoven J, Smith AM, Smyth LC, Jansson D, Scotter EL, Swanson MEV, et al. PU.1 regulates Alzheimer's disease-associated genes in primary human microglia. Mol Neurodegener. 2018;13:44.
65. Eyigoz E, Mathur S, Santamaria M, Cecchi G, Naylor M. Linguistic markers predict onset of Alzheimer's disease. eClinicalMedicine. 2020;28.
66. Yu D, Chojnowski G, Rosenthal M, Kosinski J. AlphaPulldown—a python package for protein–protein interaction screens using AlphaFold-Multimer. Bioinformatics. 2023;39:btac749.
67. Python API Reference — xgboost 1.7.6 documentation. https://xgboost.readthedocs.io/en/stable/python/python_api.html. Accessed 22 Aug 2023.
68. Rai SN, Tiwari N, Singh P, Mishra D, Singh AK, Hooshmandi E, et al. Therapeutic Potential of Vital Transcription Factors in Alzheimer's and Parkinson's Disease With Particular Emphasis on Transcription Factor EB Mediated Autophagy. Front Neurosci. 2021;15:777347.
69. Nutma E, van Gent D, Amor S, Peferoen LAN. Astrocyte and Oligodendrocyte Cross-Talk in the Central Nervous System. Cells. 2020;9:600.
70. Brown LS, Foster CG, Courtney J-M, King NE, Howells DW, Sutherland BA. Pericytes and Neurovascular Function in the Healthy and Diseased Brain. Front Cell Neurosci. 2019;13.
71. Heneka MT, Carson MJ, El Khoury J, Landreth GE, Brosseron F, Feinstein DL, et al. Neuroinflammation in Alzheimer's Disease. Lancet Neurol. 2015;14:388–405.
72. Johnson KL, Qi Z, Yan Z, Wen X, Nguyen TC, Zaleta-Rivera K, et al. Revealing protein-protein interactions at the transcriptome scale by sequencing. Mol Cell. 2021;81:4091-4103.e9.

73. Murtaza N, Cheng AA, Brown CO, Meka DP, Hong S, Uy JA, et al. Neuron-specific protein network mapping of autism risk genes identifies shared biological mechanisms and disease-relevant pathologies. Cell Rep. 2022;41:111678.
74. Hsu Y-HH, Nacu E, Liu R, Pintacuda G, Kim A, Tsafou K, et al. Using brain cell-type-specific protein interactomes to interpret genetic data in schizophrenia. 2022;:2021.10.07.21264568.
75. Singh KV, Vig L. Improved prediction of missing protein interactome links via anomaly detection. Appl Netw Sci. 2017;2:2.
76. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021;2.
77. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51:D638–46.
78. Tools for Single Cell Genomics. https://satijalab.org/seurat/. Accessed 25 Aug 2023.
79. Gupta C, Xu J, Jin T, Khullar S, Liu X, Alatkar S, et al. Single-cell network biology characterizes cell type gene regulation for drug repurposing and phenotype prediction in Alzheimer's disease. PLOS Comput Biol. 2022;18:e1010287.
80. R: The R Project for Statistical Computing. https://www.r-project.org/. Accessed 25 Aug 2023.
81. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80.
82. motifmatchr.knit. https://bioconductor.org/packages/release/bioc/vignettes/motifmatchr/inst/doc/motifmatchr.html. Accessed 25 Aug 2023.
83. JASPAR2022. Bioconductor. http://bioconductor.org/packages/JASPAR2022/. Accessed 25 Aug 2023.
84. Tan G, Lenhard B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinformatics. 2016;32:1555–6.
85. Machlab D, Burger L, Soneson C, Rijli FM, Schübeler D, Stadler MB. monaLisa: an R/Bioconductor package for identifying regulatory motifs. Bioinformatics. 2022;38:2624–5.
86. Shannon P, Richards M. MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs. 2023.
87. Coetzee SG, Coetzee GA, Hazelett DJ. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics. 2015;31:3847–9.
88. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, et al. A single-cell atlas of chromatin accessibility in the human genome. Cell. 2021;184:5985-6001.e19.
89. TxDb.Hsapiens.UCSC.hg38.knownGene. Bioconductor. http://bioconductor.org/packages/TxDb.Hsapiens.UCSC.hg38.knownGene/. Accessed 25 Aug 2023.
90. How to extract promoters positions. 2018. https://hzhou.scholar.harvard.edu/blog/how-extract-promoters-positions. Accessed 25 Aug 2023.
91. Nott A, Holtman IR, Coufal NG, Schlachetzki JCM, Yu M, Hu R, et al. Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. Science. 2019;366:1134–9.
92. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25:1841–2.
93. Index of /goldenPath/rn5/liftOver. https://hgdownload.soe.ucsc.edu/goldenPath/rn5/liftOver/. Accessed 25 Aug 2023.
94. Index of /gbdb/hg19/liftOver. https://hgdownload.soe.ucsc.edu/gbdb/hg19/liftOver/. Accessed 27 Sep 2023.

95. UCSC Genome Browser Downloads. https://hgdownload.soe.ucsc.edu/downloads.html. Accessed 25 Aug 2023.
96. Liftover to Map Genomic Coordinates Between Genome Builds with Rtracklayer Bioconductor Package. 2021.
97. Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders | Nature Neuroscience. https://www.nature.com/articles/s41593-022-01128-z. Accessed 22 Aug 2023.
98. org.Hs.eg.db. Bioconductor. http://bioconductor.org/packages/org.Hs.eg.db/. Accessed 26 Sep 2023.
99. Yu G. Chapter 1 Overview of semantic similarity analysis | Biomedical Knowledge Mining using GOSemSim and clusterProfiler.
100. Drew K, Wallingford JB, Marcotte EM. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. Mol Syst Biol. 2021;17:e10016.
101. Harmonizome. https://maayanlab.cloud/Harmonizome/. Accessed 25 Aug 2023.
102. Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human binary protein interactome. Nature. 2020;580:402–8.
103. KEGG NETWORK Database. https://www.genome.jp/kegg/network.html. Accessed 25 Aug 2023.

