# Enhancer-driven regulatory network of forebrain human development provides insights into autism

Alexandre Jourdon[1,2*], Jessica Mariani[1,2*], Feinan Wu[1,2*], Davide Capauto[1,2], Scott Norton[1], Livia Tomasini[1,2], Anahita Amiri[1], Jeremy Schreiner[1], Cindy Khanh Nguyen[1], Neal Nolan[1], Anna Szekely[3,2], James C. McPartland[1], Kevin Pelphrey[1,4], Katarzyna Chawarska[1], Pamela Ventola[1], Alexej Abyzov[5,2#], Flora M. Vaccarino[1,2,6,7,8#]

[1]Child Study Center, Yale University School of Medicine, New Haven, CT, 06520, USA
[2]Program in Neurodevelopment and Regeneration, Yale University, New Haven, CT, 06520, USA
[3]Department of Neurology, Yale University School of Medicine, New Haven, CT, 06520, USA
[4]Brain Institute, Department of Neurology, University of Virginia School of Medicine, Charlottesville, VA, 22904, USA
[5]Department of Quantitative Health Sciences, Center for Individualized Medicine, Mayo Clinic, Rochester, MN, 55905, USA
[6]Department of Neuroscience, Yale University School of Medicine, New Haven, CT, 06520, USA
[7]Kavli Institute for Neuroscience, Yale University, New Haven, CT, 06520, USA
[8]Yale Stem Cell Center, Yale University, New Haven, CT, 06520, USA

[*]These authors contributed equally
[#]Co-corresponding authors

## Abstract

Cell differentiation involves shifts in chromatin organization allowing transcription factors (TFs) to bind enhancer elements and modulate gene expression. The TF-enhancer-gene regulatory interactions that control the formation of neuronal lineages have yet to be charted in humans. Here, we mapped enhancer elements and conducted an integrative analysis of epigenomic and transcriptomic profiles across 60 days of differentiation of human forebrain organoids derived from 10 individuals with autism spectrum disorder (ASD) and their neurotypical fathers. This multi-omics profiling allowed us to build an enhancer-driven gene regulatory network (GRN) of early neural development. We validated the GRN by performing a loss-of-function experiment with FOXG1 – one of the master TFs in the development of the mammalian brain. Analysis of the constructed GRN identified regulatory hierarchies driving the specification of neuronal cell types. By analyzing differential gene expression in ASD through the GRN hierarchy, we associated the ASD transcriptomic signatures to altered activity of key TFs. We found that macrocephalic ASD was principally driven by an increased activity of BHLHE22, FOXG1, EOMES, and NEUROD2, which are major regulators of excitatory neuron fate. Normocephalic ASD, on the contrary, was driven by decreased activity of those same TFs and by an increased activity of LMX1B and FOXB1 – two upstream TF repressors of FOXG1. These findings suggest that ASD is characterized by an altered early gene regulatory program that specifies neuronal cell lineages in the fetal brain. Thus, constructing a GRN of early brain development modeled in organoids provides insights into the etiology of ASD and can guide future experimental approaches to establish its genetic causes and treatment strategies.

# Introduction

Fundamental cell fate decisions and regional specification during early human brain development underlie our larger brain size, cellular diversity, and cognitive abilities but also our susceptibility to neurodevelopmental disorders, such as autism spectrum disorders (ASD). Brain organoids derived from human induced pluripotent stem cell (iPSC) lines are a multilineage model that emulates early brain development in an individual's genetic background[1-3]. While several transcriptomic studies have charted lineage specific gene expression patterns of fetal human brain and brain organoids, less is known about how those lineages are specified by the non-coding genome, and specifically by enhancers, the main elements implicated in the modulation of gene expression[4]. Located at variable distances from gene promoters, enhancers are bound by specific transcription factors (TFs) which can modify chromatin conformation and protein-protein interactions to increase or decrease the expression of downstream target genes [5,6]. While TFs, enhancers and other types of regulatory elements have been described to some extent in the primate brain [7], understanding the global regulatory relationships between enhancers, TFs and gene expression (also called "regulome") remains challenging, particularly in a human system.

In this study we performed histone marks chromatin immunoprecipitation and sequencing (ChIP-seq) and parallel bulk and single-cell RNA sequencing (RNA-seq and scRNA-seq) analyses over a developmental time course of iPSC-derived forebrain organoids from 10 male individuals with ASD and their respective neurotypical fathers. Integrating enhancer activity with TF binding sites and gene expression on a genome-wide scale, we built a regulatory network, identifying enhancers and cognate TFs active during the progression of neurogenesis and the diversification of neuronal subtypes. By comparing enhancer activity and differential gene expression of individuals with idiopathic ASD versus their unaffected father as controls, we further delineated the regulome underlying the transcriptional alterations in macrocephalic and normocephalic ASD.

# Results

## Building a gene regulatory network based on epigenomic and transcriptomic activities during forebrain organoid development.

To characterize both epigenomic and transcriptomic states of precursor cells and early neurons, iPSC lines from 10 ASD individuals and their unaffected fathers were differentiated into forebrain organoids and histone marks-ChIP-seq, bulk RNA-seq and scRNA-seq were concurrently performed at terminal differentiation day 0 or TD0 (corresponding to 20 days after neural induction of the iPSC lines), TD30 and TD60 (**Fig. 1a**, **Table S1**). We then developed a pipeline that integrated the multi-omics features collected in parallel into a *regulatory network* of human forebrain organoid development (**Fig. 1b, Methods**). First, integrated annotation of the genome using a ChromHMM 9-state model[8] for ChIP-seq data identified putative promoters, ~360,000 **enhancer regions**, and silent chromatin (**Fig. 1c-d).** Enhancers were then linked to potential target genes by proximity to promoters or by reported distal interactions identified in human fetal brain Hi-C dataset[9-11] **(Extended Data Fig. 1a).** Altogether, 173,437 gene-linked enhancers (GLEs) were mapped across all samples (**Extended Data Fig. 1b-e, Table S2**). GLEs were then linked to TFs by the presence of a TF binding site (TFBS) motif in their sequence (**Extended Data Fig. 1e,f)**. Next, correlations between GLE activity (H3K27ac reads) and gene expression (RNA reads) across samples were used to retain the most functionally relevant links (i.e., links with significant correlation). Such refined links inform about both activating (positive correlation) and repressing (negative correlations) nature of the regulatory relationships linking TFs to enhancers, and enhancers to their downstream genes (**Fig. 1b**, **Extended Data Fig. 2a,b**).

All retained links were integrated into a TF-enhancer-gene network connecting 373 TFs, 42,035 GLEs, 6,501 protein coding genes and 1,690 long non-coding RNAs (lncRNAs) (**Table S2, T2-3**). In such network, active enhancers are pivotal elements used to chart the upstream regulatory network of any expressed gene, some with complex regulatory cascades (**Fig. 1e-f**, **Extended Data Fig. 3**). To capitalize on this characterization of enhancer-centered regulatory cascades and facilitate biological interpretability, we derived a reduced representation in the form of a gene regulatory network (GRN) where TF were directly connected to downstream gene(s) when one or several intermediate functional GLE(s) linking them was identified (**Fig.1e-f, Extended Data Fig. 2c**). We pruned 13.3% of such links by requiring a significant correlation between the expression of the TF and the downstream gene and used the obtained correlation coefficient to infer both the strength and the activating or repressing nature of the TF-to-gene regulatory action (**Methods)**. The resulting GRN consisted of 330 upstream TFs regulating 5,067 downstream transcripts, encompassing 4,034 protein-coding genes and 1,023 lncRNAs (**Fig. 1g, Table S3**).

The original set of GLEs overlapped 32% to 55% with sets of known enhancers in human [5,12,13] and the overlap was significantly higher (44% to 71%) for the subset of GLEs included in the GRN, justifying using the correlation approach for network refinement (**Fig. 1h)**. We also found that ~43% of GLEs overlap with peaks of open chromatin assessed by bulk and single cells ATAC-seq assays in human fetal brain or brain organoids datasets [7,14-16], reinforcing the confidence in our characterized enhancer elements (**Fig. 1i**).

The GRN was built around TFs and their ability to activate and repress a set of targeted genes, also known as their *regulons*. To identify how TFs structured the GRN, we evaluated the overlap between the regulons of each pair of TFs across the network and measured how similar or opposite was their inferred actions on shared targets (**Extended Data Fig 4**). This delineated an *interactome* of TFs, identifying for each TF a set of preferential collaborators and competitors that tended to regulate identical genes through either identical or separates

enhancers. Strongest interactions occurred not only between TFs with closely related binding motifs families (driven in part by similarity in targeted enhancer sites) but also across major brain regulators which are known to coregulate genes expression (**Extended Data Fig. 4a**). For instance, the interactome of the TF EOMES revealed a preferential collaboration with NEUROG2, NEUROD2 or MEF2C and a competition with FOXB1, LHX6 or ZNF823 across several targets in the GRN. Other examples included FOXP2 collaboration with ZNF548 and competition with TEAD4; or TBR1 collaboration with NFIB and ESR2 and competition with ZNF823 (**Extended Data Fig. 4b**). Altogether, the built GRN integrated enhancer identification with gene expression to chart the regulatory effects and interactions of TFs in our organoid model. We provide this network (**Table S3**) as a resource to the scientific community.

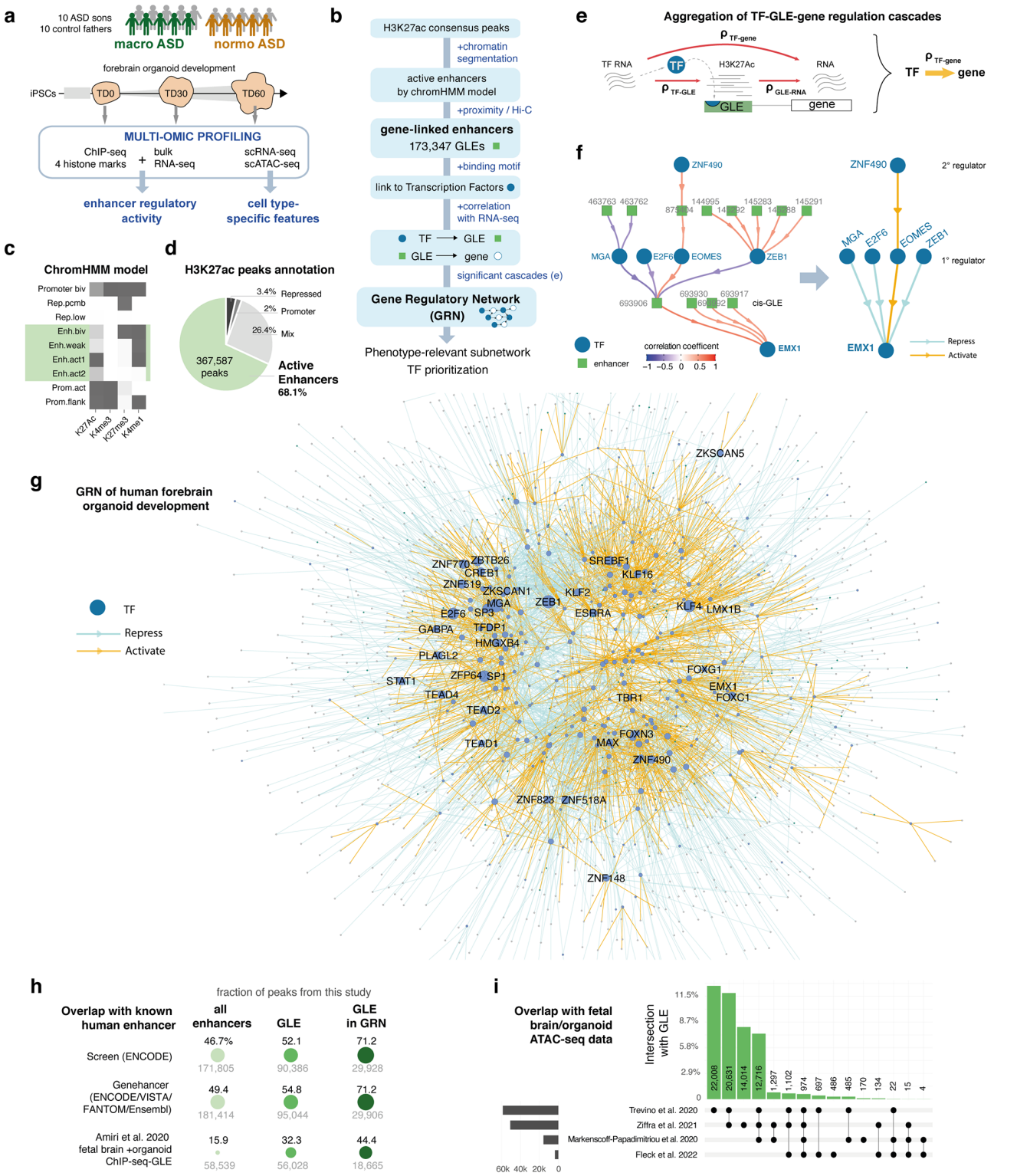



**Figure 1: Building a regulatory network during forebrain organoid development anchored around the identification of active enhancer elements.**

**a.** Study design with multi-Omic characterization of forebrain organoid development in 20 individuals including ASD individuals with or without macrocephaly ("macro ASD", head circumference > 90$^{th}$ percentile of the population, "normo ASD" otherwise).

**b.** Data integration pipeline to generate the gene regulatory network (GRN) (see **Methods**). It incorporates enhancer-gene linkages from proximity and Hi-C, transcription factor binding site (TFBS) identification, and

correlation of enhancer activity with expression of upstream TF(s) and downstream gene(s). The final step consists in the derivation of a GRN through aggregation of the TF-enhancer-gene cascades (see **e**).

**c.** The 9-states model of genome annotation built using the 4 chromatin marks ChIP-seq data from this study. Four enhancer states were then merged into an "active enhancer" state. Enhancers profiles were essentially characterized by H3K27Ac without H3K4Me3 activity.

**d.** Diagram showing the proportion of each main chromatin state identified in (**c**) for all H3K27ac consensus peaks.

**e.** Schematic of the aggregation strategy validating the intermediate role of enhancers through concordant correlation at the RNA level between TF and downstream genes (see also **Extended Data Fig. 2**). Regulatory actions were then summarized to build the TF-gene GRN, conserving only the correlation between TF and gene in the final network. Blue arrows=negative and red arrows= positive correlation ($\rho$).

**f.** Upstream enhancer regulatory network of the gene EMX1. Arrows indicate regulatory link between TFs and their bound GLEs (in green with unique ID from **Table S2**) and links from GLE to their downstream gene in *cis*. Colors of the arrows show correlation level between the linked elements. Note the four-steps distance between the TF ZNF490 and EMX1 expression. The corresponding TF-gene network is plotted on the right using TF-gene RNA level correlation as link weight (two-sided Spearman's correlation FDR corrected p-value < 0.05).

**g.** Visualization of the central component of the GRN inferred from the enhancer regulatory network, plotted by Kamada-Kawai graph layout. Genes connected by repressive regulation (i.e., $\rho_{TF-RNA} < 0$) are further apart than genes connected by activating regulation (see **Methods**). By construction, TF (blue dots) are central element of the GRN, each connecting in their regulon 1 to 1,047 downstream genes (small grey dots). Top 50 TF by regulon size (dot size) and salient TFs from our study are named.

**h** Fraction of all enhancers regions ("All Enh."), gene-linked enhancer ("GLE") and regulatory GLE ("reg-GLE", i.e., GLE significantly correlated in activity with gene expression, FDR < 0.01, **Table S2**) that overlap with reported human enhancers from 3 sources [5,12,13] . Dot shows percentage enhancer from our study overlapping with each source, corresponding number of peaks is indicated below. All overlaps increased significantly from "All Enh." to "GLE" to "reg-GLE" (p-values of one-tailed Fisher exact test < $2.2 \times 10^{-16}$).

**i.** Upset plot showing the intersection between GLE from this study and peaks identified in published bulk and single-cell ATAC-seq data from human fetal brain or brain organoid models [7,14,15,17].