Integration of multimodal data and networks



Step 1: Network regularized regression

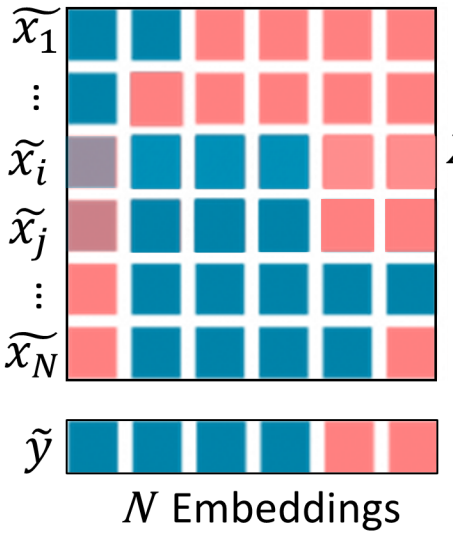
$$c^* = \underset{c}{argmin}\left\{\frac{1}{2M}\left|\left|y - Xc\right|\right|^2 + \alpha\left|\left|c\right|\right|_1 + \frac{\beta}{2N^2}\sum_{i,j=1,\ldots,N}\left(\left(\frac{c_i}{\sqrt{d_i}} - \frac{c_j}{\sqrt{d_j}}\right)^2 (w_{ij})\right)\right\}$$

c^*: regression coefficients of TFs for predicting TG expression for TF-TG regulation

Hyperparameters: β: network-constrained prior α: sparsity-prior

Step 2: Gene embeddings from network regression

$$c^* = \underset{c}{arg\,min}\left\{\frac{1}{2N}\left|\left|\tilde{y} - \tilde{X}c\right|\right|^2 + \alpha\left|\left|c\right|\right|_1\right\}$$



where

$$\tilde{X} = \sqrt{\frac{N}{M}}\Sigma^{\frac{1}{2}}U^T \qquad \tilde{y} = \sqrt{\frac{N}{M}}\Sigma^{-\frac{1}{2}}U^T X^T y$$

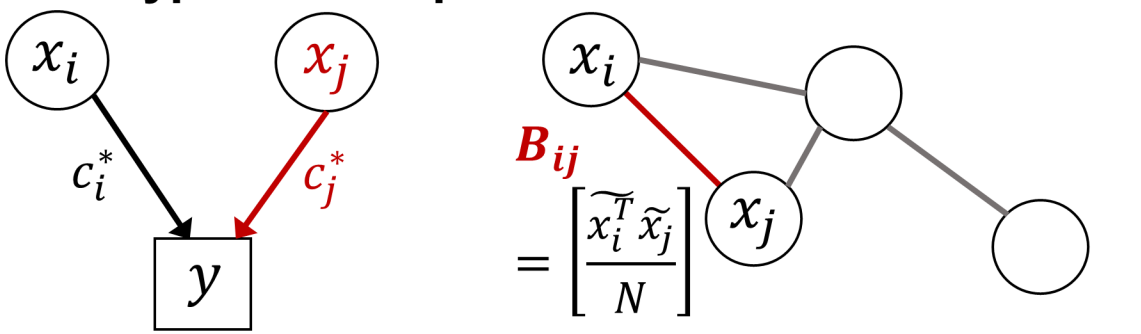
Singular Value Decomposition (SVD)

$$U\Sigma U^T = X^T X + 2\beta\frac{M}{N^2}D^T\left(W \cdot \left(N - 11^T\right)\right)D$$