## Identification of master regulators behind neurogenesis progression and cell type specification

To explore how the enhancer-centered GRN captured regulatory events of human neural development, we next sought to identify the regulatory relationships central to the progression of neurogenesis. The GRN was therefore subset to only include genes differentially expressed between TD30/60 and TD0 (i.e., late vs early, termed timeDEGs) (**Extended Data Fig. 5a-b, Table S4, T1**). Regulatory links in the GRN were then subset to include only links where change in expression of upstream TF was susceptible to drive differential expression of the downstream gene (**Fig. 2a, Methods**). This time transition subnetwork revealed two antagonizing groups of TFs driving major expression changes during early development (**Fig. 2a**, **Extended Data Fig. 6c**). To further rank TF's abilities to drive expression changes between time points, we computed for each TF in the full GRN an *activity perturbation score* that measured how its regulon (i.e., the set of its direct downstream targets) was affected between TD30/60 and TD0. We found this score to be a useful metrics to infer and rank the ability of a TF to drive observed gene expression changes between any 2 conditions, based on the TF's predicted effects on its regulon in the GRN (perturbation score metric detailed in **Extended Data Fig. 6**). Positive, negative, and near-zero perturbation scores indicated, respectively, an increased, decreased or unchanged activity of the TF on its regulon across the tested conditions. Applied to timeDEG, this TF activity perturbation analysis revealed that increased activity of TBR1, FOXP2 and POU6F1 and decreased activity of DMRTA1, FOSL1, and TEAD4 were the top drivers of gene expression changes across time transitions (**Fig. 2b**). Interestingly, some TFs with low change in expression level across time points had high activity perturbation scores (for instance ZNF490, SP1 or FOXN3). Such cases could stem from differences in posttranscriptional modulation of the TF, influencing for instance its DNA- or protein-binding abilities, although this could also result from expression changes in TFs acting on closely related regulons.

We and other have shown how organoids develop from aggregates of neural progenitors that generate a diversity of neurons over time[3,18]. Using scRNA-seq data, we have shown that our organoid method generated a diversity of forebrain neuron fates (**Fig. 2c**) and that cell diversity tended to vary between organoids from different individuals [19]. Since bulk-level gene expression in organoids is strongly driven by changes in cell type composition, we hypothesized that the GRN captured in part the regulatory events driving variations in cell fates. Indeed, we observed that many timeDEGs had cell type-specific expression in scRNA-seq data (**Fig. 2d, Extended Data Fig. 5a-c**), including many TFs drivers of time transition identified from the GRN (**Fig. 2b**). Coherently, TFs' activated regulons tended to be co-expressed in the same cells as their upstream TF, while repressed part of their regulon had often inverted expression pattern across cells (**Fig. 2d, Extended Data Fig. 5c**). This suggested that many central TFs in the GRN are driving binary choices between cell type identities or states. We therefore applied the perturbation activity score metric to cell-type specific DEGs (**Table S5**) to evaluate each TF's capacity to drive cell type-specific expression pattern and identify putative activators and repressors of specific cell fates (**Fig. 2e**). As expected, many TFs controlled the distinction between neurons and progenitor cells (e.g., FOXP2, SOX4, TFAP2E, upregulating general neuron fate and repressing RG), while some TFs were preferentially regulating genes expressed by a particular neuronal subtype. Among the latter, EOMES, BHLHE22, NEUROD2, and NFIC were the strongest activators of EN-DCP fates, while E2F6, TCF7L2 and POU2F1 were the strongest repressors of that fate, consistent with mouse loss-of-function studies [20,21].

To evaluate how cell type-specific regulatory events were captured in data derived from bulk-level ChIP-seq, we intersected our list of GLEs with peaks called from 29 scATAC-seq samples collected from the same organoid preparation. We observed that between 10-50% of all GLEs overlapped with ATAC-seq peaks identified in a cell type or shared across cells (**Fig. 2f**). Furthermore, activating enhancer-to-gene regulatory links (**Table S2, T3**) were enriched in cases where both the enhancer overlapped scATAC-seq peak and the linked gene showed specific expression in the same cell type, suggesting our bulk-derived regulatory network construction captured cell-type specific regulatory events. This enrichment was more significant for neuronal cell types and particularly for excitatory neurons (EN) (**Fig. 2g**). Altogether, these analyses suggested that identification of enhancers from

bulk epigenome along the developmental timeline captures the complex regulatory events underlying the development of a multilineage system.

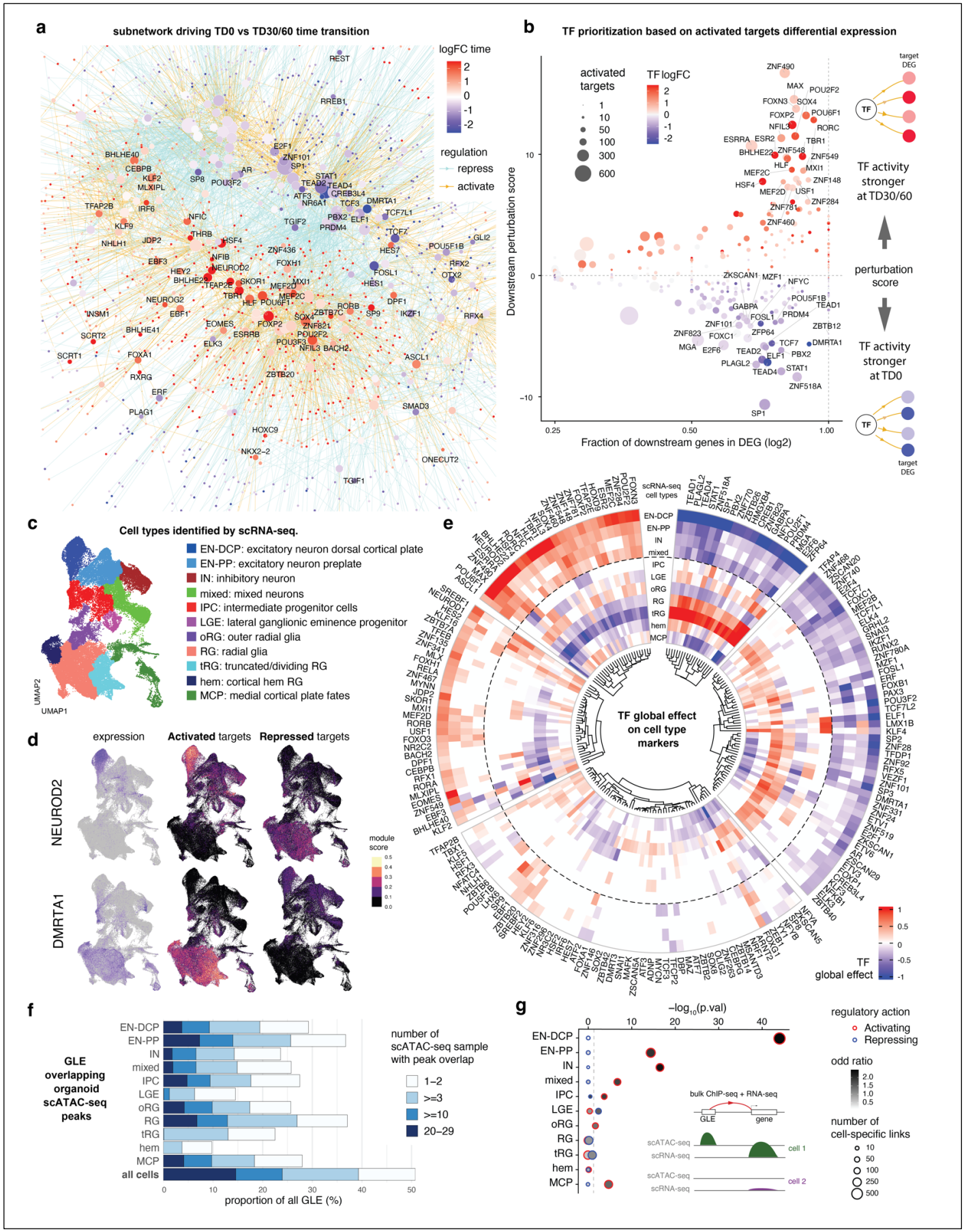


**Figure 2. Master regulators driving neurogenesis and cellular diversity in forebrain organoids**

**a.** Time transition subnetwork built from the GRN by retaining only significant timeDEGs (TD30+60 vs TD0, $|\log_2 FC| > 0.25$ and FDR < 0.01) and only concordant links connecting them: genes linked by an activating edge (yellow) need to have the same direction of change in the timeDEG (i.e., both down- or both up-regulated), and genes linked by a pressive edge (cyan) need to have opposite direction of change. Top TFs ranked by logFC are named. Kamada-Kawai graph layout with distances between genes proportional to the edge weight as in **Fig. 1g** (i.e., strong repression= more distal, strong activation= more proximal).

**b.** Dot plot showing for each TF the fraction of TF's targets that are timeDEG (x-axis) and the TF activity perturbation scores (y-axis, schematic). Dot color shows the TF's own $\log_2 FC$ in timeDEG. Dot size=regulon size in the GRN. The perturbation score ranks each TF's ability to drive the expression changes of its regulon (here, across time) and is computed from GRN prediction and DEG results for the downstream target only. As shown by yellow arrows in right side schematic, only activating relationship were used for the perturbation score in this plot. Top 25 TF by activity perturbation score are named. (see **Methods** and **Extended Data Fig. 6** for detailed description of perturbation score).

**c.** UMAP plot colored by cell types identified in 72 scRNA-seq samples from TD0, 30 and 60 organoids [19] of the same individuals used for this study.

**d**. UMAP plots colored by expression level for the TF NEUROD2 and DMRTA1 (left panels, blue=high, grey=low). To infer if TF's targets also presented cell-type specific expression, the respective downstream "activated targets" and "repressed targets" of each TF's regulon were used to derive a module expression score in the scRNA-seq data and positive module scores per cell are plotted in the UMAP (see also **Extended Data Fig. 5c**).

**e.** Circular heatmap showing the global effect of each TF on cell type-specific gene expression as quantified by scRNA-seq (see **Table S5** for cell type markers). Results for different cell types are placed as concentric circles. As in **b**, TF's global effect evaluates the TF's ability to drive expression change observed between one cell type compared to all other cells (i.e., perturbation score applied to cell-type DEGs, **Methods**). Results are grouped by k-mean clustering (k=5) and hierarchical clustering (Euclidean distance, complete method).

**f.** Fraction of all GLEs that intersect with sample-specific scATAC-seq peaks in organoids (total 29 scATAC-seq samples). "all cells" = peaks called using all cells. (see **Table S5, T2** for scATAC-seq peaks overlap results). Cell type annotations were identified by label transfer using scRNA-seq from **c** as reference.

**g.** Enrichment of cell type-specific regulatory relationships (cell-spe.links) among all enhancer-gene regulatory relationships identified (**Table S2, T3**). *Cell specific links* were defined as cases where the enhancer overlaps with scATAC-seq peaks (>= 3 samples) and the linked gene is a cell type specific marker of the same cell type. Enrichment was calculated separately for activating and repressing enhancer-gene links (red and blue circles, respectively). Activating links but not repressing links occured preferentially in the same cell type. O.R.=odd ratio and p-value of one-tail Fisher exact test are reported.

## ASD probands show dysregulation in TF regulators of excitatory and inhibitory neuron specification

We have previously shown by scRNA-seq that organoids derived from macrocephalic ASD individuals in the present cohort (macroASD) showed an increase in EN-DCP cell proportions and gene expression as compared to their unaffected fathers, whereas normocephalic ASD individuals (normoASD) showed a diametrically opposite phenotype [19]. Here, by analyzing paired ASD versus control-father DEGs by bulk RNA-seq (pDEG in **Extended Data Fig. 7**) and differentially active enhancers (pDAE in **Extended Data Fig. 8a**) we confirmed those opposite phenotypes at the bulk transcriptomic and epigenomic levels (n=4 macroASD pairs and n=5 normoASD pairs, TD0 and TD30/60 analyzed separately, see **Methods** & **Table S4.T1-2**). As in the scRNA-seq dataset, most DEGs were changed either in a cohort-specific manner or in opposite directions, with only 13 concordant pDEGs between the macroASD and normoASD cohorts (out of 641 pDEGs in either cohort at TD30/60, **Extended Data Fig. 7c**). Furthermore, bulk transcriptome data reproduced the upregulation and downregulation of gene markers of EN-DCP cells in macroASD and normoASD, respectively (**Extended Data Fig. 7b-d**). This was further supported by an upregulation of enhancers upstream of several of those EN-DCP genes in macroASD pDAE (including of EMX1, NEUROG2, NFIA, and LHX2) (**Extended Data Fig. 8a,b)**. In addition to this EN phenotype, a strong downregulation in genes related to the inhibitory neurons (IN) lineage (e.g., DLX-genes and GSX2) was observed at both TD0 and TD30/60 in macroASD pDEG (**Extended Data Fig. 7a-d**).

We then exploited our GRN to deconvolute the regulatory drivers of those opposite phenotypes. We generated a **macroASD subnetwork** by selecting TF upstream of macroASD pDEGs (**Fig. 3a**); resulting in 96 (out of 420) pDEGs being controlled by 102 TFs, of which 19 were pDEGs themselves (**Fig. 3a**). We then applied TF perturbation analysis (**Extended Data Fig. 6**) to systematically rank each TF's ability to drive differential expression of its downstream targets in ASD. Both the macroASD subnetwork and the perturbation analysis highlighted an "EN hub" upregulated in macrocephalic ASD, composed of the master regulators BHLHE22, EOMES and FOXG1, which are the strongest drivers of the macro ASD pDEG (**Fig. 3b**) and are strongly associated with EN-DCP fates (**Fig. 2e**). BHLHE22, EOMES and FOXG1 acted on several downstream TFs linked to EN-DCP fate such as TBR1, NEUROG2, ZBTB18 and BCL11A (**Fig. 2e, 3a, 3b**). While we had lower sensitivity to detect pDAEs than pDEGs, in several cases the intermediate enhancer connecting the TF to the downstream pDEG in the GRN was itself a pDAE (**Extended Data Fig. 8b**). For instance, in macroASD, the upregulation of the TF ZBTB18 could be associated with the upregulation of two of its enhancers (ID: 123747, 124166) themselves bound by the TFs BHLHE22, TBR1 and TFAP2C, three upregulated pDEGs. Similar intermediate pDAEs linked FOXG1 to BCL11A, FOXG1 to LHX2, and EOMES to EMX1 (**Extended Data Fig. 8b**). Such cases identify potential "propagation paths" of the phenotype where expression changes cause epigenomic changes in linked enhancer elements and ultimately differential expression of secondary downstream targets. In agreement with the observed decreased in IN genes in macroDEGs, the known inhibitory lineage TFs ASCL1 and LHX6 ranked among the top driver TF with decreased activity in macroASD (**Fig. 3b**).

We performed similar analysis in the normoASD cohort, obtaining a **normoASD subnetwork** which placed 85 pDEGs (out of 315 pDEG at TD30/60) under the control of 103 upstream causative TFs. In normoASD, the "EN hub" composed of the master regulators BHLHE22, EOMES and FOXG1 was strongly downregulated, causing a secondary downregulation of NEUROG2, TBR1, ZBTB18, LHX2, BCL11A, NEUROD6, and EMX1 (**Fig. 3c,d**). These TFs often co-regulated each other, with, for instance, LHX2 downstream of EOMES and FOXG1, and ZBTB18 controlled by both BHLHE22 and TBR1. Interestingly, the top three TFs with increased activity in normoASD (i.e., the up pDEGs LMX1B, FOXB1 and FOXA1, **Fig. 3d**) were direct repressors of FOXG1 and were oppositely affected in the macroASD cohort (down pDEGs). Taken together, the GRN identifies opposite dysregulation of two triads of co-regulated TFs (BHLHE22, EOMES and FOXG1 against LMX1B, FOXB1 and FOXA1) as the strongest drivers of the opposite transcriptomic phenotypes observed in our macro- and normo-ASD cohorts. Those results reinforce that, despite sharing an ASD diagnosis, organoids from these cohorts of ASD probands revealed divergent transcriptome as well as epigenome.