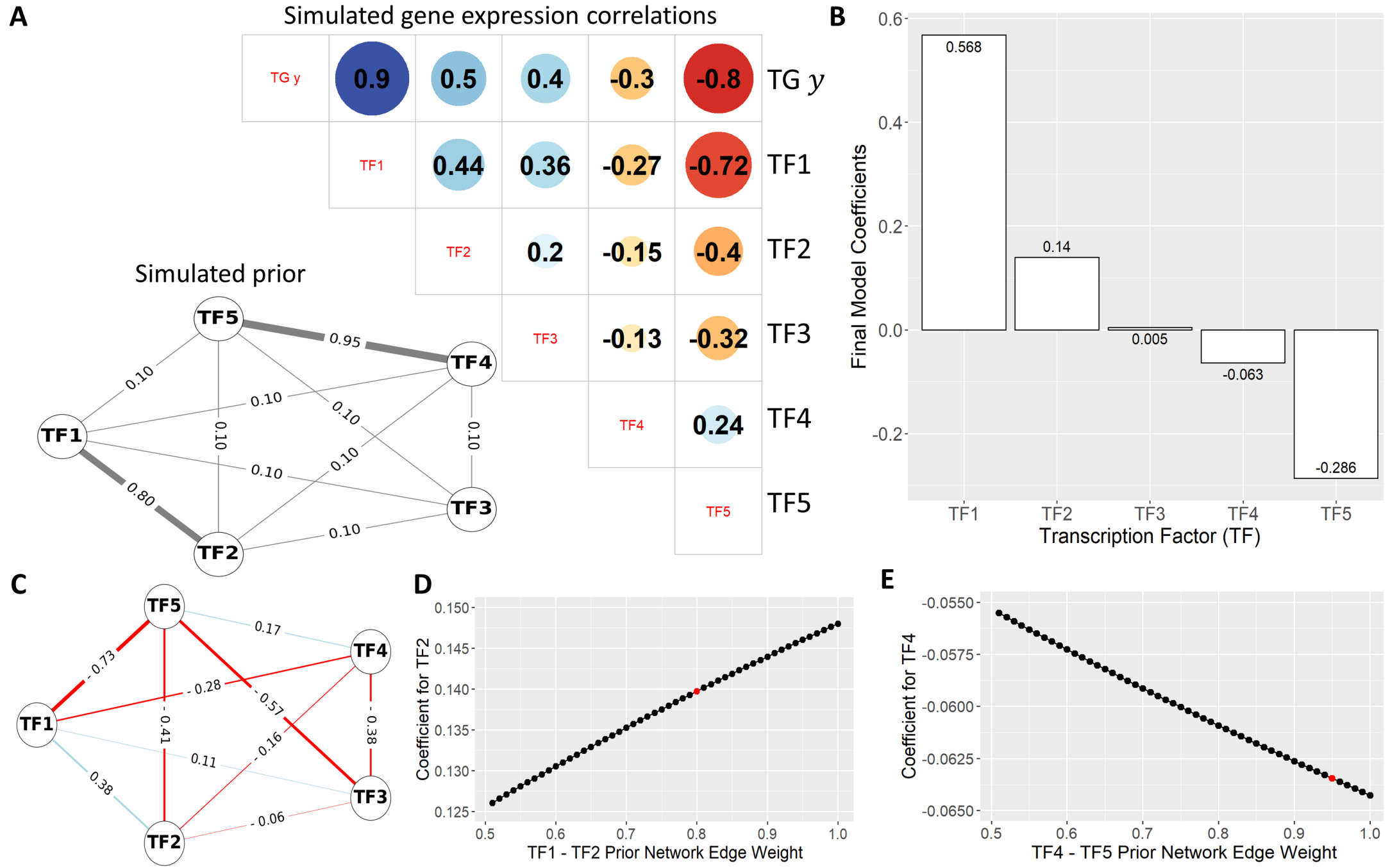
D: diagonal matrix of node degrees d

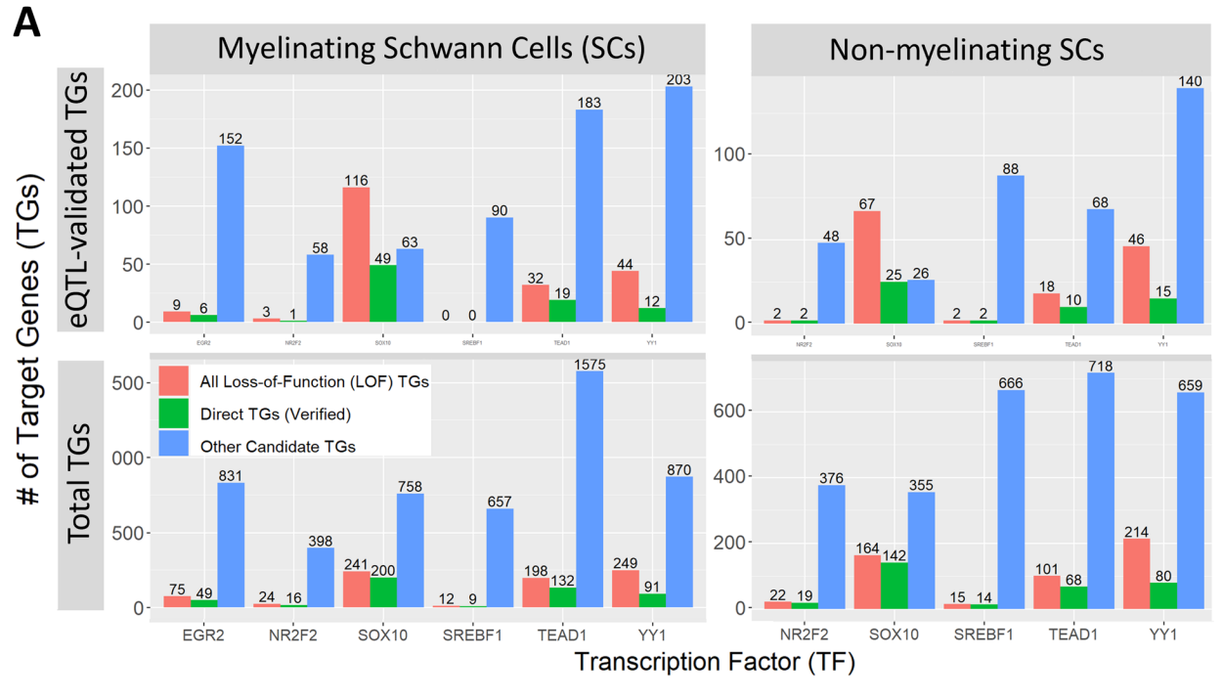
W: matrix of PPI network edge weights w

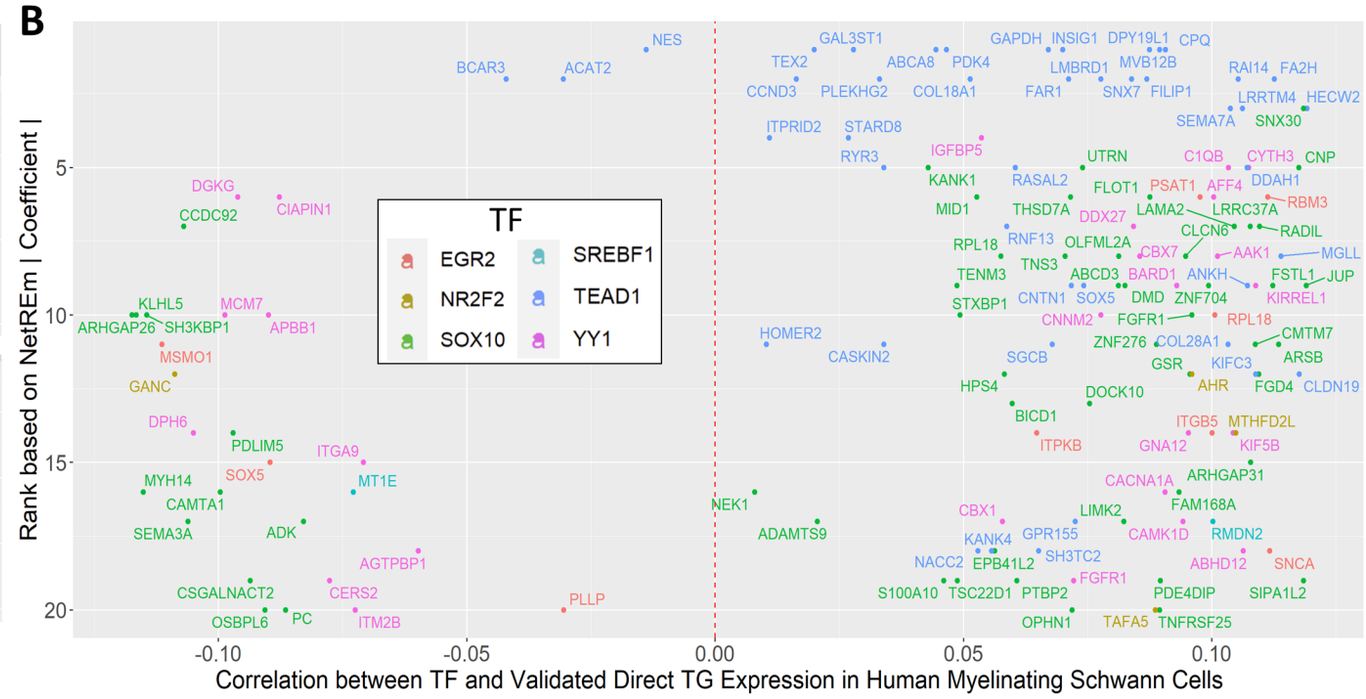
Cell-type transcription factors & coordination

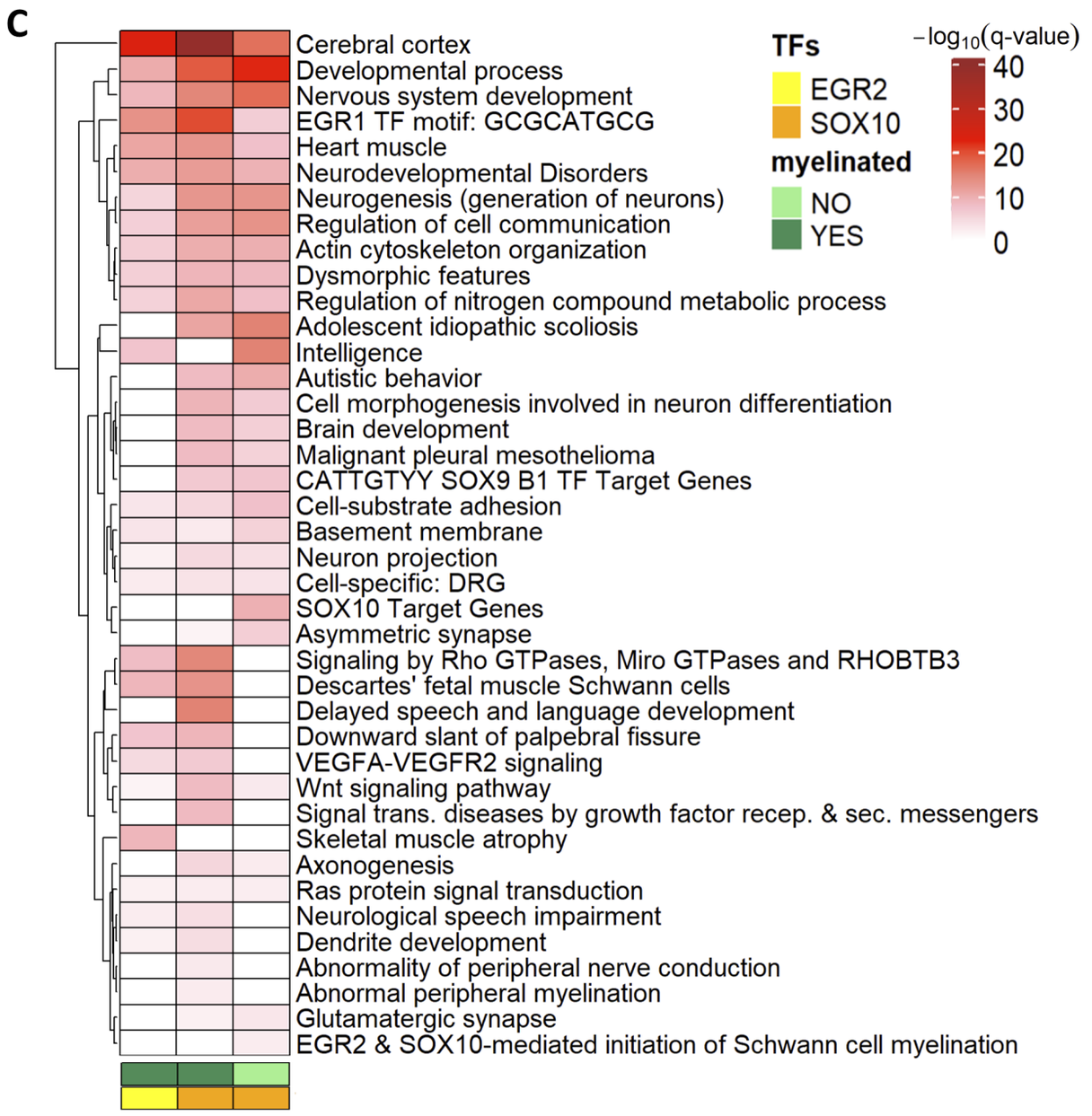


1. New cell-type TFs *2. Cell-type TF-TF Coordination*

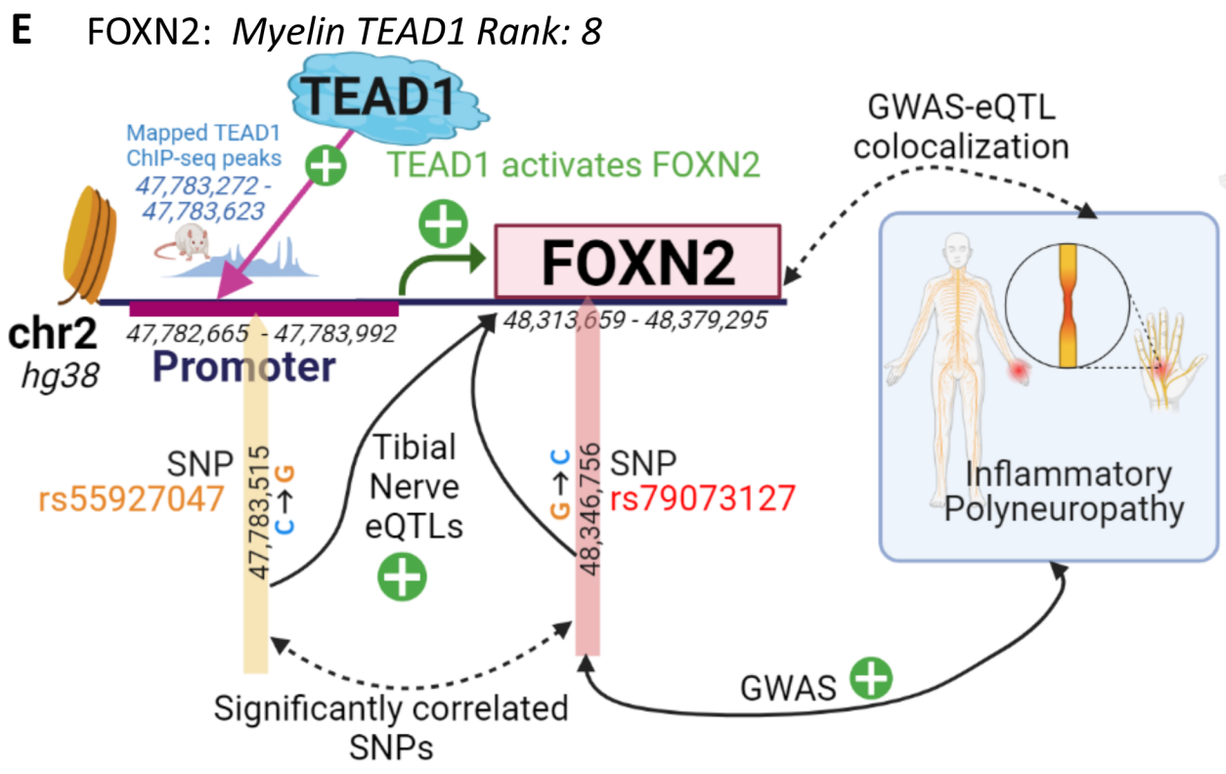

A
Simulated gene expression correlations
TG y
0.9
0.5
0.4
-0.3
-0.8
TG y
TF1
0.44
0.36
-0.27
-0.72
TF1
TF2
0.2
-0.15
-0.4
TF2
TF3
-0.13
-0.32
TF3
TF4
0.24
TF4
TF5
TF5
Simulated prior
TF5
TF4
TF1
TF2
TF3
0.95
0.10
0.80
B
Final Model Coefficients
0.568
0.14
0.005
-0.063
-0.286
Transcription Factor (TF)
C
0.17