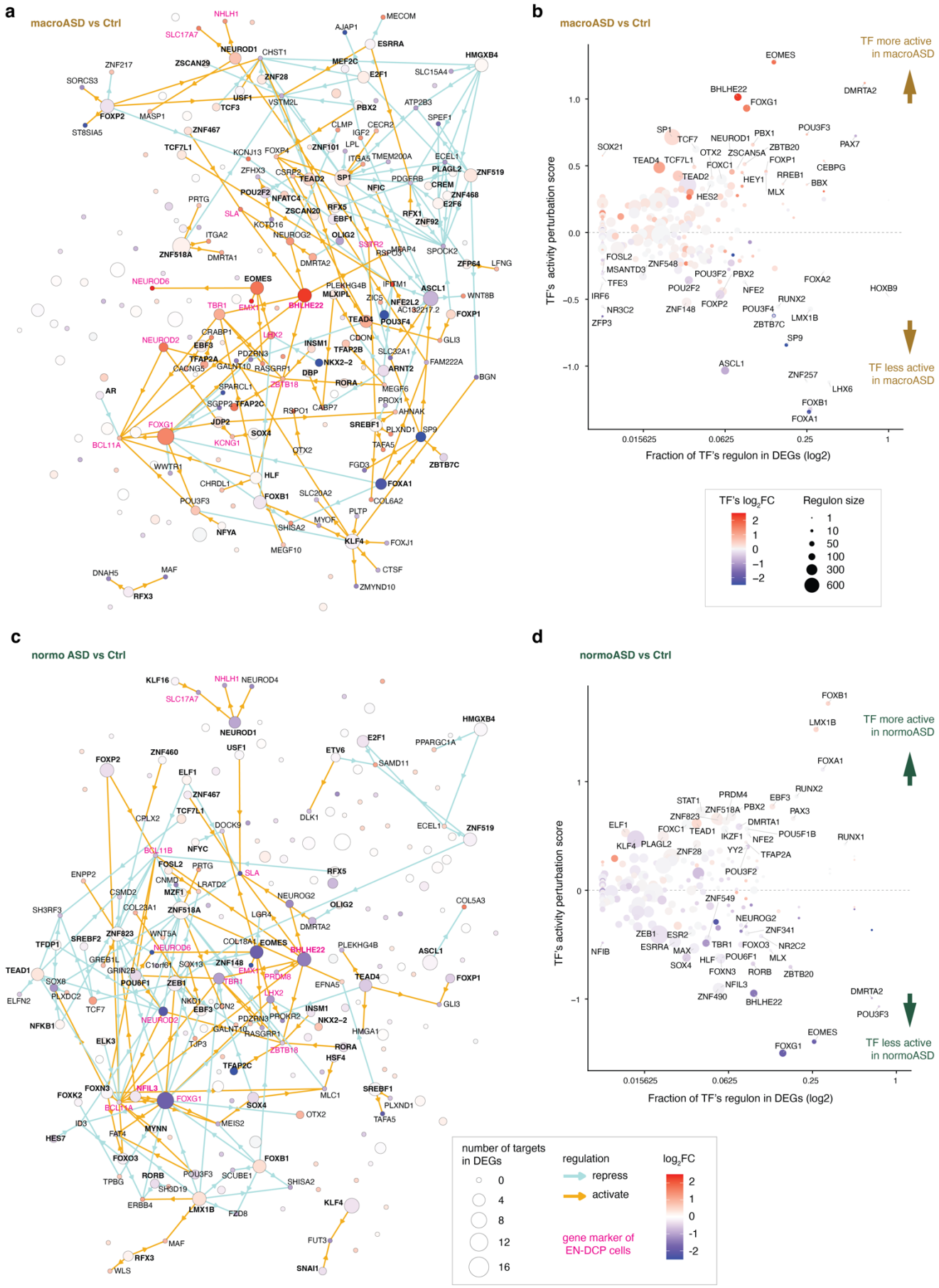

a
macroASD vs Ctrl
b
macroASD vs Ctrl
TF more active in macroASD
TF less active in macroASD
TF's activity perturbation score
Fraction of TF's regulon in DEGs (log2)
TF's log2FC
Regulon size
c
normo ASD vs Ctrl
d
normoASD vs Ctrl
TF more active in normoASD
TF less active in normoASD
number of targets in DEGs
regulation
repress
activate
log2FC
gene marker of EN-DCP cells

Figure 3. ASD probands have head-size cohort-specific alterations of the regulatory network coordinated by central transcription factors like FOXG1.

**a,c.** TF-to-gene GRN subnetwork depicting pDEGs and their upstream TFs for macroASD (a) and normoASD (c). Color of the genes represents expression change, while color of arrows represents the regulatory action (activation/repression) from the GRN (Fig. 1g). Genes are disposed identically in **a** and **c**. Dot size show number of targets in the resulting network. Gene markers of EN-DCP cells from scRNA-seq data are in red.

**b, d.** TF ranking based on their activity perturbation scores against macroASD pDEGs (**b**) and normoASD pDEGs (**d**). TFs are plotted based on the fraction of their downstream targets that are significantly pDEGs (x-axis) versus the TF's activity perturbation scores (y-axis) which evaluates each TF based on how strongly its predicted effect in the GRN (i.e., activation/repression) accounts for the observed pDEGs in its regulon. Dot colored is the log2FC value of the TF itself (ASD vs Ctrl), dot size= regulon size.

## FOXG1 loss of function validated the gene regulatory network

As FOXG1 was identified as one of the top candidate drivers of the ASD phenotypes in the GRN, we further explored its central influence on the development of forebrain organoids by focusing on its regulatory network. The upstream regulatory network comprises 14 activating GLEs in the FOXG1 *cis*-regulatory region (chr14:28,751,330–30,3825,86), which are bound by 11 repressors (including the top three TFs with increased activity in normoASD as mentioned above: FOXB1, LMX1B, and FOXA1) and 2 activators (NFIL3, ZEB1) with additional other indirect putative regulators upstream (**Fig. 4a**). In scRNA-seq data, FOXG1 was strongly expressed in RG cells, silenced in hem-RG and cells, preplate (EN-PP) and medial cortical plate (MCP), and re-expressed in postmitotic EN and IN of the dorsal cortical plate (DCP) (**Fig. 4b**). Consistently, several TFs that repressed FOXG1 (such as EBF3, FOXB1, FOXA1, ZNF823 and LMX1B) activated EN-PP genes and/or repressed EN-DCP genes (**Fig. 2e**). Downstream binding targets of the FOXG1 protein included 418 activated GLEs and 281 repressed GLEs across the genome, suggesting that FOXG1 could act as an activator or as a repressor of enhancer activity (**Extended Data Fig. 9a, Table S2**). Through those enhancers, we identified 111 downstream genes in the regulon of FOXG1, including 8 TFs (POU3F3, BCL11A, LHX2, MEIS2, MEOX1, OTX2, SOX13, ZNF503, **Extended Data Fig. 9b**, **Table S3**). FOXG1 activated targets were expressed in RG progenitor, IN (e.g., MEIS2) and EN-DCP (e.g., BCL11A, LHX2), while some of its repressed targets (e.g., OTX2) were expressed in cells with medial fates (hem-RG, and MCP cells) (**Fig. 4b, Extended Data Fig. 9c**). Finally, inside the GRN, FOXG1 competes preferentially with FOXB1, LMX1B or POU5F1B and collaborate more specifically with FOXK2, POU3F3 or FOXO3 across shared targets (**Extended Data Fig. 9d**).

To validate the FOXG1 regulon, we generated iPSC lines with FOXG1 loss of function mutations (FOXG1-/-). This was achieved by CRISPR/Cas9 genome editing to introduce frame-shifting indels (see **Methods**), resulting in generation of a premature stop codon and a truncated FOXG1 protein lacking the forkhead-DNA binding domain (**Fig. 4c**). We then assessed differences in gene expression in organoids by bulk RNA-seq at TD30/60 by comparing two homozygous FOXG1-/- edited lines to two FOXG1+/+ control lines (termed LOF-DEGs, **Table S3.T3, Methods**). LOF-DEGs comprised 250 downregulated genes and 620 upregulated genes (**Fig. 4d**). The top downregulated LOF-DEG was FOXG1 itself, suggesting that the FOXG1 protein exerts a positive autoregulatory effect on its own expression or that the FOXG1 message undergoes non-sense mediated decay. We observed that genes expressed in EN-DCP and RG were downregulated in LOF-DEG, while gene expressed in EN-PP, RG-hem and MCP, such as OTX2, were upregulated (**Fig. 4d,e**). This was concordant with the cell type specificity of many FOXG1 targets identified via the GRN (**Figs. 4b, Extended Data Fig. 10c**) and with the role of FOXG1 in cortical patterning in mouse studies[22,23], where its inactivation leads to a decrease of dorsal cortical fates and an increase in medial fates. Put together, the forebrain organoid system confirmed and reproduced the fundamental role of FOXG1 to define forebrain cortical fate by repressing medial fates, including preplate fates (i.e., EN-PP, RG-hem, MCP).

When comparing the FOXG1 LOF-DEGs to the constructed GRN, 37% of FOXG1 targets in the GRN (30 repressed and 11 activated) were significant LOF-DEGs with concordant direction of change (**Fig. 4f**). While not reaching significance for DEG, the majority of the remaining 70 FOXG1 targets in the GRN were affected by FOXG1 LOF in the expected direction. Next, the LOF-DEG results were used as an input for TF activity perturbation analysis, which yielded FOXG1 as one of the strongest drivers of the LOF-DEGs with predicted decrease in activity (**Fig. 4g**), confirming our approach for identifying driver TFs from differential expressions using the GRN (**Extended Data Fig 6**). Thus overall, the FOXG1 loss-of-function experiment validated the constructed GRN.

Finally, we observed that gene expression alterations induced by the FOXG1 LOF experiment were negatively correlated with macroASD pDEGs and positively correlated with normoASD pDEGs (**Fig. 4h**). This was attributable to top upregulated genes in macroASD being downregulated in LOF-DEGs, and to upregulated genes in normoASD being upregulated in LOF-DEGs. This similarity between ASD pDEGs and FOXG1 LOF-DEGs supported the hypothesis that FOXG1 dysregulation drives a strong part of the ASD phenotype in both cohorts.

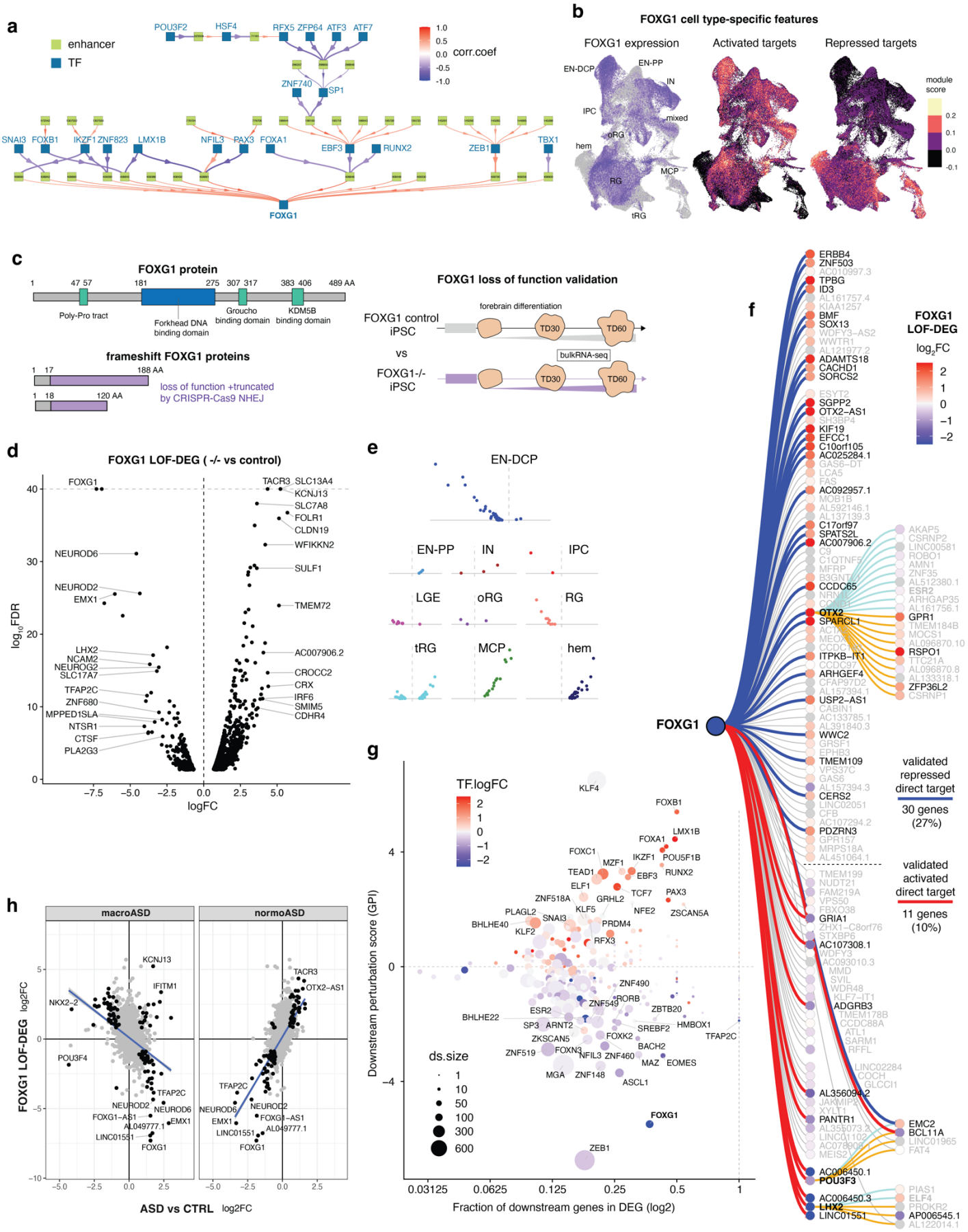

a
enhancer
TF
corr.coef
FOXG1
b
FOXG1 cell type-specific features
FOXG1 expression
Activated targets
Repressed targets
module score
c
FOXG1 protein
Poly-Pro tract
Forkhead DNA binding domain
Groucho binding domain
KDM5B binding domain
frameshift FOXG1 proteins
loss of function +truncated by CRISPR-Cas9 NHEJ
FOXG1 loss of function validation
forebrain differentiation
FOXG1 control iPSC
vs
FOXG1-/- iPSC
bulkRNA-seq
TD30
TD60
d
FOXG1 LOF-DEG ( -/- vs control)
log10FDR
logFC
e
EN-DCP
EN-PP
IN
IPC
LGE
oRG
RG
tRG
MCP
hem
f
FOXG1 LOF-DEG
log2FC
FOXG1
validated repressed direct target
30 genes (27%)
validated activated direct target
11 genes (10%)
g
TF.logFC
ds.size
Downstream perturbation score (GPI)
Fraction of downstream genes in DEG (log2)
h
macroASD
normoASD
FOXG1 LOF-DEG log2FC
ASD vs CTRL log2FC

**Figure 4: Validating the downstream targets of the master regulator FOXG1 by CRISPR inactivation**

**a.** Upstream regulators of FOXG1 including their respective cis-GLE and upstream TFs.

**b.** UMAP plots showing FOXG1 expression level in cell types juxtaposed with the positive module score expression of its downstream activated and repressed target genes.

**c.** CRISPR-induced FOXG1 loss of function experiment. Schematic representation of the normal FOXG1 protein and the two predicted out-of-frame truncated proteins generated by CRISPR/Cas9 cleavage and non-homologous end joining (NHEJ), retaining only 17-18 amino acids (AA) of the normal FOXG1 protein and depleted of all the FOXG1-functional domains, including the forkhead DNA-binding domain.

**d.** Volcano plot of FOXG1 LOF-DEGs (comparing FOXG1 LOF vs non-edited and the founder iPSC line).

**e.** Volcano plots as in **d** highlighting FOXG1 LOF-DEGs that have cell type-specific expression, highlighting the strong downregulation of EN-DCP genes and upregulation of medial fate (Hem, MCP).

**f.** Graph of FOXG1 regulon with targets colored by FOXG1 LOF-DEG results. Edges connecting FOXG1 to its targets are colored in blue for repression or red for activation when validated by FOXG1 LOF-DEG. Targets where considered validated if the predicted relationship in the GRN was concordant with the obtained LOF-DEG (FDR < 0.05 & absolute logFC > 0.25 and compatible direction of change). Direct target genes are organized from top to bottom based on their original correlation level with FOXG1 in the GRN. Dashed line separates predicted activated target (bottom) from repressed targets (top). Note that for cases where the FDR of FOXG1 LOF-DEGs does not reach significance (grey edges) the direction of change of the target in FOXG1 LOF-DEG remains concordant with the GRN prediction in a majority of cases. Secondary targets of FOXG1 are also reported (through OTX2, POU3F3 or LHX2 regulation).

**g.** Perturbation plot for FOXG1 LOF-DEG, plotting for each TF in the GRN, the fraction of its regulon in LOF-DEGs (x-axis) versus its activity perturbation score (as in **Extended Data Fig. 6a**). Log-2 fold change of the TF itself (color) and its regulon size (dot size) are indicated. Note that this analysis uses the GRN and the FOXG1 LOF-DEG result to *naively* re-identify FOXG1 as one of the main driver of the difference.

**h.** Confusion plots comparing logFC results between FOXG1 LOF-DEG and ASD vs Ctrl pDEG for macroASD and normoASD. Top genes by distance from the origin are shown. Linear regression line is shown in blue (Pearson's correlation R=-0.4 and 0.63, respectively, with both p-val < $2.2.10^{-16}$)

## Discussion

Using parallel epigenomic and transcriptomic datasets we constructed a regulome encompassing gene and enhancers in developing organoids. We leveraged the identification of proximal and distal gene-enhancer relationships and TFBS to infer regulatory action of TF on both enhancer activity and linked gene expression. The obtained network captured the driving forces behind time transitions and cell type specification during neurogenesis. The predictive capacity of this network was finally validated using the loss of function of the master regulator TF FOXG1, a central component of the network.

By using iPSCs derived from idiopathic ASD individuals we identified the TFs and dysregulated enhancers that drive the imbalance of excitatory neurons previously discovered in this ASD cohort [19]. Specifically, we suggest that a dysregulation of FOXG1, BHLHE22 and EOMES expression in opposite directions in ASD with macrocephaly and ASD with normocephaly drives the excitatory neuron imbalance in opposite directions in these ASD subtypes. Dysregulated genes in the FOXG1 LOF experiment were largely overlapping with those affected in macroASD and normoASD, strongly suggesting that FOXG1 is a central causal factor of differential gene expression in ASD. While based on a limited number of individuals, our study is a first draft of gene regulatory interactions disrupted in this disorder during early neural development.

One limitation of our GRN approach was the reliance on bulk-level data. While we show that the network captured major cell type specification, many regulatory relationships revolved around the formation of cortical EN from RG cells, the main lineage produced in forebrain organoid. We expect other subpopulation-specific regulatory interactions, behind the formation of IN cells or layer-specific cortical neurons, would be further delineated by expanding sample collections, including later stages and larger cohorts of individuals, and by incorporating single-cell level information in building of the network. Indeed, scATAC-seq data that captures open chromatin regions in each cell type overlapped only in part with ChIP-seq data, suggesting that combining both approaches could fully resolve the regulatory events occurring in each cell type. While we could delineate from the GRN major collaborators and competitors of each TF (**Extended Data Fig. 4**), further data collection would also help refine non-linear combinatorial effects occurring when multiple TFs co- regulate the same enhancers and genes.

Altogether, the regulome establishes a hierarchy between TFs and cognate enhancers that controls the dynamics of cell differentiation during typical and atypical human neurodevelopment. We show that fundamental TFs like FOXG1 can drive cascades that are extremely complex in humans, with hundreds of downstream enhancers and genes. The built regulome can have larger implications, as it allows to understand how regulators interact together to perturb brain development. For instance, using the list of identified enhancers and their upstream networks (**Table S2**), one could identify the upstream enhancers and TFs that regulate ASD risk genes (for instance the SFARI risk genes CDH3 in **Extended Data Fig. 10a**). Charting the upstream network de facto extends the genomic regions of interest for the identification of disruptive variants of any given risk gene to noncoding regions or indirectly associated coding regions, identifying paths of regulatory convergence in ASD. We showcase that principle using ASD-associated risk genes identified in SFARI, showing a clear central role of ASD risk genes such as TBR1, FOXP2 or FOXG1 to regulate other risk genes (**Extended Data Fig. 10b**). Untangling the complexity of these regulatory interactions through regulatory network analysis may reveal the paths of convergence behind the genetic heterogeneity of human developmental disorders such as ASD, schizophrenia or intellectual disabilities.

# Methods

## SAMPLE COLLECTION

### Organoid preparation

All iPSC lines and forebrain organoids preparations used in this study were identical to a previous report (see **Table S1**) [19]. Briefly, IPSC lines from controls and ASD-probands were propagated in parallel and grown in mTeSR1 media (StemCell Technologies). Forebrain organoids were generated using dual SMAD inhibition (SB431542, 10µM and LDN193189, 1µM, 7 days), followed by a proliferation phase (FGF2, 10 ng/ml + EGF, 10 ng/ml, 7days) and a terminal differentiation phase (starting at day 17, denominated TD0) in terminal differentiation media composed of Neurobasal medium supplemented with 1% N2, 2% B27 w/o vitamin A, 15 mM HEPES, Glutamax, NEAA and 55 µM 2-ME, with 10 ng/ml BDNF (R&D), 10 ng/ml GDNF (R&D). See [19] for full details.

### Chromatin immunoprecipitation

Native ChIP was carried out as previously described [5]. For chromatin pull-down, the following antibodies were used: anti-H3K4me3 (Cell Signaling, Cat#9751S), anti-H3K27ac (Active Motif, Cat# 39133), anti-H3K27me3 (Cell Signaling, cat# 9733S) and H3K4me1 (Diagenode, cat#C15410037). ChIP-libraries and sequencing were performed at the Yale Center for Genome Analysis. Paired-end DNA sequencing (2x100bp) was performed on an Illumina Hi Seq2000 to an average depth of 40M reads per sample. Organoids for ChIP-seq and bulk RNA-seq were collected in parallel from the same batch of differentiation.

### RNA extraction and sequencing

Organoids samples were collected at TD0, TD30 and TD60. Total RNA extraction was performed from at least 20 dissociated organoids per sample using Arcturus PicoPure RNA isolation kit (AppliedBiosystems, cat. 12204-01). Library preparation was performed with rRNA depletion and sequenced on an Illumina Hiseq/Novaseq to obtain an average depth of 40M reads per sample (paired-end 100bp reads) at the YCGA. Organoids for bulk RNA-seq and ChIP-seq were collected in parallel from the same batch of differentiation.

### FOXG1 LOF line

Loss-of-function (LOF) mutations in the FOXG1 gene were obtained by CRISPR-Cas9 genome editing as previously described [24]. Briefly, a guide RNAs (gRNAs) targeting the beginning of the first and only exon of the FOXG1 genes (5' CAGGCTGTTGATGCTGAACG3') followed by the PAM seq 5' AGG 3' was used to generate FOXG1 LOF iPSC lines. Cas9/gRNA mediated indels were generated in a control iPSC male line (1123-01#3) by non-homologous end joining (NHEJ), which resulted in frameshifts and premature stop codons, thus generating FOXG1 truncated proteins and presumably FOXG1 LOF alleles. Single cell-derived edited clonal lines, each containing a homozygous or heterozygous genetic mutation in the FOXG1 locus, were generated from the 1123-01#3 iPSC edited line by the limiting dilution process and screened by Sanger sequencing to identify and validate the FOXG1 mutation in each clone. Genomic DNA from single clones was extracted, PCR was performed to amplify the targeted region, and the PCR products were analyzed by Sanger-sequencing. Upon genotyping by Sanger-seq, one of the edited clones, was also analyzed by whole genome sequencing at 30x sequencing coverage [24]. Two homozygous mutations generated in different iPSC clonal lines were used in this work: the deletion of 10 base pairs (bps) or the insertion of 1 bp (Fig. 4C), both resulting in frameshift and premature stop codon, generating a 188 or 120 amino acid-truncated protein, retaining only 17 or 18 amino acids (AA) of the normal FOXG1 protein, respectively. Both versions of the truncated FOXG1 protein are depleted of all the FOXG1-functional domains, including the forkhead DNA binding domain (see schematic of the resulting FOXG1 truncated proteins in **Fig. 4C**). These FOXG1 -/- lines were compared to control FOXG1 +/+ lines, including both the founder iPSC line before editing and clonal iPSC lines where the CRISPR/Cas9-induced DNA breaks didn't result in detectable genome edits.

## DATA ANALYSIS

### Enhancer discovery

ChIP-seq reads were mapped to the UCSC hg38 reference genome using bowtie2 [25]. ChromHMM version 1.22 [8] was used to segment the genome based on the ChIP-seq profile. ChromHMM fits an emission profile for each experimental group, which we defined as organoid TD + donor ASD diagnosis. We manually determined that a Hidden Markov Model with nine hidden states provides the best fit for the data, based on how many states aligned with the roadmap epigenome emissions signature. These states were grouped into broad regulatory categories: states: 1, 8, and 9 were defined as promoter states ("Prom"), 2 and 3 as repressed states ("Rep"), and the remainder as enhancer states ("Enh") (**Fig. 1c**).

To construct a set of candidate regulatory regions that are consistent across all experimental conditions, we adapted our previous strategy [5] to merge and filter H3K27Ac ChIP-seq peaks. MACS2 [26] was used to call "original peaks" (OP) for each ChIP-seq sample against the corresponding DNA input collected during the ChIP experiment. Peaks from different samples with at least 1 bp overlap were merged using bedtools to generate H3K27Ac consensus peaks. Consensus peaks were then annotated based on unambiguous overlap with the hidden states inferred by ChromHMM. Briefly, if at least 60% of positions in a consensus peak overlapped with a particular broad category in a given stage, the consensus peak was annotated with that category at that stage; otherwise, it was annotated as "Mix". Stagewise annotations were consolidated into a global annotation for each consensus peak according to the following strategy: If the consensus peak was annotated as "Mix" in any stage, its global annotation was also "Mix". Consensus peaks that were annotated as "Prom" in one stage and "Enh" in another were also annotated as "Mix" globally. Otherwise, the global annotation took the most active annotation from any one stage, i.e. "Prom" or "Enh". "Rep" was attributed a consensus annotation only if all stages were annotated with "Rep". To increase confidence, we further required putative enhancers (i.e. consensus peaks annotated as "Enh") to be supported by at least four samples, in other words, each corresponding consensus peak was merged by OPs from at least four samples.

### Linking enhancers to genes and TF

The process of assigning putative enhancers to genes involves two strategies as previously described [5]. Briefly, transcription start sites (TSS) for all protein-coding and lncRNA genes were defined from GENCODE (v33) [27]. Enhancers within 21kb of any TSS on the linear genome were linked to the respective genes. These were supplemented with high-confidence long-range interactions between enhancers and a 1kb region around the TSS derived from Hi-C contact matrices in human fetal cortex [9] and hSC lines [10,11] with 10-kb resolution.

For transcription factor binding site analysis, homer2 [28] was used to test for enrichment of 1,201 human transcription factor position weight matrices (PWMs) from JASPAR2020 [29] among the DNA sequence of the GLEs. Transcription factor binding sites (TFBS) with an FDR-corrected enrichment p-value less than 0.01 were considered significant.

### Bulk RNA-seq analysis and differential gene expression

RNA-seq reads were aligned to the hg38 genome using STAR (v2.7) [30] and annotations from GENCODE (v33) [27]. Gene-level counts were estimated using *featureCounts* function from Subread (v2)[31]. Sequencing batch effects were removed using ComBat-seq [32]. Gene filtering (function *filterByExpr*), normalized RPKM expression values and all described differential expressed gene tests (DEGs) were generated following edgeR pipeline (v3.26)[33]. For the different DEGs presented in the study (results in **Table S4**), log-2 fold change (log2FC) and FDR-corrected p-values were computed from filtered counts with the *glmFit/glmLRT* functions, with "organoid stage" (TD0/30/60) and "family" used as covariates when appropriate (see **Table S1 for samples metadata**).

## Differential enhancer activity

Similar to RNA reads, H3K27ac read counts in all gene linked enhancers (173,400 GLEs) estimated using featureCounts v2.0.3 [31] were considered a measure of enhancer activity and tested for differential activity using edgeR (v3.4). For any given contrast between groups, enhancers with low activity were filtered out using the *filterByExpr* function in R and differential activity was evaluated using *glmFit* and *glmLRT* functions in edgeR.

## Enhancer-gene regulatory network

For each downstream gene target of a GLE, a pairwise Spearman correlation was computed between the GLE activity (H3K27ac ChIP-seq counts normalized as RPKM using EdgeR) and the expression of the target gene (RNA-seq RPKM). P-values were adjusted using the Benjamini-Hochberg method. Similarly, regulatory impact of TF on each GLE presenting the cognate TFBS was estimated as the Spearman correlation between expression of the TF transcript (RNA-seq RPKM) and the activity of the bound GLE (H3K27ac ChIP-seq RPKM). For both GLE-to-gene and TF-to-GLE, putative regulatory relationships were retained if passing FDR < 0.01 and absolute correlation coefficient > 0.25. The selected relationships were assembled into an *enhancer-gene regulatory network* (directed graph, available in **Table S2, T2-3**). Visualization and network-based computation were done using the R packages igraph (v1.4.3), tidygraph (v1.2.3) and ggraph (v2.1).

### Synthetic TF-gene-regulatory network

The synthetic TF-to-gene regulatory network was built from the original enhancer-gene regulatory network so that TF are connected to genes through at least one identified enhancer element presenting the cognate TFBS. First, cases where both the TF-GLE and the GLE-gene edges existed (so both having significant correlations) were retained to link each TF to each of its downstream genes (See **Extended Data Fig. 2** for the number of cases of regulatory cascades considered and retained). Then, Spearman's correlation between the TF and its downstream gene(s) expressions (both RNA RPKM) were calculated and retained if FDR < 0.01 and absolute correlation coefficient ($\rho > 0.25$ or $\rho < -0.25$). Edges were then retained if the sign of the TF-to-gene correlation was coherent with the TF-to-GLE and GLE-to-gene correlation signs, which is equivalent to:

$$\rho_{\text{TF-gene}} \cdot \rho_{\text{TF-GLE}} \cdot \rho_{\text{GLE-gene}} > 0$$

For plotting, the correlation between the genes was used to compute a positive edge weight in the final gene regulatory network with weight = $\rho_{\text{TF-gene}}/2 + 0.5$ (or its inverse depending on the algorithm weight definition). Networks plotted in the different figures were obtained using either Sugiyama layout, which limits the number of edge crossing in the final hierarchical plot or Kamada-Kawai layout in igraph/ggraph, which effectively repulse strongly in the final plot genes connected by a negative correlation edge (repressive) than genes connected by a positive correlation.

### TF prioritization by TF perturbation score

For integrating gene regulatory network with results obtained from differential gene expression results (DEG, by edgeR), we were inspired by the *AUCell GRN* method for scRNA-seq data (part of the SCENIC suite [34]) which evaluates the activity of the TF based on the downstream targets expression enrichment in each cells. To infer similar effect using bulk data and DEG results, we used an adaptation of the *geometric perturbation index* [35]. We exemplified the use of this **TF perturbation score** in **Extended Data Fig. 6a-b**. For a TF with N targets in the GRN, its perturbation score for a given DEG result was defined as:

$$TF.score = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \rho_{GRN.TF \to i} \times log_2 FC_{DEG.i} \; (1 - FDR_{DEG.i})$$

where for the $i^{th}$ target, $\rho_{\text{GRN.TF->i}}$ is the Spearman's correlation coefficient between the TF and the $i^{th}$ target expression level across samples as described above (corresponding to the edge's weight in our GRN, with: $-1 < \rho < 1$). The DEG's $log_2$FC value of the downstream gene is weighted by its false non-discovery rate ($1\text{-}FDR_{DEG}$) to downplay the effect of genes with low significant change without having to set a threshold on the significance.

For each edge, the product of the sign of the logFC (-1 for DOWNregulation and +1 for UPregulation) with the sign of the correlation coefficient (+1 for activation, -1 for repression) effectively evaluate if the direction of change is compatible with the action of the TF without having to treat separately activation/repression and upregulation/downregulation (See **Extended Data Fig. 6b** for examples). This generates a score with biological interpretation: a positive TF perturbation score indicates an upregulation of the TF actions described in the GRN (i.e. TF activating/repressing actions on its downstream genes), a near-zero score suggest little implication of the TF in the observed DEG results, and a negative score indicates a downregulation of the TF action. The score only considers the differential expression of the downstream targets and not of the TF itself and use the built GRN to infer the degree to which a change in the TF activity could have explained those observed differential expression.