-0.73
-0.28
-0.57
-0.41
-0.16
-0.38
0.11
0.38
-0.06
D
Coefficient for TF2
TF1 - TF2 Prior Network Edge Weight
E
Coefficient for TF4
TF4 - TF5 Prior Network Edge Weight

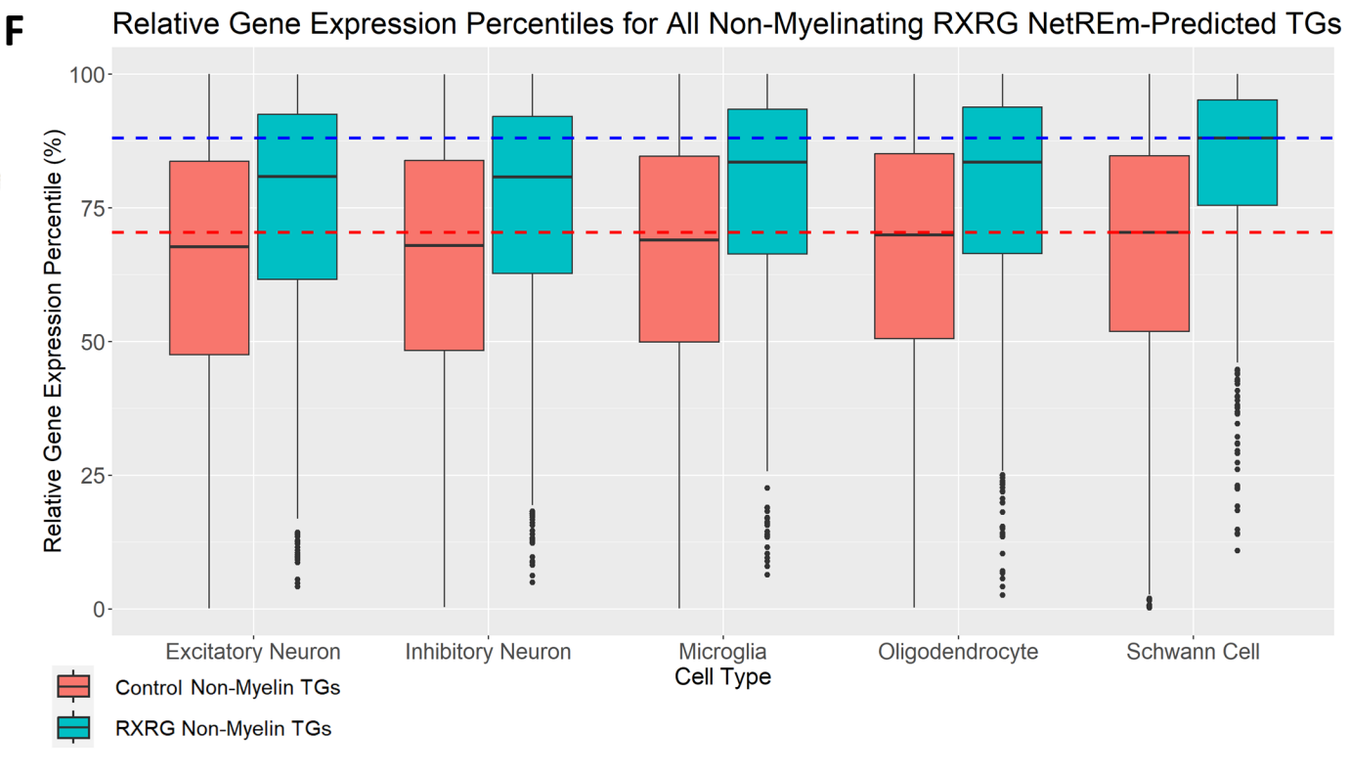

A
Myelinating Schwann Cells (SCs)
Non-myelinating SCs
eQTL-validated TGs
Total TGs
# of Target Genes (TGs)
All Loss-of-Function (LOF) TGs
Direct TGs (Verified)
Other Candidate TGs
EGR2
NR2F2
SOX10
SREBF1
TEAD1
YY1
Transcription Factor (TF)