This score was used in **Fig. 2b** with timeDEG (using in that case only activating edges with $\rho > 0$ to compute the score to facilitate interpretation, see **Extended Data Fig. 6c** showing all result); **Fig. 2e** with celltype-DEG (see below) ; **Fig.3b** with macro-ASD DEG ; **Fig.3d** with normo-DEG ; and **Fig. 4g** with FOXG1 loss of function DEG). For plots, we compared the TF perturbation score with the differential expression of the TF itself (dot blue-red color gradient), the number of target N (dot size) and the fraction of those downstream targets that meet strict DEG criteria (regardless of direction of change, cutoff: FDR < 0.05 or 0.01 and absolute $\log_2$FC > 0.25). Note that using the square root of N stems from the original geometrical definition of the metric and has here for effect to give higher score to genes with larger number of targets. For the specific case of **Fig. 2e**, we applied the TF perturbation score using cell type-specific DEGs computed in our scRNA-seq dataset. For cell type-specific DEGs (**Table S5**), the logFC and BH-adjuster p value were computed in Seurat (v3) using the *FindMarkers* function which performs differential expression test by comparing cells from each cell type to the remaining cells using a Wilcoxon's rank sum test.

### Intersection of enhancers regions external datasets

Overlap of enhancer consensus peaks identified in our data with gene enhancer external datasets were evaluated using BedTools (*intersectBed*) [36]. From this present study we used 3 different level: the full 367,587 consensus peaks annotated as "active enhancer" in any of our sample group ("all enhancers"), the 173,347 gene-linked enhancer subset (GLE, following the gene linkage step, **Table S2,T1**) and the 42,035 GLEs that show significant correlation with any linked gene or with any TF with TFBS as described above, **Table S2,T2-3**) For Screen enhancer regions from the ENCODE project [12], "candidate enhancer" coordinates in hg38 was downloaded from https://screen.wenglab.org/ (v3). For Genehancer dataset [13], the GeneHancer_v5.14 was obtained through the Genecard website. Enhancer dataset from Amiri et al. previously generated from fetal brain, iPSCs and organoids at TD0, TD11,TD30 by our lab was also used [5].

Intersection with ATAC-seq/scATAC-seq external datasets were performed using BedTools [36] considering a minimum of 20% overlap between the length of our enhancers and the length of peaks from the external dataset. For **Fig. 1**, the resulting data was plotted using the 'UpSetR' package in R [37].

### Intersection with Single cell ATAC-seq peak from parallel organoid data

To validate the enhancer activity in single cells, we intersected our list of enhancers with peaks called from 29 scATACseq samples using BedTools [36] (**Table S5**, T2). To process scATACseq data, Cell Ranger ATAC 2.0.0 [38] along with pre-built reference arc-GRCh38-2020-A-2.0.0 was applied to each sample from calling cells to generating a peak-cell count matrix. Signac vignette (https://stuartlab.org/signac/articles/pbmc_vignette.html) [39] was followed to filter the count matrix of each sample with the following cut-offs: cells containing at least 200 features, features present in at least 10 cells, log10 of fragments in peak regions per cell within two standard deviation of the mean, nucleosome signals between 2 and 4, TSS enrichment greater than 2, reads in peaks over 15% and less than 5% of peaks overlapping with black list. scATACseq data were annotated using label transfer method implemented in Seurat [40] along with our previously annotated scRNAseq datasets as reference [19]. Next peaks per library or per cell type were called using MACS2 as recommended by Seurat developers.

## TF interactome

TF-TF interactome graph based on shared gene targets between TFs in the GRN (**Table S3**). Specificity of the overlap was assessed by a connection specificity index (CSI) which is independent of the direction of the TF action (i.e. repress/activate) but evaluates if the overlap in regulon is higher for that pair of TF than with all other TFs (CSI) [41]. Collaboration vs competition was assessed by the Pearson's correlation coefficient (Inter.cor) between the TF's weights across all gene targets (negative values indicate opposite actions on the downstream genes on average).

## DATA AVAILABILITY

This study did not generate new unique reagents or DNA constructs.

Primary cell lines and iPSC lines are shared via the Infinity BiologiX LLC repository (https://ibx.bio/).

Datasets reported in this study are available through the NIMH Data Archive (NDA). The scRNA-seq and scATAC-seq data are under collection #C3957, url: https://nda.nih.gov/edit_collection.html?id=3957; the bulk RNA-seq data are under collection #C2424, url: https://nda.nih.gov/edit_collection.html?id=2424.

Code availability

All unpublished codes are described in the Methods section of the paper.

**Acknowledgments**

We are grateful to the families and children for their participation in this study. We thank Soraya Scuderi for help with the ChIP-seq protocol. We acknowledge Elise M. Cummings, Gloria Han, and Kelly Powell for help with subject recruitment and clinical characterization to produce iPSC lines and Arijit Panda for help with data upload into NDA and sharing. We thank Guilin Wang and Christopher Castaldi, and the Yale Center for Genome Analysis for library preparation, deep sequencing and Cell Ranger analysis. We thank Caihong Qiu and Jason Thomson at the Yale Stem Cell Center for the generation of the iPSC lines. We acknowledge the Yale Center for Clinical Investigation for clinical support in obtaining the biopsy specimens. We acknowledge the following grant support: National Institute of Mental Health R01 MH109648 (FMV), the Simons Foundation (Awards No. 399558 and 632742, FMV, AA). The Yale Stem Cell Center is supported in part by the Regenerative Medicine Research Fund.

**Author contributions**

F.M.V., A.Abyzov conceived the study, designed, and supervise experiments. J.McP., K. P., P.V., K.C., L.T., helped recruit patients and obtained clinical data; A.S. evaluated donor subjects and obtained skin biopsies; L.T. cultured primary cells, performed reprogramming and Q.C.'d the iPSC lines; J.M., A.J., A.Amiri generated organoid preps and processed them for bulk RNA-seq, scRNA-seq and ChIP-seq; F.W., S.N. and A.J. performed the bulk RNA-seq and ChIP seq bioinformatic analysis; F.W. performed scATAC-seq bioinformatic analyses; J.M. designed and performed the FOXG1-loss of function lines experiment; A.J., S.N., D.C. performed secondary analyses; A.J. generated display items and wrote the manuscript; all authors provided edits and comments on the manuscript.

**Competing Interests**

The authors declare no competing interests.

**Materials and Correspondence**

Correspondence and requests for materials should be addressed to Flora Vaccarino, Child Study Center, Yale University School of Medicine, 230 South Frontage Rd, New Haven CT 06520.

**References**


1 Mariani, J. *et al.* Modeling human cortical development in vitro using induced pluripotent stem cells. *Proc Natl Acad Sci U S A* **109**, 12770-12775 (2012).

2 Lancaster, M. A. *et al.* Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373-379 (2013).

3 Pasca, S. P. The rise of three-dimensional human brain cultures. *Nature* **553**, 437-445 (2018).

4 Visel, A. *et al.* A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**, 895-908 (2013).

5 Amiri, A. *et al.* Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science* **362** (2018).

6 Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet* **21**, 292-310 (2020).

7 Markenscoff-Papadimitriou, E. *et al.* A Chromatin Accessibility Atlas of the Developing Human Telencephalon. *Cell* **182**, 754-769 e718 (2020).

8 Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-216 (2012).

9 Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523-527 (2016).

10 Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294 (2013).

11 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).

12 Consortium, E. P. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).

13 Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017** (2017).

14 Ziffra, R. S. *et al.* Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* **598**, 205-213 (2021).

15 Trevino, A. E. *et al.* Chromatin accessibility dynamics in a model of human forebrain development. *Science* **367** (2020).

16 Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* (2022).

17 Fleck, J. S. *et al.* Resolving organoid brain region identities by mapping single-cell genomic data to reference atlases. *Cell stem cell* **28**, 1148-1159 e1148 (2021).

18 Ardhanareeswaran, K., Mariani, J., Coppola, G., Abyzov, A. & Vaccarino, F. M. Human induced pluripotent stem cells for modelling neurodevelopmental disorders. *Nature reviews. Neurology* **13**, 265-278 (2017).

19 Jourdon, A. *et al.* Modeling idiopathic autism in forebrain organoids reveals an imbalance of excitatory cortical neuron subtypes during early neurogenesis. *Nat Neurosci* (2023).

20 Runge, K. *et al.* Disruption of NEUROD2 causes a neurodevelopmental syndrome with autistic features via cell-autonomous defects in forebrain glutamatergic neurons. *Mol Psychiatry* (2021).

21 Joshi, P. S. *et al.* Bhlhb5 regulates the postmitotic acquisition of area identities in layers II-V of the developing neocortex. *Neuron* **60**, 258-272 (2008).

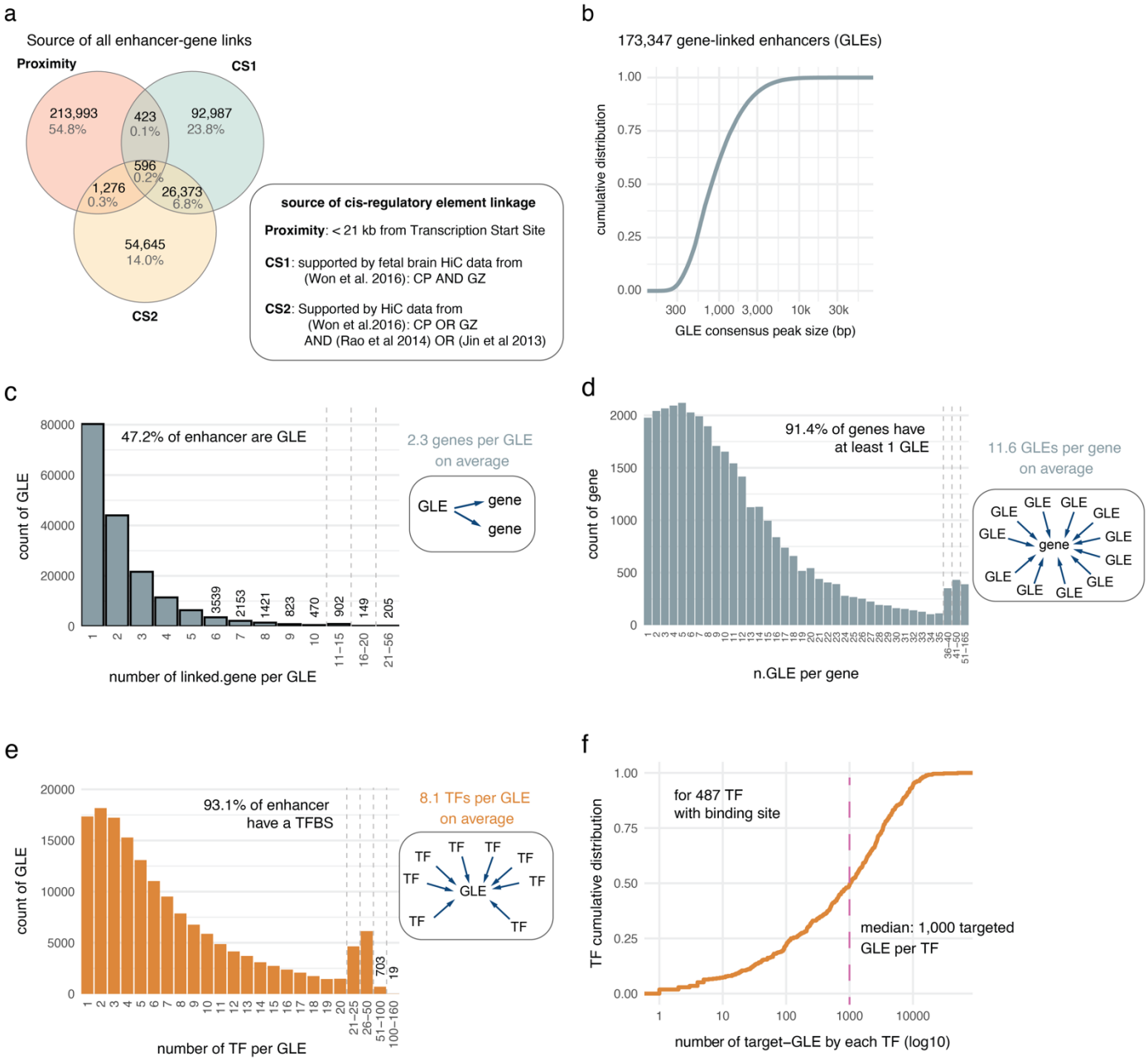
22 Muzio, L. & Mallamaci, A. Foxg1 confines Cajal-Retzius neuronogenesis and hippocampal morphogenesis to the dorsomedial pallium. *J Neurosci* **25**, 4435-4441 (2005).

23 Hanashima, C., Li, S. C., Shen, L., Lai, E. & Fishell, G. Foxg1 suppresses early cortical cell fate. *Science* **303**, 56-59 (2004).

24 Panda, A. *et al.* Clonally Selected Lines After CRISPR-Cas Editing Are Not Isogenic. *CRISPR J* **6**, 176-182 (2023).

25 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
26 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
27 Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773 (2019).
28 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).
29 Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **48**, D87-D92 (2020).
30 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
31 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).
32 Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* **2**, lqaa078 (2020).
33 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
34 Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083-1086 (2017).
35 Martin, F. *et al.* Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. *BMC Syst Biol* **6**, 54 (2012).
36 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
37 Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938-2940 (2017).
38 Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**, 925-936 (2019).
39 Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**, 1333-1341 (2021).
40 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529 (2021).
41 Fuxman Bass, J. I. *et al.* Using networks to measure similarity between genes: association index selection. *Nat Methods* **10**, 1169-1176 (2013).

**Extended Data Fig. 1**

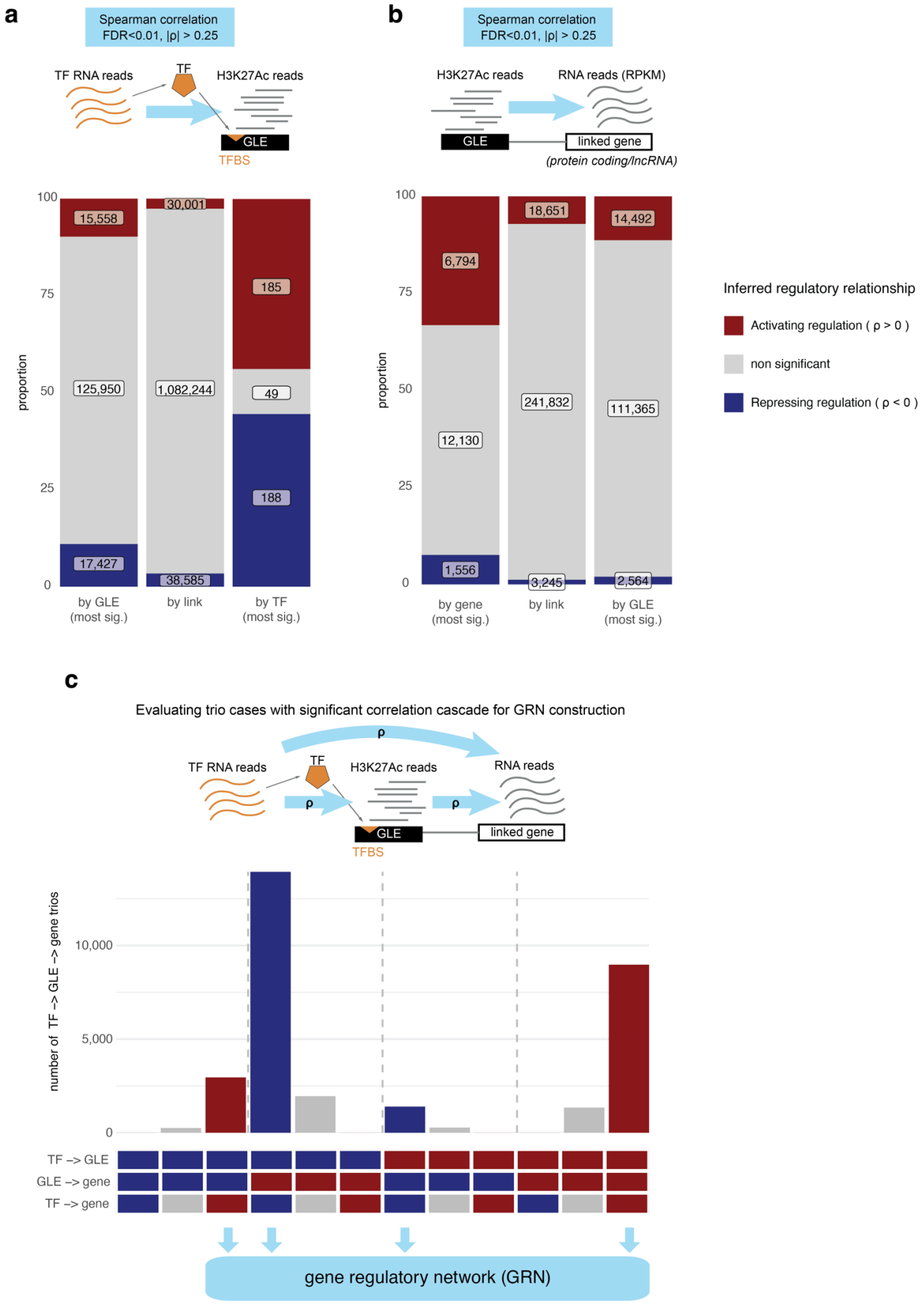



**Extended Data Fig. 1: General metrics related to the linkage between genes, gene-linked enhancers (GLE) and TFs**

**a)** Venn Diagram showing the sources used to link each GLE to downstream gene(s). See boxed description: CS1= confident set 1 supported by Hi-C linkages from both cortical plate (CP) and germinal zone (GZ) datasets from 3 human fetal brains (Won et al 2016, ref #9); CS2= confidence set 2, supported by CP OR GZ set (Won et al 2016, ref#9), AND by either Rao et al 2014 (ref# 11) OR Jin et al 2013 (ref # 10) human stem cell data. **b)** Cumulative distribution of the sizes of all GLE consensus peaks. Wider GLE peaks were obtained by merging sets of overlapping peaks throughout samples (see **Methods**). **c)** Histogram showing the distribution of

number of linked genes per GLE. **d)** Histogram showing the distribution of number of linked GLE per gene. **e)** Histogram showing the distribution of number of unique TF targeting each GLE (though the detection of TFBS). **f)** Cumulative distribution of the number of targeted GLE per TF. Pink line=median. X-axis is in log10 scale.