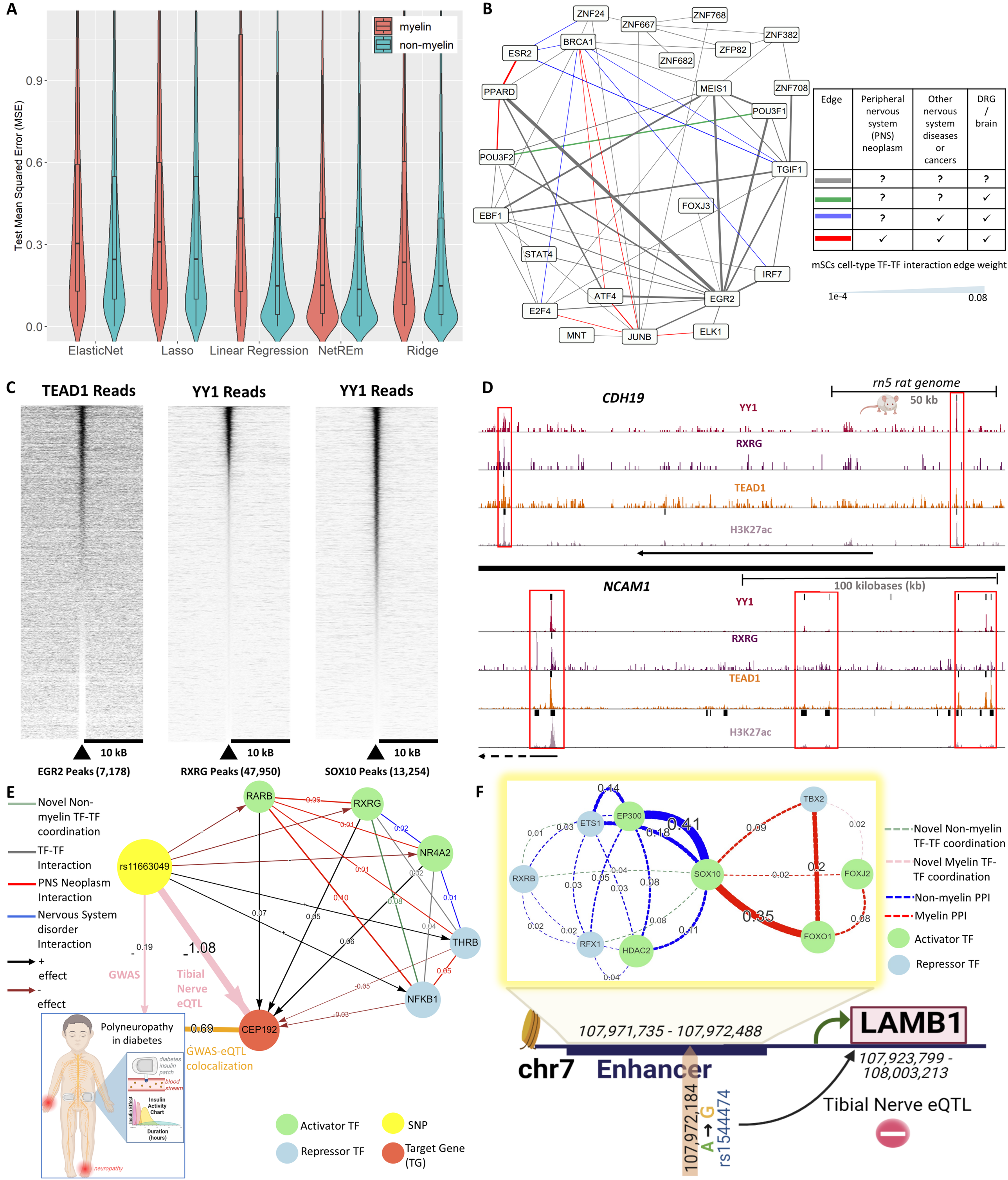

B
Rank based on NetREm | Coefficient |
TF
EGR2
NR2F2
SOX10
SREBF1
TEAD1
YY1
Correlation between TF and Validated Direct TG Expression in Human Myelinating Schwann Cells