## Extended Data Fig. 2



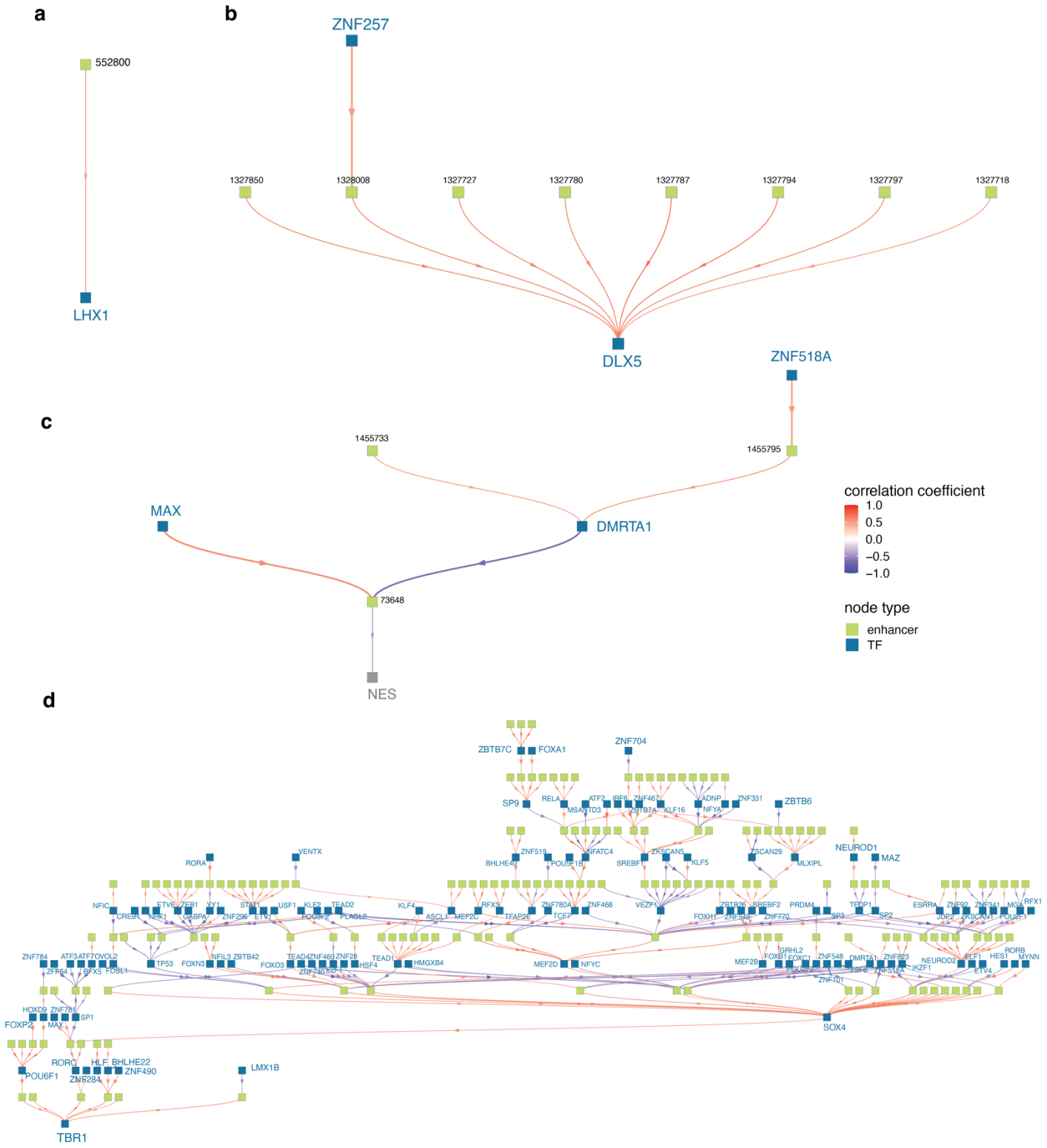
**Extended Data Fig. 2: Inferring regulatory relationship by correlating enhancer activity and gene expression for each TF-GLE and GLE-gene linkage**

**a.** Bar plots showing the classification of TF-to-GLE regulatory relationships as evaluated by significant correlation between GLE's H3K27ac reads level and TF's RNA reads level (two-sided Spearman's correlation coefficient $|\rho| > 0.25$ and FDR-corrected p-value < 0.01). "by GLE (most sig.)" = status of GLE according to its most significant correlation (since GLE can be connected to multiple TFs and vice-versa, only the link with the highest significance is considered here to categorize the element as activated or repressed); "by link" = status of each unique link (i.e,. TF-enhancer pair), "by TF (most sig.)" = status of each TF according to their most significant correlation.

**b.** Bar plots showing the proportion of GLE-to-gene regulatory relationships as evaluated by significant correlation between H3K27ac reads level and linked transcript RNA reads level (Spearman's correlation coefficient $|\rho| > 0.25$ and FDR-corrected p-value < 0.01). "by gene (most sig.)" = status of gene according to their most significant correlation (i.e., activated or repressed); "by link" = status of each unique GLE-gene link evaluated; "by GLE (most sig.)" = status of each GLE according to their most significant correlation.

**c.** Bar plot showing the type of regulatory cascade identified for TF-GLE-target gene trios. Trios are separated by regulatory direction for each step, including the presence of an overall significant correlation between the transcript RNA level of the TF and the putative downstream gene in bottom row, as indicated by colors below the plot (red=activating, blue=repressing as in **a**, **b**). Note the absence of trios with *inconsistent* relationships. An example of inconsistent relationship would be a case where the TF is negatively correlated with the GLE and the GLE is positively correlated with the downstream gene but where the TF is positively correlated with the downstream gene suggesting an inconsistency in the cascade. For building the gene regulatory network (GRN) linking directly TF to genes, only trios with consistent correlations cascades were retained (bottom arrows), removing the cases where the TF and the downstream genes expression levels were not correlated with each other.

**Extended Data Fig. 3**

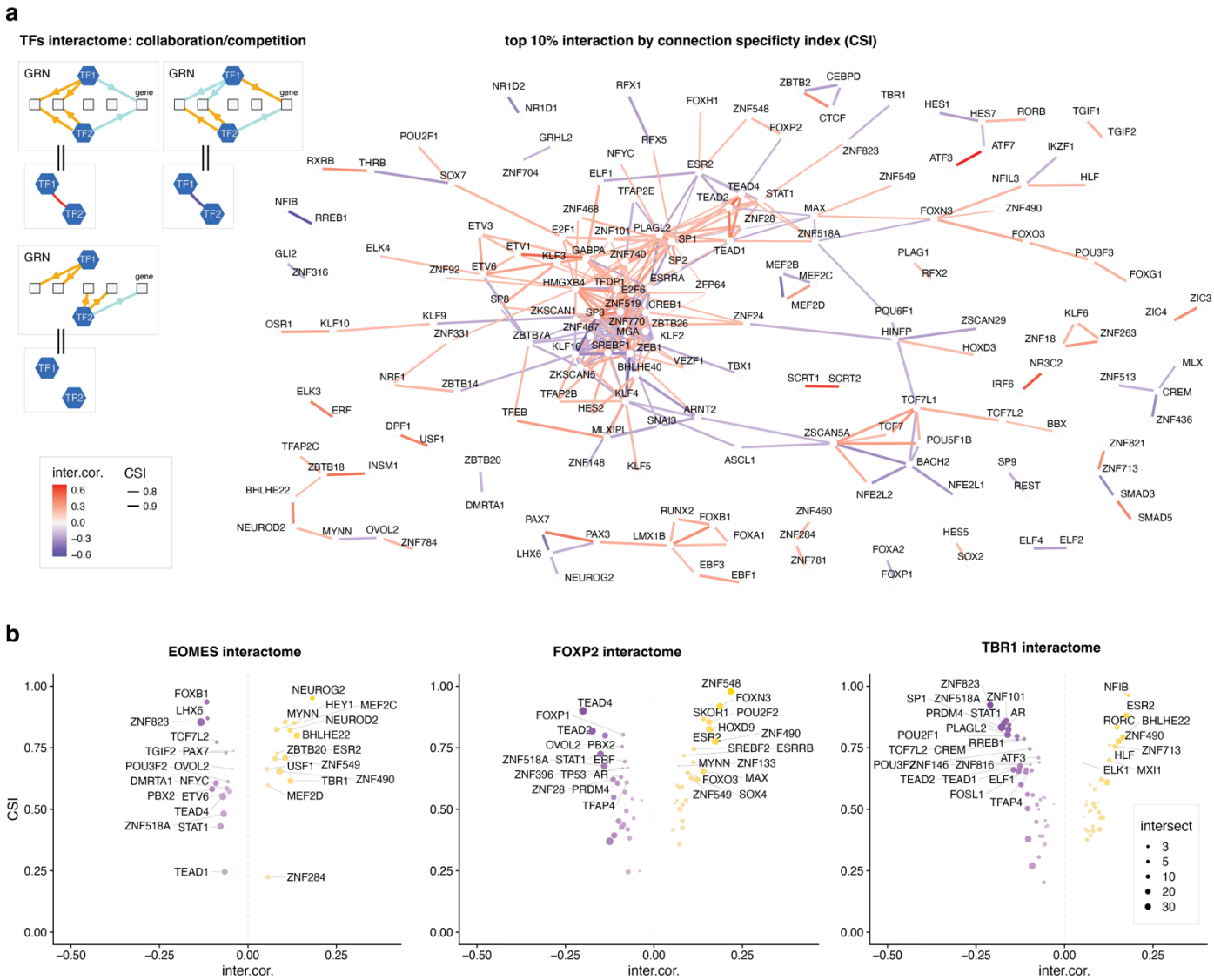


**Extended Data Fig. 3: Examples of local enhancer regulatory network upstream for specific genes**

**a-d.** Examples of detailed upstream enhancer-gene regulatory networks focused on *LHX1* (**a**), *DLX5* (**b**), *NES* (*NESTIN*, **c**) and TBR1(**d**). Networks **(a)-(d)** are organized by increased level of complexity. From each starting gene (bottom square), upstream cis-regulatory GLE(s) (green squares, labelled by unique ID from **Table S2**) are indicated in the 2$^{nd}$ layer and relationship to the target gene(s) is shown by descending arrows (activating relationships in red and repressing in blue as evaluated by correlation coefficient). When an upstream binding TF have been identified, GLEs are then connected to upstream TFs (blue squares) identified in the network and connected by a descending edge with arrow. If upstream TFs have been themselves associated with further upstream GLEs in their own cis-regulatory region, those are plotted on an additional layer. Further TF-GLE and GLE-to-gene are then plotted in additional layers, propagating therefore the network in the upstream direction until no upstream node is identified in the network (the plot is therefore meant to be read from bottom to top, i.e., in the upstream direction). Note that in this network visualization, TF blue squares represent both the TF *gene transcript* and the TF protein as a unique element. If the gene of interest is itself a TF (e.g., TBR1, LHX1 or DLX5), its downstream network is not plotted here. For simplification of the *TBR1*-focused network, only downstream edges are displayed and arrow connecting genes upward are not displayed (i.e., putative feedback loop in the system). Note that the complexity of the *TBR1* upstream network stems in part to its relationship with SOX4 that has himself a complex upstream network.

Through this analysis of upstream nodes, the enhancer regulatory network allows to identify all TFs and non-coding enhancer regions putatively associated with the expression of a gene of interest during organoid neural development and their putative hierarchy in regulatory relationships.

# Extended Data Fig. 4

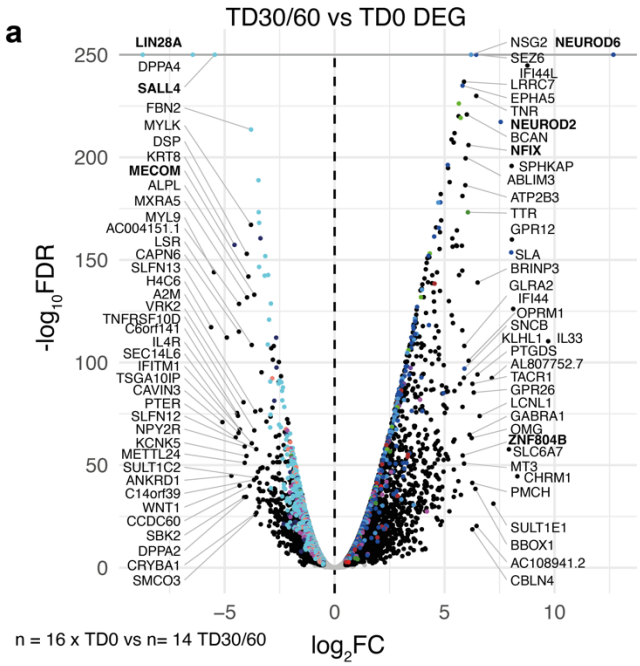


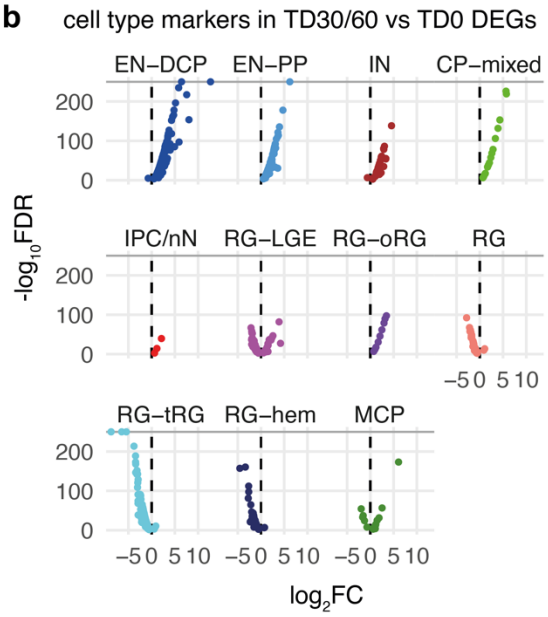
**Extended Data Fig. 4: Charting a TF interactome delineating collaborations and competitions between TFs as evaluated by regulon overlap.**