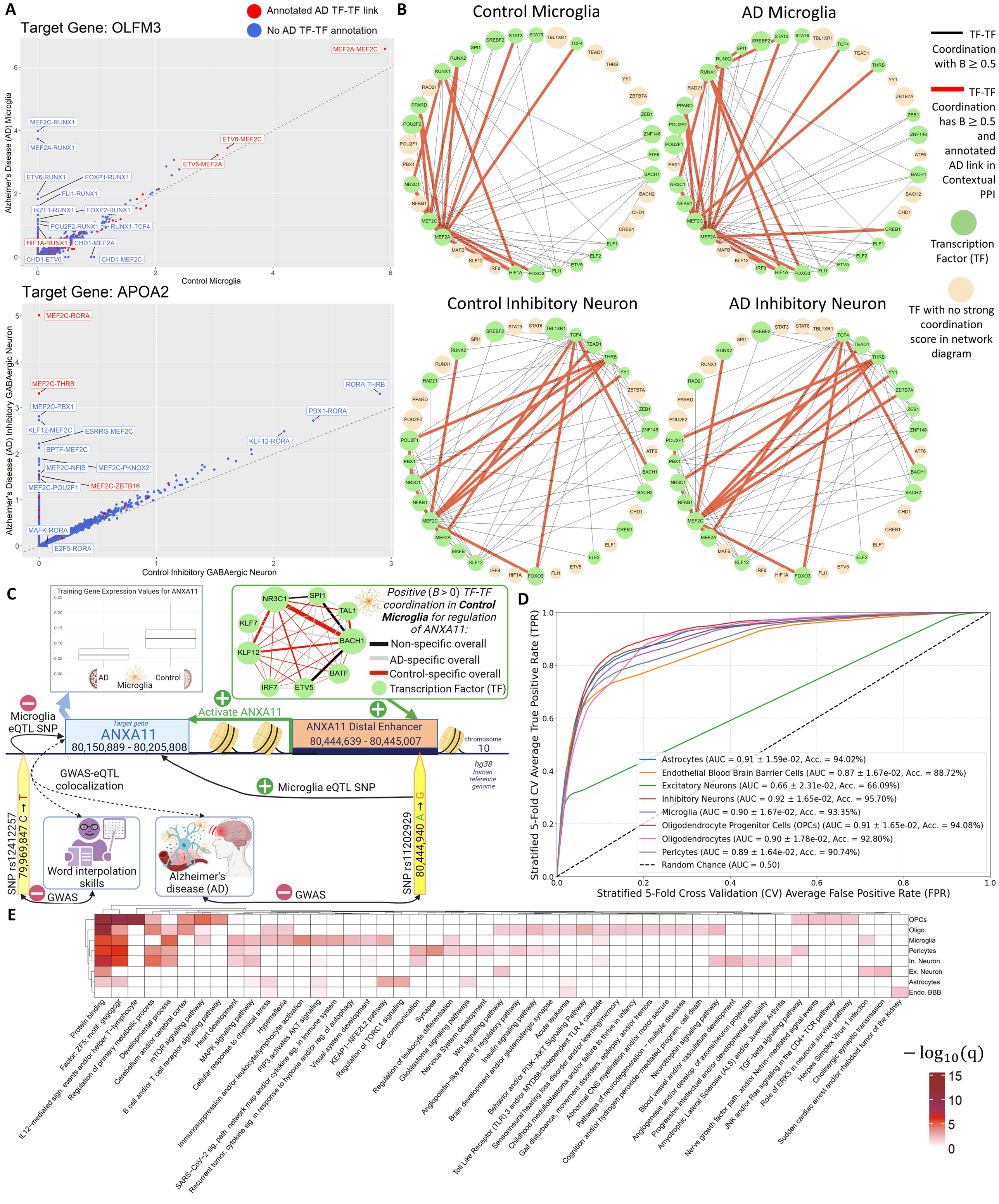

C
Cerebral cortex
Developmental process
Nervous system development
EGR1 TF motif: GCGCATGCG
Heart muscle
Neurodevelopmental Disorders
Neurogenesis (generation of neurons)
Regulation of cell communication
Actin cytoskeleton organization
Dysmorphic features
Regulation of nitrogen compound metabolic process
Adolescent idiopathic scoliosis
Intelligence
Autistic behavior
Cell morphogenesis involved in neuron differentiation
Brain development
Malignant pleural mesothelioma
CATTGTYY SOX9 B1 TF Target Genes
Cell-substrate adhesion
Basement membrane
Neuron projection
Cell-specific: DRG
SOX10 Target Genes
Asymmetric synapse
Signaling by Rho GTPases, Miro GTPases and RHOBTB3
Descartes' fetal muscle Schwann cells
Delayed speech and language development
Downward slant of palpebral fissure
VEGFA-VEGFR2 signaling
Wnt signaling pathway
Signal trans. diseases by growth factor recep. & sec. messengers
Skeletal muscle atrophy
Axonogenesis
Ras protein signal transduction
Neurological speech impairment
Dendrite development
Abnormality of peripheral nerve conduction
Abnormal peripheral myelination
Glutamatergic synapse
EGR2 & SOX10-mediated initiation of Schwann cell myelination
TFs
EGR2
SOX10
myelinated
NO
YES
−log10(q-value)

D
GPR157: Myelin SOX10 Rank: 2
SOX10 Sciatic Nerve (SN) Reads: Run 204
K27Ac Immunoprecipitation Sorted Tag Directory
CEP95: Myelin SOX10 Rank: 2
SOX10 Sciatic Nerve (SN) Reads: Run 204
K27Ac Immunoprecipitation Sorted Tag Directory
SCAMP5: Myelin SOX10 Rank: 3
SOX10 Sciatic Nerve (SN) Reads: Run 204
K27Ac Immunoprecipitation Sorted Tag Directory


E
FOXN2: Myelin TEAD1 Rank: 8
TEAD1
Mapped TEAD1 ChIP-seq peaks
47,783,272 - 47,783,623
TEAD1 activates FOXN2
FOXN2
chr2
hg38
47,782,665 - 47,783,992
Promoter
48,313,659 - 48,379,295
SNP
rs55927047
47,783,515
C→G
Tibial Nerve eQTLs
G→C
48,346,756
SNP
rs79073127
GWAS-eQTL colocalization
Inflammatory Polyneuropathy
GWAS
Significantly correlated SNPs


F
Relative Gene Expression Percentiles for All Non-Myelinating RXRG NetREm-Predicted TGs
Relative Gene Expression Percentile (%)
Excitatory Neuron
Inhibitory Neuron
Microglia
Oligodendrocyte
Schwann Cell
Cell Type
Control Non-Myelin TGs
RXRG Non-Myelin TGs


A
myelin
non-myelin
Test Mean Squared Error (MSE)
0.0
0.3
0.6
0.9
ElasticNet
Lasso
Linear Regression
NetREm
Ridge
B
ZNF24
ZNF667
ZNF768
ZNF382
BRCA1
ESR2
ZFP82
ZNF682
PPARD
MEIS1
ZNF708
POU3F1
POU3F2
TGIF1
EBF1
FOXJ3
STAT4
IRF7
ATF4
EGR2
E2F4
MNT
JUNB
ELK1
Edge
Peripheral nervous system (PNS) neoplasm
Other nervous system diseases or cancers
DRG / brain
?
?
?
?
?
✓
?
✓
✓
✓
✓
✓
mSCs cell-type TF-TF interaction edge weight
1e-4
0.08
C
TEAD1 Reads
YY1 Reads
YY1 Reads
10 kB
10 kB
10 kB
EGR2 Peaks (7,178)
RXRG Peaks (47,950)
SOX10 Peaks (13,254)
D
rn5 rat genome
CDH19
50 kb
YY1
RXRG
TEAD1
H3K27ac
NCAM1
100 kilobases (kb)
YY1
RXRG
TEAD1
H3K27ac
E
Novel Non-myelin TF-TF coordination
TF-TF Interaction
PNS Neoplasm Interaction
Nervous System disorder Interaction
+ effect
- effect
rs11663049
RARB
RXRG
NR4A2
THRB
NFKB1
CEP192
0.19
1.08
GWAS
Tibial Nerve eQTL
0.69
GWAS-eQTL colocalization
Polyneuropathy in diabetes
diabetes insulin patch
blood stream
Insulin Activity Chart
Insulin Effect
Duration (hours)
neuropathy
Activator TF
Repressor TF
SNP
Target Gene (TG)
F
EP300
ETS1
RXRB
SOX10
TBX2
FOXJ2
FOXO1
RFX1
HDAC2
0.41
0.35
0.2
Novel Non-myelin TF-TF coordination
Novel Myelin TF-TF coordination
Non-myelin PPI
Myelin PPI
Activator TF
Repressor TF
107,971,735 - 107,972,488
LAMB1
chr7
Enhancer
107,923,799 - 108,003,213
Tibial Nerve eQTL
107,972,184
A → G
rs1544474


A
Annotated AD TF-TF link
No AD TF-TF annotation
Target Gene: OLFM3
Alzheimer's Disease (AD) Microglia
Control Microglia
MEF2A-MEF2C
MEF2C-RUNX1
MEF2A-RUNX1
ETV6-MEF2C
ETV6-MEF2A
ETV6-RUNX1
FOXP1-RUNX1
FLI1-RUNX1
IKZF1-RUNX1
FOXP2-RUNX1
POU2F2-RUNX1
RUNX1-TCF4
HIF1A-RUNX1
CHD1-MEF2A
CHD1-ETV6
CHD1-MEF2C
Target Gene: APOA2
Alzheimer's Disease (AD) Inhibitory GABAergic Neuron
Control Inhibitory GABAergic Neuron
MEF2C-RORA
MEF2C-THRB
MEF2C-PBX1
KLF12-MEF2C
ESRRG-MEF2C
BPTF-MEF2C
MEF2C-NFIB
MEF2C-PKNOX2
MEF2C-POU2F1
MEF2C-ZBTB16
MAFK-RORA
E2F5-RORA
RORA-THRB
PBX1-RORA
KLF12-RORA
B
Control Microglia
AD Microglia
Control Inhibitory Neuron
AD Inhibitory Neuron
TF-TF Coordination with B ≥ 0.5
TF-TF Coordination has B ≥ 0.5 and annotated AD link in Contextual PPI
Transcription Factor (TF)
TF with no strong coordination score in network diagram
C
Training Gene Expression Values for ANXA11
AD
Control
Microglia
Positive (B > 0) TF-TF coordination in Control Microglia for regulation of ANXA11:
Non-specific overall
AD-specific overall
Control-specific overall
Transcription Factor (TF)
Microglia eQTL SNP
Activate ANXA11
Target gene
ANXA11
80,150,889 - 80,205,808
ANXA11 Distal Enhancer
80,444,639 - 80,445,007
chromosome 10
hg38 human reference genome
GWAS-eQTL colocalization
Microglia eQTL SNP
SNP rs12412257 79,969,847 C → T
SNP rs11202929 80,444,940 A → G
Word interpolation skills
Alzheimer's disease (AD)
GWAS
GWAS
D
Stratified 5-Fold CV Average True Positive Rate (TPR)
Stratified 5-Fold Cross Validation (CV) Average False Positive Rate (FPR)
Astrocytes (AUC = 0.91 ± 1.59e-02, Acc. = 94.02%)
Endothelial Blood Brain Barrier Cells (AUC = 0.87 ± 1.67e-02, Acc. = 88.72%)
Excitatory Neurons (AUC = 0.66 ± 2.31e-02, Acc. = 66.09%)
Inhibitory Neurons (AUC = 0.92 ± 1.65e-02, Acc. = 95.70%)
Microglia (AUC = 0.90 ± 1.67e-02, Acc. = 93.35%)
Oligodendrocyte Progenitor Cells (OPCs) (AUC = 0.91 ± 1.65e-02, Acc. = 94.08%)
Oligodendrocytes (AUC = 0.90 ± 1.78e-02, Acc. = 92.80%)
Pericytes (AUC = 0.89 ± 1.64e-02, Acc. = 90.74%)
Random Chance (AUC = 0.50)
E
OPCs
Oligo.
Microglia
Pericytes
In. Neuron
Ex. Neuron
Astrocytes
Endo. BBB
Protein binding
Factor: ZF5; motif: gsgcgcgr
IL12-mediated sign. events and/or helper T-lymphocyte
Regulation of primary metabolic process
Developmental process
Cerebellum and/or cerebral cortex
mTOR signaling pathway
B cell and/or T cell receptor signaling pathway
Heart development
MAPK signaling pathway
Cellular response to chemical stress
Hyperreflexia
Immunosuppression and/or leukocyte/lymphocyte activation
PIP3 activates AKT signaling
SARS-CoV-2 sig. path. network map and/or cytokine sig. in immune system
Recurrent tumor, cytokine sig. in response to hypoxia and/or reg. of autophagy
Visual system development
KEAP1-NFE2L2 pathway
Regulation of TORC1 signaling
Cell communication
Synapse
Regulation of leukocyte differentiation
Glioblastoma signaling pathways
Nervous System development
Wnt signaling pathway
Angiopoietin-like protein 8 regulatory pathway
Insulin signaling pathway
Acute leukemia
Brain development and/or glutamatergic synapse
Behavior and/or PI3K-AKT Signaling Pathway
Toll Like Receptor (TLR) 3 and/or MYD88-independent TLR 4 cascade
Sensorineural hearing loss disorder and/or failure to thrive in infancy
Childhood medulloblastoma and/or learning/memory
Gait disturbance, movement disorders, epilepsy, and/or tremors
Abnormal CNS myelination and/or motor seizure
Pathways of neurodegeneration - multiple diseases
Cognition and/or hydrogen peroxide-mediated program. cell death
Neurotrophin signaling pathway
Blood vessel and/or vasculature development
Angiogenesis and/or develop. of axons/neuron projection
Progressive intellectual and/or developmental disability
Amyotrophic Lateral Sclerosis (ALS) and/or Juvenile Arthritis
TGF-beta signaling pathway
Nerve growth factor path. and/or Netrin-mediated signal events
JNK and/or Ras signaling in the CD4+ TCR pathway
Role of ERK5 in neuronal survival pathway
Herpes Simplex Virus 1 infection
Cholinergic synaptic transmission
Sudden cardiac arrest and/or rhabdoid tumor of the kidney
$-\log_{10}(q)$