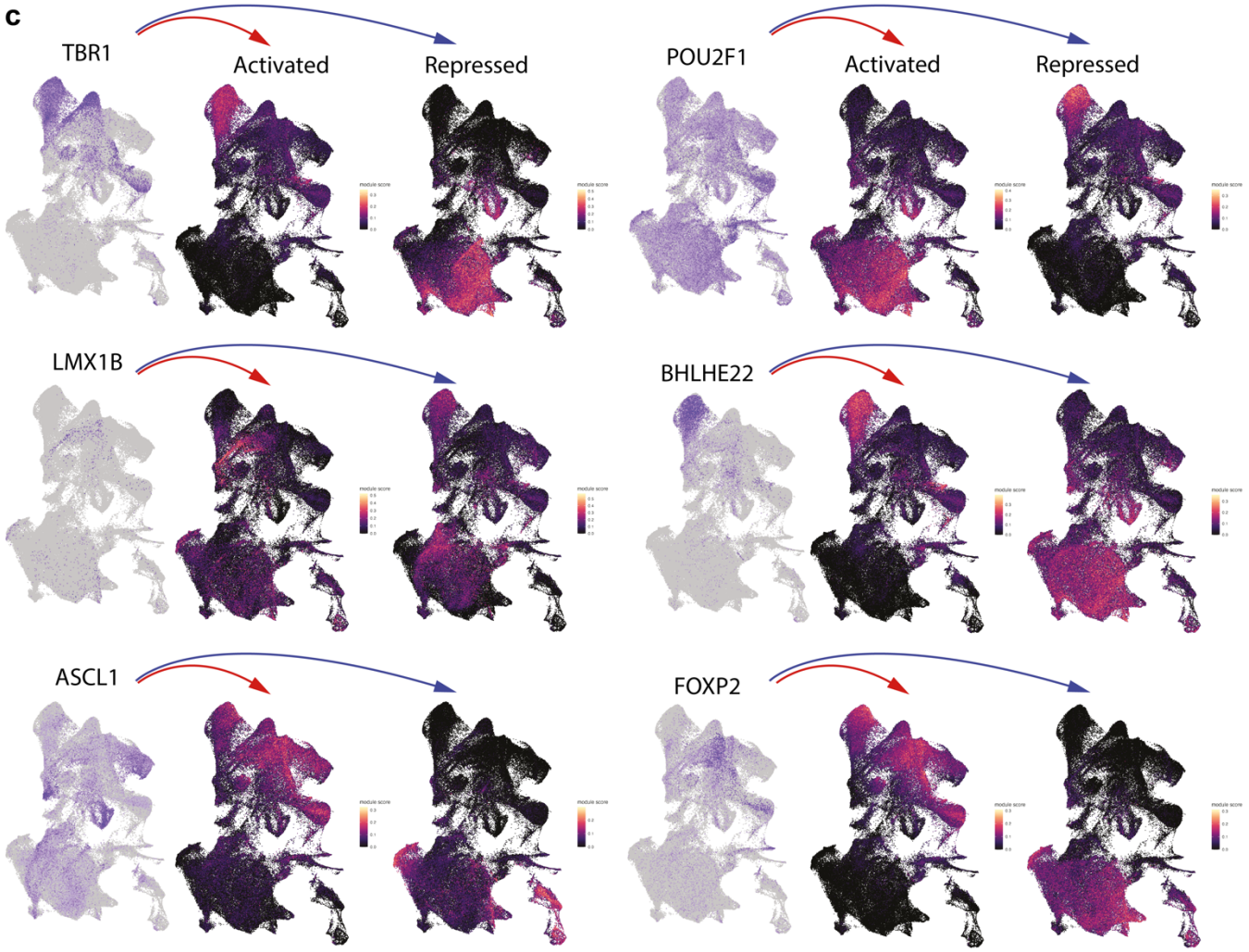
**a.** TF-TF putative interactome graph based on shared gene targets between TFs in the GRN (**Table S3**). Left side schematic shows the principle behind interactome evaluation, with the goal to further summarize the GRN (yellow/cyan arrows) into a TF-TF interactome network (red/blue edges). Specificity of the overlap was assessed by a connection specificity index (CSI) which is independent of the direction of the TF action (i.e. repress/activate) but evaluates if the overlap in regulon is higher for that pair of TF than with all other TFs (CSI, Bass et al 2013, ref. #41). Collaboration vs competition was assessed by the Pearson's correlation coefficient (Inter.cor) between the TF's weights across all gene targets (negative values indicate opposite actions on the downstream genes on average). Graph displays the top10 % specific interactions by CSI. While interactions between closely related TFs (e.g., SCRT1/SCRT2, TCF7L1/TCF7L2 pairs) could be related to a similarity in binding motif and an overlap in targeted enhancers (data not shown), interactome of TFs also included collaborators and competitors that can bind through different enhancers to influence the same genes.

**c.** Putative interactome focused on 3 specific TFs (EOMES, FOXP2, TBR1) listing top 30 most specific collaborators (right, positive inter.cor. score as defined in b)) and competitors (left, negative inter.cor. score) as ranked by Connection Specificity Index (as defined in b). Note that interaction can include both upstream and downstream TFs in the GRN (as interaction is evaluated by overlap in shared targets).

**Extended Data Fig. 5.**



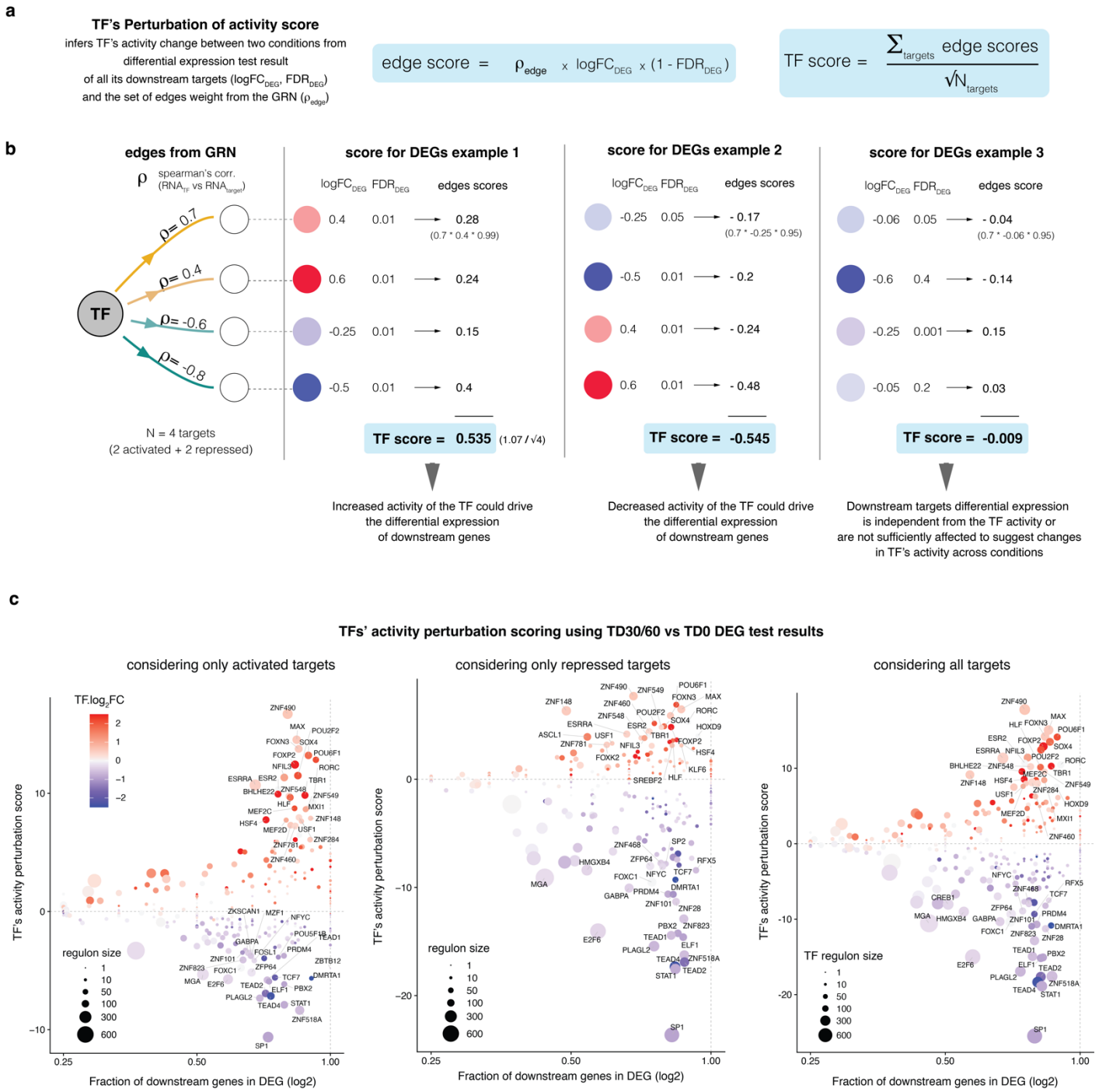



**Extended Data Fig. 5: Differential expression between early (TD0) and late stages (TD30/60) confirms the progression of neurogenesis and its relationship with cell type specification.**

- **a.** Volcano plot of the timeDEG (TD30+60 vs TD0). Dots are colored when identified as cell type-specific markers as defined by scRNA-seq analysis (cell type markers, **Table S5**). Gene names are annotated for the top 40 timeDEG as ranked by $\log_2$FC (if meeting FDR < 0.01 & $|\log_2 FC| > 2.5$) with TF in bold.
- **b.** Isolation of cell type-specific markers from **a**. Note the different distribution of the direction of change (zero marked by vertical line) between neuronal (top) and outer radial glia (RG-oRG) cell type markers (upregulated over time) compared to early radial glia-specific markers (i.e., RG, RG-tRG, RG-hem) (downregulated over time).
- **c.** Single cell expression patterns for TFs and their activated and repressed targets. Sets of 3 UMAP plots per TF showing TF expression level (left) and module scores of their downstream activated (middle) and repressed targets (right). scRNA-seq data from 70 organoid samples was used (2000 cells per library with samples from TD0, TD30 and TD60 as described in Jourdon et al. 2023) {Jourdon, 2023 #13264}. Presented TFs had at least 15 activated/repressed targets with detectable expression level by scRNA-seq. For TF expression level, normalized expression per cell using is used (SCT -normalized counts from Seurat, blue=high and grey=low); to represent average expression for each set of targets, module score was derived (from Seurat R package, dark=low to purple/yellow=high). Note the broadly mutually exclusive expression patterns of repressed vs activated targets for each TF regulon.

**Extended Data Fig. 6: Ranking TFs involvement in differential gene expression based on downstream targets using geometric perturbation index**



**Extended Data Fig. 6: Using activity perturbation score as a metric to evaluate activity changes of each TF between any two conditions based solely on expression changes in their regulons.**

**a**. For each TF, an *activity perturbation score* can be computed using GRN's edges correlation values with all downstream targets included in its regulon (Spearman's correlation coefficients in our GRN) and from the result of differential expression test of the downstream targets (i.e., log2 fold change penalized by significance, as measured by the FDR-corrected p-value estimated by EdgeR differential expression test). Note that to compute the score, the DEG results do not need to be filtered for significance as significance is considered in the equation (1-FDR); the TF itself does not need to be differentially expressed and its own DEG results are not considered in the score. The perturbation score was derived from the geometric perturbation index proposed by Martin et al 2012 {Martin, 2012 #13865}. See **Methods.**

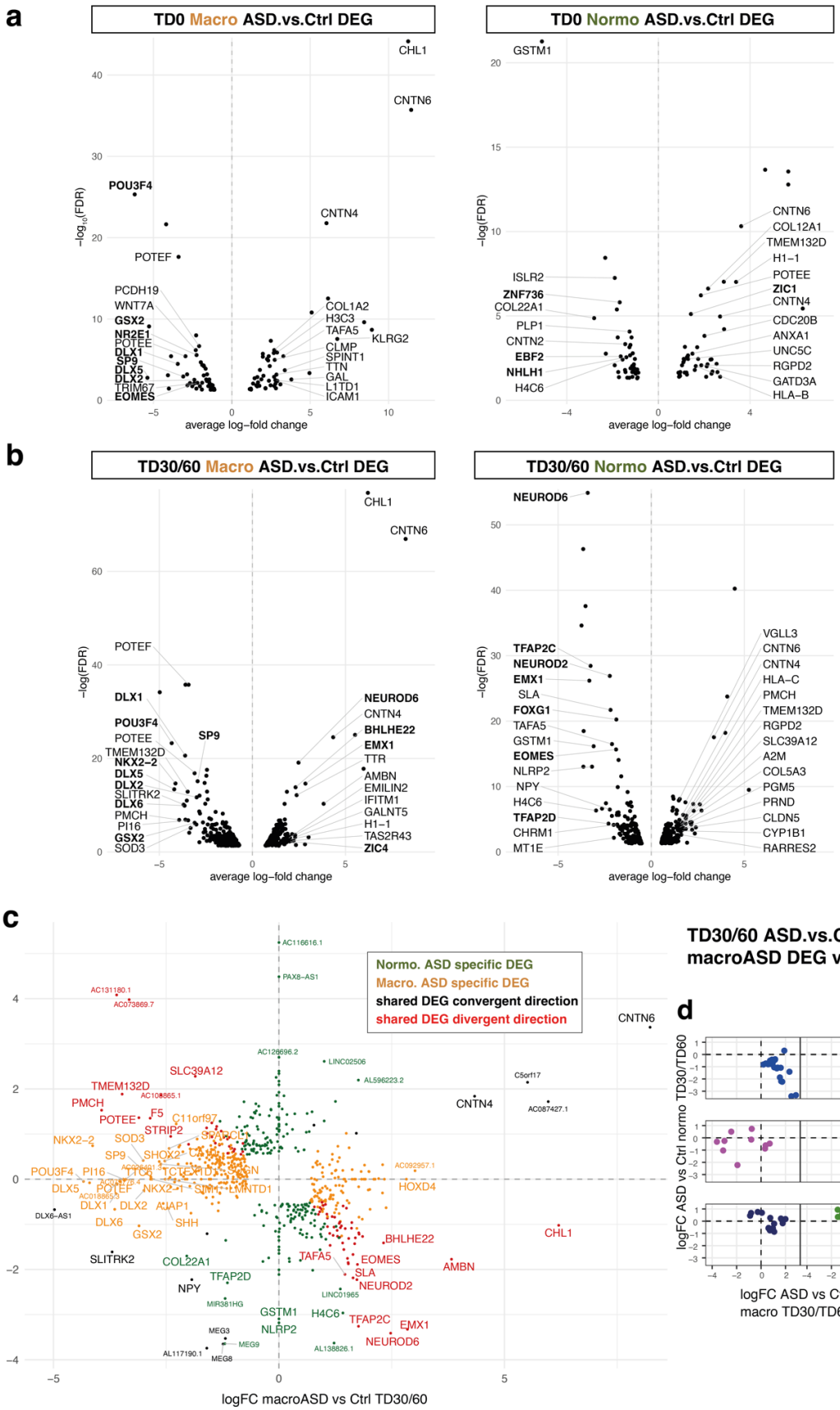
**b.** Theoretical examples for 3 different sets of DEG results using a simplified TF regulon where the TF is connected to 2 activated and 2 repressed targets. Computation of the perturbation score and interpretations are shown.

**c.** For a given DEG test result such as the timeDEG (TD30/60 vs TD0), the perturbation score for each TF can be computed considering only activating (left), repressive (middle) or both types of regulatory action in their regulon (right). The perturbation score (y axis) is plotted against the fraction of the TF regulon that meet significance in the DEG results (FDR < 0.05, |log2FC| > 0.25) and the expression change of the TF itself is plotted by color.

Since most TF have both activating and repressive relationships they are scored in all cases and show similar ranking overall. Note that the scores here are highly concordant since the GRN was built with the same data as those used for time DEGs, but the ranking of perturbation scores extracted the most relevant TFs driving the TD0 to TD30/60 transition.

Overall, the perturbation score constitutes a versatile way to compare TF's ability to drive expression changes while exploiting both repressive and activating regulatory actions inside a complex GRN. Although there could be many different complementary or alternative metrics to derive this information, we found that the perturbation score presented here had the advantage to generate clear interpretative results. The score was used in **Fig. 2b** with timeDEG ; **Fig. 2e** with cell type-DEG ; **Fig.3b** with macro-ASD DEG ; **Fig.3d** with normo-DEG ; and **Fig. 4g** with FOXG1 loss-of-function DEG results).

**Extended Data Fig. 7.**



**Extended Data Fig. 7: Differential gene expression between ASD probands and controls**

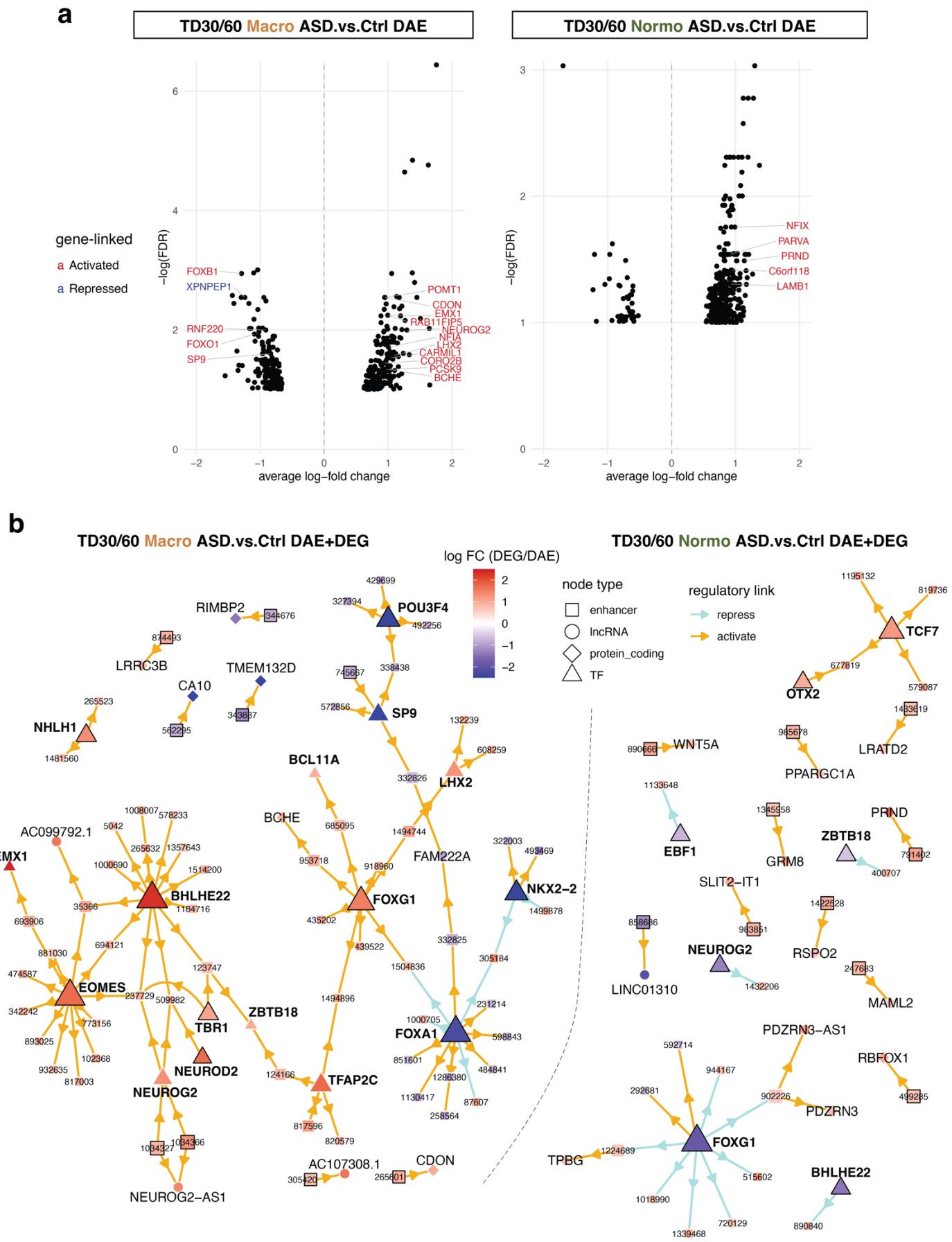
**a-b.** Volcano plots of DEG results per cohort (Macrocephalic=macro, Normocephalic=normo) and stage (TD0 or TD30/60 = using both TD30 and TD60 samples, see **Method**). Top 15 protein coding DEGs (ranked by $\log_2$FC for each direction) and with FDR < 0.01 are annotated with TF in bold.

**c.** Confusion scatter plot of macroASD DEGs (x axis) vs normoASD DEGs (y axis) $\log_2$FC results at TD30/60 showing concordance (upper right/bottom left quadrants, black color), discordance (top left/bottom right quadrants, red color) or cohort specific (green/yellow dots) DEGs among the two cohorts. Only genes that were DEG (i.e., passing $|\log_2 FC| > 0.25$ & FDR < 0.05 in either cohort) were plotted. When a gene was not tested in a cohort, a dummy $\log_2$FC value of zero for that cohort was used for plotting. Gene names are indicated when located at 2.5 unit from the origin (i.e. square root of $(\text{logFC.Macro}^2 + \text{logFC.Normo}^2) > 2.5$), with lncRNA in a smaller font. Color indicates the DEG category when comparing Macro and Normo results (red=shared (showing either concordant directions or discordant directions); normoASD-specific = green ; macroASD specific = yellow).

**d.** Confusion scatter plots of DEG results at TD30/60 plotted as in **c** for gene with cell type-specific expression in scRNA data (i.e. cluster maker of cell types (**Table S5**)) as indicated by color. Cell type specific expression was evaluated for each of the 11 main cell types. Results show that genes expressed in EN-DCP were increased in Macro-ASD and decreased in Normo-ASD and that IN related genes were specifically downregulated in Macro-ASD.

Cell type abbreviations: EN-DCP : excitatory neuron-dorsal cortical plate ; EN-PP : excitatory neuron-preplate ; IN : inhibitory neuron ; CP-mixed : mixed neurons ; IPC/nN: intermediate progenitor cell /newborn neurons ; RG-LGE: lateral ganglionic eminence progenitor ; RG: radial glia ; oRG outer radial glia ; RG-tRG: truncated RG or dividing RG ; RG-hem: cortical hem-like progenitors ; MCP: medial cortical plate cells

**Extended Data Fig. 8.**

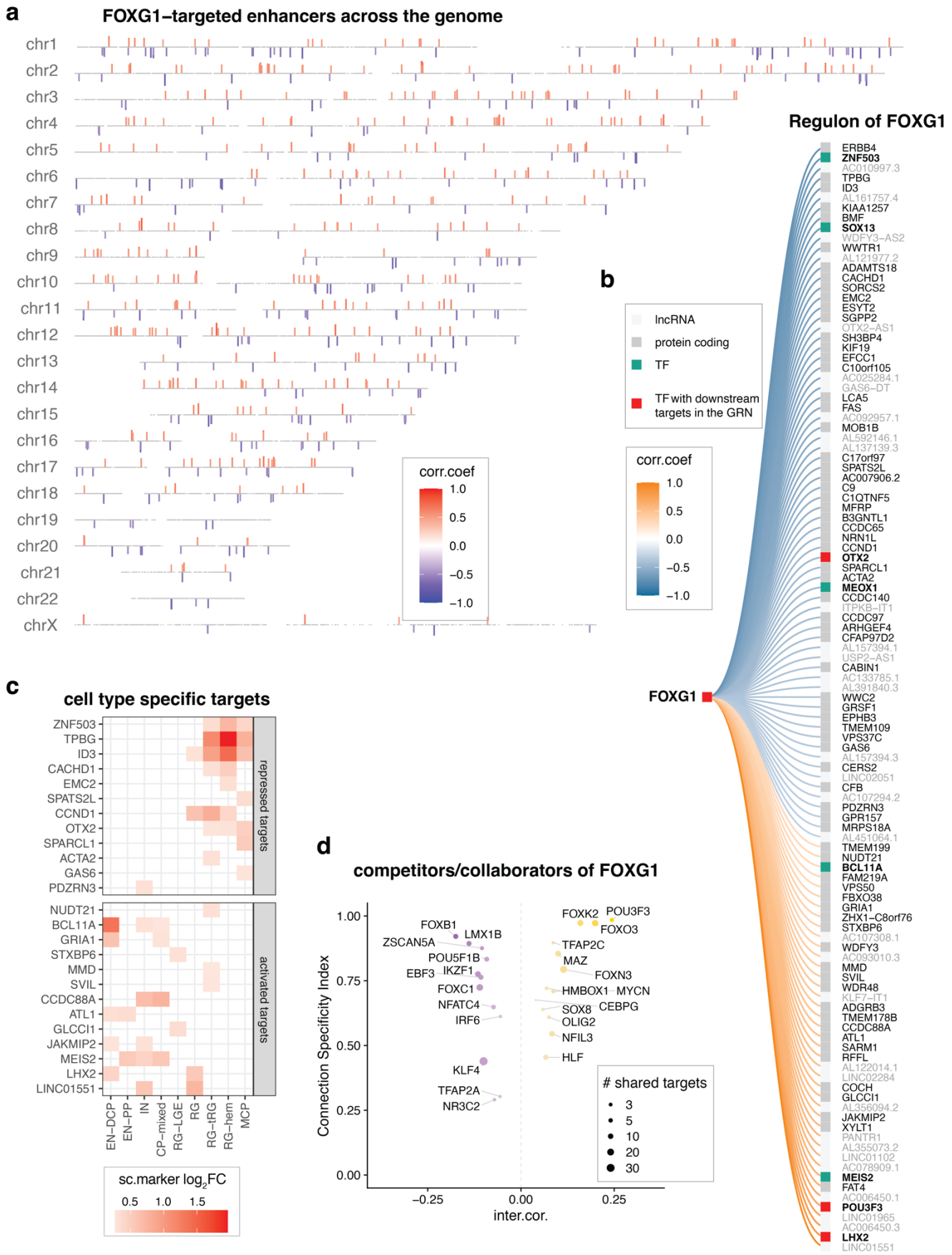


**Extended Data Fig. 8: Differential activity of gene linked enhancer (DAE) and linked genes between ASD proband and control by head-size cohort and time point.**

**a.** Volcano plots of differentially active enhancers at TD30/60 in ASD probands vs control fathers. DAE are annotated by their linked protein coding gene when included in the regulome (i.e., with significant correlation). The relationship between the DAE with the linked gene(s) is indicated by colors: red=activating (Act.reg.) or blue=repressing (Rep.reg.) corresponding, respectively, to a positive or negative correlation between H3K27ac and transcript RNA reads across the full dataset. Note that an enhancer can be linked to multiple genes and a gene can be linked to multiple enhancers (e.g., TMEM132D). Full DAE results by gene-link enhancer ID are available in **Table S4, T2**.

**b**. DAE/DEG enhancer regulatory network. The full enhancer network was subset to plot only links connecting DAE to DEG (such link can be between a TF and a downstream GLE or between a GLE and a downstream gene target). Upstream-most elements (i.e., with no other upstream differentially active element in the network) are outlined with a black line. Note that for enhancer, a threshold of FDR < 0.1 was used.

**Extended Data Fig. 9.**



**Extended Data Fig. 9: FOXG1 downstream network in forebrain organoids**

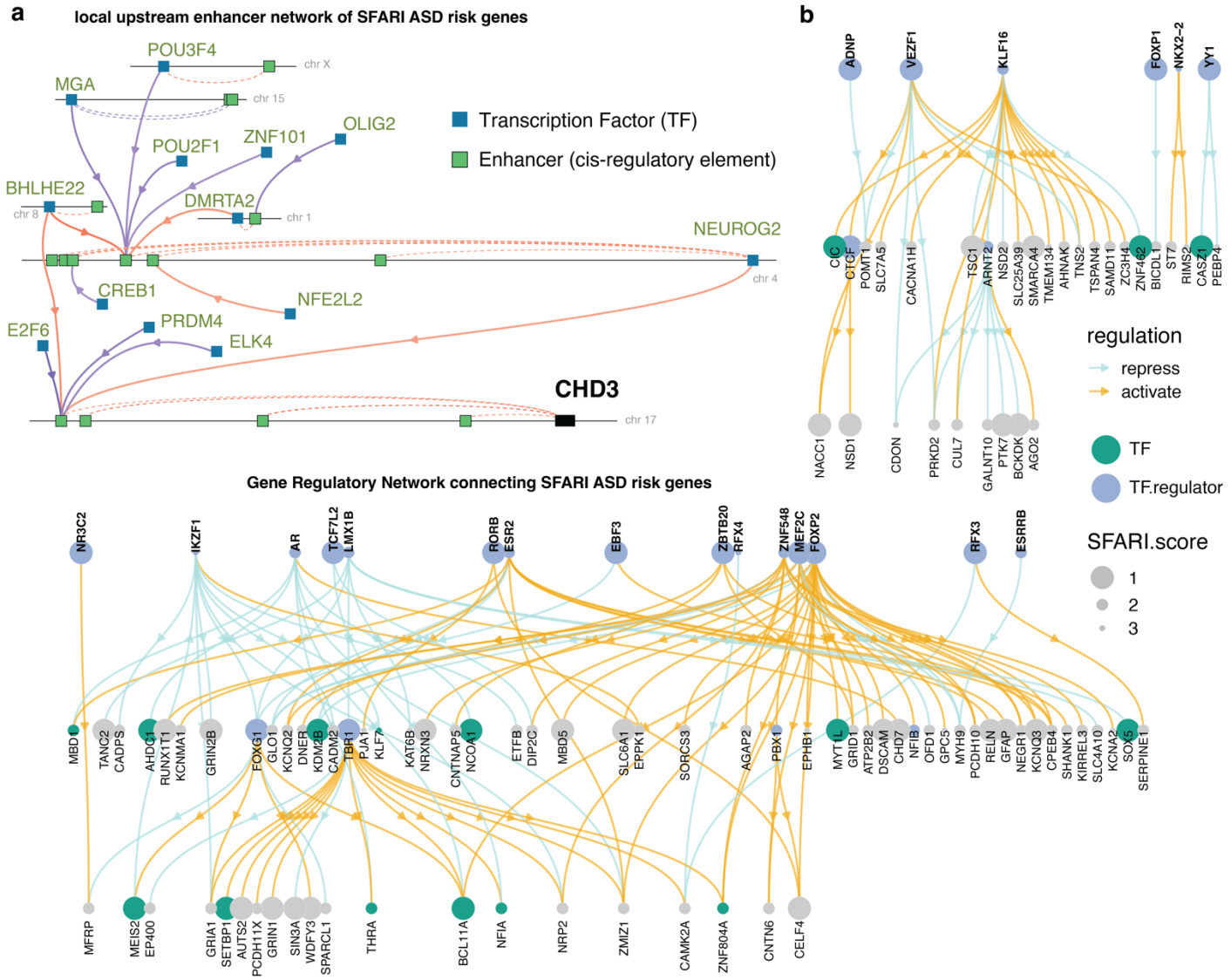
**a.** Bar plot showing the genomic location of enhancers targeted by FOXG1 per chromosome, with color and bar size proportional to the correlation coefficient between the enhancer activity (H3K27ac RPKM level) and FOXG1 transcript level (RNA RPKM) across samples. Grey trace shows the location of all gene-linked enhancer as background.

**b.** Full downstream network of FOXG1 ranked by correlation strength (edge color). The type of the downstream target is indicated by color (e.g., lncRNA, TF, etc.)

**c**. Gene targets of FOXG1 that have cell type-specific expression (marker.log2FC= DEG between the cell type and all other cells in scRNA-seq (**Table S5**), using here only positive log2FC as marker of cell type-specific expression).

**d.** Dot plot of FOXG1 interactome showing TFs that are collaborators (gold) and competitors (purple) of FOXG1 across the GRN (see Fig. S4) with the specificity of the interaction ranked by CSI (y-axis, corresponding to the fraction of all other TFs that have lesser regulon overlaps than the one observed for the evaluated pair).

**Extended Data Fig. 10.**




**Extended Data Fig. 10: ASD risk genes noncoding enhancer regions and SFARI genes in the GRN**

**a.** Example of upstream enhancer subnetwork for the ASD risk gene and chromatin modifier CHD3. Enhancers (green boxes) are connected to their downstream genes by dashed lines. TF (blue boxes) are connected to their downstream enhancer targets by arrows (red=activation, blue=repression). When in turn TF have themselves an enhancer controlling their expression, it is indicated on the linear genome. Enhancers are positioned at scale in the linear genome (their width is not represented, chromosome number is indicated, see Table S2 T4 for positions). Note the 6 degrees-separation between CHD3 expression and the TF OLIG2.

**b.** GRN restricted to links connecting 348 ASD risk genes (risk score indicated by size, with score=1 more confidently associated to ASD as defined by SFARI database). Risk genes were preferentially associated in the GRN (Fisher exact test two-sided p-value = 0.0015), notably because they included 33 highly interconnected TFs (in blue = TF.regulator). Note how some TF risk genes such as BCL11A are downstream of important regulators such as FOXP2, TBR1 or FOXG1. Also note, that other known risk genes among the available ~1000 SFARI genes were either not included in the GRN or not connected to another risk genes, thus are not present in this graph.